Jan Beran · Yuanhua Feng
Sucharita Ghosh · Rafal Kulik

Long-Memory Processes

Probabilistic Properties and Statistical Methods

Springer

Long-Memory Processes

Jan Beran • Yuanhua Feng • Sucharita Ghosh •
Rafal Kulik

Long-Memory Processes

Probabilistic Properties and Statistical Methods

Springer

Jan Beran
Dept. of Mathematics and Statistics
University of Konstanz
Konstanz, Germany

Sucharita Ghosh
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland

Yuanhua Feng
Faculty of Business Administration
and Economics
University of Paderborn
Paderborn, Germany

Rafal Kulik
Dept. of Mathematics and Statistics
University of Ottawa
Ottawa, Ontario, Canada

ISBN 978-3-642-35511-0 ISBN 978-3-642-35512-7 (eBook)
DOI 10.1007/978-3-642-35512-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013936942

Mathematics Subject Classification (2010): 62Mxx, 62M09, 62M10, 60G18, 60G22, 60G52, 60G60, 91B84

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To our families

Preface

Long-memory, or more generally fractal, processes are known to play an important role in many scientific disciplines and applied fields such as physics, geophysics, hydrology, economics, finance, climatology, environmental sciences, biology, medicine, telecommunications, network engineering, to name a few. There are several reasons for the ubiquitous occurrence of processes in the realm of long memory. First of all, hyperbolic scaling occurs naturally (up to modifications by slowly varying functions) in limit theorems for partial sums, since, under very general conditions, the limiting processes are necessarily self-similar. One may in fact say that in the world of stochastic processes, self-similar processes play the same fundamental role as stable distributions (including the normal) in the world of finite-dimensional distributions. Hyperbolic scaling phenomena are also an essential ingredient in statistical physics (a related notion is, for example, the so-called renormalization group). This is, at least partially, connected with the role of self-similar processes in limit theorems. Another reason for the occurrence of long-memory phenomena is aggregation. This, together with heterogeneity, is a frequent explanation of long-range dependence in an economic context. In telecommunications and computer networks, distributional properties of waiting times can lead to similar results. Finally, there is also a connection to fractals (though not always direct, depending on more specific distributional assumptions).

Although the notion of long memory and related topics can be traced far back into the early 20th or even the late 19th century, it is probably fair to say that the subject has been brought to the attention of a wider mathematical audience (and, in particular, probabilists and statisticians) by the pioneering work of Mandelbrot and his coworkers. A similar pathbreaking role can be attributed to Granger in economics, to Dobrushin (and before, to Kolmogorov) in physics and, even earlier, to Hurst in hydrology. These early contributions motivated a number of eminent probabilists to develop a theory of stochastic processes in the realm of stochastic self-similarity, scaling laws and nonstandard limit theorems. The development of statistical methods followed. An overview of the state of the art in the early 1990s can be found, for instance, in Beran (1994a). Other books and monographs on the topic, most of them with a special focus on certain areas of application or specific methods or processes,

are, for instance, Park and Willinger (2000), Dehling et al. (2002), Embrechts and Maejima (2002), Robinson (2002), Doukhan et al. (2003), Rangarajan and Ding (2003), Teyssière and Kirman (2005), Bertail et al. (2006), Samorodnitsky (2006), Palma (2007) and Giraitis et al. (2012).

Since the appearance of the first monograph on statistical methods for long-memory processes in the early 1990s, there has been an enormous development. One now has a much better understanding of the probabilistic foundations and statistical principles, various new techniques have been introduced to derive limit theorems and other fundamental results, and a large variety of flexible statistical methods are now available, including parametric, nonparametric, semiparametric and adaptive inference for stationary, nonstationary, locally stationary and nonlinear processes. This book grew out of the need to summarize the main results in this rapidly expanding area. Due to the progress in the last two decades, a more systematic account of theory and methods can be given. The aim is to cover both, probabilistic and statistical aspects, in as much detail as possible (given a limited number of pages), while at the same time including a broad range of topics. Because of the enormous number of theoretical and an even more overwhelming quantity of applied papers in this area, it was not possible to include all interesting results, and we apologize in advance to all authors whose contributions we could not mention. Apart from the mathematical theory, practical aspects of data analysis are discussed and illustrated by examples from various fields of application. We hope that this book will be useful to researchers interested in mathematical aspects of long-memory processes as well as to readers whose focus is on practical data analysis.

We would like to thank Todd Mitchell (JISAO, University of Washington) for the Sahel rainfall index series (data source: National Oceanic and Atmospheric Administration Global Historical Climatology Network (version 2), at the National Climatic Data Center of NOAA), the Federal Office of the Environment (FOEN), Switzerland and Hintermann & Weber, AG, Switzerland, for the species count data, to Giovanni Galizia and Martin Strauch (Department of Biology, University of Konstanz) for calcium imaging data, and to Bimal Roy and Sankhya, B, for the permission to reproduce figures from Ghosh (2009). Also, other online data bases where time series are available for free download are gratefully acknowledged, including in particular R.J. Hyndman's Time Series Data Library; the River Discharge Database of The Center of Sustainability and Global Environment, Gaylord Nelsen Institute for Environmental Studies, University of Wisconsin-Madison; the Climate Explorer of the Royal Netherlands Meteorological Institute; the Physionet databank funded by the National Institute of Health; the NASA Ozone Processing Team.

J.B. would like to thank the University of Konstanz for granting him a sabbatical with the sole purpose of working on this book and Paul Embrechts and colleagues at RiskLab (ETH) for their hospitality during that sabbatical. Thanks go in particular to Bikram Das, Marius Hofert, Georg Mainik, Artem Sapozhnikov, Alain-Sol Sznitman, Hans Herrmann and Paul Embrechts for stimulating discussions; to Martin Schützner, Dieter Schell, Yevgen Shumeyko, Arno Weiershäuser and Dirk Ocker for years of fruitful collaboration; and to Volker Bürkel for reading parts of a preliminary manuscript.

Y.F. would like to thank the University of Paderborn and the Faculty of Business Administration and Economics for great support. In particular, the Faculty of Business Administration and Economics kindly provided the financial support for an additional 3-year assistant position for supporting teaching at the Professorship of Econometrics and Quantitative Methods of Empirical Economic Research, so that Y.F. could pay more attention to writing this book. Some Ph.D. students and students in the research group have helped by collecting and handling some related data. Special thanks go to colleagues from the Department of Economics and the Center of International Economics, in particular, Thomas Gries and Manfred Kraft, for collaboration, helpful discussions and support.

S.G. would like to thank the Forest Resources and Management unit of the WSL for unstinting support and the colleagues from the WSL IT unit for seeing through that the computing jobs ran without a glitch. Swiss National Science Foundation funded two 3-year projects in the domain of climate change, supporting Ph.D. students. Substantive collaboration with Christoph Frei, ETH Zurich and MeteoSwiss, Switzerland, on Swiss precipitation dynamics, Brigitta Amman, Willy Tinner (palaeoecology) and Jakob Schwander (physics), University of Bern, on statistical topics related to rapid climate change, and background information and data on vascular plant species richness in Switzerland provided by Matthias Plattner, Hintermann and Weber, AG, Switzerland, and Federal Office of the Environment, Switzerland, are gratefully acknowledged. Additional thanks go to her former students Dana Draghicescu, Patricia Menéndez and Hesam Montazeri for many hours of interesting discussions and joint work, the Statistics groups at ETH Zurich and EPF Lausanne, in particular, Stefan Morgenthaler, Hans Künsch and Werner Stahel for collegiality and collaboration, and Otto Wildi, WSL, for inspirational discussions in particular during editorial collaboration on another book.

R.K. would like to thank his Ph.D. supervisor, Ryszard Szekli of the University of Wrocław, Poland, for introducing him into the world of long memory. Special thanks to Raluca Balan, André Dabrowski and Gail Ivanoff of the University of Ottawa, and Miklós Csörgő and Barbara Szyszkowicz of Carleton University for giving him the opportunity to spend fruitful years as a postdoctoral fellow. R.K. would also like to thank Paweł Lorek, Marc Raimondo, Philippe Soulier and Cornelia Wichelhaus for fruitful collaboration in the field of long-range dependence.

We would like to thank our families for their support. J.B. and S.G. would like, in particular, to thank Céline Beran and Sir Hastings for all those long walks and some of the best afternoons in Ebmatingen near Greifensee that anyone can dream of. Y.F. would in particular like to thank Suju, Zhongyi and Katharina for great understanding, attention and steady family support during the long march in the last years. Finally, R.K. would like to thank Kasia, Basia and Ewa "*za cierpliwość*".

Konstanz, Germany — Jan Beran
Paderborn, Germany — Yuanhua Feng
Birmensdorf, Switzerland — Sucharita Ghosh
Ottawa, Canada — Rafal Kulik

Contents

Chapter 1
Definition of Long Memory

1.1 Historic Overview

A long time before suitable stochastic processes were available, deviations from independence that were noticeable far beyond the usual time horizon were observed, often even in situations where independence would have seemed a natural assumption. For instance, the Canadian–American astronomer and mathematician Simon Newcomb (Newcomb 1895) noticed that in astronomy errors typically affect whole groups of consecutive observations and therefore drastically increase the "probable error" of estimated astronomical constants so that the usual $\sigma/\sqrt{n}$-rule no longer applies. Although there may be a number of possible causes for Newcomb's qualitative finding, stationary long-memory processes provide a plausible "explanation". Similar conclusions were drawn before by Peirce (1873) (see also the discussion of Peirce's data by Wilson and Hilferty (1929) and later in the book by Mosteller and Tukey (1977) in a section entitled "How $\sigma/\sqrt{n}$ can mislead"). Newcomb's comments were confirmed a few years later by Pearson (1902), who carried out experiments simulating astronomical observations. Using an elaborate experimental setup, he demonstrated not only that observers had their own personal bias, but also each individual measurement series showed persisting serial correlations. For a discussion of Pearson's experiments, also see Jeffreys (1939, 1948, 1961), who uses the term "internal correlation". Student (1927) observes the "phenomenon which will be familiar to those who have had astronomical experience, namely that analyses made alongside one another tend to have similar errors; not only so but such errors, which I may call semi-constant, tend to persist throughout the day, and some of them throughout the week or the month. . . . Why this is so is often quite obscure, though a statistical examination may enable the head of the laboratory to clear up large sources of error of this kind: it is not likely that he will eliminate all such errors. . . . The chemist who wishes to impress his clients will therefore arrange to do repetition analyses as nearly as possible at the same time, but if he wishes to diminish his real error, he will separate them by as wide an interval of time as possible." Since, according to Student, it is difficult to remove the error even by careful statistical examination, simple trends are probably not what he had in mind. Instead, a second-order

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_1, © Springer-Verlag Berlin Heidelberg 2013

property such as slowly decaying autocorrelations may come close to his notion of "semi-constant errors". For spatial data, the Australian agronomer Smith (1938) found in so-called uniformity trials an empirical law for wheat yield variation across space that contradicts the assumption of independence or summable correlations since the standard deviation of the sample mean converges to zero at a slower rate than the square root of the plot size. These findings were later taken up by Whittle (1956, 1962), who proposed space-time models based on stochastic partial differential equations exhibiting hyperbolically decaying spatial correlations and thereby a possible explanation of Fairfield Smith's empirical law. In hydrology, Hurst (1951) discovered an empirical law while studying the long-term storage capacity of reservoirs for the Nile (also see Hurst et al. 1965). Built on his empirical findings, Hurst recommended to increase the height of the planned Aswan High Dam far beyond conventional forecasts. Feller (1951) showed that Hurst's findings are incompatible with the assumption of weak dependence or finite moments. Later Mandelbrot coined the terms "Noah effect" for long-tailed distributions and Joseph- or Hurst-effect for "long-range dependence". The latter refers to Genesis 41, 29–30, where the "seven years of great abundance" and "seven years of famine" may be interpreted as an account of strong serial correlations. The approach of Mandelbrot and his coworkers lead to a new branch of mathematics that replaced conventional geometric objects by "fractals" and "self-similarity" (e.g. Mandelbrot 1965, 1967, 1969, 1971, 1977, 1983; Mandelbrot and van Ness 1968; Mandelbrot and Wallis 1968a, 1968b, 1969a, 1969b, 1969c) and popularized the topic in many scientific fields, including statistics. In economics, the phenomenon of long memory was discovered by Granger (1966). Simultaneously with Hosking (1981), Granger and Joyeux (1980) introduced fractional ARIMA models that greatly improved the applicability of long-range dependence in statistical practice. In geology, Matheron developed the field of geostatistics using, in particular, processes and statistical techniques for modelling spatial long memory (see e.g. Matheron 1962, 1973; Solo 1992). From the mathematical point of view, the basic concepts of fractals, self-similarity and long-range dependence existed long before the topic became fashionable; however, their practical significance had not been fully recognized until Mandelbrot's pioneering work. For instance, the Hausdorff dimension, which plays a key role in the definition of fractals, was introduced by Hausdorff (1918) and studied in detail by Abram Samoilovitch Besicovitch (e.g. Besicovitch 1929; Besicovitch and Ursell 1937). In the 17th century, Leibnitz (1646–1716) considered recursive self-similarity, and about one hundred years later, Karl Weierstrass described a function that is continuous but nowhere differentiable. The first fractal is attributed to the Czech mathematician Bernard Bolzano (1781–1848). Other early fractals include the Cantor set (Cantor 1883; but also see Smith 1875; du Bois-Reymond 1880 and Volterra 1881), the Koch snowflake (von Koch 1904), Wacław Sierpiński's triangle (Sierpinksi 1915) and the Lévy curve (Lévy 1938). (As a precaution, it should perhaps be mentioned at this place that, although fractal behaviour is often connected with long-range dependence, it is by no means identical and can, in some situations, even be completely separated from the dependence structure; see Chap. 3, Sect. 3.6.) Mathematical models for long-memory type behaviour in physics have

been known for some time in the context of turbulence (see e.g. Kolmogorov 1940, 1941). Power-law correlations have been known to be connected with critical phenomena, for instance in particle systems such as the Ising model (Ising 1924) and the renormalization group (see e.g. Cassandro and Jona-Lasinio 1978, also see the review paper by Domb 1985 and references therein). The study of critical phenomena in physics goes even much further back in history (Berche et al. 2009), to Baron Charles Cagniard de la Tour (1777–1859), who called a critical point in the phase transition "l'état particulier". With respect to unusual limit theorems for dependent observations, Rosenblatt (1961) seems to be among the first ones to derive a noncentral limit theorem where the limiting process is non-Gaussian due to nonsummable correlations and nonlinearity. This seminal paper led to further developments in the 1970s and 1980s (see e.g. Davydov 1970a, 1970b; Taqqu 1975, 1979; Dobrushin and Major 1979). The literature on statistical methods for long-memory processes until the early 1990s is summarized in Beran (1994a).

1.2 Data Examples

In this section we discuss some data examples with typical long-memory behaviour. On the way, a few heuristic methods for detecting and assessing the strength of long-range dependence will be introduced (see Sect. 5.4).

Classical areas where long-range dependence occurs frequently are dendrochronology and hydrology. We will therefore start with examples from these fields. Yearly tree ring measurements usually stretch over hundreds of years, and long memory often occurs in a rather 'pure' form, in the sense that a hyperbolic behaviour of the autocorrelations and the spectral density holds for almost all lags and frequencies respectively. Therefore, tree ring series are often used as prime examples of strong dependence and self-similarity. Consider for instance Fig. 1.1 (the data source is Hyndman, Time Series Data Library, http://robjhyndman.com/TSDL). The following typical features can be observed:

(a) *Spurious trends and cycles, and self-similarity*: The observed series exhibit local trends and periodicities that appear to be spurious, however, because they disappear again and are of varying length and frequency. Furthermore, these features and the overall visual impression of the time series remain the same when considering aggregated data, with disjoint adjacent blocks of observations being averaged (see Fig. 1.2). This is an indication of stochastic 'self-similarity', which is the property that rescaling time changes the (joint) probability distribution by a scaling factor only.

(b) *Slow hyperbolic decay*: The sample autocorrelations

$$\hat{\rho}(k) = \frac{1}{n} \sum_{i=1}^{n-|k|} (x_i - \bar{x})(x_{i+|k|} - \bar{x})$$

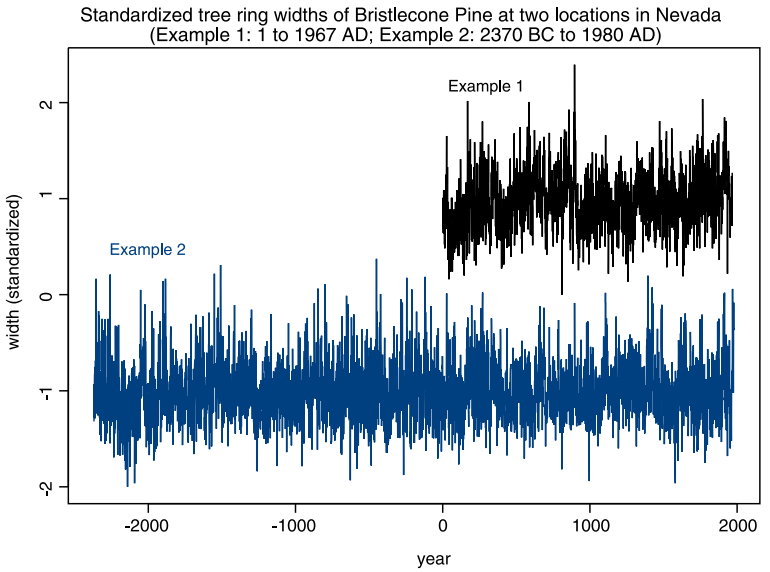

Fig. 1.1 Two typical tree ring series

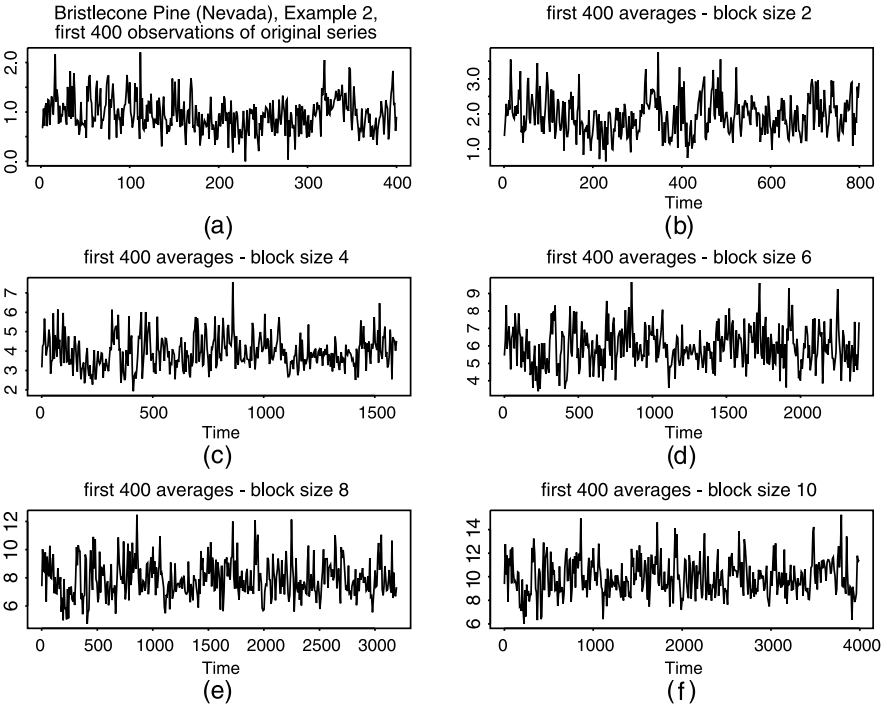

Fig. 1.2 (**a**) Tree ring series, Example 1; (**b**)–(**f**) aggregated series $\bar{x}_t = m^{-1}(x_{(t-1)m+1} + \cdots + x_{tm})$ $(t = 1, 2, \ldots, 400)$ with blocks lengths equal to 2, 4, 6, 8 and 10 respectively

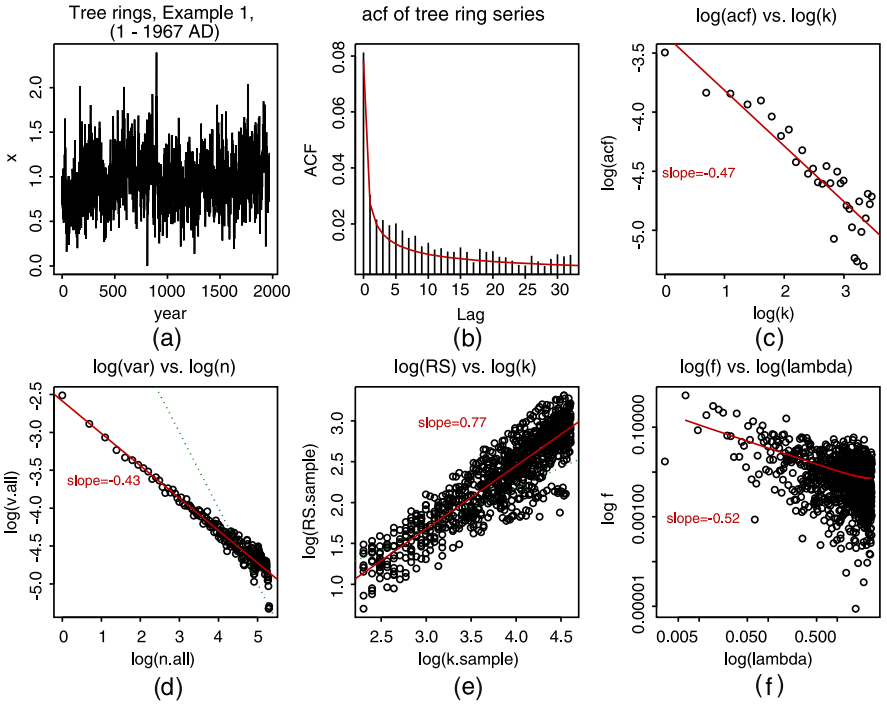

Fig. 1.3 Tree ring example 1: (**a**) observed yearly series; (**b**) empirical autocorrelations $\hat{\rho}(k)$; (**c**) $\log \hat{\rho}(k)$ vs. $\log k$; (**d**) $\log s_m^2$ vs. $\log m$; (**e**) $\log R/S$ vs. $\log k$; (**f**) $\log I(\lambda)$ vs. $\log \lambda$

(with $\bar{x} = n^{-1} \sum x_i$) decay slowly with increasing lag k. More specifically, the decay of $\hat{\rho}(k)$ appears to be hyperbolic with a rate $k^{-\alpha}$ (for some $0 < \alpha < 1$), implying nonsummability. This phenomenon is called long memory, strong memory, long-range dependence, or long-range correlations. This is illustrated in Fig. 1.3(c), where $\log \hat{\rho}(k)$ is plotted against $\log k$. The points are scattered around a straight line of the form $\log \hat{\rho}(k) \approx \text{const} + \beta_\rho \log k$ with $\beta_\rho \approx -0.5$. Similarly, the variance of the sample mean appears to decay to zero at a slower rate than n^{-1}. This can be seen empirically in Fig. 1.3(d) with $\log s_m^2$ plotted against $\log m$, where s_m^2 is the sample variance of means based on disjoint blocks of m observations, i.e.

$$s_m^2 = \frac{1}{n_m - 1} \sum_{i=1}^{n_m} (\bar{x}_{(i-1)m,m} - \bar{x})^2,$$

where

$$\bar{x}_{t,m} = \frac{1}{m} \sum_{j=1}^{m} x_{t+j}$$

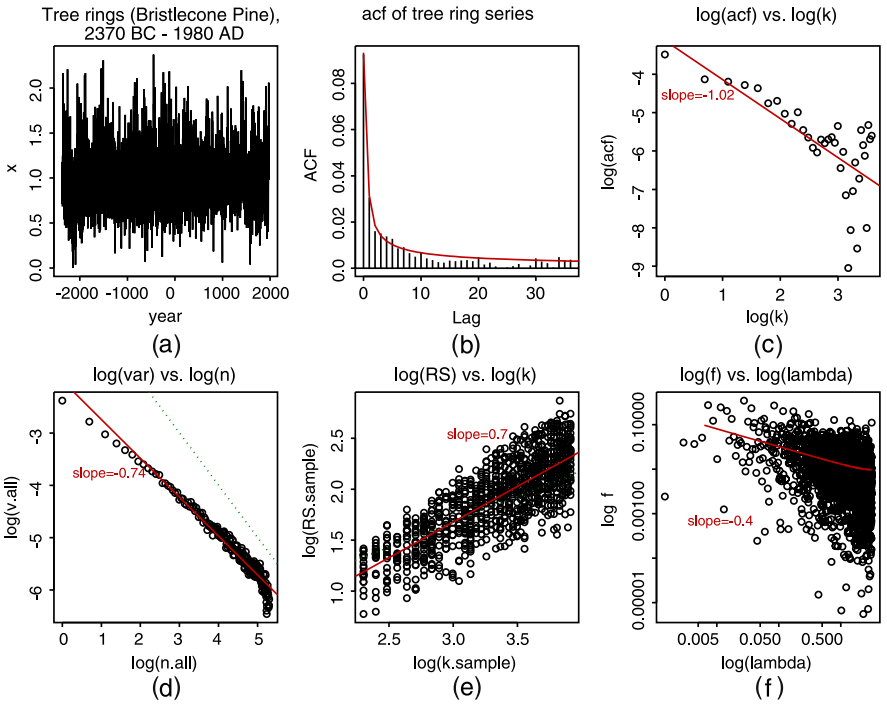

Fig. 1.4 Tree ring example 2: (**a**) observed yearly series; (**b**) empirical autocorrelations $\hat{\rho}(k)$; (**c**) $\log \hat{\rho}(k)$ vs. $\log k$; (**d**) $\log s_m^2$ vs. $\log m$; (**e**) $\log R/S$ vs. $\log k$; (**f**) $\log I(\lambda)$ vs. $\log \lambda$

and $n_m = [n/m]$. The fitted slope in Fig. 1.3(d) is close to $\beta_{s^2} = -0.4$, suggesting s_m^2 being proportional to $m^{-0.4}$, which is much slower than the usual rate of m^{-1}. A further statistic that is sometimes used to detect long-range dependence is the so-called R/S-statistic displayed in Fig. 1.3(f). The R/S-statistic is defined by

$$R/S(t,m) = \frac{R(t,m)}{S(t,m)},$$

where

$$R(t,m) = \max_{1 \le i \le m} \left(y_{t+i} - y_t - \frac{i}{m}(y_{t+m} - y_t) \right) - \min_{1 \le i \le m} \left(y_{t+i} - y_t - \frac{i}{m}(y_{t+m} - y_t) \right),$$

$$y_u = \sum_{i=1}^{u} x_i,$$

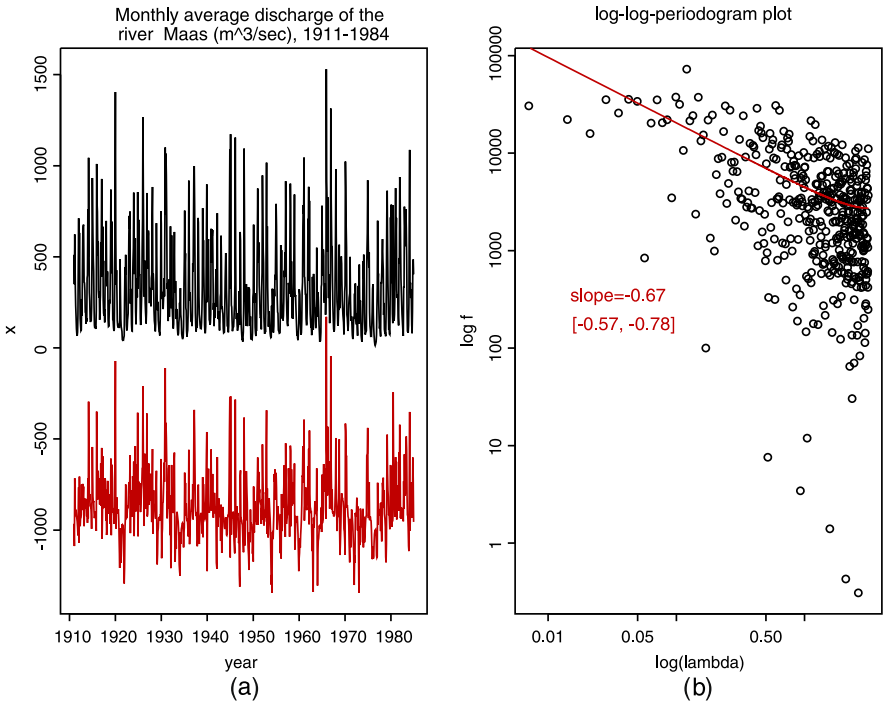

Fig. 1.5 (**a**) Monthly average discharge of the river Maas (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**)

and

$$S(t,m) = \sqrt{\frac{1}{m}\sum_{i=t+1}^{t+m}(x_i - \bar{x}_{t,m})^2}.$$

This definition originates from hydrology (see e.g. Hurst 1951), where R corresponds to the optimal capacity of a reservoir when outflow is linear, with x_i denoting the inflow at time i. Figure 1.3(f) shows $R/S(t,m)$ versus m, plotted in log-log-coordinates. Again, we see a linear relationship between $\log R/S$ (as a function of m) and $\log m$, with a slope close to $\beta_{R/S} = 0.8$. In contrast, under independence or short-range dependence, one expects a slope of 0.5 (see Sect. 5.4.1). Finally, Fig. 1.3(f) displays the logarithm of the periodogram $I(\lambda)$ (as an empirical analogue of the spectral density f) versus the log-frequency. Again an essentially linear relationship can be observed. The negative slope is around $\beta_f = -0.5$, suggesting the spectral density having a pole at the origin of the order $\lambda^{-0.5}$. Similar results are obtained for Example 2 in Figs. 1.4(a) through (f). The slopes for the log-log plots of $\hat{\rho}(k)$, s_m^2, R/S and $I(\lambda)$ are this time $\beta_\rho \approx -1$, $\beta_{s^2} \approx -0.7$, $\beta_{R/S} \approx 0.7$ and $\beta_f \approx -0.4$ respectively.

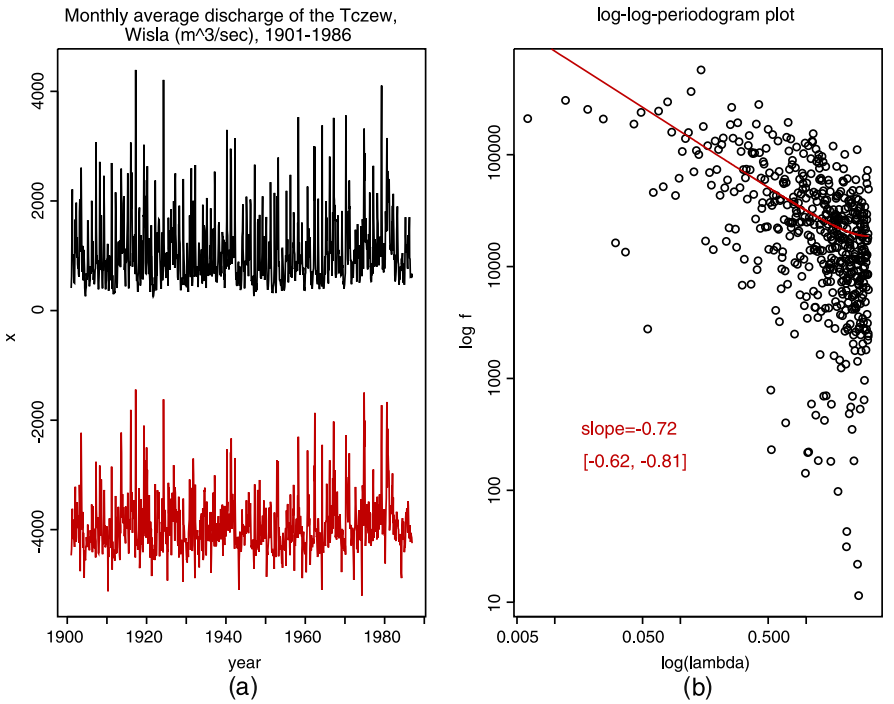

Fig. 1.6 (**a**) Monthly average discharge of the river Wisła at Tczew (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**)

Next, we consider river flow data. Figures 1.5(a), 1.6(a), 1.7(a) and 1.8(a) show the average monthly river discharge (in $m^3/\,sec$) for four rivers from different parts of the world: (1) Maas at the Lith station (The Netherlands); (2) Wisła at Tczew (Poland); (3) Tejo at V.V. de Rodao (Portugal) and (4) White River at Mouth Near Ouray, Utah (USA). The data are from the River Discharge Database of The Center of Sustainability and Global Environment, Gaylord Nelsen Institute for Environmental Studies, University of Wisconsin-Madison. Since these are monthly data, there is a strong seasonal component. To obtain an idea about the dependence structure for large lags, a seasonal effect is first removed by subtracting the corresponding monthly means (i.e. average January temperature, average February temperature etc.). The original and the deseasonalized data are shown in the upper and lower part of each time series picture respectively. For each of the deseasonalized series, the points in the log-log-periodogram (all figures (b)) are scattered nicely around a straight line for all frequencies.

The data examples shown so far may be somewhat misleading because one may get the impression that discovering long memory can be done easily by fitting a straight line to the observed points in an appropriate log-log-plot. Unfortunately, the situation is more complicated, even if one considers river flows only. For instance, Figs. 1.9, 1.10 and 1.11 show log-log-plots for the Danube at four different

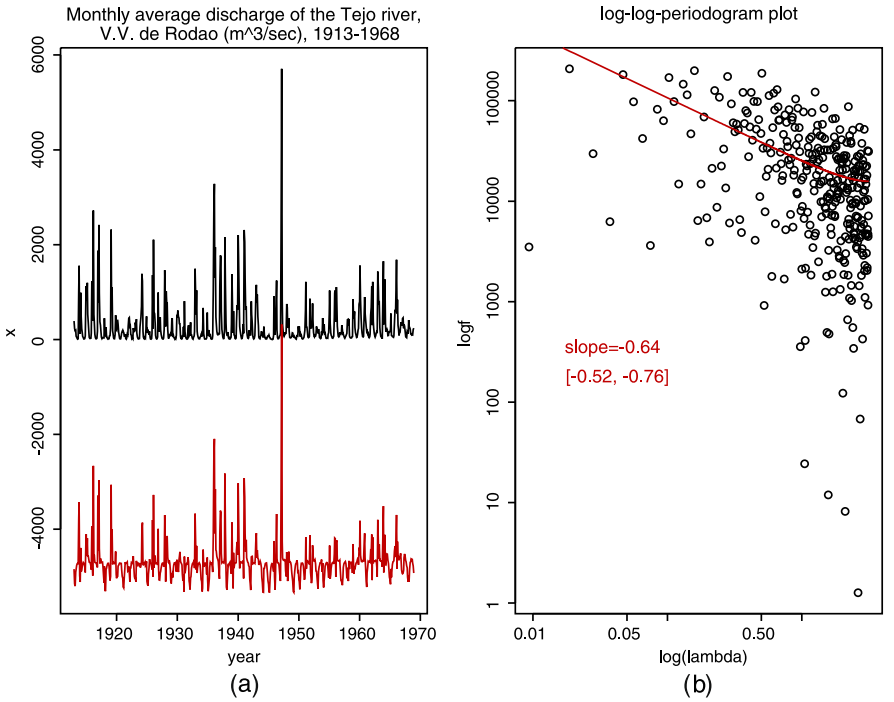

Fig. 1.7 (**a**) Monthly average discharge of the river Tejo at V.V. de Rodao (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**)

stations: (1) Bratislava (Slovakia); (2) Nagymaros (Hungary); (3) Drobeta-Turnu Severin (Romania); (4) Ceatal Izmail (Romania). Consider first the measurements in Bratislava. The points in the log-log-plots no longer follow a straight line all the way. It is therefore not clear how to estimate the 'ultimate' slopes (i.e. the asymptotic slopes as $m, k \to \infty$ and $\lambda \to 0$ respectively). Fitting a straight line to all points obviously leads to a bad fit in the region of interest (i.e. for k and m large, and λ small). This is one of the fundamental problems when dealing with long-memory (and, as we will see later, also so-called antipersistent) series: the definition of 'long memory' is an asymptotic one and therefore often difficult to detect and quantify for finite samples. A substantial part of the statistical literature on long-memory processes is concerned with this question (this will be discussed in particular in Chap. 5). In contrast to the straight lines in Figs. 1.9(b) and (c), the fitted spectral density in Fig. 1.9(d) is based on a more sophisticated method that combines maximum likelihood estimation (MLE) with the Bayesian Information Criterion (BIC) for fractional ARIMA models. This and related data adaptive methods that allow for deviations from the straight line pattern will be discussed in Chap. 5 (Sects. 5.5 to 5.10) and Chap. 7 (Sects. 7.4.5 and 7.4.6).

Analogous observations can be made for the other Danube series. To save space, only the log-log-periodogram plots are shown (Figs. 1.10, 1.11). Note that the MLE

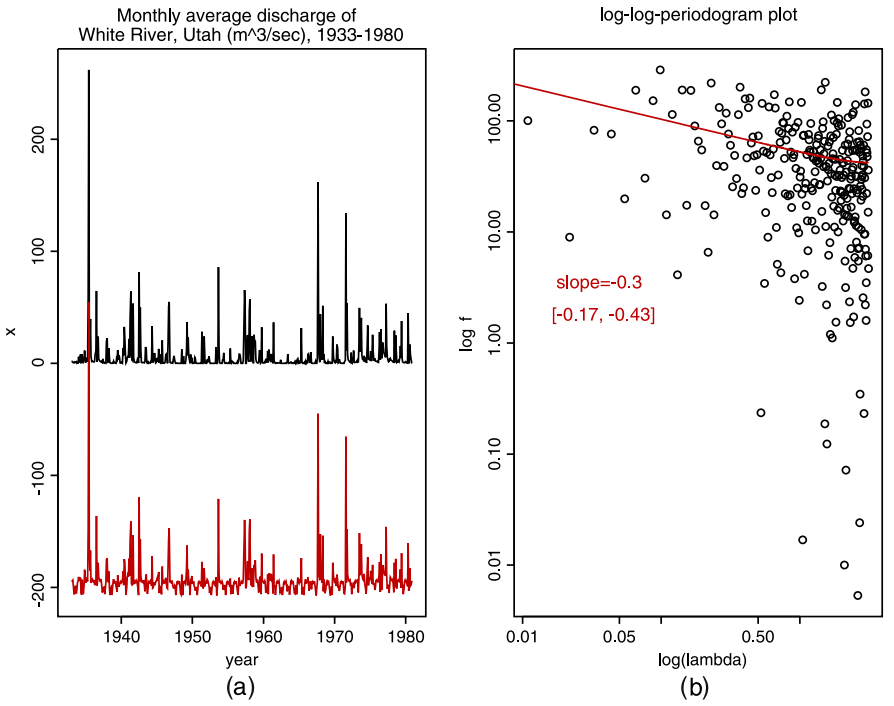

Fig. 1.8 (**a**) Monthly average discharge of White River, Utah (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**)

estimates of β_f (-0.25, -0.31, -0.25, -0.29) are all very similar. It seems that a value around -0.25 to -0.3 is typical for the Danube in these regions. On the other hand, the slope changes as one moves upstream. For instance, at Hofkirchen in Germany (lower panel in Sect. 1.11), long memory appears to be much stronger with $\beta_f \approx -0.75$, and a straight line fits all the way.

An even more complex river flow series are monthly measurements of the Nile river at Dongola in Sudan, displayed in Fig. 1.12. Seasonality is very strong here, and subtracting seasonal means does not remove all of it (see Figs. 1.12(a), (b)). A possible reason is that the seasonal effect may change over time; it may be non-linear, or it may be stochastic. The MLE fit combined with the BIC captures the remaining seasonality quite well. This model assumes seasonality (remaining after previous subtraction of the deterministic one) to be stochastic.

The data examples considered so far could be modelled by stationary processes. Often stationarity is not a realistic assumption, or it is at least uncertain. This makes identification of stochastic long memory even more difficult, because typical long-memory features may be confounded with nonstationary components. Identifying and assessing possible long-memory components is however essential for correct inference about the non-stationary components. A typical example is the assessment of global warming. Figure 1.13(a) shows yearly average temperatures in cen-

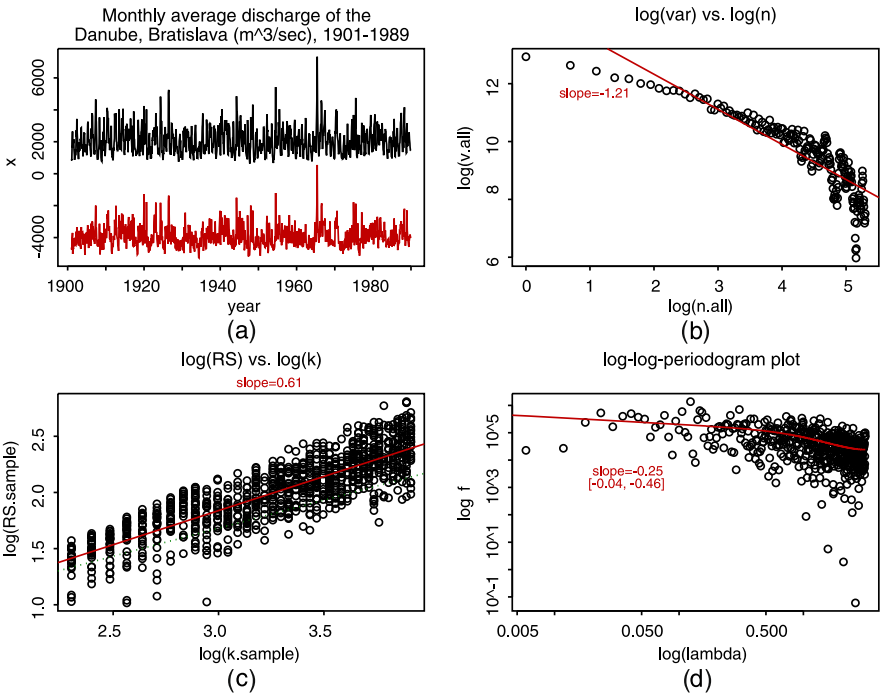

Fig. 1.9 Monthly average discharge of the Danube at Bratislava (*upper series*: original; *lower series*; deseasonalized) and various log-log-plots for the deseasonalized series

tral England for the years 1659 to 2010 (Manley 1953, 1974; Parker et al. 1992; Parker and Horton 2005). The data were downloaded using the Climate Explorer of the Royal Netherlands Meteorological Institute. The main question here is whether there is evidence for a systematic increase. The simplest way of answering this question is to fit a straight line and test whether the slope, say β_1, is positive. The dependence structure of the regression residuals has an influence on testing whether β_1 is significantly larger than zero. As will be shown later, if the observations are given by $y_t = \beta_0 + \beta_1 t + e_t$ with e_t being stationary with long-range dependence such that $\rho(k) \sim c|k|^{2d-1}$ (as $|k| \to \infty$) for some $d \in (0, \frac{1}{2})$, then the variance of the least squares estimator of β_1 increases by a constant times the factor n^{2d} compared to the case of uncorrelated or weakly dependent residuals (see Sect. 7.1). This means that correct confidence intervals are wider by a factor proportional to n^d. The difference can be quite substantial. For example, the estimate of d for the Central England series is about 0.2. For the given data size, we thus have a factor of $n^d = 704^{0.2} \approx 3.7$. It is therefore much more difficult to obtain a significant result for β_1 than under independence. Complicating the matter further, one may argue that the trend, if any, may not be linear so that testing for β_1 leads to wrong conclusions. Furthermore, the observed series may even be nonstationary in the sense of random walk (or unit roots). As will be discussed in Chap. 7 (Sects. 7.4.5 and 7.4.6), there is a method (so-

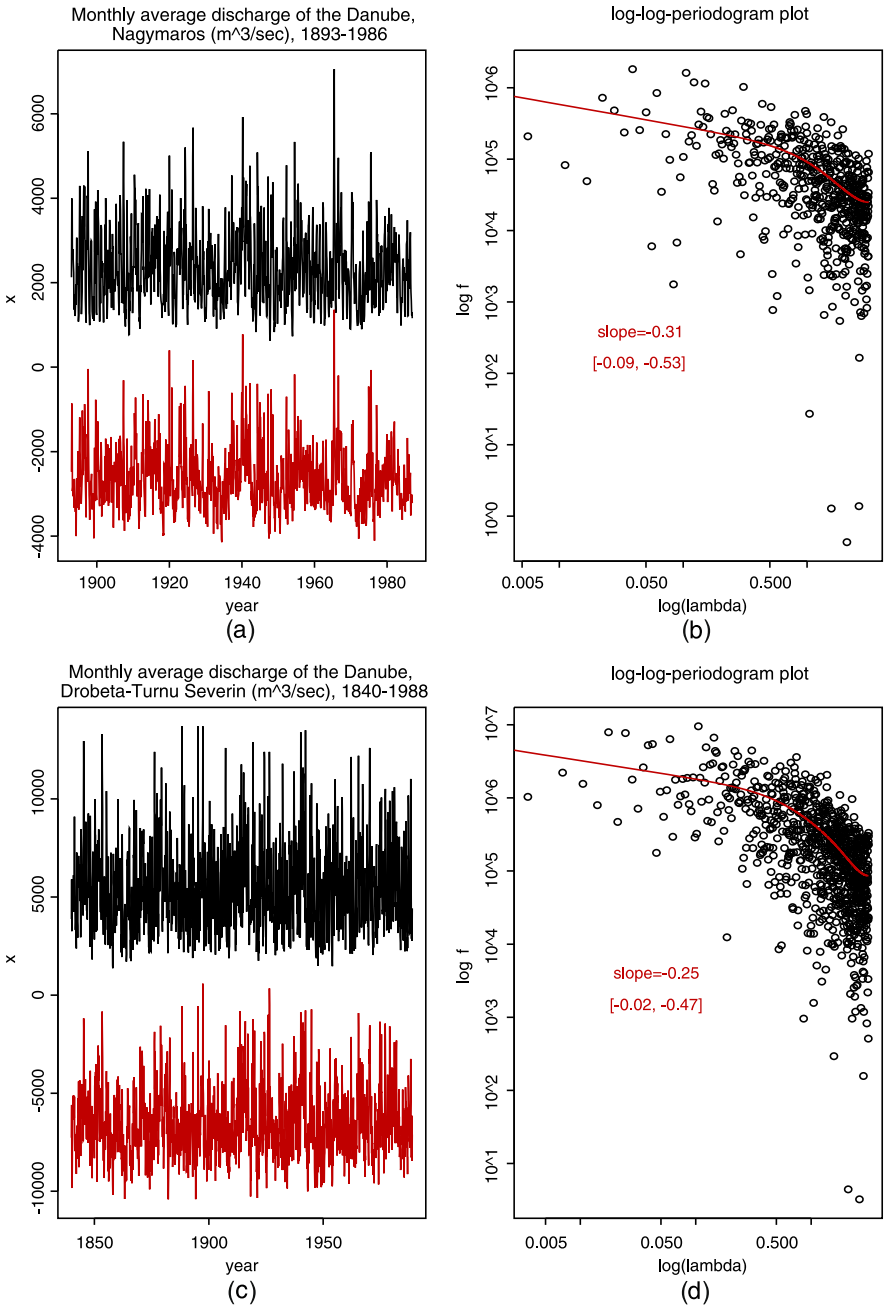

Fig. 1.10 (**a**) Monthly average discharge of the Danube at Nagymaros (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**); (**c**) monthly average discharge of the Danube at Drobeta-Turnu (*upper series*: original; *lower series*; deseasonalized); (**d**) log-log-periodogram of the deseasonalized series in (**c**)

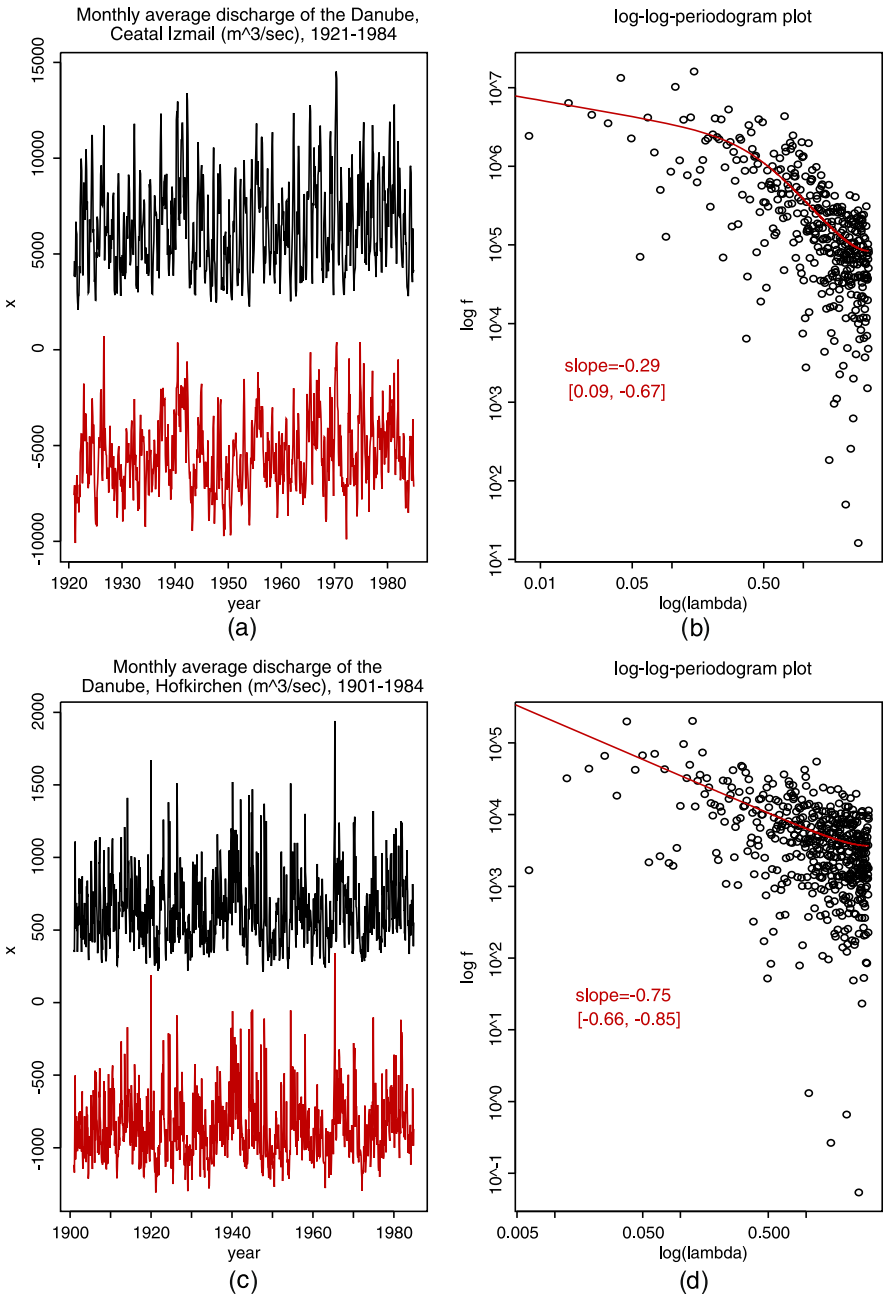

Fig. 1.11 (**a**) Monthly average discharge of the Danube at Ceatal Izmail (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**); (**c**) monthly average discharge of the Danube at Hofkirchen (*upper series*: original; *lower series*; deseasonalized); (**d**) log-log-periodogram of the deseasonalized series in (**c**)

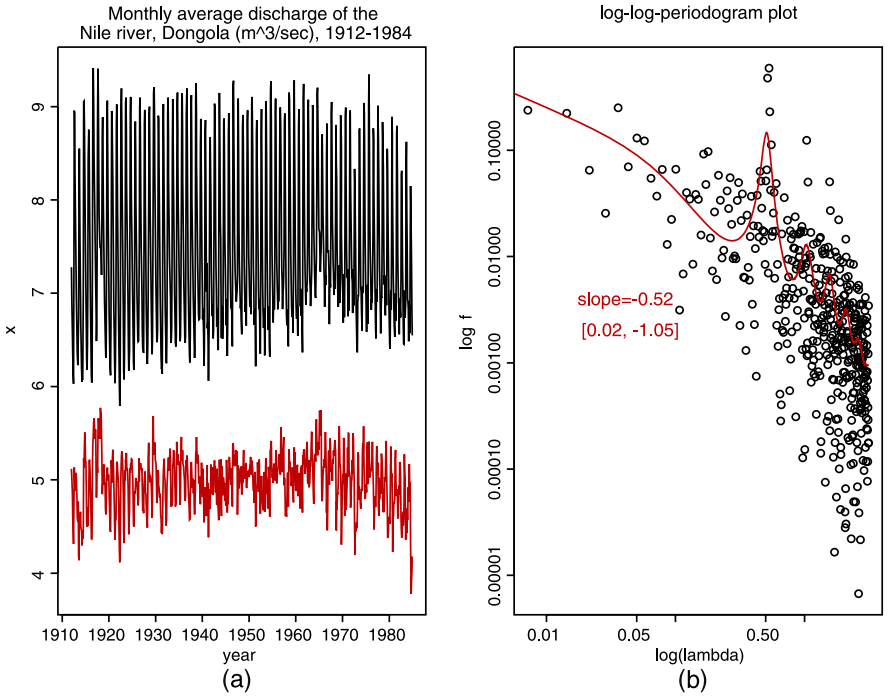

Fig. 1.12 (**a**) Monthly average discharge of the Nile river at Dongola (*upper series*: original; *lower series*; deseasonalized); (**b**) log-log-periodogram of the deseasonalized series in (**a**)

called SEMIFAR models) that incorporates these possibilities using nonparametric trend estimation, integer differencing and estimation of the dependence parameters. Clearly, the more general a method is, the more difficult it becomes to obtain significant results. Nevertheless, the conclusion based on SEMIFAR models is that the trend is increasing and significantly different from a constant.

Another series with a clear trend function is displayed in Fig. 1.14. The measurements are monthly averaged length-of-day anomalies (Royal Netherlands Meteorological Institute). Overall, one can see that there is a slight decline together with a cyclic movement. The fitted line was obtained by kernel smoothing. As will be seen in Chap. 7, the crucial ingredient in kernel smoothing is the bandwidth. A good choice of the bandwidth depends on the dependence structure of the residuals. For the data here, the residuals have clear long memory. In fact, the estimated long-memory parameter is very close to the boundary of nonstationarity so that the possibility of a spectral density proportional to λ^{-1} (as $\lambda \to 0$) cannot be excluded. Processes with this property are also called $1/f$-noise (which, in our notation, should actually be called $1/\lambda$-noise because f stands for frequency).

In the previous examples, the trend function is obviously smooth. Quite different time series are displayed in Figs. 1.15(a) and (d). The data were downloaded from the Physionet databank funded by the National Institute of Health (Goldberger et al.

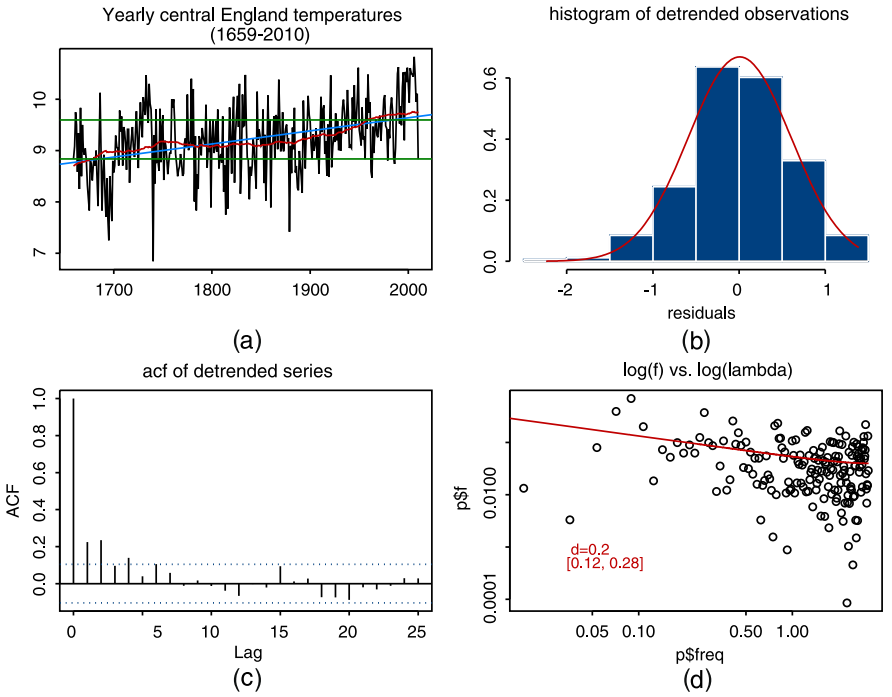

Fig. 1.13 (**a**) Yearly mean Central England temperatures together with a fitted least squares line and a nonparametric trend estimate; (**b**) histogram of residuals after subtraction of the nonparametric trend function; (**c**) acf of residuals; (**d**) log-log-periodogram of residuals

2000). The upper series in Fig. 1.15(a) shows consecutive stride intervals (stride-to-stride measures of footfall contact times) of a healthy individual, whereas the upper series in Fig. 1.15(d) was obtained for a patient suffering from Parkinson's disease. The complete data set consists of patients with Parkinson's disease ($N = 15$), Huntington's disease ($N = 20$) and amyotrophic lateral sclerosis ($N = 13$), as well as a control group ($N = 16$) (Hausdorff et al. 1997, 2000). Both series in Figs. 1.15(a) and (d) contain a spiky, somewhat periodic but also irregular, component. A natural approach to analysing such data is to decompose them into a 'spiky' component and the rest. Here, kernel smoothing is not appropriate because it tends to blur sharp peaks. Instead, wavelet thresholding (see e.g. Donoho and Johnstone 1995) separates local significant spikes from noise more effectively. The series plotted below the original ones are the trend functions fitted by standard minimax thresholding using Haar wavelets, the series at the bottom and, enlarged, in Figs. 1.15(b) and (e) are the corresponding residuals. The log-log-periodogram plots for the residual series and fitted fractional ARIMA spectral densities in Figs. 1.15(c) and (f) indicate long memory. A comparison of Figs. 1.15(c) and (f) shows that the slope β_f is less steep for the Parkinson patient. Indeed, using different techniques, Hausdorff et al. (1997, 2000) found evidence for β_f being closer to zero for patients suffering form Parkinson's disease (and other conditions such as Huntington's disease or

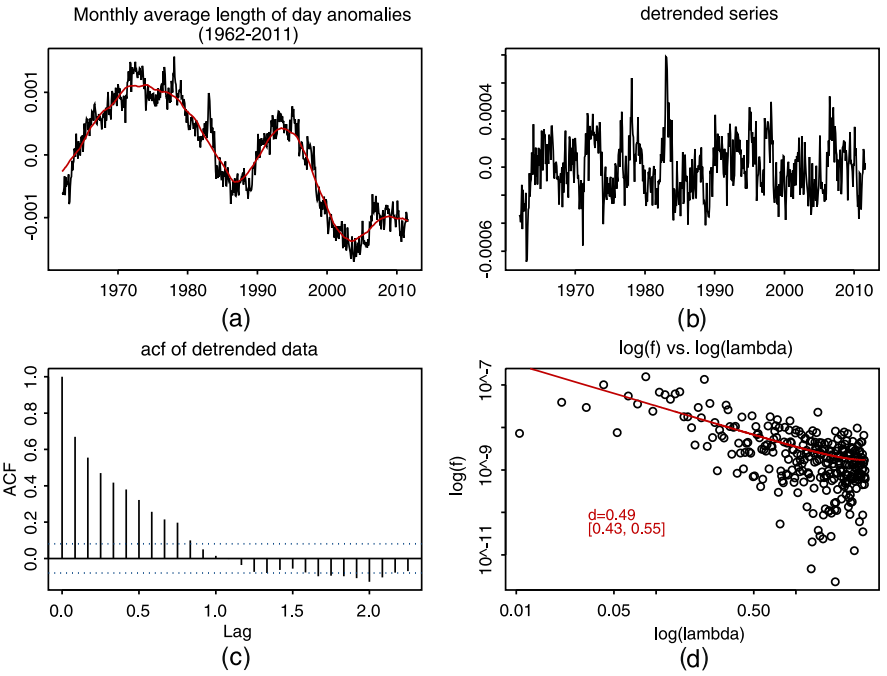

Fig. 1.14 (**a**) Monthly averaged length-of-day anomalies (in seconds); (**b**) residuals after subtraction of the nonparametric trend function; (**c**) acf of residuals; (**d**) log-log-periodogram of residuals

Amytrophic Lateral Sclerosis). Applying the approach described here to all available data confirms these findings. Boxplots of estimated values of β_f (Fig. 1.16) show a tendency for β_f to be closer to zero for the Parkinson patients. It should be noted, however, that the results may depend on the way tuning constants in wavelets thresholding were chosen. In view of the presence of long memory in the residuals, a detailed study of wavelet-based trend estimation under long-range dependence is needed. This will be discussed in more detail in Chap. 7 (Sect. 7.5).

A different kind of nonstationarity is typical for financial time series. Figure 1.17(a) shows daily values of the DAX index between 3 January 2000 and 12 September 2011. The series is nonstationary, but the first difference looks stationary (Fig. 1.17(b)), and the increments are uncorrelated (Fig. 1.17(c)). In this sense, the data resemble a random walk. However, there is an essential difference. Consider, as a measure of instantaneous volatility, the transformed series $Y_t = |\log X_t - \log X_{t-1}|^{\frac{1}{4}}$ (see Ding and Granger 1996; Beran and Ocker 1999). Figure 1.17(d) shows that there is a trend in the volatility series Y_t. Moreover, even after removing the trend, the series exhibits very slowly decaying correlations and a clearly negative slope in the log-log-periodogram plot (Figs. 1.17(e) and (f)). This is very much in contrast to usual random walk.

A completely different application where a trend and long memory are present is displayed in Figs. 1.18(a) through (d). These data were provided to us by Giovanni

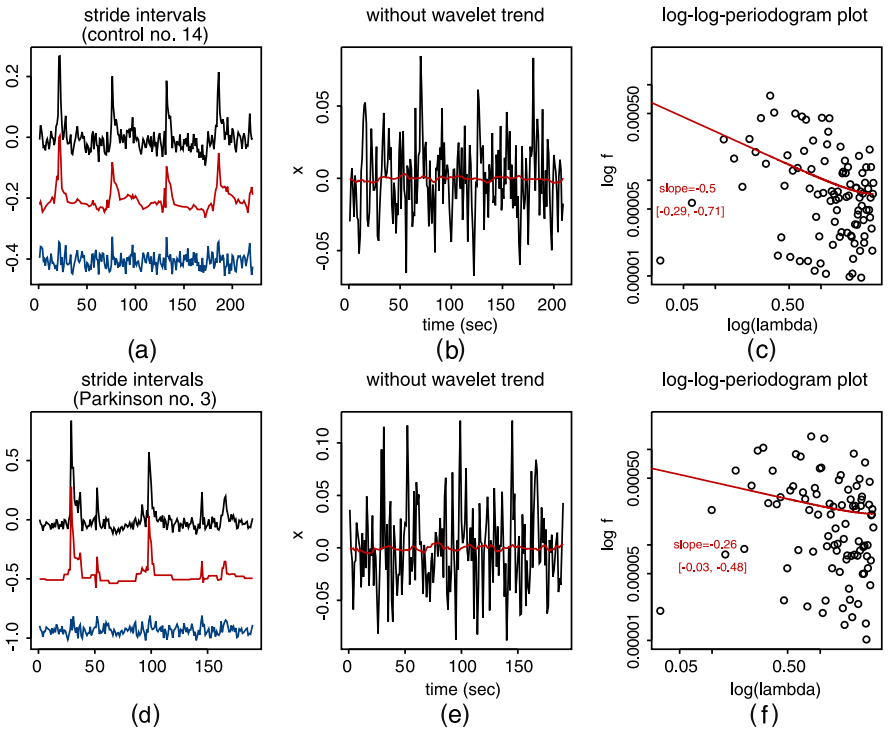

Fig. 1.15 Consecutive stride intervals for (**a**) a healthy individual and (**d**) a patient with Parkinson's disease. The original data are plotted *on top*, the trend functions fitted by minimax wavelet thresholding are given in the *middle*, and the series at the *bottom* correspond to the residuals. The residuals are also plotted separately in (**b**) and (**e**), the corresponding log-log-periodograms in Figs. (**c**) and (**f**) respectively

et al. (Department of Biology, University of Konstanz) and are part of a long-term project on olfactory coding in insects (see, Joerges et al. 1997; Galán et al. 2006; Galizia and Menzel 2001). The original observations consisted of optical measurements of calcium concentration in the antennal lobe of a honey bee. It is known that stimuli (odors) lead to characteristic activity patterns across spherical functional units, the so-called glomeruli, which collect the converging axonal input from a uniform family of receptor cells. It is therefore expected that, compared to a steady state, the between-glomeruli-variability of calcium concentration is higher during a response to an odor. This is illustrated in Fig. 1.18(a). For each time point t (with time rescaled to the interval [0, 1]), an empirical entropy measure X_t was calculated based on the observed distribution of calcium concentration across the glomeruli. The odor was administered at the 30th of $n = 100$ time points. The same procedure was carried out under two different conditions, namely without and with adding a neurotransmitter. The research hypothesis is that adding the neurotransmitter enhances the reaction, in the sense that the initial relative increase of the entropy curve is faster. Because of the known intervention point t_0 and the specific shape of a typ-

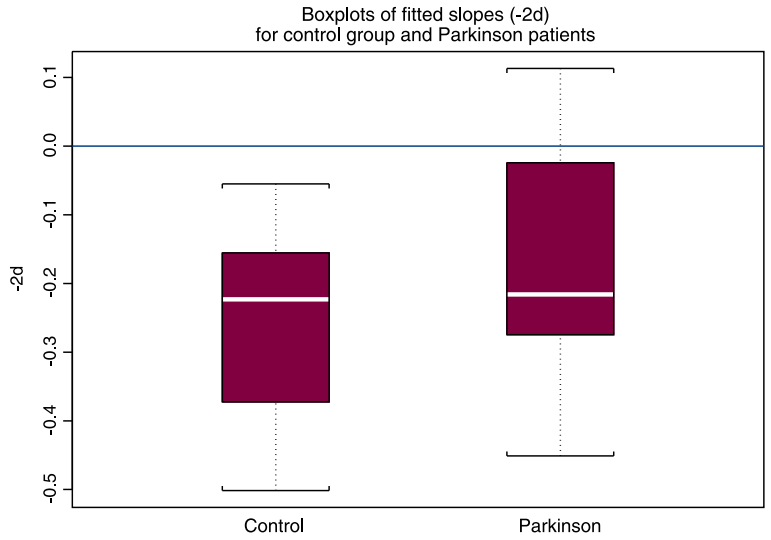

Fig. 1.16 Boxplots of slopes in the log-log-periodogram plot for the control group (*left*) and for a group of patients suffering from Parkinson's disease (*right*)

ical response curve, a good fit can be obtained by a linear spline function with one fixed knot η_0 at t_0 and two subsequent free knots η_1, $\eta_2 > t_0$. The quantity to compare (between the measurements "without" and "with" neurotransmitter) is the slope β belonging to the truncated variable $(t - \eta_0)_+$. The distribution of the least squares estimate of β depends on the dependence structure of the residual process. For the bee considered in Fig. 1.18, the residuals exhibit clear long memory in the first case (no neurotransmitter), whereas long memory is not significant in the second case. For the collection of bees considered in this experiment, long memory, short memory and antipersistence could be observed. How to calculate confidence intervals for β and other parameters in this model will be discussed in Chap. 7 (Sect. 7.3).

An example of spatial long memory is shown in Fig. 1.19. The data in (a) correspond to differences between the maximal and minimal total column ozone amounts within the period from 1 to 7 January 2006, measured on a grid with a resolution of 0.25 degrees in latitude and longitude. The measurements were obtained by the Ozone Monitoring Instrument (OMI) on the Aura 28 spacecraft (Collection 3 OMI data; for details on the physical theory used in assessing ozone amounts, see e.g. Vasilkov et al. 2008; Ahmad et al. 2004; data source: NASA's Ozone Processing Team, http://toms.gsfc.nasa.gov). Figures 1.19(c) and (d) display values of the periodograms in log-log-coordinates when looking in the horizontal (East–West) and vertical direction (North–South) of the grid respectively. Both plots indicate long-range dependence. The solid lines were obtained by fitting a fractional ARIMA lattice process (see Chap. 9, Sects. 9.2 and 9.3). This is a simple model that allows for different long-range, short-range and antipersistent dependence structures in the horizontal and vertical direction. A formal test confirms indeed that long-

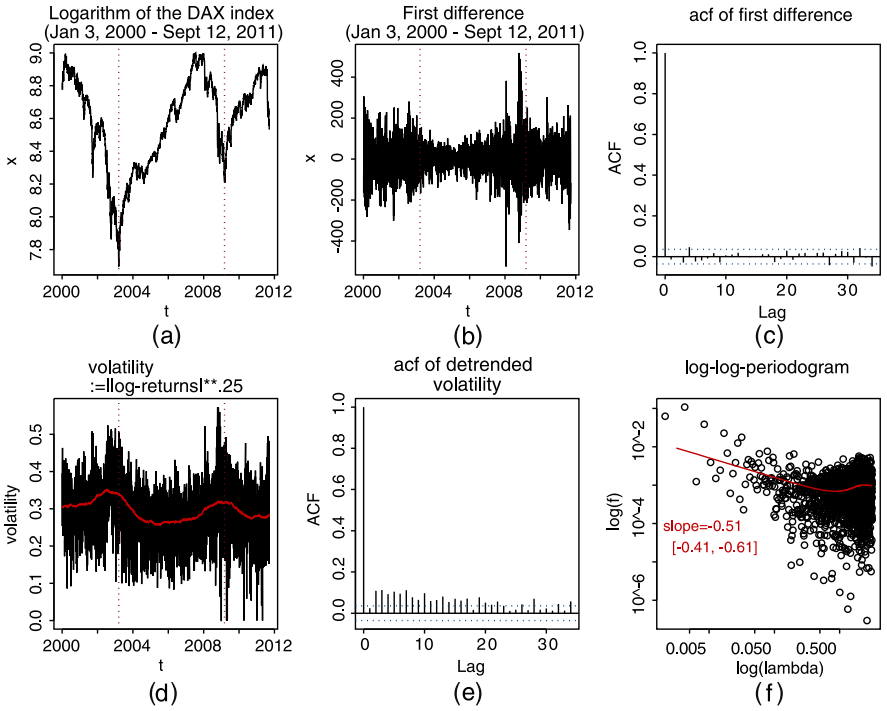

Fig. 1.17 Daily values of the DAX index between 3 January 2000 and 12 September 2011: (**a**) logarithm of original series; (**b**) differenced series (log-returns); (**c**) acf of the series in (**b**); (**d**) $Y_t = |\log X_t - \log X_{t-1}|^{\frac{1}{4}}$ together with a fitted nonparametric trend function; (**e**) acf of Y_t after detrending; (**f**) log-log-periodogram of Y_t after detrending

range dependence in the North–South direction is stronger than along East–West transects.

1.3 Definition of Different Types of Memory

1.3.1 Second-Order Definitions for Stationary Processes

Consider a second-order stationary process X_t ($t \in \mathbb{Z}$) with autocovariance function $\gamma_X(k)$ ($k \in \mathbb{Z}$) and spectral density $f_X(\lambda) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_X(k) \exp(-ik\lambda)$ ($\lambda \in [-\pi, \pi]$). A heuristic definition of *linear* long-range dependence, short-range dependence and antipersistence is given as follows: X_t has (a) long memory, (b) short memory or (c) antipersistence if, as $|\lambda| \to 0$, $f_X(\lambda)$ (a) diverges to infinity, (b) converges to a finite constant, or (c) converges to zero respectively. Since $2\pi f_X(\lambda) = \sum \gamma_X(k)$, this is essentially (in a sense specified more precisely below) equivalent to (a) $\sum \gamma_X(k) = \infty$, (b) $0 < \sum \gamma_X(k) < \infty$ and (c) $\sum \gamma_X(k) = 0$.

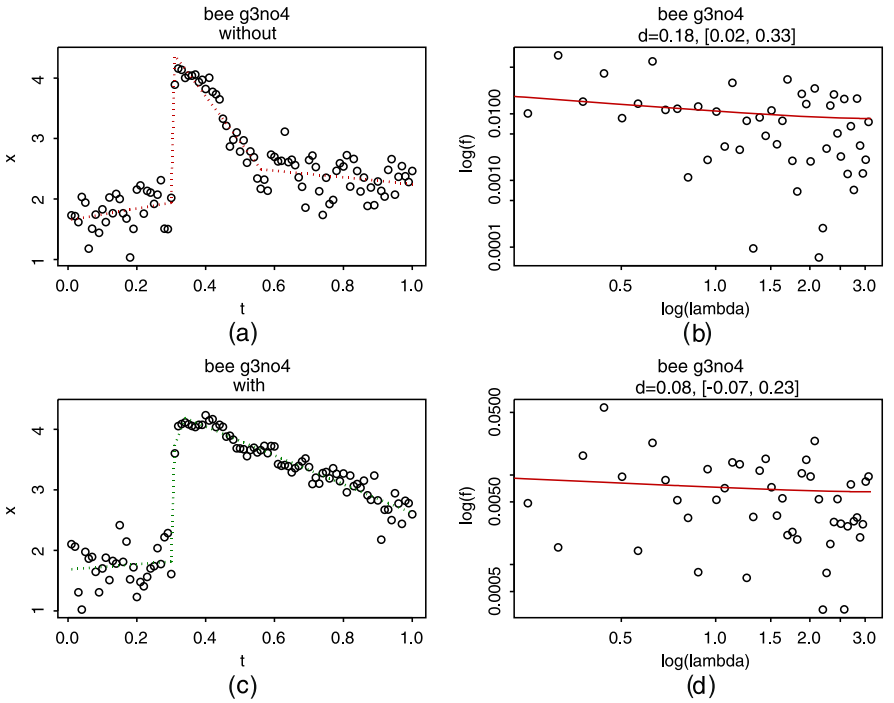

Fig. 1.18 Empirical entropy of calcium concentrations in the antennal lobe of a honey bee exposed to hexanol: (**a**) original series without neurotransmitter and linear splines fit; (**b**) log-log-periodogram of residuals; (**c**) original series with neurotransmitter and linear splines fit; (**d**) log-log-periodogram of residuals

In the following more formal definitions will be given. First, the notion of slowly varying functions is needed (Karamata 1930a, 1930b, 1933; Bajšanski and Karamata 1968/1969; Zygmund 1968; also see e.g. Seneta 1976; Bingham et al. 1989; Sedletskii 2000). Here and throughout the book, the notation $a_n \sim b_n$ $(n \to \infty)$ for two real- or complex-valued sequences a_n, b_n will mean that the ratio a_n/b_n converges to one. Similarly for functions, $g(x) \sim h(x)$ $(x \to x_0)$ will mean that $g(x)/h(x)$ converges to one as x tends to x_0.

First, we need to define so-called slowly varying functions. There are two slightly different standard definitions by Karamata and Zygmund respectively.

Definition 1.1 A function $L : (c, \infty) \to \mathbb{R}$ $(c \geq 0)$ is called slowly varying at infinity in Karamata's sense if it is positive (and measurable) for x large enough and, for any $u > 0$,

$$L(ux) \sim L(x) \quad (x \to \infty).$$

The function is called slowly varying at infinity in Zygmund's sense if for x large enough, it is positive and for any $\delta > 0$, there exists a finite number $x_0(\delta) > 0$ such

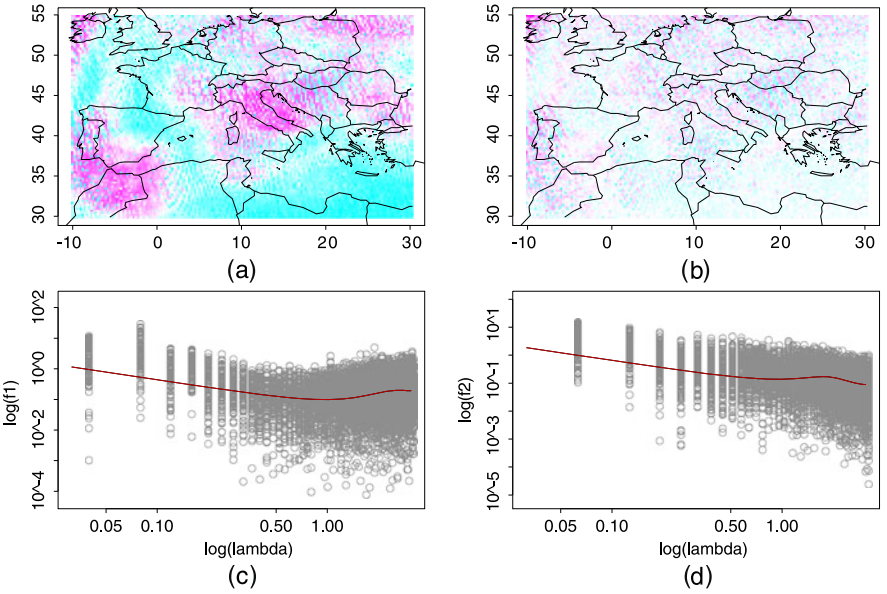

Fig. 1.19 Daily total column ozone amounts from the Ozone Monitoring Instrument (OMI) on the Aura 28 spacecraft: (**a**) maximum minus minimum of observed ozone levels measured between 1–7 January 2006, plotted on a grid with a resolution of 0.25 degrees in latitude and longitude; (**b**) residuals after fitting a FARIMA lattice model; (**c**) and (**d**) log-log-periodogram of the data in (**a**) in the horizontal and vertical directions respectively

that for $x > x_0(\delta)$, both functions $p_1(x) = x^{\delta} L(x)$ and $p_2(x) = x^{-\delta} L(x)$ are monotone.

Similarly, L is called slowly varying at the origin if $\tilde{L}(x) = L(x^{-1})$ is slowly varying at infinity.

A standard formal definition of different types of *linear* dependence structures is given as follows.

Definition 1.2 Let X_t be a second-order stationary process with autocovariance function $\gamma_X(k)$ $(k \in \mathbb{Z})$ and spectral density

$$f_X(\lambda) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_X(k) \exp(-ik\lambda) \quad \big(\lambda \in [-\pi, \pi]\big).$$

Then X_t is said to exhibit (linear) (a) long-range dependence, (b) intermediate dependence, (c) short-range dependence, or (d) antipersistence if

$$f_X(\lambda) = L_f(\lambda) |\lambda|^{-2d},$$

where $L_f(\lambda) \geq 0$ is a symmetric function that is slowly varying at zero, and (a) $d \in (0, \frac{1}{2})$, (b) $d = 0$ and $\lim_{\lambda \to 0} L_f(\lambda) = \infty$, (c) $d = 0$ and $\lim_{\lambda \to 0} L_f(\lambda) = c_f \in (0, \infty)$, and (d) $d \in (-\frac{1}{2}, 0)$ respectively.

Note that the terminology "short-range dependence" (with $d = 0$) is reserved for the case where $L_f(\lambda)$ converges to a finite constant c_f. The reason is that if $L_f(\lambda)$ diverges to infinity, then the autocovariances are not summable although $d = 0$. This case resembles long-range dependence, though with a slower rate of divergence. For a discussion of models with "intermediate" dependence, see for instance Granger and Ding (1996). In principle, any of the usual notions of "slowly varying" may be used in the definition of L_f. The most common ones are the definitions by Karamata and Zygmund given above. The two theorems below show that Karamata's definition is more general. First, we need the definition of regularly varying functions and two auxiliary results.

Definition 1.3 A measurable function $g : \mathbb{R}_+ \to \mathbb{R}$ is called regularly varying (at infinity) with exponent α if $g(x) \neq 0$ for large x and, for any $u > 0$,

$$\lim_{x \to \infty} \frac{g(ux)}{g(x)} = u^\alpha.$$

The class of such functions is denoted by $\mathrm{Re}(\alpha)$.

Similarly, a function g is called regularly varying at the origin with exponent α if $\tilde{g}(x) = g(x^{-1}) \in \mathrm{Re}(-\alpha)$. We will denote this class by $\mathrm{Re}_0(\alpha)$.

Slowly varying functions are regularly varying functions with $\alpha = 0$. For regularly varying functions, integration leads to the following asymptotic behaviour.

Lemma 1.1 *Let $g \in \mathrm{Re}(\alpha)$ with $\alpha > -1$ and integrable on $(0, a)$ for any $a > 0$. Then $\int_0^x g(t)\,dt \in \mathrm{Re}(\alpha + 1)$, and*

$$\int_0^x g(t)\,dt \sim \frac{x g(x)}{\alpha + 1} \quad (x \to \infty).$$

Note that this result is just a generalization of the integration of a power $x^{-\alpha}$, where we have the exact equality $\int_0^x t^{-\alpha}\,dt = x^{1-\alpha}/(\alpha + 1)$. Lemma 1.1 is not only useful for proving the theorem below, but also because asymptotic calculations of variances of sample means can usually be reduced to approximations of integrals by Riemann sums. An analogous result holds for $\alpha < -1$:

Lemma 1.2 *Let $g \in \mathrm{Re}(\alpha)$ with $\alpha < -1$ and integrable on (a, b) for any $0 < a \leq b < \infty$. Then $\int_x^\infty g(t)\,dt \in \mathrm{Re}(\alpha + 1)$, and*

$$\int_x^\infty g(t)\,dt \sim -\frac{x g(x)}{\alpha + 1} \quad (x \to \infty).$$

Now it can be shown that slowly varying functions in Karamata's sense can be characterized as follows.

Theorem 1.1 *L is slowly varying at infinity in Karamata's sense if and only if*

$$L(x) = c(x) \exp\left\{ \int_1^x \frac{\eta(t)}{t}\, dt \right\} \quad (x \geq 1),$$

where $c(\cdot)$ and $\eta(\cdot)$ are measurable functions such that

$$\lim_{x\to\infty} c(x) = c \in (0, \infty),$$

$$\lim_{x\to\infty} \eta(x) = 0$$

and $\eta(\cdot)$ is locally integrable.

Proof First, we show that the representation above yields a slowly varying function. Let $s > 0$, $s \in [a, b]$, and write

$$\psi_s(x) := \frac{L(sx)}{L(x)} = \frac{c(sx)}{c(x)} \exp\left(\int_x^{sx} \frac{\eta(t)}{t}\, dt \right).$$

Since $c(x) \to c$ and $\eta(t) \to 0$, we have for sufficiently large x, and arbitrary $\varepsilon > 0$,

$$(1-\varepsilon) \exp\bigl(-\varepsilon \max\bigl(|\log a|, |\log b|\bigr)\bigr) \leq \psi_s(x) \leq (1+\varepsilon) \exp\bigl(\varepsilon \max\bigl(|\log a|, |\log b|\bigr)\bigr).$$

Letting $\varepsilon \to 0$, we obtain the slowly varying property.

Assume now that L is slowly varying. Define

$$\tilde{\eta}(s) := \frac{sL(s)}{\int_0^s L(t)\, dt}.$$

Then with $U(s) = \int_0^s L(t)\, dt$,

$$\int_1^x \frac{\tilde{\eta}(s)}{s}\, ds = \int_1^x \frac{L(s)}{U(s)}\, ds = \int u^{-1}\, du = \log\bigl(cU(x)\bigr),$$

where the last integration is over ($c = \int_0^1 L(t)\, dt$, $U(x) = \int_0^x L(t)\, dt$). Thus,

$$U(x) = c \exp\left(\int_1^x \frac{\tilde{\eta}(t)}{t}\, dt \right),$$

and consequently, taking derivatives on both sides of the latter expression, we have

$$L(x) = c \frac{\tilde{\eta}(x)}{x} \exp\left(\int_1^x \frac{\tilde{\eta}(t)}{t}\, dt \right) = c\tilde{\eta}(x) \exp\left(\int_1^x \frac{\tilde{\eta}(t) - 1}{t}\, dt \right).$$

Thus, L has the required representation. It remains to show that $\eta(x) = \tilde{\eta}(x) - 1 \to 0$ and $\tilde{\eta}(x) \to 1$. This follows directly from Karamata's theorem (Lemma 1.1) and the definition of $\tilde{\eta}(x)$. □

On the other hand, for Zygmund's definition one can show the following:

Theorem 1.2 *L is slowly varying in Zygmund's sense if and only if there is an $x_0 \in [1, \infty)$ such that*

$$L(x) = c \exp\left\{ \int_1^x \frac{\eta(t)}{t}\, dt \right\} \quad (x \geq x_0),$$

where c is a finite positive constant, and $\eta(\cdot)$ is a measurable function such that $\lim_{x\to\infty} \eta(x) = 0$.

In terms of regularly varying functions the definition of long-range dependence and antipersistence can be rephrased as follows: long memory and antipersistence means that $f \in \mathrm{Re}_0(-2d)$ with $d \in (0, \frac{1}{2})$ and $d \in (-\frac{1}{2}, 0)$ respectively. Since slowly varying functions are dominated by power functions, $f(\lambda) = L_f(\lambda)|\lambda|^{-2d}$ implies that for $d > 0$, the spectral density has a hyperbolic pole at the origin, whereas it converges to zero for $d < 0$. In contrast, under short-range dependence, $f(\lambda)$ converges to a positive finite constant. Alternative terms for long-range dependence are persistence, long memory or strong dependence. Instead of "(linear) long-range dependence", one also uses the terminology "slowly decaying correlations", "long-range correlations" or "strong correlations". This is justified by the following equivalence between the behaviour of the spectral density at the origin and the asymptotic decay of the autocovariance function (see e.g. Zygmund 1968; Lighthill 1962; Beran 1994a; Samorodnitsky 2006):

Theorem 1.3 *Let $\gamma(k)$ $(k \in \mathbb{Z})$ and $f(\lambda)$ $(\lambda \in [-\pi, \pi])$ be the autocovariance function and spectral density respectively of a second-order stationary process. Then the following holds*:

(i) *If*

$$\gamma(k) = L_\gamma(k)|k|^{2d-1},$$

where $L_\gamma(k)$ is slowly varying at infinity in Zygmund's sense, and either $d \in (0, \frac{1}{2})$, or $d \in (-\frac{1}{2}, 0)$ and $\sum_{k\in\mathbb{Z}} \gamma(k) = 0$, then

$$f(\lambda) \sim L_f(\lambda)|\lambda|^{-2d} \quad (\lambda \to 0)$$

with

$$L_f(\lambda) = L_\gamma\big(\lambda^{-1}\big)\pi^{-1}\Gamma(2d)\sin\left(\frac{\pi}{2} - \pi d\right). \tag{1.1}$$

(ii) *If*

$$f(\lambda) = L_f(\lambda)|\lambda|^{-2d} \quad (0 < \lambda < \pi),$$

where $d \in (-\frac{1}{2}, 0) \cup (0, \frac{1}{2})$, *and* $L_f(\lambda)$ *is slowly varying at the origin in Zygmund's sense and of bounded variation on* (a, π) *for any* $a > 0$, *then*

$$\gamma(k) \sim L_\gamma(k)|k|^{2d-1} \quad (k \to \infty),$$

where

$$L_\gamma(k) = 2L_f\left(k^{-1}\right)\Gamma(1-2d)\sin \pi d. \tag{1.2}$$

Note that in the case of antipersistence the autocovariances are absolutely summable but $|\gamma(k)|$ still converges at a hyperbolic rate that can be rather slow, compared for instance with an exponential decay. Also note that $d = 0$ is not included in the theorem because the condition $\gamma(k) = L_\gamma(k)|k|^{-1}$ would imply that $\gamma(k)$ is not summable. In principle (possibly under additional regularity conditions), this would correspond to intermediate dependence with $f(\lambda)$ diverging at the origin like a slowly varying function (see Definition 1.2). To obtain short-range dependence in the sense of Definition 1.2, the summability of $\gamma(k)$ is a minimal requirement. For instance, an exponential decay defined by $|\gamma(k)| \le ca^k$ (with $0 < c < \infty, 0 < a < 1$) together with $\sum_{k\in\mathbb{Z}} \gamma(k) = c_f > 0$ implies $f(\lambda) \sim c_f$ as $\lambda \to 0$. A general statement including all four types of dependence structures can be made however with respect to the sum of the autocovariances:

Corollary 1.1 *If*

$$f(\lambda) = L_f(\lambda)|\lambda|^{-2d} \quad (0 < \lambda < \pi),$$

where $d \in (-\frac{1}{2}, \frac{1}{2})$, *and* $L_f(\lambda) = L(\lambda^{-1})$ *is slowly varying at the origin in Zygmund's sense and of bounded variation on* (a, π) *for any* $a > 0$, *then the following holds. For* $-\frac{1}{2} < d < 0$,

$$\sum_{k=-\infty}^{\infty} \gamma(k) = 2\pi f(0) = 0,$$

whereas for $0 < d < \frac{1}{2}$,

$$\sum_{k=-\infty}^{\infty} \gamma(k) = 2\pi \lim_{\lambda\to 0} f(\lambda) = \infty.$$

Moreover, for $d = 0$, *we have*

$$0 < \sum_{k=-\infty}^{\infty} \gamma(k) = 2\pi f(0) = 2\pi c_f < \infty$$

if $0 < \lim_{\lambda\to 0} L_f(\lambda) = c_f < \infty$ *and*

$$\sum_{k=-\infty}^{\infty} \gamma(k) = 2\pi \lim_{\lambda\to 0} f(\lambda) = \infty$$

if $\lim_{\lambda\to 0} L_f(\lambda) = \infty$.

From these results one can see that characterizing linear dependence by the spectral density is more elegant than via the autocovariance function because the equation $f(\lambda) = L_f(\lambda)|\lambda|^{-2d}$ is applicable in all four cases (long-range, short-range, intermediate dependence and antipersistence).

Example 1.1 Let X_t be second-order stationary with Wold decomposition

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j},$$

where ε_t are uncorrelated zero mean random variables, $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t) < \infty$, and

$$a_j = (-1)^j \binom{-d}{j} = (-1)^j \frac{\Gamma(1-d)}{\Gamma(j+1)\Gamma(1-d-j)}$$

with $-1/2 < d < 1/2$. Then a_j are the coefficients in the power series representation

$$A(z) = (1-z)^{-d} = \sum_{j=0}^{\infty} a_j z^j.$$

Therefore, the spectral density of X_t is given by

$$\begin{aligned} f_X(\lambda) &= \frac{\sigma_\varepsilon^2}{2\pi}\left|A\left(e^{-i\lambda}\right)\right|^2 = \frac{\sigma_\varepsilon^2}{2\pi}\left|1-e^{-i\lambda}\right|^{-2d} = \frac{\sigma_\varepsilon^2}{2\pi}\left|2(1-\cos\lambda)\right|^{-d} \\ &\sim \frac{\sigma_\varepsilon^2}{2\pi}|\lambda|^{-2d} \quad (\lambda \to 0). \end{aligned}$$

Thus, we obtain short-range dependence for $d = 0$ (and in fact uncorrelated observations), antipersistence for $-\frac{1}{2} < d < 0$ and long-range dependence for $0 < d < \frac{1}{2}$. If the innovations ε_t are independent, then X_t is called a fractional ARIMA(0, d, 0) process (Granger and Joyeux 1980; Hosking 1981; see Chap. 2, Sect. 2.1.1.4).

Example 1.2 Let X_t be second-order stationary with spectral density

$$f_X(\lambda) = \log\left|\frac{\pi}{\lambda}\right| = L_f(\lambda).$$

This is a case with intermediate dependence. The autocovariance function is given by

$$\operatorname{var}(X_t) = \gamma_X(0) = 2\left(\pi \log \pi - \int_0^\pi \log \lambda \, d\lambda\right) = 2\pi,$$

and for $k > 0$,

$$\begin{aligned}\gamma_X(k) &= 2\int_0^\pi \cos k\lambda \cdot (\log \pi - \log \lambda)\, d\lambda = -2\int_0^\pi \cos k\lambda \cdot \log \lambda \, d\lambda \\ &= \frac{2}{k}\int_0^\pi \frac{\sin k\lambda}{\lambda}\, d\lambda = \frac{2}{k} Si(\pi k),\end{aligned}$$

where $Si(\cdot)$ is the sine integral function. For $k \to \infty$, we obtain the Dirichlet integral

$$\lim_{k\to\infty} Si(\pi k) = \int_0^\infty \frac{\sin \lambda}{\lambda}\, d\lambda = \frac{\pi}{2},$$

so that

$$\gamma_X(k) \sim \pi k^{-1} \quad (k \to \infty),$$

$$\rho_X(k) \sim \frac{1}{2} k^{-1} \quad (k \to \infty),$$

and

$$\sum_{k=-(n-1)}^{n-1} \gamma_X(k) \sim 2\pi \log n \quad (n \to \infty).$$

The behaviour of the spectral density at the origin also leads to a simple universal formula for the variance of the sample mean $\bar{x} = n^{-1} \sum_{t=1}^n X_t$:

Corollary 1.2 *Suppose that* $f(\lambda) \sim L_f(\lambda)|\lambda|^{-2d}$ $(\lambda \to 0)$ *for some* $d \in (-\frac{1}{2}, \frac{1}{2})$, *where* $L_f(\lambda) = L(\lambda^{-1})$ *is slowly varying at zero in Zygmund's sense and of bounded variation on* (a, π) *for any* $a > 0$. *Furthermore, assume that in the case of* $d = 0$ *the slowly varying function* L_f *is continuous at the origin. Then*

$$\operatorname{var}(\bar{x}) \sim v(d) f\left(n^{-1}\right) n^{-1} \quad (n \to \infty)$$

with

$$v(d) = \frac{2\Gamma(1-2d)\sin(\pi d)}{d(2d+1)} \quad (d \neq 0)$$

and

$$v(0) = \lim_{d\to 0} v(d) = 2\pi.$$

Proof We have

$$\begin{aligned}\operatorname{var}(\bar{x}) &= n^{-1}\sum_{k=-(n-1)}^{n-1}\left(1-\frac{|k|}{n}\right)\gamma(k)\\ &= n^{-1}\sum_{k=-(n-1)}^{n-1}\gamma(k) - n^{-1}\sum_{k=-(n-1)}^{n-1}\frac{|k|}{n}\gamma(k)\end{aligned}$$

with

$$\gamma(k) \sim L_\gamma(k)|k|^{2d-1}.$$

For $0 < d < \frac{1}{2}$, this implies

$$\begin{aligned}\operatorname{var}(\bar{x}) &\sim 2L_\gamma(n)n^{-1}\left[\sum_{k=1}^{n-1}k^{2d-1} - n^{-1}\sum_{k=1}^{n-1}k^{2d}\right]\\ &= 2L_\gamma(n)n^{2d-1}\left[\sum_{k=1}^{n-1}\left(\frac{k}{n}\right)^{2d-1}n^{-1} - \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)^{2d}n^{-1}\right]\\ &\sim 2L_\gamma(n)n^{2d-1}\left[\int_0^1 x^{2d-1}\,dx - \int_0^1 x^{2d}\,dx\right]\\ &= 2L_\gamma(n)n^{2d-1}\left[\frac{1}{2d} - \frac{1}{2d+1}\right] = \frac{L_\gamma(n)n^{2d-1}}{d(2d+1)}.\end{aligned}$$

Using Theorem 1.3, we can write this as

$$\frac{L_\gamma(n)n^{2d-1}}{d(2d+1)} = \frac{2\Gamma(1-2d)\sin(\pi d)}{d(2d+1)}L_f\left(n^{-1}\right)n^{2d-1} = \nu(d)L_f\left(n^{-1}\right)n^{2d-1}.$$

Thus,

$$\operatorname{var}(\bar{x}) \sim \nu(d)L_f\left(n^{-1}\right)n^{2d-1} \sim \nu(d)f\left(n^{-1}\right)n^{-1}.$$

For $d = 0$ and $0 < L_f(0) = c_f < \infty$, we have

$$0 < \sum_{k=-\infty}^{\infty}\gamma(k) = 2\pi f(0) < \infty,$$

so that $|k|\gamma(k)$ is Cesaro summable with limit zero. Hence,

$$\lim_{n\to\infty} n^{-1}\sum_{k=-(n-1)}^{n-1}\frac{|k|}{n}\gamma(k) = 0$$

and

$$\operatorname{var}(\bar{x}) \sim n^{-1} \sum_{k=-(n-1)}^{n-1} \gamma(k) \sim 2\pi f(0) n^{-1}.$$

Thus, we may write

$$\operatorname{var}(\bar{x}) \sim \nu(0) L_f(0) n^{-1} \sim \nu(0) f\left(n^{-1}\right) n^{-1},$$

where

$$\nu(0) = \lim_{d \to 0} \nu(d) = \lim_{d \to 0} \frac{2\sin(\pi d)}{d} = 2\pi.$$

Finally, for $-\frac{1}{2} < d < 0$, we have $\sum_{k \in \mathbb{Z}} \gamma(k) = 0$, so that

$$\begin{aligned}
\operatorname{var}(\bar{x}) &= n^{-1} \sum_{k=-(n-1)}^{n-1} \gamma(k) - n^{-1} \sum_{k=-(n-1)}^{n-1} \frac{|k|}{n} \gamma(k) \\
&= -2n^{-1} \sum_{k=n}^{\infty} \gamma(k) - n^{-1} \sum_{k=-(n-1)}^{n-1} \frac{|k|}{n} \gamma(k) \\
&\sim 2L_\gamma(n) n^{-1} \left[-\sum_{k=n}^{\infty} k^{2d-1} - n^{-1} \sum_{k=1}^{n-1} k^{2d} \right] \\
&= 2L_\gamma(n) n^{2d-1} \left[-\sum_{k=n}^{\infty} \left(\frac{k}{n}\right)^{2d-1} n^{-1} - \sum_{k=1}^{n-1} \left(\frac{k}{n}\right)^{2d} n^{-1} \right] \\
&\sim 2L_\gamma(n) n^{2d-1} \left[-\int_1^{\infty} x^{2d-1}\, dx - \int_0^1 x^{2d}\, dx \right] \\
&= 2L_\gamma(n) n^{2d-1} \left[\frac{1}{2d} - \frac{1}{2d+1} \right] = \nu(d) L_f\left(n^{-1}\right) n^{2d-1} \\
&\sim \nu(d) f\left(n^{-1}\right) n^{-1}.
\end{aligned}$$

□

Corollary 1.2 illustrates that knowledge about the value of d is essential for statistical inference. If short memory is assumed but the actual value of d is larger than zero, then confidence intervals for $\mu = E(X_t)$ will be too narrow by an increasing factor of n^d, and the asymptotic level of tests based on this assumption will be zero. This effect is not negligible even for small sample sizes. Table 1.1 shows simulated rejection probabilities (based on 1000 simulations) for the t-test at the nominal 5 %-level of significance. The numbers are based on 1000 simulations of a fractional ARIMA(0, d, 0) process with $d = 0.1$, 0.2, 0.3 and 0.4 respectively (see Chap. 2, Sect. 2.1.1.4, for the definition of FARIMA models).

The second-order definitions of long-range dependence considered here can be extended to random fields with a multivariate index t. A complication that needs

Table 1.1 Simulated rejection probabilities (under the null hypothesis) for the t-test at the nominal 5 %-level of significance. The results are based on 1000 simulations of a fractional ARIMA(0, d, 0) process with $d = 0.1$, 0.2, 0.3 and 0.4 respectively

n	$d = 0.1$	0.2	0.3	0.4
10	0.10	0.21	0.33	0.53
50	0.16	0.38	0.55	0.72
100	0.20	0.42	0.62	0.78

to be addressed for two- or higher-dimensional indices is however that dependence may not be isotropic (see e.g. Boissy et al. 2005; Lavancier 2006, 2007; Beran et al. 2009). This will be discussed in Chap. 9. A further important extension includes multivariate spectra with power law behaviour at the origin that may differ for the different components of the process (see e.g. Robinson 2008).

1.3.2 Volatility Dependence

The characterization of nonlinear long memory is more complicated in general since there are many ways in which nonlinearity can occur. In econometric applications, the main focus is on dependence in volatility in the sense that X_t are uncorrelated but the squares X_t^2 are correlated. The definitions of long memory given above can then be carried over directly by simply considering X_t^2 instead of X_t. A more difficult, and partially still open, issue is how to define concrete statistically convenient models that are stationary with existing fourth moments and long-range correlations in X_t^2 (see e.g. Robinson 1991; Bollerslev and Mikkelsen 1996; Baillie et al. 1996a; Ding and Granger 1996; Beran and Ocker 2001; Giraitis et al. 2000a, 2004, 2006; Giraitis and Surgailis 2002). This is discussed in detail in Sect. 2.1.3. A very simple model that is well defined and obviously exhibits long-range dependence can be formulated as follows.

Proposition 1.1 *Let ε_t ($t \in \mathbb{Z}$) be i.i.d. random variables with $E(\varepsilon_t) = 0$ and* $\mathrm{var}(\varepsilon_t) = 1$. *Define*

$$X_t = \sigma_t \varepsilon_t$$

with $\sigma_t = \sqrt{v_t}$, $v_t \geq 0$ independent of ε_s ($s \in \mathbb{Z}$) and such that

$$\gamma_v(k) = cov(v_t, v_{t+k}) \sim c \cdot |k|^{2d-1}$$

for some $0 < d < \frac{1}{2}$. Then for $k \neq 0$,

$$\gamma_X(k) = 0,$$

whereas

$$\gamma_{X^2}(k) = cov\big(X_t^2, X_{t+k}^2\big) = \gamma_v(k) \sim c \cdot |k|^{2d-1} \quad (k \to \infty).$$

Proof Since $E(X_t) = E(\sigma_t)E(\varepsilon_t) = 0$, we have for $k \neq 0$,

$$\gamma_X(k) = E(X_t X_{t+k}) = E(\sigma_t \sigma_{t+k})E(\varepsilon_t \varepsilon_{t+k}) = 0.$$

Moreover, for $k \neq 0$,

$$\begin{aligned}\gamma_{X^2}(k) &= E\big(\sigma_t^2 \sigma_{t+k}^2\big)E\big(\varepsilon_t^2 \varepsilon_{t+k}^2\big) - E\big(\sigma_t^2 \varepsilon_t^2\big)E\big(\sigma_{t+k}^2 \varepsilon_{t+k}^2\big) \\ &= E(v_t v_{t+k}) - E(v_t)E(v_{t+k}) \\ &= \gamma_v(k) \sim c \cdot |k|^{2d-1} \quad (k \to \infty).\end{aligned}$$

□

The main problem with this model is that σ_t and ε_t are not directly observable. One would however like to be able to separate the components σ_t and ε_t even though only their product X_s $(s \leq t)$ is observed. This is convenient, for instance, when setting up maximum likelihood equations for estimating parameters that specify the model (see e.g. Giraitis and Robinson 2001). One therefore often prefers to assume a recursive relation between v_t and past values of X_t. The difficulty that arises then is to prove the existence of a stationary solution and to see what type of volatility dependence is actually achieved. For instance, in the so-called ARCH(∞) model (Robinson 1991; Giraitis et al. 2000a) one assumes

$$\sigma_t^2 = v_t = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2$$

with $b_j \geq 0$ and $\sum b_j < \infty$. As it turns out, however, long-range dependence—defined in the second-order sense as above—cannot be obtained. This and alternative volatility models with long-range dependence will be discussed in Sect. 2.1.3.

1.3.3 Second-Order Definitions for Nonstationary Processes

For nonstationary processes, Heyde and Yang (1997) consider the variance

$$V_m = \operatorname{var}\big(X_t^{(m)}\big) \tag{1.3}$$

of the aggregated process

$$X_t^{(m)} = X_{tm-m+1} + \cdots + X_{tm} \tag{1.4}$$

and the limit

$$V = \lim_{m \to \infty} D_m^{-1} V_m, \tag{1.5}$$

where

$$D_m = \sum_{i=tm-m+1}^{tm} E\big(X_i^2\big). \tag{1.6}$$

The process $X_t^{(m)}$ ($t \in \mathbb{Z}$) is then said to exhibit long memory if $V = \infty$. This definition is applicable both to second-order stationary processes and to processes that need to be differenced first. Note that the block mean variance $m^{-2}V_m$ is also called Allan variance (Allan 1966; Percival 1983; Percival and Guttorp 1994).

1.3.4 Continuous-Time Processes

The definition of long memory and antipersistence based on autocovariances can be directly extended to continuous-time processes.

Definition 1.4 Let $X(t)$ ($t \in \mathbb{R}$) be a stationary process with autocovariance function $\gamma_X(u) = cov(X(t), X(t+u))$ and spectral density $f_X(\lambda)$ ($\lambda \in \mathbb{R}$). Then $X(t)$ is said to have long memory if there is a $d \in (0, \frac{1}{2})$ such that

$$\gamma_X(u) = L_\gamma(u)u^{2d-1}$$

as $u \to \infty$, or

$$f_X(\lambda) = L_f(\lambda)|\lambda|^{-2d}$$

as $\lambda \to 0$, where L_γ and L_f are slowly varying at infinity and zero respectively. Similarly, $X(t)$ is said to be antipersistent if these formulas hold for some $d \in (-\frac{1}{2}, 0)$ and, in case of the formulation via γ_X, the additional condition

$$\int_{-\infty}^{\infty} \gamma_X(u)\,du = 0$$

holds.

Note that, as in discrete time, the definition of long-range dependence given here implies $\int \gamma_X(u)\,du = \infty$. A more general definition is possible by using the conditions $\int \gamma_X(u)\,du = \infty$ and $\int \gamma_X(u)\,du = 0$ only. However, the first condition would then also include the possibility of intermediate dependence.

Finally note that an alternative definition can also be given in terms of the variance of the integrated process $Y(t) = \int_0^t X(s)\,ds$. This is analogous to a nonlinear growth of the variance of partial sums for discrete time processes.

Definition 1.5 Let $Y(t) = \int_0^t X(s)\,ds$ and assume that $\mathrm{var}(Y(t)) < \infty$ for all $t \geq 0$. Then Y (and X) is said to have long-range dependence if

$$\mathrm{var}\big(Y(t)\big) = L(t)t^{2d+1}$$

for some $0 < d < \frac{1}{2}$, where L is slowly varying at infinity. Moreover, Y (and X) is said to be antipersistent if

$$\mathrm{var}\big(Y(t)\big) = L(t)t^{2H} = L(t)t^{2d+1}$$

for some $-\frac{1}{2} < d < 0$, where L is slowly varying at infinity.

This definition means that the growth of the variance of $Y(t)$ is faster than linear under long-range dependence and slower than linear for antipersistent processes. The connection between the two definitions is given by

$$\operatorname{var}\big(Y(t)\big) = \int_0^t \left(\int_0^t \gamma_X(s-r)\,dr \right) ds = 2\int_0^t (t-u)\gamma_X(u)\,du.$$

If $\gamma_X(u) = cov(X(t), X(t+u)) \sim L_\gamma(u)|u|^{2d-1}$, where $d \in (0, \frac{1}{2})$ and L_γ is slowly varying at infinity (i.e. $X(t)$ has long memory in the sense of Definition 1.4), then application of Lemma 1.1 leads to

$$\operatorname{var}\big(Y(t)\big) \sim \frac{1}{d(2d+1)} L_\gamma(t) t^{2d+1}.$$

Thus, $X(t)$ has also long memory in the sense of Definition 1.5. The analogous connection holds for antipersistence, taking into account the additional condition $\int \gamma_X(u)\,du = 0$.

For nonnegative processes, the expected value often grows at a linear rate. Typical examples are counting processes or renewal processes with positive rewards (see Sects. 2.2.4 and 4.9). Long-range dependence and antipersistence can therefore also be expressed by comparing the growth of the variance with the growth of the mean.

Definition 1.6 Let $Y(t) = \int_0^t X(s)\,ds \geq 0$ and assume that $\operatorname{var}(Y(t)) < \infty$ for all $t \geq 0$. Then Y (and X) is said to have long-range dependence if

$$\lim_{t\to\infty} \frac{\operatorname{var}(Y(t))}{E[Y(t)]} = +\infty.$$

Similarly, Y (and X) is said to be antipersistent if

$$\lim_{t\to\infty} \frac{\operatorname{var}(Y(t))}{E[Y(t)]} = 0.$$

1.3.5 Self-similar Processes: Beyond Second-Order Definitions

Another classical way of studying long memory and antipersistence is based on the relationship between dependence and self-similarity.

Definition 1.7 A stochastic process $Y(u)$ $(u \in \mathbb{R})$ is called *self-similar* with self-similarity parameter $0 < H < 1$ (or H-self-similar) if for all $c > 0$, we have

$$\big(Y(cu), u \in \mathbb{R}\big) \stackrel{d}{=} \big(c^H Y(u), u \in \mathbb{R}\big),$$

where $\stackrel{d}{=}$ denotes equality in distribution.

Self-similar processes are a very natural mathematical object to look at because they are the only possible weak limits of appropriately normalized and centered partial sums $S_n(u) = \sum_{t=1}^{[nu]} X_t$ $(u \in [0,1])$ based on stationary and ergodic sequences X_t $(t \in \mathbb{Z})$ (Lamperti 1962, 1972). If a process $Y(u)$ $(u \in \mathbb{R})$ is H-self-similar with stationary increments (so-called H-SSSI), then the discrete-time increment process $X_t = Y(t) - Y(t-1)$ $(t \in \mathbb{Z})$ is stationary. Note also that $Y(0) =_d c^H Y(0)$ for any arbitrarily large $c > 0$, so that necessarily $Y(0) = 0$ almost surely.

To see how the self-similarity parameter H is related to long memory, we first consider a case where the second-order definition of long memory is applicable. If second moments exist, then the SSSI-property implies, for $u \geq v > 0$,

$$\gamma_Y(u,u) = \operatorname{var}\big(Y(u)\big) = u^{2H}\gamma_Y(1,1) = u^{2H}\sigma^2$$

and

$$\operatorname{var}\big(Y(u) - Y(v)\big) = \operatorname{var}\big(Y(u-v)\big) = \sigma^2(u-v)^{2H}.$$

Since $\operatorname{var}(Y(u) - Y(v)) = \gamma_Y(u,u) + \gamma_Y(v,v) - 2\gamma_Y(u,v)$, this means that the autocovariance function is equal to

$$\gamma_Y(u,v) = \frac{\sigma^2}{2}\big[|u|^{2H} + |v|^{2H} - |u-v|^{2H}\big] \quad (u,v \in \mathbb{R}).$$

By similar arguments, the autocovariance function of the increment process X_t $(t \in \mathbb{Z})$ is given by

$$\gamma_X(k) = cov(X_t, X_{t+k}) = \frac{\sigma^2}{2}\big[|k-1|^{2H} + |k+1|^{2H} - 2|k|^{2H}\big] \quad (k \in \mathbb{N}). \quad (1.7)$$

By Taylor expansion in $x = k^{-1}$ around $x = 0$ it follows that, as k tends to infinity,

$$\gamma_X(k) \sim \sigma^2 H(2H-1)k^{2H-2}.$$

In the notation of Definition 1.2 we therefore have $L_\gamma(k) = \sigma^2 H(2H-1)$,

$$H = d + \frac{1}{2},$$

and X_t $(t \in \mathbb{Z})$ has long memory if $\frac{1}{2} < H < 1$. Also note that for the variance of $S_n = \sum_{t=1}^n X_t$, self-similarity implies

$$\operatorname{var}(S_n) = \operatorname{var}\big(Y(n) - Y(0)\big) = n^{2H}\sigma^2,$$

so that, for $H > \frac{1}{2}$, the variance grows at a rate that is faster than linear. For $H = \frac{1}{2}$, all values of $\gamma_X(k)$ are zero except for $k = 0$, so that X_t $(t \in \mathbb{Z})$ is an uncorrelated sequence. For $0 < H < \frac{1}{2}$, $\gamma_X(k)$ is summable, so that, in contrast to the case with

$H > \frac{1}{2}$, the sum over all covariances can be split into three terms,

$$\begin{aligned}
&\sum_{k=-\infty}^{\infty} \left[|k-1|^{2H} + |k+1|^{2H} - 2|k|^{2H}\right] \\
&\quad = \sum_{k=-\infty}^{\infty} |k-1|^{2H} + \sum_{k=-\infty}^{\infty} |k+1|^{2H} - 2\sum_{k=-\infty}^{\infty} |k|^{2H} \\
&\quad = \sum_{k=-\infty}^{\infty} |k|^{2H} + \sum_{k=-\infty}^{\infty} |k|^{2H} - 2\sum_{k=-\infty}^{\infty} |k|^{2H} = 0.
\end{aligned}$$

In other words, $0 < H < \frac{1}{2}$ implies antipersistence. The simplest SSSI process with finite second moments is a Gaussian process, the so-called fractional Brownian motion (fBm), usually denoted by B_H. Note that B_H is the only Gaussian SSSI-process because apart from the variance σ^2, the first two moments are fully specified by the SSSI-property. The corresponding increment sequence X_t ($t \in \mathbb{R}$ or $\mathbb{Z}$) is called fractional Gaussian noise (FGN).

To see how to extend the relationship between the self-similarity parameter H and long-range dependence beyond Gaussian processes, we first look at an explicit time-domain representation of fractional Gaussian motion. The definition and existence of fBm follow directly from the definition of its covariance function. The difference between standard Brownian motion (with $H = \frac{1}{2}$) and fractional Brownian motion with $H \neq \frac{1}{2}$ can be expressed by a moving average representation of $B_H(u)$ on the real line, which is a weighted integral of standard Brownian motion. For $H \neq \frac{1}{2}$, we have

$$B_H(u) = \int_{-\infty}^{\infty} Q_{u,1}(x; H)\, dB(x), \tag{1.8}$$

where

$$Q_{u,1}(x; H) = c_1\left[(u-x)_+^{H-\frac{1}{2}} - (-x)_+^{H-\frac{1}{2}}\right] + c_2\left[(u-x)_-^{H-\frac{1}{2}} - (-x)_-^{H-\frac{1}{2}}\right],$$

and c_1, c_2 are deterministic constants. This representation is not unique since it depends on the choice of c_1 and c_2. A causal representation of fBm is obtained if we choose $c_2 = 0$ and

$$c_1 = \frac{\sqrt{\Gamma(2H+1)\sin(\pi H)}}{\Gamma(H+\frac{1}{2})} = \left\{\int_0^{\infty}\left[(1+s)^{H-\frac{1}{2}} - s^{H-\frac{1}{2}}\right]^2 ds + \frac{1}{2H}\right\}^{-\frac{1}{2}}.$$

One can verify that the kernel $Q_{u,1}(\cdot, H)$ has the following property: for all $0 \le v < u$, $x \in \mathbb{R}$,

$$Q_{u,1}(x; H) - Q_{v,1}(x; H) = Q_{u-v,1}(x - v; H), \tag{1.9}$$

$$Q_{cu,1}(cx; H) = c^{H-1/2} Q_{u,1}(x; H). \tag{1.10}$$

The first property reflects stationarity of increments. The second property leads to self-similarity with self-similarity parameter H. It should be mentioned at this point that representation (1.8) is not valid for an fBm on [0, 1].

As we have seen above, if the second moments are assumed to exist, then the definition of self-similarity fully determines the autocorrelation structure. This leads to a direct definition of Gaussian self-similar processes. The existence and construction of *non-Gaussian* self-similar processes is less straightforward because the autocorrelation structure is not enough. One way of obtaining a large class of non-Gaussian self-similar processes is to extend the integral representation (1.8) to *multiple Wiener–Itô integrals* (see e.g. Major 1981). This can be done as follows. For $q \geq 1$ and $0 < H < 1$, we define the processes

$$Z_{H,q}(u) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Q_{u,q}(x_1, \ldots, x_q; H)\, dB(x_1) \cdots dB(x_q) \tag{1.11}$$

where the kernel $Q_{u,q}$ is given by

$$Q_{u,q}(x_1, \ldots, x_q) = \int_0^u \left(\prod_{i=1}^{q} (s - x_i)_+^{-(\frac{1}{2} + \frac{1-H}{q})} \right) ds.$$

All kernels have the two properties guaranteeing stationarity of increments and self-similarity. The self-similarity property is of the form

$$Q_{cu,q}(cx_1, \ldots, cx_q; H) = c^{H - \frac{q}{2}} Q_{u,q}(x_1, \ldots, x_q; H).$$

The exponent $-q/2$ instead of $-1/2$ is due to the fact that dB occurs q times in the product. More explicitly, we can see that the scaling property of $Q_{u,q}$ implies self-similarity with parameter H as follows:

$$\begin{aligned}
Z_{H,q}(cu) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Q_{ct,q}(x_1, \ldots, x_q; H)\, dB(x_1) \cdots dB(x_q) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Q_{ct,q}\left(c\frac{x_1}{c}, \ldots, c\frac{x_q}{c}; H\right) dB\left(c\frac{x_1}{c}\right) \cdots dB\left(c\frac{x_q}{c}\right) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} c^{H - \frac{q}{2}} Q_{u,q}(y_1, \ldots, y_k; H) c^{\frac{q}{2}} dB(y_1) \cdots dB(y_q) \\
&= c^H \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Q_{u,q}(y_1, \ldots, y_q; H)\, dB(y_1) \cdots dB(y_k) = c^H Z_{H,q}(u).
\end{aligned}$$

For $q > 1$, the process $Z_{H,q}(u)$ $(u \in \mathbb{R})$ is no longer Gaussian and is called *Hermite* process on $\mathbb{R}$. Sometimes one also uses the terminology Hermite–Rosenblatt process, though "Rosenblatt process" originally refers to the case with $q = 2$ only (see Taqqu 1975).

Equation (1.11) also leads to a natural extension to self-similar processes with long memory and nonexisting second moments. This can be done by replacing

Brownian motion by a process whose second moments do not exist. Note that Brownian motion is just a special example of the much larger class of Lévy processes. These are defined by the property that they have stationary independent increments and vanish at zero almost surely. The nonexistence of second moments can be achieved by assuming that the Lévy process is a symmetric α-stable (SαS) process $Z_\alpha(\cdot)$ for some $0 < \alpha < 2$. This means that every linear combination $Y = \sum_{j=1}^{m} c_i Z_\alpha(u_i)$ has a symmetric α-stable distribution with characteristic function $\varphi(\omega) = E[\exp(i\omega Y)] = \exp(-a|\omega|^\alpha)$. In particular, S$\alpha$S Lévy processes are self-similar with self-similarity parameter $H_{Lévy} = 1/\alpha$. Hence, we note that unlike in the Gaussian case of fBm, here self-similarity does not have anything to do with long memory. Furthermore, symmetric α-stable Lévy processes arise as limits of appropriately standardized partial sums $S_{[nu]} = \sum_{i=1}^{[nu]} X_t$, where X_t are i.i.d. and have symmetric heavy tails with tail index α in the sense that

$$\lim_{x\to-\infty} |x|^\alpha P(X < -x) = \lim_{x\to+\infty} x^\alpha P(X > x) = C_1 \tag{1.12}$$

for some $0 < \alpha < 2$ and a suitable constant C_1 (see e.g. Embrechts et al. 1997; Embrechts and Maejima 2002, and Sect. 4.3). In particular, the process $S_{[nu]}$ has to be standardized by $d^{-1}(n)$, where $d(n) = n^{H_{Lévy}} = n^{1/\alpha}$. Therefore, for sequences X_t with tail index $\alpha < 2$, the self-similarity parameter $H = H_{Lévy} = 1/\alpha$ is the analogue to $H = \frac{1}{2}$ in the case of finite second moments. If, on the other hand, a nondegenerate limit of $d^{-1}(n)S_{[nu]}$ is obtained for standardizations $d(n)$ proportional to n^H with $H > 1/\alpha$, then the memory (in the sequence X_t) is so strong that partial sums diverge faster than for Lévy processes. This is analogous to $H > \frac{1}{2}$ in the case of finite second moments. Therefore, long memory is associated with the condition $H > 1/\alpha$. (Note that for $\alpha = 2$, we are back to finite second moments, so that we obtain the previous condition $H > \frac{1}{2}$.) In analogy to the case of finite second moments we may also define the fractional parameter $d = H - 1/\alpha$. Long memory is then associated with $d > 0$. Note also that, since the self-similarity parameter is by definition in the interval $(0, 1)$, long memory cannot be achieved for $\alpha < 1$.

As we will see in Sect. 4.3, in general the limit of $d^{-1}(n)S_{[nu]}$ is a *Linear Fractional stable motion* defined by

$$\tilde{Z}_{H,\alpha}(u) = \int_{-\infty}^{\infty} Q_{u,1}(x; H, \alpha)\, dZ_\alpha(x) \tag{1.13}$$

with

$$Q_{u,1}(x; H, \alpha) = c_1\big[(u-x)_+^{H-1/\alpha} - (-x)_+^{H-1/\alpha}\big] + c_2\big[(u-x)_-^{H-1/\alpha} - (-x)_-^{H-1/\alpha}\big] \tag{1.14}$$

and $H > 1/\alpha$. This definition is obviously analogous to (1.8) for fractional Brownian motion. Moreover, the definition is valid for $H \in (0, 1)$, $H \neq 1/\alpha$.

1.3.6 Other Approaches

1.3.6.1 Different Dependence Measures

For processes with infinite second moments, long-range dependence has to be measured by other means than autocorrelations, the spectral density or the variance of cumulative sums. For instance, the variance V_m defined in (1.3) can be replaced by

$$\hat{V}_m = \frac{X_t^{(m)}}{\sum_{i=tm-m+1}^{tm} X_i^2} \tag{1.15}$$

(also see Hall 1997). An alternative dependence measure is for example the so-called codifference (Samorodnitsky and Taqqu 1994). Suppose that X_t $(t \in Z)$ have a symmetric distribution. Then the codifference is defined by

$$\tau_X(k) = \log \frac{E[e^{i(X_{t+k}-X_t)}]}{E[e^{iX_{t+k}}]E[e^{-iX_t}]}. \tag{1.16}$$

Note that τ_X can also be defined in continuous time. For Gaussian processes, $\tau_X(k)$ coincides with the autocovariance function $\gamma_X(k)$.

1.3.6.2 Extended Memory

Granger (1995) and Granger and Ding (1996) consider a different property characterizing long-term effects of observations from the remote past.

Definition 1.8 Let X_t be a stochastic process defined for $t \in \mathbb{Z}$ or $t \in \mathbb{N}$ and such that $E(X_t^2) < \infty$ for all t. Consider the prediction

$$\hat{X}_{t+k} = E[X_{t+k} \mid X_s, s \le t].$$

Then X_t is said to have extended memory if there is no constant $c \in \mathbb{R}$ such that $\hat{X}_{t+k} \to_p c$ as $k \to \infty$.

Example 1.3 Consider a random walk process defined by $X_t = \sum_{s=1}^{t} \varepsilon_s$ $(t \ge 1)$ where ε_t are i.i.d. $N(0, \sigma_\varepsilon^2)$ distributed with $\sigma_\varepsilon^2 > 0$. Then

$$\hat{X}_{t+k} = X_t$$

for all $k \ge 1$, so that $\hat{X}_{t+k}$ does not converge to a constant but is instead $N(0, t\sigma_\varepsilon^2)$-distributed for all k. Thus, random walk has extended memory. Similarly, for $Y_t = \exp(X_t) = \exp(\sum_{s=1}^{t} \varepsilon_s)$, we have

$$\hat{Y}_{t+k} = Y_t E\left[\exp\left(\sum_{j=1}^{k} \varepsilon_{t+j}\right)\right] = Y_t \exp\left(\frac{1}{2}\sigma_\varepsilon^2 k\right).$$

Again, Y_t does not converge to a constant, but instead $P(\hat{Y}_{t+k} \to \infty) = 1$. This illustrates that extended memory also captures nonlinear dependence. The reason is that the conditional expected value and not just the best linear forecast is considered. More generally, any strictly monotone transformation $G(X_t)$ has extended memory (see e.g. Granger and Ding 1996 and references therein). In contrast, for $|\varphi| < 1$, the equation $X_t = \varphi X_{t-1} + \varepsilon_t$ $(t \in \mathbb{Z})$ has a unique stationary causal solution $X_t = \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j}$, and

$$\hat{X}_{t+k} = \varphi^k X_t \underset{p}{\to} 0.$$

More generally, for a purely stochastic invertible linear second-order stationary process with Wold representation $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ and i.i.d. ε_t, we have

$$\hat{X}_{t+k} = E[X_{t+k} \mid X_s, s \le t] = \sum_{j=0}^{\infty} a_{j+k} \varepsilon_{t-j},$$

so that

$$\operatorname{var}(\hat{X}_{t+k}) = E\big(\hat{X}_{t+k}^2\big) = \sum_{j=k}^{\infty} a_j^2 \underset{k\to\infty}{\to} 0.$$

Since $\hat{X}_{t+k}$ converges to zero in the L^2-norm and in probability, the process X_t does not have extended memory.

1.3.6.3 Long Memory as Phase Transition

The approach in this section was initiated by G. Samorodnitsky, see Samorodnitsky (2004, 2006). Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability measures that describe the finite-dimensional distributions of a stationary stochastic process $\mathbf{X} = (X_t)$ $(t \in \mathbb{Z}$ or $t \in \mathbb{R})$. We assume that as θ varies over the parameter space Θ, the marginal distribution of X_t does not change. Consider a measurable functional $\phi = \phi(\mathbf{X})$. Its behaviour may be different for different choices of θ. Now, assume that the parameter space Θ can be decomposed into $\Theta_1 \cup \Theta_2$ such that the behaviour of the functional does not change too much as long as $\theta \in \Theta_1$, but changes significantly when we cross the boundary between Θ_1 and Θ_2. Furthermore, the behaviour changes as θ varies across Θ_2. This way, we can view the models with $\theta \in \Theta_1$ as short-memory models and those with $\theta \in \Theta_2$ as long-memory models. One has to mention here that this notion of LRD does not look at one particular parameter (in contrast to the case of a finite variance where θ can be thought of as an exponent of a hyperbolic decay of covariances). Instead, it is tied to each particular functional. It may happen that a particular model is LRD for one functional but not for another. In other words, if we have two functionals ϕ_1 and ϕ_2, the decomposition of the parameter space may be completely different, i.e. $\Theta = \Theta_1(\phi_1) + \Theta_2(\phi_1)$ and $\Theta = \Theta_1(\phi_2) + \Theta_2(\phi_2)$ with $\Theta_1(\phi_1) \neq \Theta_1(\phi_2)$.

Example 1.4 (Partial Sums) Denote by $L_f(\lambda)$ a function that is slowly varying at the origin in Zygmund's sense. Let $\mathbf{X} = (X_t, t \in \mathbb{Z})$ be a stationary Gaussian sequence with spectral density $f_X(\lambda) \sim L_f(\lambda)\lambda^{-2d}$ (as $\lambda \to 0$) but $f_X(0) \neq 0$, and assume that $d =: \theta \in [-\infty, \frac{1}{2})$. (Here $d = -\infty$ is interpreted as the case of i.i.d. random variables.) For the functional $\phi_1(\mathbf{x}) = \sum_{t=1}^n x_t$, the parameter space may be decomposed into $(0, \frac{1}{2}) \cup \{0\} \cup (-\infty, 0]$. For the sub-space $(0, \frac{1}{2})$, the rate of convergence changes for different choices of θ. In other words, according to Samorodnitsky's definition, $\mathbf{X}$ is ϕ_1-LRD for $\theta \in (0, \frac{1}{2})$ since then the partial sum has to be scaled by $L_\gamma^{-1/2}(n)n^{-d-\frac{1}{2}}$ to obtain a nondegenerate limit. Otherwise, if $\theta \in [-\infty, 0]$, then the scaling is $n^{-1/2}$. If instead, we consider the functional $\phi_2(\mathbf{x}) = \sum_{t=1}^n (x_t^2 - 1)$, then the parameter space is decomposed into $(\frac{1}{4}, \frac{1}{2}) \cup \{\frac{1}{4}\} \cup [-\infty, \frac{1}{4})$. The process $\mathbf{X}$ is ϕ_2-LRD for $\theta \in (\frac{1}{4}, \frac{1}{2})$. We refer to Chap. 4 for a detailed discussion of limit theorems for partial sums.

Example 1.5 (Maxima) Let $\mathbf{X} = (X_t, i \in \mathbb{Z})$ be as in Example 1.4, but we consider the functional $\phi_3(\mathbf{x}) = \max_{t=1}^n x_t$. The limiting behaviour of maxima of Gaussian sequences with nonsummable autocovariances or autocovariances that sum up to zero is the same as under independence. Thus, according to Samorodnitsky's definition, $\mathbf{X}$ is not max-LRD. We refer to Sect. 4.10 for limit theorems for maxima.

However, the main reason to consider the "phase transition" approach is to quantify long-memory behaviour for stationary stable processes. In particular, if $X_t = Z_{H,\alpha}(t) - Z_{H,\alpha}(t-1)$, where $Z_{H,\alpha}(\cdot)$ is a Linear Fractional Stable motion (1.13), then, due to self-similarity, $n^{-H}\sum_{t=1}^n X_t$ equals in distribution $Z_{H,\alpha}(1)$, where $H = d + 1/\alpha$. On the other hand, if X_t are i.i.d. symmetric α-stable, then $n^{-1/\alpha}\sum_{t=1}^n X_t$ equals in distribution an α-stable random variable. Hence, the phase transition from short memory to long memory occurs at $H = 1/\alpha$. A similar transition occurs in the case of ruin probabilities.

Example 1.6 (Ruin Probabilities) As in Mikosch and Samorodnitsky (2000), assume again that $X_t = Z_{H,\alpha}(t) - Z_{H,\alpha}(t-1)$, where $Z_{H,\alpha}(\cdot)$ is a Linear Fractional Stable motion. The authors consider the rate of decay of ruin probabilities

$$\psi(u) = P\left(\sum_{t=1}^n X_t > cn + u \text{ for some } n \in \mathbb{N}\right)$$

as u tends to infinity. As it turns out, for $H > 1/\alpha$, $\psi(u)$ is proportional to $u^{-(\alpha-\alpha H)}$, whereas for $0 < H \leq 1/\alpha$, the decay is of the order $u^{-(\alpha-1)}$. Thus, for $H > 1/\alpha$, the decay is slower, which means that the probability of ruin is considerably larger than for $H \leq 1/\alpha$. Moreover, the decay depends on H for $H > 1/\alpha$, whereas this is not the case for $H \leq 1/\alpha$. It is therefore natural to say that X_t has long memory if $H > 1/\alpha$ and short memory otherwise.

Example 1.7 (Long Strange Segments) Another possibility of distinguishing between short- and long-range dependent ergodic processes is to consider the rate at

which so-called long strange segments grow with increasing sample size (Ghosh and Samorodnitsky 2010; Mansfield et al. 2001; Rachev and Samorodnitsky 2001). Suppose that X_t is a stationary process with $\mu = E(X_t) = 0$ and the ergodic property in probability holds (i.e. the sample mean converges to μ in probability). Given a measurable set A, one defines

$$R_n(A) = \sup\{j - i : 0 \leq i < j \leq n, \bar{x}_{i:j} \in A\},$$

where

$$\bar{x}_{i:j} = (j-i)^{-1} \sum_{t=i+1}^{j} X_t$$

is the sample mean of observations $X_{i+1}, \ldots, X_j$. In other words, the random number $R_n(A) \in \mathbb{N}$ is the maximum length of a segment from the first n observations whose average is in A. Why such segments are called "strange" can be explained for sets A that do not include the expected value $\mu = 0$. Since the sample mean converges to zero, one should not expect too long runs that are bounded away from zero. It turns out, however, that for long-memory processes, the maximal length of such runs tends to be longer than under short memory, in the sense that R_n diverges to infinity at a faster rate.

The phase transition approach leads also to much more general stationary stable processes. It turns out that stationary stable processes can be decomposed into a dissipative and a conservative flow. The conservative flow part is usually associated with long memory. We refer to Samorodnitsky (2002, 2004, 2005, 2006), Racheva-Iotova and Samorodnitsky (2003), Resnick and Samorodnitsky (2004) for further details and examples.

Chapter 2
Origins and Generation of Long Memory

In this chapter we discuss typical methods for constructing long-memory processes. Many models are motivated by probabilistic and statistical principles. On the other hand, sometimes one prefers to be lead by subject specific considerations. Typical for the first approach is the definition of linear processes with long memory, or fractional ARIMA models. Subject specific models have been developed for instance in physics, finance and network engineering. Often the occurrence of long memory is detected by nonspecific, purely statistical methods, and subject specific models are then developed to explain the phenomenon. For example, in economics aggregation is a possible reason for long-range dependence, in computer networks long memory may be due to certain distributional properties of interarrival times. Often long memory is also linked to fractal structures.

2.1 General Probabilistic Models

2.1.1 Linear Processes with Finite Second Moments

2.1.1.1 General Definition of Linear Processes

The simplest time series models are linear processes. Given independent identically distributed variables ε_t ($t \in \mathbb{Z}$), a causal linear process (or causal linear sequence, infinite moving average) is defined by

$$X_t = \mu + \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = \mu + A(B)\varepsilon_t \tag{2.1}$$

$$= \mu + \left(\sum_{j=0}^{\infty} a_j B^j \right) \varepsilon_t \quad (t \in \mathbb{Z}) \tag{2.2}$$

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_2, © Springer-Verlag Berlin Heidelberg 2013

with B denoting the backshift operator defined by $B\varepsilon_t = \varepsilon_{t-1}$. Here, "causal" refers to the fact that X_t does not depend on any future values of ε_t. For simplicity of notation and without loss of generality, we will assume in the following that $\mu = 0$. In order that X_t is well defined, convergence of the infinite series has to be guaranteed in a suitable way. If X_t has to have finite second moments, then we need to impose that $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t) < \infty$ and $\sum_{j=0}^{\infty} a_j^2 < \infty$. Also, since ε_{t-j} are supposed to model random mean-adjusted deviations ("innovations") at time t, it is assumed that $E(\varepsilon_t) = 0$. Under these conditions, the series is convergent in the $L^2(\Omega)$-sense, i.e. for each t, there is a random variable X_t such that

$$\lim_{n\to\infty} \left\| X_t - \sum_{j=0}^{n} a_j \varepsilon_{t-j} \right\|^2_{L^2(\Omega)} = \lim_{n\to\infty} E\left[\left(X_t - \sum_{j=0}^{n} a_j \varepsilon_{t-j} \right)^2 \right] = 0.$$

We will also call X_t an L^2-linear process.

2.1.1.2 Ergodicity

The first essential question one has to ask before thinking of statistical methods is whether the ergodic property with constant limit holds, i.e. for instance if the sample mean $\bar{x} = n^{-1} \sum_{i=1}^{n} X_t$ converges to $\mu = E(X_t)$ in a well-defined way. If almost sure convergence is required, then the fundamental result to answer this question is Birkhoff's ergodic theorem (Birkhoff 1931, also see e.g. Breiman 1992, Chap. 6). It states that $\bar{x}$ converges almost surely to μ if X_t is strictly stationary, $E(|X_t|) < \infty$ and X_t is ergodic. The last property, ergodicity, means that for tail events ("asymptotic events"), measurable with respect to the σ-algebra generated by X_t, the probability is either zero or one, but never anything in between (for an exact definition, see e.g. Walters 1989). In general, ergodicity may not be easy to check. However, a simple sufficient condition is that, for each t, the process can be written almost surely as $X_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots)$ where $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is a measurable function (see e.g. Stout 1974, Theorem 3.5.8).

For linear processes defined in $L^2(\Omega)$, we have the L^2-representation $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$, so that, by Jensen's inequality,

$$E\big(|X_t|\big) \le \sqrt{E\big(X_t^2\big)} < \infty.$$

Moreover, since $Y_k = \sum_{j=0}^{k} a_j \varepsilon_{t-j}$ $(k = 0, 1, 2, \ldots)$ is a martingale with $\sup_k E[Y_k^2] < \infty$, Doob's martingale convergence theorem (see e.g. Breiman 1992, Chap. 5) implies that the equality $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ also holds almost surely. Thus, $X_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots)$ a.s. with $f(u_1, u_2, \ldots) = \sum_{j=0}^{\infty} a_j u_j$, so that X_t is ergodic. Moreover, the almost sure representation guarantees that X_t is not only second order but also strictly stationary. Birkhoff's ergodic theorem is therefore applicable for all linear processes defined in $L^2(\Omega)$.

2.1.1.3 Long Memory, Short Memory, Antipersistence

For linear processes in $L^2(\Omega)$, long-range dependence, short memory and antipersistence may be defined via the autocovariance function or the spectral density. Since X_t has the spectral density

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\sum_{j=0}^{\infty} a_j e^{-ij\lambda}\right|^2 \tag{2.3}$$

and autocovariances

$$\gamma_X(k) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} a_j a_{j+k}, \tag{2.4}$$

it is easy to see (see Lemmas 2.1, 2.2, 2.3 below) how to specify the coefficients a_j to obtain different types of dependence structures. In the following we will consider three cases, with L_a denoting a function that is slowly varying at infinity in Zygmund's sense.

- Long Memory:

$$a_j = L_a(j)\, j^{d-1} \quad \left(0 < d < \frac{1}{2}\right). \tag{2.5}$$

- Antipersistence:

$$a_j = L_a(j)\, j^{d-1} \quad \left(-\frac{1}{2} < d < 0\right) \quad \text{and} \quad \sum_{j=0}^{\infty} a_j = 0. \tag{2.6}$$

- Short Memory:

$$\sum_{j=0}^{\infty} |a_j| < \infty \quad \text{and} \quad \sum_{j=0}^{\infty} a_j \neq 0. \tag{2.7}$$

The long-memory condition implies $\sum a_j = \infty$. This could also be used as a definition of a long-range dependent linear process, but for most practical applications, the more specific condition (2.5) is general enough, even if L_a is confined to slowly varying functions that converge to a finite constant c_a. The same applies to the condition for antipersistence. The explanation why the three cases are associated with long memory, antipersistence and short memory respectively is given in the following three lemmas. The proofs will be given later in Sect. 4.2.

Lemma 2.1 *Let X_t ($t \in \mathbb{Z}$) be an L^2-linear process such that (2.5) holds. Denote by $\gamma_X(k) = cov(X_t, X_{t+k})$ the autocovariance function of X_t. Then*

$$\gamma_X(k) \sim L_\gamma(k) k^{2d-1} \quad (k \to \infty)$$

with

$$L_\gamma(k) = \sigma_\varepsilon^2 L_a^2(k) \int_0^\infty x^{d-1}(x+1)^{d-1}\,dx$$
$$= \sigma_\varepsilon^2 L_a^2(k) B(1-2d, d),$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ *is the Beta function. Moreover,*

$$f_X(\lambda) \sim L_f(\lambda)|\lambda|^{-2d} \quad (\lambda \to 0)$$

with

$$L_f(\lambda) = L_\gamma\big(\lambda^{-1}\big)\pi^{-1}\Gamma(2d)\sin\left(\frac{\pi}{2} - \pi d\right).$$

Lemma 2.2 *Let* X_t $(t \in \mathbb{Z})$ *be an* L^2*-linear process such that* (2.7) *holds and also* $\sum j|a_j| < \infty$*. Then*

$$\sum_{k=-\infty}^{\infty} \big|\gamma_X(k)\big| < \infty, \qquad \sum_{k=-\infty}^{\infty} \gamma_X(k) \neq 0,$$

and f_X *is continuous at the origin.*

Lemma 2.3 *Let* X_t $(t \in \mathbb{Z})$ *be an* L^2*-linear process such that* (2.6) *holds. Then*

$$\gamma_X(k) \sim L_\gamma(k) k^{2d-1} \quad (k \to \infty)$$

with

$$L_\gamma(k) = L_a^2(k) \int_0^\infty x^{d-1}\big[(x+1)^{d-1} - 1\big]\,dx.$$

Moreover,

$$f_X(\lambda) \sim L_f(\lambda)|\lambda|^{-2d} \quad (\lambda \to 0)$$

with

$$L_f(\lambda) = L_\gamma\big(\lambda^{-1}\big)\pi^{-1}\Gamma(2d)\sin\left(\frac{\pi}{2} - \pi d\right).$$

Note that the additional condition $\sum j|a_j| < \infty$ in Lemma 2.2 is not necessary, but was chosen here to make the proof simple. For most short-memory processes considered in statistical applications, the asymptotic decay of a_j is exponential, so that this condition holds. Also note that for $d < 0$, the integral $\int_0^\infty x^{d-1}[(x+1)^{d-1} - 1]\,dx$ is finite, whereas $\int_0^\infty x^{d-1}(x+1)^{d-1}\,dx = \infty$ because of the pole of the order $x^{d-1} \gg x^{-1}$ at zero (with "$\gg$" meaning that $\lim_{x\to 0}|x^{d-1}/x^{-1}| = \infty$).

2.1.1.4 Fractional ARIMA Models

A particularly useful class of linear processes that includes all three dependence structures is obtained by extending classical ARMA and ARIMA processes. Due to their simplicity and flexibility, ARMA and ARIMA processes are probably the most popular class of linear models in time series analysis. They were introduced and popularized by Box and Jenkins (1970). A stationary causal ARMA(p, q) process is defined by the equation

$$\varphi(B)X_t = \psi(B)\varepsilon_t, \tag{2.8}$$

where ε_t are assumed to be i.i.d. with zero mean and finite variance σ_ε^2, and $\varphi(z) = 1 - \sum_{j=1}^{p} \varphi_j z^j$ and $\psi(z) = \sum_{j=0}^{q} \psi_j z^j$ are polynomials with no common roots and all roots outside the unit circle. Box and Jenkins extended this definition also to integrated processes. For $d \in \{1, 2, \ldots\}$, an ARIMA(p, d, q) process is defined recursively by $Y_0 = 0$ and

$$(1 - B)^d Y_t = X_t \quad (t \geq 1), \tag{2.9}$$

where X_t is given by (2.8). Note that (2.8) can also be included by setting $d = 0$. For $d \geq 1$, the process Y_t is nonstationary, but it can be transformed into a stationary ARMA process by taking the dth difference $(1 - B)^d Y_t$. For instance, if $p = q = 0$ and $d = 1$, then Y_t is a random walk process, and the first difference yields the i.i.d. sequence ε_t. In econometrics, Y_t defined by (2.9) is also called integrated of order d, or $I(d)$.

In order to obtain a model that is somewhere between an ARMA and an ARIMA process, Granger and Joyeux (1980) and Hosking (1981) proposed to allow for the possibility of noninteger values of d. The resulting processes are called fractional ARIMA(p, q) processes (also FARIMA(p, d, q) or ARFIMA(p, d, q)). Formally, this can be justified as follows. Let $d \in (-\frac{1}{2}, \frac{1}{2})$ and denote by ε_t i.i.d. zero mean random variables with finite variance σ_ε^2. Consider the series expansions

$$A(z) = (1 - z)^{-d} = \sum_{j=0}^{\infty} a_j z^j,$$

$$A^{-1}(z) = (1 - z)^{d} = \sum_{j=0}^{\infty} b_j z^j$$

for $|z| \leq 1$, $z \neq 1$ with

$$a_j = \binom{-d}{j}(-1)^j = \frac{\Gamma(-d+1)}{\Gamma(j+1)\Gamma(-d-j+1)}(-1)^j,$$

$$b_j = \binom{d}{j}(-1)^j = \frac{\Gamma(d+1)}{\Gamma(j+1)\Gamma(d-j+1)}(-1)^j.$$

Using Stirling's formula and the property $\Gamma(x+1) = x\Gamma(x)$, one can see that, as $j \to \infty$, the coefficients in $X_t = \sum a_j \varepsilon_{t-j}$ and $\sum b_j X_{t-j} = \varepsilon_t$ are of the forms

$$\begin{aligned} a_j &= \frac{\Gamma(-d+1)}{\Gamma(j+1)\Gamma(-d-j+1)}(-1)^j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} \\ &\sim \frac{1}{\Gamma(d)} j^{d-1} \end{aligned}$$

and

$$\begin{aligned} b_j &= \frac{\Gamma(d+1)}{\Gamma(j+1)\Gamma(d-j+1)}(-1)^j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \\ &\sim \frac{1}{\Gamma(-d)} j^{-d-1}. \end{aligned}$$

Since d is in the interval $(-\frac{1}{2}, \frac{1}{2})$, this implies that the functions $g(\lambda) = A(e^{-i\lambda})$ and $h(\lambda) = A^{-1}(e^{-i\lambda})$ are in the $L^2(F_\varepsilon)$ space of functions on $[-\pi, \pi]$ with norm

$$\|g\|^2 = \int_{-\pi}^{\pi} \left|g(\lambda)\right|^2 dF_\varepsilon(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \int_{-\pi}^{\pi} \left|g(\lambda)\right|^2 d\lambda < \infty.$$

Here $F_\varepsilon(\lambda) = \sigma_\varepsilon^2/(2\pi) \int_{-\pi}^{\lambda} d\nu$ denotes the spectral distribution of ε_t. Therefore, $A(B)$ and $A^{-1}(B)$ are valid filters in the sense of $L^2(\Omega)$-convergence, and the linear process

$$X_t = A(B)\varepsilon_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{2.10}$$

is a well-defined, stationary and invertible process with spectral representation

$$X_t = \int e^{it\lambda} A\left(e^{-i\lambda}\right) dM_\varepsilon(\lambda)$$

(with M_ε denoting the spectral measure of ε_t) and autoregressive representation

$$A^{-1}(B)X_t = (1-B)^d X_t = \varepsilon_t. \tag{2.11}$$

This defines a FARIMA$(0, d, 0)$ process for all values $-\frac{1}{2} < d < \frac{1}{2}$. Including the AR- and MA-filters $\varphi(B)$ and $\psi(B)$ does not change the asymptotic rate of convergence of a_j and b_j, so that by the same arguments the equation

$$(1-B)^d \varphi(B) X_t = \psi(B)\varepsilon_t \tag{2.12}$$

has a unique stationary invertible solution which is called an ARFIMA(p, d, q) or FARIMA(p, d, q) process. Applying the filter (or fractional differencing operator) $(1-B)^d$, the process

$$Z_t = (1-B)^d X_t = \varphi^{-1}(B)\psi(B)\varepsilon_t \tag{2.13}$$

is an ordinary stationary ARMA(p,q) process. In contrast to standard integrated processes (i.e. $I(d)$ with integer $d \geq 1$), the FARIMA(p,d,q) process is stationary as long as d is in the interval $(-\frac{1}{2}, \frac{1}{2})$. Fractionally integrated processes can be obtained in an analogous manner by setting $Y_0 = 0$ and $(1-B)^m Y_t = X_t$ with $m \in \{1,2,\ldots\}$ and X_t satisfying (2.12). This means that there are two differencing parameters, the integer parameter m needed to make the process stationary and the fractional parameter d needed in addition to m to obtain a standard ARMA process. Note that, since d is confined to the open interval $(-\frac{1}{2}, \frac{1}{2})$, it is possible to recover both parameters from their sum $d_{\text{total}} = d + m$ by $m = [d_{\text{total}} + \frac{1}{2}]$ and $d = d_{\text{total}} - m$. Thus, Y_t may be called a FARIMA(p, d_{total}, q) process.

Definition (2.10) implies that the spectral density of a stationary FARIMA$(0,d,0)$ process is equal to

$$\begin{aligned} f_X(\lambda) &= \frac{\sigma_\varepsilon^2}{2\pi}\left|A\left(e^{-i\lambda}\right)\right|^2 = \frac{\sigma_\varepsilon^2}{2\pi}\left|1-e^{-i\lambda}\right|^{-2d} \\ &= L_f(\lambda)|\lambda|^{-2d} \end{aligned}$$

with $L_f(\lambda) \sim \sigma_\varepsilon^2/(2\pi)$ as $\lambda \to 0$. More generally, applying the filters $\varphi(B)$ and $\psi(B)$, the spectral density of a stationary FARIMA(p,d,q) process is equal to

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\psi(e^{-i\lambda})}{\varphi(e^{-i\lambda})}\right|^2 \left|1-e^{-i\lambda}\right|^{-2d} \tag{2.14}$$

$$= f_{\text{ARMA}}(\lambda)\left|1-e^{-i\lambda}\right|^{-2d} = L_f(\lambda)|\lambda|^{-2d}, \tag{2.15}$$

where $f_{\text{ARMA}}(\lambda)$ is the spectral density of the corresponding ARMA(p,q) process, and the slowly varying function L_f is given by

$$L_f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\psi(e^{-i\lambda})}{\varphi(e^{-i\lambda})}\right|^2 \sim c_{f,\text{ARMA}} = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\psi(1)}{\varphi(1)}\right|^2 \quad (\lambda \to 0).$$

Exact explicit formulas for autocovariances and autocorrelations are complicated in general, though in principle they follow directly from the Wold representation (2.10). The formulas are simple however for a FARIMA$(0,d,0)$ process (the essential formula solving the respective integrals can be found in Gradshteyn and Rhyzhik 1965, p. 372), with

$$\gamma_X(k) = \sigma_\varepsilon^2 \frac{(-1)^k \Gamma(1-2d)}{\Gamma(1+k-d)\Gamma(1-k-d)}. \tag{2.16}$$

In particular,

$$\gamma_X(0) = \sigma_\varepsilon^2 \frac{\Gamma(1-2d)}{\Gamma^2(1-d)},$$

and, for the autocorrelation $\rho_X(k) = \gamma_X(k)/\gamma_X(0)$, we have

$$\rho_X(k) = \sigma_\varepsilon^2 \frac{(-1)^k \Gamma^2(1-d)}{\Gamma(1+k-d)\Gamma(1-k-d)} = \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(1+k-d)}. \tag{2.17}$$

The asymptotic behavior of $\gamma_X(k)$ for a FARIMA(p,d,q) process follows from the behavior of f_X at the origin given in (2.15) and Theorem 1.3:

$$\gamma_X(k) \sim c_\gamma |k|^{2d-1} \quad (k \to \infty)$$

with

$$c_\gamma = 2c_{f,\mathrm{ARMA}} \Gamma(1-2d) \sin \pi d.$$

Note that this result can also be obtained directly by considering the asymptotic decay of the coefficients a_j. In particular, for a FARIMA$(0,d,0)$ process, we can use the identity

$$\frac{\sin \pi d}{\pi} = \frac{1}{\Gamma(1-d)\Gamma(d)}$$

to obtain

$$\gamma_X(k) \sim \left(\frac{\sigma_\varepsilon^2}{\pi} \Gamma(1-2d) \sin \pi d \right) |k|^{2d-1} = \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} |k|^{2d-1}$$

and

$$\rho_X(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)} |k|^{2d-1}.$$

A further useful result by Hosking (1981) is that the partial correlations of a FARIMA$(0,d,0)$ process are given by

$$\beta_{kj} = -\binom{k}{j} \frac{\Gamma(j-d)\Gamma(k-d-j+1)}{\Gamma(-d)\Gamma(k-d+1)}$$

and asymptotically, as $j,k \to \infty$ and $j/k \to 0$,

$$\beta_{kj} \sim \frac{1}{\Gamma(-d)} j^{-d-1}.$$

For $j = k$, we have $\beta_{kk} = d/(k-d)$. Recall that $\hat{X}_{n+1} = \sum_{j=1}^n \beta_{nj} X_{n+1-j}$ is the optimal linear prediction of X_{n+1} given $X_1, \ldots, X_n$ (also see Chap. 8).

One may ask at this point why $d = -\frac{1}{2}$ and $\frac{1}{2}$ were excluded. The reason can be seen in the asymptotic behaviour of b_j. For $d = -\frac{1}{2}$, the coefficients b_j are proportional to $j^{-\frac{1}{2}}$, so that $\sum b_j^2 = \infty$, and $A^{-1}(e^{-i\lambda})$ is no longer in $L^2(F_\varepsilon)$. This means that X_t is no longer invertible, even though the process $X_t = A(B)\varepsilon_t$ is well defined. The same comments apply to $d = -\frac{1}{2} + m$ where m is a positive integer, since the mth difference of X_t is not invertible, and to $d = -\frac{1}{2} + m$ with m a negative integer, since there X_t is the mth difference of a noninvertible process.

2.1.1.5 Other Fractionally Differenced Processes, FEXP Processes, Fractional Gaussian Noise

Equation (2.13) can be extended by replacing $Z_t = \varphi^{-1}(B)\psi(B)\varepsilon_t$ by any L^2-linear short-memory process. The interpretation is that the fractional differencing filter $(1-B)^d$ removes the long-memory component, and the rest can be anything (linear) with short memory. Similarly, flexible models for the spectral density can be obtained by replacing f_{ARMA} in (2.14) by any continuous, or more generally any short-memory, density function f_{short}.

Example 2.1 Let

$$f_{\mathrm{short}}(\lambda) = \exp\bigl(\eta_0 + \eta_1 \cos(\lambda) + \cdots + \eta_p \cos(p\lambda)\bigr).$$

Then

$$\begin{aligned} f_X(\lambda) &= f_{\mathrm{short}}(\lambda)\bigl|1 - e^{-i\lambda}\bigr|^{-2d} \\ &= \exp\left(\sum_{j=0}^{p+1} \theta_j g_j(\lambda)\right) \end{aligned}$$

with

$$\begin{aligned} &g_0(\lambda) \equiv 1, \qquad g_1(\lambda) = \log\bigl|1 - e^{-i\lambda}\bigr|, \\ &g_2(\lambda) = \cos(\lambda), \qquad \ldots, \qquad g_{p+1}(\lambda) = \cos(p\lambda), \end{aligned}$$

and

$$\theta_0 = \eta_0, \qquad \theta_1 = -2d, \qquad \theta_2 = \eta_1, \qquad \ldots, \qquad \theta_{p+1} = \eta_p$$

is a so-called fractional exponential model of order p or FEXP(p) model introduced in Beran (1993) and Robinson (1994a). The short-memory version with $d = 0$ is discussed in Bloomfield (1973).

Example 2.2 Let $B_H(u)$ ($u \in \mathbb{R}$) be a fractional Brownian motion (see Sect. 1.3.5), and $\xi_t = B_H(t) - B_H(t-1)$ ($t \in \mathbb{Z}$) be a discrete-time fractional Gaussian noise (fGn). Self-similarity of B_H implies a very specific autocovariance structure of ξ_t:

$$\gamma_\xi(k) = \frac{\sigma^2}{2}\bigl(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\bigr)$$

(see (1.7)). Therefore, for most observed data, fractional Gaussian noise is not flexible enough. A much more flexible class is obtained by defining

$$X_t = \varphi^{-1}(B)\psi(B)\xi_t. \tag{2.18}$$

This means that after passing X_t through the ARMA filter $\varphi(B)\psi^{-1}(B)$ we obtain fractional Gaussian noise. In other words, long-range dependence is modelled

by the correlation structure of fGn, whereas the short-memory part is captured by ARMA-type dependence. This is an attractive alternative to usual FARIMA processes because the variance of the sample mean of fGn has the simple form

$$\operatorname{var}\left(n^{-1}\sum_{t=1}^{n}\xi_t\right) = \operatorname{var}\bigl(n^{-1}B_H(n)\bigr) = \sigma^2 n^{2H-2}$$

and fractional Brownian motion is the asymptotic limit of normalized sums (see Chap. 4).

2.1.2 Linear Processes with Infinite Second Moments

Linear processes with infinite second moments can be defined as in (2.1), however with $L^2(\Omega)$-convergence replaced by almost sure limits. Let ε_t $(t \in \mathbb{Z})$ be a sequence of i.i.d. random variables such that

$$F_\varepsilon(-x) = P(\varepsilon \le -x) \sim (1-p)x^{-\alpha}, \qquad \bar{F}_\varepsilon(x) = P(\varepsilon > x) \sim px^{-\alpha} \quad (x \to \infty), \tag{2.19}$$

where $p \in (0,1)$, and the tail index α is in the interval $(0,2)$. More generally, one may replace p and $1-p$ by slowly varying functions. If $\alpha > 1$, then we assume $E(\varepsilon) = 0$. For $\alpha \le 1$, the expected value does not exist, so that instead ε is assumed to have a distribution that is symmetric around zero. The following lemma formulates sufficient conditions on the coefficients a_j so that the linear process is well defined.

Lemma 2.4 *Let ε_t be i.i.d. with distribution function $F_\varepsilon(x) = P(\varepsilon \le x)$ satisfying (2.19). Moreover, assume that*

$$\sum_{j=0}^{\infty} |a_j|^\delta < \infty$$

for some $0 < \delta < \min(1,\alpha)$ and define

$$X_{t,n} = \sum_{j=0}^{n} a_j \varepsilon_{t-j}, \qquad Y_{t,n} = \sum_{j=0}^{n} |a_j \varepsilon_{t-j}|.$$

Then there are strictly stationary processes X_t and Y_t such that

$$P\Bigl(\lim_{n\to\infty} X_{t,n} = X_t\Bigr) = 1$$

and

$$P\Bigl(\lim_{n\to\infty} Y_{t,n} = Y_t\Bigr) = 1.$$

We then write

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}, \qquad Y_t = \sum_{j=0}^{\infty} |a_j \varepsilon_{t-j}|$$

with equality in the a.s. sense.

Proof (See e.g. Brockwell and Davis 1991.) We use the $L^\delta(\Omega)$-norm defined by $\|X\|_\delta = \{E(|X|^\delta)\}^{\frac{1}{\delta}}$. Note that $E(|\varepsilon|^\delta) = \|\varepsilon\|_\delta^\delta < \infty$. Applying Minkowski's inequality, we have

$$E\left(|X_{t,n}|^\delta\right) = \|X_{t,n}\|_\delta^\delta \le \left(\sum_{j=0}^{\infty} \|a_j \varepsilon_{t-j}\|_\delta\right)^\delta = \left\{\|\varepsilon\|_\delta \sum_{j=0}^{\infty} |a_j|\right\}^\delta < \infty.$$

Hence, $|X_{t,n}|$ (and also $Y_{t,n}$) converges almost surely to a finite limit, which implies that the same is true for $X_{t,n}$. □

Note that for $1 < \alpha \le 2$, $E(|\varepsilon|) < \infty$, so that convergence is even achieved if $\delta = 1$. The conditions in this lemma are needed to obtain the a.s. convergence of $Y_{t,n}$. The problem is however that these assumptions exclude the coefficients a_j that would correspond to what may be called long memory. More specifically, in analogy to the case of finite variance, consider

$$a_j = L_a(j) j^{d-1}.$$

The conditions in Lemma 2.4 imply that d must be such that $\min(1, \alpha)(d-1) < -1$, i.e. $d < 1 - 1/\min(1, \alpha)$. This would exclude positive values of d. Fortunately, the convergence of $Y_{t,n}$ is not a necessary condition for the convergence of $X_{t,n}$. If only the convergence of $X_{t,n}$ is what we are looking for, then the assumption on the coefficients can be relaxed as follows (cf. Kokoszka and Taqqu 1995a, 1995b, 1996).

Lemma 2.5 *Let ε_t be i.i.d. with distribution function $F_\varepsilon(x) = P(\varepsilon \le x)$ satisfying (2.19) and density $f_\varepsilon(x) = F_\varepsilon'(x)$. Assume that, for some $\delta > 0$,*

$$\sum_{j=0}^{\infty} |a_j|^{\alpha-\delta} < \infty. \tag{2.20}$$

Then there is a strictly stationary process such that

$$P\left(\lim_{n\to\infty} X_{t,n} = X_t\right) = 1.$$

We then write

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$$

with equality in the a.s. sense.

For coefficients of the form $a_j = L_a(j)j^{d-1}$, we thus obtain the condition $\alpha(d-1) < -1$, i.e. $d < 1 - \alpha^{-1}$. Thus, for $0 < \alpha \leq 1$, positive values of d are still excluded. However, for $1 < \alpha < 2$, positive values of d may be chosen from the interval $(0, 1 - \alpha^{-1})$. As α approaches the limiting case $\alpha = 2$, the upper bound reaches $\frac{1}{2}$, which is the same as for linear processes with finite variance. This is to be expected, because symmetric α-stable random variables with $\alpha = 2$ are normally distributed.

A further property of X_t is that the linear process inherits the tail index α (see Sect. 4.10), since under condition (2.20),

$$\lim_{x\to\infty} \frac{P(|X_t| > x)}{P(|\varepsilon_t| > x)} = \sum_{j=0}^{\infty} |a_j|^{\alpha}. \tag{2.21}$$

Note that for $\alpha \in (0, 1]$, (2.20) implies the summability of the coefficients, and (2.21) follows from Davis and Resnick (1985). If $\alpha \in (1, 2)$, then (2.21) is a statement of Theorem 2.2 in Kokoszka and Taqqu (1996). We thus see that for $0 < \alpha \leq 1$, neither $E(X_t)$ nor $\mathrm{var}(X_t)$ exists whatever the coefficients a_j are (unless they are all zero). On the other hand, for $1 < \alpha < 2$, $E(X_t)$ is finite, and the variance is defined but infinite.

So far, we associated the case where $d > 0$ with "long-range dependence" only by pure analogy with the finite variance case. It remains to be shown that there is indeed long-range dependence for $d > 0$. A meaningful notion of long memory can be given by considering measures of dependence applicable to infinite-variance variables, as introduced in Sect. 1.3.6.1. For instance, for the codifference (see Eq. (1.16)), the following result follows from Kokoszka and Taqqu (1995b).

Theorem 2.1 *Let ε_t be i.i.d. symmetric α-stable random variables with $1 < \alpha \leq 2$, and $a_j = L_a(j)j^{d-1}$ with $d \in (0, 1 - \alpha^{-1})$. Then, as $k \to \infty$, the codifference between X_t and X_{t+k} is of the form*

$$\tau(X_t, X_{t+k}) \sim C_\tau \cdot k^{\alpha(d-1)+1},$$

where C_τ is a finite constant.

Note, in particular, that in the case of normally distributed innovations ($\alpha = 2$) we obtain the well-known formula for the autocovariance function $\gamma(k) \sim Ck^{2d-1}$. The influence of the tail index α shows that the decay of τ becomes slower the smaller α is, i.e. the heavier the tail is. For further fundamental results on linear long-memory processes with heavy tails, see e.g. Avram and Taqqu (1986), Kasahara et al. (1988), Astrauskas et al. (1991), Samorodnitsky and Taqqu (1994), Kokoszka (1996), Kokoszka and Taqqu (1993, 1995a, 1995b, 1996, 2001), Kokoszka and Mikosch (1997), Koul and Surgailis (2001), Mansfield et al. (2001), Rachev and Samorodnitsky (2001), Thavaneswaran and Peiris (2001), Samorodnitsky (2002, 2004, 2006), Surgailis (2002), Racheva-Iotova and Samorodnitsky (2003), Pipiras and Taqqu (2000b), Levy and Taqqu (2005), Stoev and Taqqu (2005a, 2005b), Beran

et al. (2012), also see Samorodnitsky and Taqqu (1994) and Embrechts and Maejima (2002) and references therein.

2.1.3 Nonlinear Processes—Volatility Models

2.1.3.1 Introduction

In financial applications, observations such as (log-)returns are often uncorrelated but not independent. More specifically, often the squared observations (or other powers) exhibit long-range dependence (Whistler 1990, Crato and de Lima 1993, Dacorogna et al. 1993, Ding et al. 1993, Baillie et al. 1996a, Andersen and Bollerslev 1997a, 1997b, Breidt et al. 1998, Granger 1998, Lobato and Savin 1998, Robinson and Zaffaroni 1998, Bollerslev and Mikkelsen 1999, Ray and Tsay 2000, Barndorff-Nielsen and Shephard 2001, Beran and Ocker 2001, Arteche 2004, Deo et al. 2006b, Granger and Hyung 2004, Morana and Beltratti 2004, Harvey 2007, Corsi 2009, Scharth and Medeiros 2009). In financial language this is interpreted as strong dependence in volatility, in particular in the sense that high volatilities tend to cluster. This lead to the development of models that are nonlinear in the sense that the conditional variance depends on the past and possibly also on time itself. For short-memory volatility dependence, there is an extended literature initiated by the pathbreaking work of Engle (1982) and Bollerslev (1986), who introduced ARCH(p) and GARCH(p, q) models respectively. Apart from applied work, there is an enormous literature that describes mathematical properties such as stationarity, tail behaviour, dependence, estimation and limit theorems for GARCH and related models. However, GARCH(p, q) models cannot explain the empirical observation that often dependence in volatility is rather strong and long-lasting while the process still seems to be stationary. The question is therefore how to either extend GARCH models or to define new models in order to incorporate long-range dependence. The first natural extension is the so-called ARCH(∞) process. The general framework was introduced in Robinson (1991). Stationarity and dependence properties were studied in Kokoszka and Leipus (2000), Giraitis et al. (2000c), Kazakevičius and Leipus (2002, 2003), Giraitis et al. (2006) and Douc et al. (2008) among others. At first sight this extension seems to be analogous to the modification of ARMA(p, q) models to MA(∞)-processes with non-summable weights (see Sect. 2.1.1.4). However, as it turns out, a stationary ARCH(∞) sequence with finite variance must have summable weights, and this rules out long memory. In analogy to IGARCH processes, one can however define IARCH(∞) and FIGARCH models (Baillie et al. 1996a), which necessarily have an infinite variance. The existence of a strictly stationary solution was proved in Douc et al. (2008). However, dependence properties including the interpretation of long memory are not clear.

Since the ARCH(∞) model cannot capture long memory in volatility, an alternative is the so-called LARCH(∞) process introduced by Robinson (1991). Its stationarity and dependence properties were studied by Giraitis et al. (2000b, 2003, 2004),

estimation and limit theorems were considered in Giraitis et al. (2000c), Berkes and Horváth (2003), Beran (2006), Beran and Feng (2007), Beran and Schützner (2009). Furthermore, Giraitis and Surgailis (2002) considered bilinear models consisting of a combination of long memory in the mean with long memory in volatility described by a LARCH(∞) structure. Since the conditional scaling process σ_t in LARCH(∞) models can be negative, Surgailis (2008) introduced a so-called LARCH$_+(\infty)$ process where $\sigma_t > 0$ is guaranteed. This process can also capture heavy-tailed behaviour.

Studying properties of GARCH(p, q), ARCH(∞) or LARCH(∞) processes can be mathematically quite demanding. In contrast, establishing existence, stationarity and dependence properties is generally quite easy for so-called "stochastic volatility" models. The first model of this type is the EGARCH process introduced by Nelson (1990) and extended to the long-memory setting (under the name FIEGARCH) by Bollerslev and Mikkelsen (1996). Independently, Breidt et al. (1998) introduced a slightly different long-memory stochastic volatility process (also called LMSV). Further extensions can be found in Robinson and Zaffaroni (1997, 1998). For stationarity and asymptotic properties, see e.g. Harvey (1998) and Surgailis and Viano (2002), for extensions with heavy tails, see Davis and Mikosch (2001) and Kulik and Soulier (2011, 2012, 2013). For reviews in the econometric context, see e.g. Baillie (1996) and Henry and Zaffaroni (2003).

To be more specific, we start with an informal definition of volatility models. Following Giraitis et al. (2006), the notion of stochastic volatility usually stands for models of the form

$$X_t = \sigma_t \varepsilon_t, \tag{2.22}$$

where ε_t ($t \in \mathbb{Z}$) are i.i.d. random variables with mean zero and unit variance, and σ_t is a (usually positive) measurable function of the past values ε_s, X_s ($s \le t-1$) and possibly some additional, unobservable information. Furthermore, ε_s ($s \ge t$) is independent of ε_s, X_s ($s \le t-1$). It follows that

$$E(X_t \mid \sigma_s, \varepsilon_s, s \le t-1) = 0$$

and

$$\mathrm{Var}(X_t \mid \sigma_s, \varepsilon_s, s \le t-1) = \sigma_t^2.$$

It should be noted, however, that actually no standard terminology exists. For instance, in the context of pricing of derivatives, "stochastic volatility" often refers to the special case where the sequences σ_t ($t \in \mathbb{Z}$) and ε_t ($t \in \mathbb{Z}$) are mutually independent. If this is not the case, then one talks of "stochastic volatility with leverage".

We now discuss the most important models in more detail.

2.1.3.2 GARCH (Generalized Autoregressive Conditionally Heteroscedastic) Models

The best known volatility model is the ARCH(p) process and its generalization, the GARCH(p, q) process, introduced by Engle (1982) and Bollerslev (1986) respec-

tively, and studied by Nelson (1990) and Bougerol and Picard (1992), among others (see Berkes et al. 2003 for an overview). By GARCH(p,q) one means the model defined by (2.22) where the conditional variance $\sigma_t^2 = E[X_t^2|X_s, \varepsilon_s, s \le t-1]$ is given by

$$\sigma_t^2 = \beta_0 + \sum_{j=1}^{p} \alpha_i \sigma_{t-j}^2 + \sum_{j=1}^{q} \beta_j X_{t-j}^2. \tag{2.23}$$

The GARCH(p,q) equation (2.23) for the conditional variance can be written as

$$\big(1-\alpha(B)\big)\sigma_t^2 = \beta_0 + \beta(B)X_t^2, \tag{2.24}$$

where $\alpha(z) = \sum_{j=1}^{p} \alpha_i z^j$ and $\beta(B) = \sum_{j=1}^{q} \beta_j z^j$. Alternatively, we can write (2.24) as

$$\big(1-\alpha(B)-\beta(B)\big)\sigma_t^2 = \beta_0 + \beta(B)Z_t, \tag{2.25}$$

where $Z_t = X_t^2 - \sigma_t^2$ $(t \in \mathbb{Z})$ are uncorrelated. If $(1-\alpha(B))$ can be inverted, then we obtain an explicit representation of σ_t^2 as a function of past observations (Nelson and Cao 1992):

$$\sigma_t^2 = \big(1-\alpha(1)\big)^{-1}\beta_0 + \big(1-\alpha(B)\big)^{-1}\beta(B)X_t^2 =: b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2, \tag{2.26}$$

where b_j $(j \ge 1)$ are the coefficients in the power series expansion of $\beta(z)/(1-\alpha(z))$:

$$b_j = \frac{1}{j!}\frac{d^j}{dz^j}\left(\frac{\sum_{j=0}^{q}\beta_j z^j}{1-\sum_{j=1}^{p}\alpha_i z^j}\right)\bigg|_{z=0}. \tag{2.27}$$

It can be shown that the asymptotic decay of the coefficients b_j is exponential (see e.g. Kokoszka and Leipus 2000), so that autocovariances of σ_t^2 (and hence X_t^2) are summable.

2.1.3.3 IGARCH (Integrated GARCH) Processes

Consider first an ARCH(1) model, $X_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = b_0 + b_1 X_{t-1}^2$ with $b_1 = 1$, $b_0 > 0$ and $E[\varepsilon_1^2] = 1$. Suppose that X_t is a solution such that $E[X_t^2]$ does not depend on time t. This is called an IARCH(1) process. Since by assumption the second moment does not depend on time, we have $E[X_t^2] = E[\sigma_t^2] = b_0 + E[\sigma_t^2]$. Now $b_0 \neq 0$, so that the last equation implies that the variance of X_t is infinite. Note also that, as in Eq. (2.25), we can write

$$(I-B)X_t^2 = b_0 + Z_t \tag{2.28}$$

with $Z_t = X_t^2 - \sigma_t^2 = X_t^2 - E[X_t^2|X_s, \varepsilon_s, s \le t-1]$ uncorrelated. This resembles the equation for a random walk with drift. However, in contrast to random walk,

a strictly stationary solution (though with an infinite variance) exists under suitable conditions. The IARCH(1) definition can be generalized to IGARCH(p, q) models with parameters α_i, β_j in (2.23) satisfying the unit root condition $\sum_{j=1}^{p} \alpha_j + \sum_{j=1}^{q} \beta_j = 1$.

2.1.3.4 ARCH(∞) Processes

Using general coefficients b_j in the representation (2.26) of the conditional variance leads to the following definition.

Definition 2.1 Let $b_j \geq 0$ $(j = 0, 1, 2, \ldots)$ and suppose that ε_t $(t \in \mathbb{Z})$ are i.i.d. zero mean random variables. Then X_t $(t \in \mathbb{Z})$ is called an ARCH(∞) process if it is a second-order and/or strictly stationary solution of

$$X_t = \sigma_t \varepsilon_t, \tag{2.29}$$

$$\sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2. \tag{2.30}$$

Usually, it is also assumed that the first two moments of ε_t are finite and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t) = 1$. The reason for the latter assumption is identifiability since statistically the parameter σ_ε^2 is not distinguishable from b_0. A more general definition is given by Robinson (1991).

Definition 2.2 Let $b_j \geq 0$ $(j = 0, 1, 2, \ldots)$ and ξ_t $(t \in \mathbb{Z})$ be i.i.d. nonnegative random variables. Then a process Y_t $(t \in \mathbb{Z})$ is called an ARCH(∞) process if it is a second-order and/or strictly stationary solution of

$$\begin{aligned} Y_t &= v_t \xi_t, \\ v_t &= b_0 + \sum_{j=1}^{\infty} b_j Y_{t-j}. \end{aligned} \tag{2.31}$$

Note that the second definition includes the first one by setting $Y_t = X_t^2$, $v_t = \sigma_t^2$ and $\xi_t = \varepsilon_t^2$. Another possibility is for instance $Y_t = |X_t|^\alpha$, $v_t = \sigma_t^\alpha$ and $\xi_t = |\varepsilon_t|^\alpha$ for some $\alpha > 0$.

Many general results on ARCH(∞) models can be found in Kokoszka and Leipus (2000), Giraitis et al. (2000a), Kazakevičius and Leipus (2002, 2003) and Giraitis et al. (2006). The first question to be answered is under which conditions a stationary solution exists. Above, we essentially answered the question for GARCH processes. More general results for arbitrary ARCH(∞) models are discussed for instance in Giraitis et al. (2000a). The basic idea is to first obtain a Volterra expansion by recursion (initially without checking its validity formally) and then to check mathematically that (or rather under which assumptions) this is indeed a solution

and in how far it is unique. The Volterra expansion—if convergent—is obtained as follows:

$$\begin{aligned} Y_t &= \xi_t \left(b_0 + \sum_{j=1}^{\infty} b_j \xi_{t-j} v_{t-j} \right) \\ &= \xi_t \left(b_0 + \sum_{j_1=1}^{\infty} b_{j_1} \xi_{t-j} \left(b_0 + \sum_{j_2=1}^{\infty} b_{j_2} \xi_{t-j_1-j_2} v_{t-j_1-j_2} \right) \right) \\ &= \cdots = \xi_t b_0 \sum_{l=0}^{\infty} \sum_{j_1,\ldots,j_l=1}^{\infty} b_{j_1} \ldots b_{j_l} \xi_{t-j_1} \ldots \xi_{t-j_1-\cdots-j_l}, \end{aligned} \tag{2.32}$$

where the inner sum is understood as 1 for $l = 0$. A more concise notation is

$$Y_t = b_0 \sum_{l=0}^{\infty} M_l(t), \tag{2.33}$$

where $M_0(t) = \xi_t$, and

$$\begin{aligned} M_l(t) &= \sum_{j_1,\ldots,j_l=1}^{\infty} b_{j_1} \cdots b_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_l} \\ &= \sum_{j_l < j_{l-1} < \cdots < j_1 < t}^{\infty} b_{t-j_1} b_{j_1-j_2} \cdots b_{j_{l-1}-j_l} \xi_t \xi_{j_1} \cdots \xi_{j_l}. \end{aligned} \tag{2.34}$$

The following theorem establishes sufficient conditions under which X_t is a stationary solution of (2.31) with finite expected value (see Kokoszka and Leipus 2000, Giraitis et al. 2000a).

Theorem 2.2 *Under the assumptions*

$$\mu_\xi = E(\xi_t) < \infty \tag{2.35}$$

and

$$\mu_\xi \sum_{j=1}^{\infty} b_j < 1, \tag{2.36}$$

(2.33) *is a strictly stationary solution of* (2.31). (*If* $b_0 = 0$, *then* $X_t = 0$ *almost surely.*) *Moreover,* $E(Y_t) < \infty$, *and* Y_t *is unique in the class of nonanticipatory solutions, where nonanticipatory means that* Y_t *is independent of* ξ_s $(s \geq t + 1)$. *If in addition*

$$\mu_{\xi^2} = E\left(\xi_t^2\right) < \infty \tag{2.37}$$

and

$$\mu_{\xi^2}^{1/2} \sum_{j=1}^{\infty} b_j < 1, \tag{2.38}$$

then Y_t is also a unique second order stationary solution.

Remark 2.1 It should be mentioned that condition (2.38) is sufficient but not necessary for the existence of a second-order stationary solution Y_t (see Giraitis et al. 2006).

Proof All ξ's in $M_l(t)$ are independent, so that

$$\begin{aligned} E\big[M_l(t)\big] &= \sum_{j_1,\ldots,j_l=1}^{\infty} b_{j_1} \cdots b_{j_l} E(\xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_l}) \\ &= \mu_\xi^{l+1} \sum_{j_1,\ldots,j_l=1}^{\infty} b_{j_1} \cdots b_{j_l} = \mu_\xi \left(\mu_\xi \sum_{j=1}^{\infty} b_j \right)^l. \end{aligned}$$

Since by assumption $0 \le \mu_\xi \sum_{j=1}^{\infty} b_j < 1$, we have

$$0 \le E(Y_t) = \mu_\xi b_0 \sum_{l=0}^{\infty} \left(\mu_\xi \sum_{j=1}^{\infty} b_j \right)^l < \infty.$$

Since $Y_t \ge 0$, this also implies that Y_t is finite almost surely and hence well defined. Moreover, Y_t is clearly strictly stationary and a solution of (2.31). For uniqueness, we refer to Giraitis et al. (2000a).

The only result that remains to be proven is that the condition $\mu_{\xi^2}^{1/2} \sum_{j=1}^{\infty} b_j < 1$ implies the finiteness of $E(Y_t^2)$. First of all, note that

$$\gamma_Y(k) = cov(Y_0, Y_k) = b_0^2 \sum_{r,s=0}^{\infty} cov\big(M_r(0), M_s(k)\big) = b_0^2 \sum_{r,s=0}^{\infty} \gamma_M(k; r, s)$$

and, using the second part of (2.34),

$$\begin{aligned} \gamma_M(k; r, s) = & \sum_{\substack{j_s < \cdots < j_1 < k \\ l_r < \cdots < l_1 < 0}} b_{k-j_1} b_{j_1-j_2} \cdots b_{j_{s-1}-j_s} b_{-l_1} b_{l_1-l_2} \cdots b_{l_{r-1}-l_r} \\ & \times cov(\xi_k \xi_{j_1} \cdots \xi_{j_s}, \xi_0 \xi_{l_1} \cdots \xi_{l_r}). \end{aligned}$$

The result then follows by relatively simple but tedious algebra (for details, see Giraitis et al. 2000a). Specifically, it can be shown that

$$0 \le \gamma_Y(k) \le \mu_{\xi^2} b_0^2 \sum_{l=0}^{\infty} D^l,$$

where

$$D = \mu_{\xi^2}^{1/2} \sum_{j=1}^{\infty} b_j < 1.$$

This implies the finiteness of $\gamma_Y(k)$, $k \geq 0$. □

Note that the nonexistence of a stationary solution Y_t with finite mean under the condition $\mu_\xi \sum_{j=1}^{\infty} b_j = 1$ and $b_0 > 0$ can easily be seen by inverting the defining equation. Assume that v_t $(t \geq 0)$ is stationary and let $\mu_v = E(v_t)$. Then taking the expected value in (2.31) leads to

$$\begin{aligned} E(Y_t) &= \mu_\xi E(v_t) = \mu_\xi \mu_v \\ &= \mu_\xi \left(b_0 + \sum_{j=1}^{\infty} b_j \mu_\xi E(v_{t-j}) \right) \\ &= \mu_\xi \left(b_0 + \mu_\xi \mu_v \sum_{j=1}^{\infty} b_j \right). \end{aligned}$$

Since $\mu_\xi \neq 0$, this means

$$\left(1 - \mu_\xi \sum_{j=1}^{\infty} b_j \right) \mu_v = b_0,$$

so that $\mu_\xi \sum_{j=1}^{\infty} b_j = 1$ (and $\mu_v < \infty$) implies $b_0 = 0$, which is a contradiction to the assumption $b_0 > 0$.

Example 2.3 For the standard ARCH(∞) process in Definition 2.1, Theorem 2.2 means that a unique strictly stationary solution with finite mean $E(X_t^2)$ is given by

$$X_t^2 = b_0 \sum_{l=0}^{\infty} \sum_{j_1,\ldots,j_l=1}^{\infty} b_{j_1} \cdots b_{j_l} \varepsilon_t^2 \varepsilon_{t-j_1}^2 \cdots \varepsilon_{t-j_1-\cdots-j_l}^2,$$

whenever

$$\sigma_\varepsilon^2 = E\left(\varepsilon_t^2\right) < \infty$$

and

$$\sigma_\varepsilon^2 \sum_{j=1}^{\infty} b_j < 1.$$

Note that X_t itself is then also second-order stationary. Under the usual specification $\sigma_\varepsilon^2 = 1$, this means

$$\sum_{j=1}^{\infty} b_j < 1.$$

Moreover, the process X_t^2 is also second-order stationary if

$$E\left(\varepsilon_t^4\right) < \infty$$

and

$$\sqrt{E\left(\varepsilon_t^4\right)} \sum_{j=1}^{\infty} b_j < 1.$$

For instance, if ε_t are standard normal, then we have the conditions $\sum b_j < 1$ and $\sqrt{3}\sum b_j < 1$, or, in other words, just the condition

$$\sum_{j=1}^{\infty} b_j < \frac{1}{\sqrt{3}} \approx 0.577.$$

The second question that has to be addressed is how fast $\gamma_Y(k)$ converges to zero. Based on the following theorem, it follows that Y_t is a short-memory process:

Theorem 2.3 *If*

$$D = \mu_{\xi^2}^{1/2} \sum_{j=1}^{\infty} b_j < 1, \tag{2.39}$$

then

$$0 \le \sum_{k=-\infty}^{\infty} \gamma_Y(k) < \infty. \tag{2.40}$$

Proof An extended computation in Giraitis et al. (2000a) yields

$$\sum_{k=0}^{\infty} \gamma_M(k; r, s) \le \mu_{\xi^2} D^{r+s} (r+1)(s+1).$$

Thus,

$$\begin{aligned} \sum_{k=-\infty}^{\infty} \gamma_Y(k) &= b_0^2 \sum_{k=-\infty}^{\infty} \sum_{r,s=0}^{\infty} \gamma_M(k; r, s) \\ &\le 2\mu_{\xi^2} b_0^2 \sum_{r,s=0}^{\infty} D^{r+s} (r+1)(s+1) \end{aligned}$$

$$= 2\mu_{\xi^2} b_0^2 \left[\sum_{r=0}^{\infty} D^r (r+1) \right]^2 < \infty$$

since $0 < D < 1$. □

The exact rate at which $\gamma_Y(k)$ converges to zero is determined by the coefficients b_j as follows.

Theorem 2.4 *If*

$$D = \mu_{\xi^2}^{1/2} \sum_{j=1}^{\infty} b_j < 1 \tag{2.41}$$

and either (a) $b_j = C\varphi^j$ $(j \to \infty)$ *for some* $\varphi \in (0,1)$ *and* $C \geq 0$, *or* (b) $b_j \sim Cj^{-\alpha}$ $(j \to \infty)$ *for some* $\alpha > 1$, *then there exists a constant* $0 < C_2 < \infty$ *such that*

$$\gamma_Y(k) \leq C_2 \big[\varphi(1+C)\big]^k \tag{2.42}$$

in case (a) *and*

$$\gamma_Y(k) \sim C_2 k^{-\alpha} \tag{2.43}$$

in case (b).

Proof For the exponential case, we refer to Kokoszka and Leipus (2000). For the hyperbolic case, the proof is given in Giraitis et al. (2000a). Here, we just sketch the idea of the proof of (b) briefly. Recall that

$$\gamma_Y(k) = cov(Y_0, Y_k) = b_0^2 \sum_{r,s=0}^{\infty} \gamma_M(k; r, s).$$

The result then follows from the inequality

$$\gamma_M(k; r, s) \leq C^* k^{-\alpha} (r+1)^{\alpha+2} (s+1)^{\alpha+2} D^{r+s}$$

(and a similar lower bound) with C^* suitably chosen since then, for suitable $0 < \tilde{C}, C < \infty$,

$$\gamma_Y(k) \leq \tilde{C} k^{-\alpha} \sum_{r,s=0}^{\infty} (r+1)^{\alpha+2} (s+1)^{\alpha+2} D^{r+s}$$

$$= \tilde{C} k^{-\alpha} \left[\sum_{r=0}^{\infty} (r+1)^{\alpha+2} D^r \right]^2 \leq C k^{-\alpha}.$$

The main technical difficulty is to prove the inequality for $\gamma_M(k; r, s)$. Again we omit the elaborate though in principle not too difficult proof (see Giraitis et al. 2000a). □

The results in Theorems 2.3 and 2.4 imply that long-range dependence cannot be achieved under the given assumptions. However, one may come very close to the case of intermediate memory since α may be arbitrarily close to 1.

2.1.3.5 IARCH(∞) and FIGARCH Models

As we noted in the case of the standard ARCH(∞) model (see Example 2.3), the existence of a second-order stationary solution of

$$X_t = \sigma_t \varepsilon_t \tag{2.44}$$

and

$$\sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2, \tag{2.45}$$

with finite variance requires $\sum_{j=1}^{\infty} b_j < 1$. In particular, long memory is ruled out. If $\sum_{j=1}^{\infty} b_j = 1$, then in analogy to IGARCH processes, X_t is called an IARCH(∞) process and has necessarily an infinite variance. A particular example is the so-called FIGARCH$(0, d, 0)$ process. To motivate the definition, recall (2.28), i.e. $(I - B)X_t^2 = b_0 + Z_t$. Replacing $(I - B)$ by the fractional differencing operator $(I - B)^d$ $(d \in (0, 1/2))$, we obtain $(I - B)^d X_t^2 = b_0 + (X_t^2 - \sigma_t^2)$. Equivalently, a FIGARCH$(0, d, 0)$ process is defined as the solution of Eqs. (2.44) and

$$\sigma_t^2 = b_0 + \big(I - (I - B)^d\big) X_t^2, \tag{2.46}$$

where ε_t $(t \in \mathbb{Z})$ are i.i.d. zero-mean unit-variance random variables. A FIGARCH$(0, d, 0)$ process has the representation (2.45) with coefficients b_j defined by $\sum_{j=1}^{\infty} b_j B^j = I - (I - B)^d$. From the properties of $(1 - B)^d$ it follows that $b_j \sim c j^{-(d+1)}$ and $\sum_{j=1}^{\infty} b_j = 1$. The FIGARCH$(0, d, 0)$ model and its more general version FIGARCH(p, d, q) were introduced in Baillie et al. (1996a) without proving their existence. Sufficient conditions for the existence of a stationary solution of (2.44) and (2.45) were given in Douc et al. (2008).

Theorem 2.5 *Let* $\mu_p = E[|\xi_0|^{2p}] < \infty$ *and* $A_p = \sum_{j=1}^{\infty} b_j^p$. *If*

$$\mu_p A_p < 1,$$

then a stationary solution of (2.44)–(2.45) *exists and is given by the infinite Volterra series. Furthermore,* $E[|X_1|^{2p}] < \infty$.

We can see that for $p = 1$, Theorem 2.5 includes the statement of Theorem 2.2. We give a short proof since it is very similar to the proof of Theorem 2.2.

Proof Writing the formal Volterra expansion

$$\sigma_t^2 = b_0 \sum_{l=0}^{\infty} \sum_{j_1,\ldots,j_l=1}^{\infty} b_{j_1} \cdots b_{j_l} \varepsilon_{t-j_1}^2 \cdots \varepsilon_{t-j_1-\cdots-j_l}^2,$$

and applying the independence of the ε_t's and the inequality $(a+b)^p \le a^p + b^p$, we obtain

$$E\left[\sigma_t^{2p}\right] \le b_0^p \left[1 + \sum_{l=1}^{\infty} (\mu_p A_p)^l\right] = \frac{b_0^p}{1 - A_p \mu_p}.$$

This shows the finiteness of $E[\sigma_t^{2p}]$ and of $E[|X_t|^{2p}]$. □

Theorem 2.5 is particularly interesting for the case with $A_1 = \mu_1 = 1$ as in the IARCH(∞) and FIGARCH($0, d, 0$) models. If at the same time $A_p \mu_p < 1$ for some $p \in (0, 1)$, then Theorem 2.5 implies that a stationary solution with a finite moment of order $2p$, but necessarily an infinite variance, exists. The question is then whether and under which conditions it is possible to have $A_p \mu_p < 1$ in spite of the condition $A_1 = 1$. A partial answer is provided in Douc et al. (2008). They show that a sufficient condition is

$$\sum_{j=1}^{\infty} b_j \log(b_j) + E\left[\varepsilon^2 \log\left(\varepsilon^2\right)\right] < \infty.$$

In particular, the FIGARCH($0, d, 0$) coefficients fulfill this condition.

The next question is in which sense FIGARCH processes exhibit long-range dependence. Currently, this is still an open problem. In particular, it is not clear if d is linked in any way to long-memory properties of the sequence.

2.1.3.6 LARCH(∞) Models

As mentioned above, second-order stationary ARCH(∞) processes cannot capture long memory in volatility. Robinson (1991) introduced the so-called linear ARCH (LARCH) process defined by

$$X_t = \varepsilon_t \sigma_t, \tag{2.47}$$

$$\sigma_t = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}, \tag{2.48}$$

where ε_t are i.i.d. zero mean random variables with $\sigma_\varepsilon^2 = E(\varepsilon_t^2) = 1$. The model is again of the form (2.22), and hence $E(X_t \mid X_s, s < t) = 0$. Furthermore, X_t is a martingale difference. The essential modification compared to ARCH(∞)-processes is that σ_t instead of σ_t^2 is expressed as a linear function of X_t (instead of X_t^2). A rigorous treatment of probabilistic aspects, such as stationarity and moment assumptions,

was given in Giraitis et al. (2000c, 2004). As we will see below, the conditional variance σ_t^2 in a LARCH(∞) model may exhibit long memory, which is in contrast to ARCH(∞) models. On the other hand, σ_t can become negative, so that it may be more difficult to interpret it directly as volatility.

The first question that needs to be addressed is again whether a stationary solution exists. The following notation will be used:

$$\|b\|_p = \left(\sum_{j=1}^{\infty} |b_j|^p\right)^{\frac{1}{p}},$$

$$\mu_p = E\big[\varepsilon_t^p\big], \qquad |\mu|_p = E\big[|\varepsilon_t|^p\big],$$

where $p \in \mathbb{N}$. By repeated iteration of (2.47) the candidate for a solution can be written formally as

$$\sigma_t = b_0\left(1 + \sum_{k=1}^{\infty} \sum_{j_1,\ldots,j_k=1}^{\infty} b_{j_1} \cdots b_{j_k} \varepsilon_{t-j_1} \cdots \varepsilon_{t-j_1-\cdots-j_k}\right). \tag{2.49}$$

Equivalently,

$$\sigma_t = b_0\left(1 + \sum_{k=1}^{\infty} \sum_{j_k<\cdots<j_1=t}^{\infty} b_{t-j_1} \cdots b_{j_{k-1}-j_k} \varepsilon_{j_1} \cdots \varepsilon_{j_k}\right). \tag{2.50}$$

Whether the expression on the right-hand side is well defined in the sense of mean squared convergence is easy to check because the set $A = \{\varepsilon_{s_1} \cdots \varepsilon_{s_k} : s_1 < \cdots < s_k, k \geq 1\}$ is a an orthogonal system in $L^2(\Omega)$. Hence,

$$\operatorname{var}(\sigma_t) = b_0^2 \sum_{k=1}^{\infty} \sum_{j_1,\ldots,j_k=1}^{\infty} b_{j_1}^2 \cdots b_{j_k}^2 = b_0^2 \sum_{k=1}^{\infty} \|b\|_2^{2k} = \frac{b_0^2 \|b\|_2^2}{1 - \|b\|_2^2}.$$

Since $E[\sigma_t] = b_0$ we also have

$$E\big[\sigma_t^2\big] = \frac{b_0^2}{1 - \|b\|_2^2}. \tag{2.51}$$

This means that $\|b\|_2^2 < 1$ is a necessary and sufficient condition for the L^2-convergence of the series. By construction, σ_t defined by this Volterra expansion solves Eqs. (2.47) and (2.48). Note also that, in analogy to ARCH(∞) processes, b_0 should be assumed to be nonzero because otherwise $\sigma_t = 0$ almost surely. The main difference compared to ARCH(∞) processes is that only a condition on the summability of b_j^2 is required. Therefore, the absolute values $|b_j|$ need not be summable. This is the key to obtaining long-range dependence in volatility. Note also that the coefficients b_j need not be positive.

These results can be summarized as follows (Giraitis et al. 2004):

Theorem 2.6

(i) *A nonanticipative solution* X_t *of* (2.47) *and* (2.48) *with* $\sup_t E(X_t^2) < \infty$ *exists if and only* $\|b\|_2 < 1$. *Moreover, this solution is unique, it is given by* (2.49), *and it is strictly and second-order stationary.*

(ii) *If* $b_0 = 0$ *and* $\|b\|_2 < \infty$, *then* $\sigma_t = 0$ *a.s. is a unique solution of* (2.47) *and* (2.48).

The second question is at what rate the autocovariance functions of σ_t and X_t^2 respectively decay to zero. Here it is much easier to obtain the answer than for ARCH processes. The reason is that, since X_t are uncorrelated,

$$\sigma_t = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j} \tag{2.52}$$

is the Wold representation of σ_t. Thus,

$$\gamma_\sigma(k) = cov(\sigma_t, \sigma_{t+k}) \tag{2.53}$$

$$= \sigma_X^2 \sum_{j=1}^{\infty} b_j b_{j+k} = E\left(\sigma_t^2\right) \sum_{j=1}^{\infty} b_j b_{j+k} \tag{2.54}$$

$$= \frac{b_0^2}{1 - \|b\|_2^2} \sum_{j=1}^{\infty} b_j b_{j+k} \tag{2.55}$$

and

$$\rho_\sigma(k) = \frac{\sum_{j=1}^{\infty} b_j b_{j+k}}{\|b\|_2^2}.$$

The long-range dependence for σ_t then follows the same way as for linear processes (see Lemma 2.1):

Corollary 2.1 *Suppose that*

$$b_j \sim c_b j^{d-1} \quad (j \to \infty)$$

with $d \in (0, \frac{1}{2})$ *and* $0 < c_b < \infty$. *Then*

$$cov(\sigma_t, \sigma_{t+k}) \sim c_\sigma^2 k^{2d-1} \quad (k \to \infty),$$

where

$$c_\sigma^2 = c_b^2 E\left[\sigma_0^2\right] B(1-2d, d)$$

$$= \frac{c_b^2 b_0^2}{1 - \|b\|^2} B(1-2d, d).$$

A somewhat more involved proof shows that the long-memory property of σ_t carries over to X_t^2. More generally, Giraitis et al. (2000c) derive correlations for powers of X_t as follows:

Theorem 2.7 *Assume that* $\mu_{2p} = E[\varepsilon_t^{2p}] < \infty$ *for some* $p \in \mathbb{N}$ *and*

$$\big(4^p - 2p - 1\big)\mu_{2p}^{1/p}\|b\|_2^2 < 1.$$

Moreover, suppose that

$$b_j \sim c_b j^{d-1} \quad (j \to \infty)$$

with $d \in (0, \frac{1}{2})$, $0 < c_b < \infty$, *and let*

$$C(m) = \mu_m \frac{m E[\sigma_0^m]}{b_0} c_\sigma.$$

Then for $m = 2, \ldots, p$,

$$\gamma_{X^m}(k) = cov\big(X_t^m, X_{t+k}^m\big) \sim C^2(m)|k|^{2d-1}$$

as $k \to \infty$.

Proof The proof in Giraitis et al. (2000c) is quite involved, so that we omit details. The general idea is as follows: Setting $y_{t,m} := (\varepsilon_t^m - \mu_m)\sigma_t^m$, we have the orthogonal decomposition

$$X_t^m = \mu_m \sigma_t^m + y_{t,m}. \tag{2.56}$$

Since $X_t = \sigma_t \varepsilon_t$ and ε_t is independent of the past, we have, for $k > 0$,

$$\begin{aligned} cov(y_{t,m}, y_{t+k,m}) &= E\big[\big(\varepsilon_t^m - \mu_m\big)\sigma_t^m\big(\varepsilon_{t+k}^m - \mu_m\big)\sigma_{t+k}^m\big] \\ &= E\big[\big(\varepsilon_t^m - \mu_m\big)\sigma_t^m\sigma_{t+k}^m\big]E\big[\varepsilon_{t+k}^m - \mu_m\big] = 0 \end{aligned}$$

and also

$$\begin{aligned} cov\big(\sigma_t^m, y_{t+k,m}\big) &= E\big[\sigma_t^m\big(\varepsilon_{t+k}^m - \mu_m\big)\sigma_{t+k}^m\big] \\ &= E\big[\sigma_t^m\sigma_{t+k}^m\big]E\big[\varepsilon_{t+k}^m - \mu_m\big] = 0. \end{aligned}$$

This leads to the decomposition

$$\begin{aligned} cov\big(X_t^m, X_{t+k}^m\big) &= cov\big(\mu_m\sigma_t^m + y_{t,m}, \mu_m\sigma_{t+k}^m + y_{t+k,m}\big) \\ &= \mu_m^2 cov\big(\sigma_t^m, \sigma_{t+k}^m\big) + \mu_m cov\big(y_{t,m}, \sigma_{t+k}^m\big). \end{aligned}$$

Under the assumption that $b_j \sim c_b j^{d-1}$ $(j \to \infty)$, it can then be shown that, as k tends to infinity,

$$cov\big(\sigma_{t+k}^m, y_{t,m}\big) = o\big(k^{2d-1}\big)$$

and that

$$cov\left(\sigma_t^m - \frac{mE[\sigma_0^m]}{b_0}\sigma_t, \sigma_{t+k}^m - \frac{mE[\sigma_0^m]}{b_0}\sigma_{t+k}\right) = o\left(k^{2d-1}\right). \tag{2.57}$$

From (2.57) one then concludes that, as $k \to \infty$,

$$cov\left(\sigma_t^m, \sigma_{t+k}^m\right) \sim \left(\frac{mE[\sigma_0^m]}{b_0}\right)^2 cov(\sigma_t, \sigma_{t+k}).$$

Applying Corollary 2.1 yields

$$cov\left(X_t^m, X_{t+k}^m\right) \sim \mu_m^2 \left(\frac{mE[\sigma_0^m]}{b_0}\right)^2 c_\sigma k^{2d-1}. \qquad \square$$

This result is quite remarkable since the asymptotic rate at which autocorrelations of X_t^m decay does *not* depend on m, only the constant changes. This is very much in contrast to results on nonlinear transformations applied to linear processes (see e.g. Corollary 3.6). Note also that the condition $(4^p - 2p - 1)\mu_{2p}^{1/p}\|b\|_2^2 < 1$ makes sure that the first $2p$ moments of X_t exist. A more general result on the existence of moments with weaker assumptions is given in Giraitis et al. (2004). The sufficient conditions used in their proofs are the following:

Condition 2.1 (M_3) $|\mu|_3 < \infty$ *and*

$$|\mu|_3^{1/3}\|b\|_3 + 3\theta\|b\|_2 < 1,$$

where θ *is such that*

$$3\theta^2 - 3\theta - 1 = 0.$$

Condition 2.2 (M_{2p}) $|\mu|_{2p} < \infty$ *and*

$$\sum_{j=2}^{2p} \binom{2p}{j} \|b\|_j^j |\mu|_j < 1.$$

Theorem 2.8 *Suppose that* (M_r) *holds where either* $r = 3$ *or* $r = 2p$ *for* $p \geq 2$. *Then*

$$E\left[|\sigma_t|^r\right] < \infty, \qquad E\left[|X_t|^r\right] < \infty.$$

With increasing r the conditions (M_r) imply stronger restrictions on the coefficients b_j. In the original derivation of the strictly and second-order stationary solution, only $E(\varepsilon_t^2) < \infty$ and $\|b\|_2 < 1$ were assumed. Thus, no assumption that links b_j and the moments of ε_t is needed. This is not the case for higher moments.

Somewhat simple but much stronger conditions that imply (M_r) can be given as follows. For (M_3), this is

$$(\tilde{M}_3) \quad \|b\|_2 < \frac{1}{|\mu|_3^{1/3} + 3.81}.$$

Condition (M_4) follows from

$$(\tilde{M}_4) \quad \|b\|_2 < \frac{1}{\sqrt{|\mu|_4 + 4|\mu_3| + 6}},$$

and for $p \geq 3$, one may impose the sufficient condition

$$(\tilde{M}_{2p}) \quad \|b\|_2 < \frac{1}{\sqrt{\sum_{j=2}^{2p} \binom{2p}{j} |\mu_j|}}.$$

To show that these conditions imply the previous ones, one observes first that $\theta \approx 1.27$. This implies $\|b\|_r < 1$ (and hence $\|b\|_r^k \leq \|b\|_r$ for $k \geq 1$) for each of the norms involved in the inequalities and $\|b\|_r \leq \|b\|_2$. Since the right-hand side of these inequalities is smaller than one, these are much more restrictive assumptions than the initial inequality $\|b\|_2 < 1$. It should be noted at the same time that the conditions linking moments of ε_t and the coefficients b_j do not restrict the range of possible rates at which b_j converges to zero. The reason is that as long as the norm $\|b\|_r$ is finite, it can be can be made arbitrarily small by multiplying b_j with a suitable constant.

Example 2.4 Let ε_t be i.i.d. $N(0,1)$ distributed. Then $\mu_{2k+1} = 0$, $\mu_{2k} = (2k-1)(2k-3)\cdots 1$ $(k \geq 1)$ and $|\mu|_3 = \sqrt{8/\pi}$. Condition $(\tilde{M}_3)$ can then be written as

$$\|b\|_2 < \frac{1}{(8/\pi)^{\frac{1}{6}} + 3.81} \approx 0.2008.$$

Consider, for instance, $b_j = j^{-2/3} = j^{d-1}$ with $d = 1/3$. Then

$$\|b\|_2 = \sqrt{\sum_{j=1}^{\infty} j^{-4/3}} \approx 1.8976.$$

Thus, in order that a stationary solution with a finite second moment exists, we need to divide b_j at least by a factor of about 1.9. This result is independent of the distribution of ε_t. On the other hand, if we want that the third moment of X_t is finite and we know that ε_t are $N(0,1)$ distributed, then we need to divide by a factor

$$c > 1.8976/0.2008 \approx 9.45.$$

Obviously this is a much stronger restriction.

It can be shown that the moment conditions (M_{2p}) are weaker than the condition $(4^p - 2p - 1)\mu_{2p}^{1/p}\|b\|_2^2 < 1$ used in Theorem 2.7. (It may be conjectured that (M_{2p}) is sufficient to establish the decay of covariances in Theorem 2.7.)

2.1.3.7 LARCH$_+(\infty)$ Processes

As mentioned above, the 'volatility' σ_t in the LARCH(∞) process is not necessarily positive. Since one would like to interpret σ_t as a standard deviation, various suggestions how to make σ_t positive have been discussed in the literature. Here, we describe the approach proposed by Surgailis (2008). Recall that a LARCH process can be written as

$$X_t = \varepsilon_t \sigma_t = b_0 \varepsilon_t + \varepsilon_t \sum_{j=1}^{\infty} b_j X_{t-j},$$

where ε_t $(t \in \mathbb{Z})$ are i.i.d. zero mean random variables with unit variance. Consider now two mutually independent i.i.d. sequences η_t and ξ_t $(t \in \mathbb{Z})$ with zero mean and unit variance, and modify X_t as follows:

$$X_t = b_0 \eta_t + \xi_t \sum_{j=1}^{\infty} b_j X_{t-j} =: b_0 \eta_t + \xi_t A_t. \tag{2.58}$$

More generally, one may also include the possibility of a correlation $\rho = cor(\xi_t, \eta_t)$ between ξ_t and η_t (see Surgailis 2008). Note that for $\rho = 1$, one is back to the original LARCH model. Here, we focus on the simpler case with $\rho = 0$. Note that it is not clear immediately that X_t can be written in the "standard" volatility form $X_t = \sigma_t \varepsilon_t$. This will be shown below.

To derive the stationary solution, we write the formal Volterra expansion

$$X_t = b_0 \left(\eta_t + \xi_t \sum_{k=1} \sum_{j_1, \ldots, j_k} b_{j_1} \cdots b_{j_k} \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_{k-1}} \eta_{t-j_1-\cdots-j_k} \right).$$

This expansion implies immediately that X_t are uncorrelated and $E(X_t) = 0$. Furthermore, since $E(\eta_1^2) = E(\xi_1^2) = 1$, we obtain

$$\mathrm{var}(X_t) = b_0^2 + b_0^2 \frac{\|b\|_2^2}{1 - \|b\|_2^2}.$$

Again, as in the standard LARCH(∞) case, it can be shown that $\|b\|_2 < 1$ is necessary for the existence of a unique second-order stationary solution.

Let $\mathcal{F}_t$ be the sigma field generated by ξ_s, η_s $(s \le t)$. Now, we will show that X_t in (2.58) can be written as $X_t = \sigma_t \varepsilon_t$, where σ_t is $\mathcal{F}_{t-1}$-measurable, and ε_t $(t \in \mathbb{Z})$ is a martingale difference with respect to $\mathcal{F}_t$. For a moment, we do not impose any

form on σ_t, except for measurability. For uniqueness, we will impose the additional condition $E[\varepsilon_t^2|\mathcal{F}_s, s<t]=1$. Define

$$\varepsilon_t = \frac{b_0\eta_t + \xi_t A_t}{\sigma_t}.$$

Since (ξ_t, η_t) is independent of $\mathcal{F}_{t-1}$, we have the martingale difference property

$$E[\varepsilon_t \mid \mathcal{F}_{t-1}] = \sigma_t^{-1}\big(b_0 E[\eta_t] + A_t E[\xi_t]\big) = 0.$$

Furthermore,

$$\begin{aligned} E\big[\varepsilon_t^2 \,\big|\, \mathcal{F}_s, s<t\big] &= E\left[\frac{b_0^2\eta_t^2}{\sigma_t^2}\,\middle|\,\mathcal{F}_s, s<t\right] \\ &\quad + 2E\left[\frac{b_0\eta_t\xi_t A_t}{\sigma_t^2}\,\middle|\,\mathcal{F}_s, s<t\right] + E\left[\frac{\xi_t^2 A_t^2}{\sigma_t^2}\,\middle|\,\mathcal{F}_s, s<t\right] \\ &= \frac{b_0^2 + A_t^2}{\sigma_t^2}, \end{aligned}$$

so that the imposed condition $E[\varepsilon_t^2|\mathcal{F}_s, s<t]=1$ yields

$$\sigma_t^2 = b_0^2 + A_t^2$$

(which is clearly measurable with respect to $\mathcal{F}_{t-1}$). We summarize these findings in the following theorem, which is a simplified version of Surgailis (2008).

Theorem 2.9 *Assume that $\sum_{j=1}^{\infty} b_j^2 < 1$. Then X_t in (2.58) has the unique strictly and second-order stationary solution given by*

$$X_t = \sigma_t \varepsilon_t,$$

where

$$\sigma_t = \sqrt{b_0^2 + A_t^2}$$

and

$$\varepsilon_t = \frac{b_0\eta_t + \xi_t A_t}{\sigma_t}$$

with

$$A_t = \sum_{j=1}^{\infty} b_j X_{t-j},$$

and ε_t $(t \in \mathbb{Z})$ is a martingale difference with respect to $\mathcal{F}_t$ such that $E[\varepsilon_t^2|\mathcal{F}_s, s<t]=1$.

Thus, the LARCH$_+(\infty)$ process can be written in the form (2.22), with $\sigma_t > 0$ interpretable directly as the conditional standard deviation of X_t. However, in contrast to the LARCH(∞) model, the conditional variance σ_t^2 is not defined explicitly. Instead, it follows implicitly from the construction of the model. Moreover, the random variables ε_t $(t \in \mathbb{Z})$ are no longer i.i.d. Note finally that the advantage of the LARCH$_+(\infty)$ model is that one can model heavy tails by choosing η_t to be regularly varying.

2.1.3.8 SV (Stochastic Volatility) Models

The mathematical difficulty with defining volatility models by recursive equations such as (2.23), (2.30) or (2.48) is that it is not clear a priori whether a solution (in particular a stationary solution) exists. Moreover, it is difficult to design recursive models with long memory. An alternative approach where existence is much easier to show and long memory is easy to generate is to define a process explicitly as a function of existing processes. This may be done, for instance, as follows.

Definition 2.3 Let ε_t $(t \in \mathbb{Z})$ be a sequence of i.i.d. random variables with $E(\varepsilon_t) = 0$, independent of a stationary sequence σ_t $(t \in \mathbb{Z})$. Then

$$X_t = \sigma_t \varepsilon_t$$

is called a stochastic volatility (SV) model. If σ_t is a long-memory process, then X_t $(t \in \mathbb{Z})$ is called a long-memory stochastic volatility model (LMSV).

To allow for more generality, we may consider SV models with leverage.

Definition 2.4 Let ε_t $(t \in \mathbb{Z})$ be a sequence of i.i.d. random variables with $E(\varepsilon_t) = 0$ and σ_t $(t \in \mathbb{Z})$ a stationary sequence. Moreover, assume that ε_t is independent of $\{\sigma_s, s \le t\}$. Then

$$X_t = \sigma_t \varepsilon_t$$

is called a stochastic volatility (SV) model with leverage.

As we mentioned in the introduction to this section, there is no consensus on what is meant by a "stochastic volatility" model. In time series analysis all models of the form $X_t = \sigma_t \varepsilon_t$ are called informally volatility models. In particular, a GARCH(p, q) process is a volatility process. In financial mathematics (and in particular in option pricing) "stochastic volatility" rather means the presence of two independent "noise components"; one being the noise sequence ε_t and the other one coming in via the definition of σ_t. This is the case for instance in Definition 2.3. On the other hand, if dependence between sequences σ_t and ε_t is introduced, then we have a leverage effect, that is, a dependence between previous returns X_{t-1} and future volatilities σ_t. In particular, Definition 2.3 excludes ARCH(∞) processes, whereas Definition 2.4 is very general and includes Definition 2.3 as well as

ARCH(∞) models. However, this becomes clear only after it has been established that a solution of the ARCH(∞)-equation exists such that σ_t has the properties above.

The point of Definitions 2.3 and 2.4 is that we may start with any i.i.d. sequence ε_t and any stationary process σ_t. Thus, if we have such processes already, then the existence of X_t is guaranteed. For instance, a simple explicit model is obtained by setting $\sigma_t = V(\zeta_t)$ where V is a positive function and $\zeta_t = \sum_{j=1}^{\infty} a_j \eta_{t-j}$ is a linear process with (η_t, ε_t) $(t \in \mathbb{Z})$ being a sequence of i.i.d. random vectors. In particular, if $V(x) = \exp(x)$ and $\eta_t = g(\varepsilon_t)$ with a deterministic function g, then the model is called an EGARCH model. The letter "E" stands for exponential (see Nelson 1990). Its long-memory modification, with a_j being FARIMA(p, d, q) weights, the so-called Fractionally Integrated Exponential GARCH (FIEGARCH) model, was introduced in Bollerslev and Mikkelsen (1996). The special case where ζ_t $(t \in \mathbb{Z})$ is Gaussian with long memory and independent of ε_t $(t \in \mathbb{Z})$ was considered in Breidt et al. (1998). They called the process a long-memory stochastic volatility model (LMSV), as we did in Definition 2.3.

Due to the simple explicit form, it is relatively easy to characterize the dependence structure of SV models with or without leverage. First, from the definitions it is obvious that X_t is a martingale difference. For instance, if $v_\varepsilon = E(\varepsilon_t^2) < \infty$ and $v_\sigma = E(\sigma_t^2) < \infty$, then, for $k \geq 1$,

$$\begin{aligned}\gamma_X(k) &= E[\varepsilon_t \varepsilon_{t+k} \sigma_t \sigma_{t+k}] = E\big[E(\varepsilon_t \varepsilon_{t+k} \sigma_t \sigma_{t+k} | \varepsilon_s, \sigma_s, s \leq t)\big] \\ &= E[\varepsilon_{t+k}] E[E\varepsilon_t \sigma_t \sigma_{t+k}] = 0.\end{aligned}$$

Moreover, if $E(\varepsilon_t^4) < \infty$, then

$$\begin{aligned}E\big(X_t^2 X_{t+k}^2\big) &= v_\varepsilon E\big(\varepsilon_t^2 \sigma_t^2 \sigma_{t+k}^2\big), \\ E\big(X_t^2\big) &= v_\varepsilon E\big(\sigma_t^2\big),\end{aligned}$$

and

$$\gamma_{X^2}(k) = v_\varepsilon \big[E\big(\varepsilon_t^2 \sigma_t^2 \sigma_{t+k}^2\big) - v_\varepsilon E^2\big(\sigma_t^2\big)\big]. \tag{2.59}$$

In particular, if X_t is an SV model as in Definition 2.3, then

$$\gamma_{X^2}(k) = v_\varepsilon^2 cov\big(\sigma_0^2, \sigma_k^2\big). \tag{2.60}$$

To obtain more explicit results, one needs to assume a more specific structure of σ_t. For instance, if $\sigma_t = V(\zeta_t) = \exp(\zeta_t)$, $E(\eta_t^2) < \infty$ and $a_j = L_a(j) j^{d-1}$ $(0 < d < \frac{1}{2})$, then the long-memory property of the linear process ζ_t is inherited by σ_t, σ_t^2 and X_t^2, because the exponential function has Appell rank 1 (this is explained in more detail in Sects. 3.3 and 4.2.5). For the volatility component of X_t, we obtain

$$\begin{aligned}\gamma_\sigma(k) &= cov\big(\exp(\zeta_t), \exp(\zeta_{t+k})\big), \\ \gamma_{\sigma^2}(k) &= cov\big(\exp(2\zeta_t), \exp(2\zeta_{t+k})\big),\end{aligned}$$

so that, as $k \to \infty$,

$$\gamma_\sigma(k) \sim \text{const} \cdot \gamma_\zeta(k) \sim \text{const} \cdot L_a^2(k) k^{2d-1},$$
$$\gamma_{\sigma^2}(k) \sim \text{const} \cdot \gamma_\zeta(k) \sim \text{const} \cdot L_a^2(k) k^{2d-1}.$$

Combining the last approximations with (2.60), for the LMSV model, we obtain

$$\gamma_{X^2}(k) \sim \text{const} \cdot L_a^2(k) k^{2d-1}. \tag{2.61}$$

In the case of an SV model with leverage, (2.59) and the result for $\gamma_{\sigma^2}(k)$ do not yield (2.61) immediately. Nevertheless, the asymptotic formula for $\gamma_{X^2}(k)$ is still valid, as shown in Harvey (1998) and Surgailis and Viano (2002).

Both Definitions 2.3 and 2.4 of stochastic volatility models also allow for modelling heavy tails in X_t (and hence also in X_t^2) by using heavy-tailed innovations ε_t, i.e.

$$F_\varepsilon(-x) = P(\varepsilon \le -x) \sim (1-p)x^{-\alpha}, \qquad \bar{F}_\varepsilon(x) = P(\varepsilon > x) \sim px^{-\alpha} \quad (x \to \infty), \tag{2.62}$$

where $p \in (0, 1)$, and the tail index α is in the interval $(1, \infty)$. Since the mean exists, we may use the assumption $E(\varepsilon) = 0$ (see also the discussion at the beginning of Sect. 2.1.2). If the distribution of σ_t has lighter tails such that $E(\sigma^{\alpha+\delta}) < \infty$ for some $\delta > 0$, then the process X_t inherits the tail index from ε_t, i.e. (2.62) holds for X_t with the same value of α as for ε_t. More specifically, we have by Breiman's lemma (see Resnick 2007, Proposition 7.5) that, as $x \to \infty$,

$$P(X_t > x) \sim pE\big(\sigma^\alpha\big)x^{-\alpha}, \qquad P(X_t < -x) \sim (1-p)E\big(\sigma^\alpha\big)x^{-\alpha}. \tag{2.63}$$

Such heavy-tailed SV models were considered in Davis and Mikosch (2001) and Kulik and Soulier (2011, 2012, 2013).

2.1.3.9 FARIMA Processes with GARCH Innovations (FARIMA-GARCH)

Linear long memory can also be combined with dependence in volatility. For instance, Beran and Feng (2001a) consider FARIMA-GARCH models

$$(1-B)^d \varphi(B)\big(Y_i - g(t_i)\big) = \psi(B)e_i,$$

where e_i $(i \in \mathbb{Z})$ is a stationary GARCH process, $t_i = i/n$, and $g(t)$ is a nonparametric trend function (also see Ling and Li 1997 for a similar model and Baillie et al. 1996b for some applications in finance). Giraitis and Surgailis (2002) define a class of bilinear ARCH-type models with the possibility of having linear long-range dependence as well as long memory in volatility.

2.1.3.10 Multivariate Extensions

The extension of volatility models with long-range dependence to multivariate time series is very important for financial applications. There is therefore a rapidly growing econometric literature on multivariate fractional volatility models. For recent references, see e.g. Kirman and Teyssière (2002), Chiriac and Voev (2010), Fleming and Kirby (2011), Bollerslev et al. (2012).

2.1.4 Counting Processes

2.1.4.1 Introduction

Assume that τ_j ($j \in \mathbb{Z}$) is a strictly increasing sequence such that $\tau_{-1} < 0 \le \tau_0 < \tau_1$. This sequence can be interpreted as arrival times of a customer to a queueing system, moments of claims from an insurance policy, initiation times of transmissions from a source, etc. The increments $X_j = \tau_j - \tau_{j-1}$ ($j \in \mathbb{Z}$) are often called interarrival times or interpoint distances. By definition, X_j are strictly positive. If X_j are i.i.d. with common marginal distribution $F(x) = P(X_1 \le x)$, then X_j ($j \in \mathbb{Z}$) is called a renewal sequence (on the real line), and τ_j are called renewal epochs. We will assume that $\mu = E[X_1] < \infty$. By definition, the distribution of $\tau_0 - \tau_{-1} = X_0$ is the same as that of $X_1, X_2, \ldots$. The associated counting process is defined by

$$N(t) = \#\{j : \tau_j \le t\} = \sum_{j=0}^{\infty} 1\{\tau_j \in [0,t]\} \quad (t \ge 0).$$

In other words, $N(t) = 0$ if $\tau_0 > t$ and $N(t) = k$ ($k \ge 1$) if $\tau_{k-1} \le t < \tau_k$. Recall that the counting process is stationary if for each collection of Borel sets $B_1, \ldots, B_k$ and any $t \ge 0$, the distribution of $(N(B_1 + t), \ldots, N(B_k + t))$ does not depend on t. Here, $B + t = \{x + t : x \in B\}$, and $N(B)$ counts number of points τ_j in the set B. This can be achieved by placing "0" uniformly between τ_{-1} and τ_0. The resulting distribution of the distance between 0 and τ_0 is given by

$$\begin{aligned} P(\tau_0 > x) &= \frac{1}{\mu} \int_x^{\infty} \bar{F}(u)\, du = \frac{1}{\mu} \int_x^{\infty} \big(1 - F(u)\big)\, du \\ &=: \bar{F}^{(0)}(x) = 1 - F^{(0)}(x). \end{aligned}$$

Unless stated otherwise, we will use the term "renewal process" quite loosely, to describe either the interpoint distances X_j, the renewal epochs τ_j, or the counting process $N(t)$.

Alternatively, if one wants to start with the definition of renewal epochs on $[0, \infty)$, one can define the renewal epochs sequence as $\tau_0 \sim F^{(0)}$, where $F^{(0)}$ is

an initial distribution, and

$$\tau_j = \tau_0 + \sum_{k=1}^{j} X_k$$

with X_j ($j \in \mathbb{N}$) being i.i.d. with common distribution F. A similar discussion is applicable to any stationary sequence X_j of strictly positive random variables with finite mean. The associated sequence τ_j ($j \in \mathbb{Z}$) is referred to as points of a stationary point process. Moreover, the stationary renewal process N is associated with two renewal functions U, $\tilde{U}$ defined by

$$U(t) = 1 + \sum_{k=1}^{\infty} F^{k*}(t)$$

with $F^{k*}(t) = P(X_1 + \cdots + X_k \le t)$ denoting the kth convolution of the distribution F, and

$$\tilde{U}(t) = \sum_{k=0}^{\infty} P(\tau_k \le t) = E\big[N(t)\big] = \mu^{-1} t =: \lambda t.$$

For $U(t)$, one has $U(t)/t \to \lambda$ as $t \to \infty$, but $U(t) = \lambda t = \tilde{U}(t)$ for all t holds only if $N(t)$ is a Poisson process with rate λ. Furthermore, $N(t)/t \to \lambda$ in probability as $t \to \infty$. The quantity λ is called the intensity or rate of the renewal process N. For any stationary point process, we have $E[N(t)] = \lambda t$. For more details on renewal theory, see e.g. Resnick (1992), Chap. 3, and Daley and Vere-Jones (1988, 2007).

2.1.4.2 Long Memory in Counts

Consider a stationary point process τ_j ($j \in \mathbb{Z}$). If the associated counting process N is LRD in the sense of Definition 1.6, then following Daley and Vesilo (1997), we say that the process has Long-Range count Dependence (LRcD). Here, "c" stands for "count".

Recall that for a stationary point process, we have $E[N(t)] = \lambda t$. Therefore, LRcD is equivalent to

$$\lim_{t \to \infty} \frac{\mathrm{var}(N(t))}{E[N(t)]} = +\infty.$$

The ratio $V/E = \mathrm{var}(N(t))/E[N(t)]$ is often referred to as the *index of dispersion* Cox and Isham (1980), *Fano Factor* or normalized variance (Lowen and Teich 2005, p. 66).

It is usually difficult to establish the LRcD property of the counting process associated with a stationary sequence X_j. The following theorem gives necessary and sufficient conditions for a stationary renewal process to have the LRcD property.

Note that this long memory does not have anything to do with *dependence* properties of the underlying renewal sequence of interpoint distances X_j. Instead, long memory in counts is generated by heavy tails of F, the distribution of X_j.

Theorem 2.10 *A stationary renewal process N is LRcD if and only if* $\operatorname{var}(X_1) = +\infty$.

Proof We present the proof of sufficiency only. The complete proof can be found in Daley (1999). Implicitly it was established in Feller (1971, pp. 331–332). Let X be a generic random variable of a sequence of interpoint distances. The finiteness of the second moment of the positive random variable X can be described by a condition on F^0 or F. Indeed,

$$\begin{aligned} E\big[X^2\big] = +\infty \quad &\Longleftrightarrow \quad \int_0^\infty \bar{F}^0(t)\,dt = +\infty \\ &\Longleftrightarrow \quad \int_0^\infty t\bar{F}(t)\,dt = +\infty. \end{aligned} \tag{2.64}$$

Recall the renewal function $U(t) = 1 + \sum_{k=1}^\infty F^{k*}(t)$. Now, the variance function of $N(t)$ fulfills

$$\operatorname{var}\big(N(t)\big) = \lambda \int_0^t \big(2\big(U(s) - \lambda s\big) + 1\big)\,ds,$$

where $U(t)$ is the renewal function (Daley and Vere-Jones 1988, p. 77). The function

$$\tilde{V}(t) = \operatorname{var}\big(N(t)\big) + \lambda t = 2\lambda \int_0^t \big(U(s) - \lambda s\big)\,ds$$

fulfills the renewal equation

$$\tilde{V}(t) = \int_0^t ds \int_s^\infty \bar{F}(v)\,dv + \int_0^t \tilde{V}(t-u)\,dF(u).$$

Appealing to the solution of the general renewal equation and using $U(t) \geq \mu^{-1}t$, we obtain

$$\begin{aligned} \tilde{V}(t) &= 2\lambda^2 \int_0^t dU(u) \int_0^{t-u} ds \int_s^\infty \bar{F}(v)\,dv \\ &= 2\lambda^2 \int_0^t ds \int_0^{t-s} dU(u) \int_s^\infty \bar{F}(v)\,dv \\ &= 2\lambda^2 \int_0^t U(t-s)\,ds \int_s^\infty \bar{F}(v)\,dv \\ &= 2\lambda^2 \int_0^t U(s)\,ds \int_{t-s}^\infty \bar{F}(v)\,dv \end{aligned}$$

$$= 2\lambda^2 \int_0^\infty \bar{F}(v)\,dv \int_{(t-v)_+}^t U(s)\,ds$$
$$\geq \lambda^3 \left[\int_0^t v(2t-v)\bar{F}(v)\,dv + t^2 \int_t^\infty \bar{F}(v)\,dv \right]. \tag{2.65}$$

Therefore, dividing both sides of (2.65) by t, we have

$$\tilde{V}(t)/t = \lambda^3 \left[\int_0^t v\left(2 - \frac{v}{t}\right)\bar{F}(v)\,dv + t \int_t^\infty \bar{F}(v)\,dv \right]$$
$$\geq \lambda^3 \left[\int_0^t v\bar{F}(v)\,dv + t \int_t^\infty \bar{F}(v)\,dv \right]$$

because $0 < v/t < 1$. Letting $t \to +\infty$, the right-hand side of the inequality tends to $+\infty$ if $E[X_1^2] = +\infty$. □

Example 2.5 Consider a stationary renewal process such that $P(X_1 > x) = x^{-\alpha}L(x)$, $\alpha \in (1,2)$. Then

$$U(x) - x/\mu \sim \frac{x^{2-\alpha}L(x)}{\mu^2(\alpha-1)(2-\alpha)},$$

see Teugels (1968) or Daley and Vesilo (1997). Using the representation

$$\tilde{V}(t) = \mathrm{Var}\big[N(t)\big] + \lambda t = 2\lambda \int_0^t \big(U(s) - \lambda s\big)\,ds$$

and Lemma 1.1, we conclude

$$\tilde{V}(t) \sim V(t) \sim 2\lambda^3 \frac{t^{3-\alpha}L(t)}{(\alpha-1)(2-\alpha)(3-\alpha)} = 2\lambda^3 \frac{t^{2H}L(t)}{(\alpha-1)(2-\alpha)(3-\alpha)}.$$

Thus, the renewal process is LRcD with Hurst parameter $H = (3-\alpha)/2 \in (1/2, 1)$.

The theorem can be extended to counting processes such that the interpoint distances have some (weak) dependence structure together with some monotonicity properties (see Kulik and Szekli 2001).

Theorem 2.10 implies that one way of generating long memory in a counting process is via heavy tails of interarrival distances. On the other hand, LRcD can be generated by long memory of interpoint distances. This is illustrated by the following example.

Example 2.6 (Long-Memory Markov Chain) Daley et al. (2000) considered the following long-range dependent sequence. Let $\{v_n\}_{n\geq 1}$ be an increasing sequence of

positive real numbers and $\{\pi_n\}_{n\geq 1}$ a probability sequence. For the transition probability matrix

$$q_{ij} = \begin{cases} 1 & \text{if } j = i-1, i \geq 2, \\ 1-p & \text{if } j \neq i, i \geq 2, \ p \in (0,1), \\ \pi_j & \text{otherwise,} \end{cases}$$

consider a stationary Markov chain $\{J_n\}_{n\geq 1}$ with the stationary distribution $\pi_i = P(J_1 = i)$, $i \geq 1$. Then X_j $(j \geq 1)$ defined by $X_j = v_{J_j}$ forms a stationary sequence. Under suitable conditions on v_n and π_n, the authors showed that X_j has a finite variance and non-summable covariances. Therefore the sequence is LRD. Furthermore, they showed that the associated counting process has the LRcD property.

Example 2.7 (Long-Memory Stochastic Duration) Consider two independent sequences ε_j, σ_j $(j \in \mathbb{Z})$ where ε_j is an uncorrelated process. We assume that ε_j are strictly positive, so that $E[\varepsilon_0] > 0$. Then $X_j = \varepsilon_j \sigma_j$ inherits the dependence structure from σ_j since

$$cov(X_0, X_k) = E(X_0 X_k) - E(X_0)E(X_k) = E^2[\varepsilon_0] cov(\sigma_0, \sigma_k).$$

Assume for instance that $\sigma(y) = \exp(y)$ and

$$\sigma_j = \sigma(\zeta_j), \qquad \zeta_j = \sum_{k=1}^{\infty} a_k \xi_{j-k},$$

where ζ_j is a long-memory Gaussian process with $a_k \sim c_a k^{d-1}$. The model $X_j = \varepsilon_j \exp(\zeta_j)$ was introduced in Deo et al. (2007). The sequence X_j $(j \in \mathbb{Z})$ has long memory that propagates to the counting process $N(t)$. It was shown in Deo et al. (2009) that $\text{var}(N(t)) \sim Ct^{2d+1}$.

It should be mentioned that the asymptotic behaviour of the variance of the counting process in Example 2.7 was not obtained from a direct computation (which may be impossible), but rather from the limiting behaviour of $N(t)$. More generally, if the partial sums $\sum_{j=1}^{[nt]} X_j$ converge to a fractional Brownian motion, then the associated counting process also converges weakly to fBm (see Sect. 4.2.6). Under additional conditions (such as uniform integrability), one can derive the behaviour of $\text{var}(N(t))$ from $\text{var}(B_H(t))$.

It should be noted that this approach does not work for the LRcD renewal process described above. There, the partial sums $\sum_{j=1}^{[nt]} X_j$ converge weakly to a Lévy process, which implies the (finite-dimensional) convergence of the counting process to a Lévy process. Since the limiting Lévy process has infinite variance, we cannot conclude anything about $\text{var}(N(t))$.

So far we saw that LRcD can be generated either by heavy tails or by long memory of interpoint distances. There are however also other possibilities.

Example 2.8 (Mixed Poisson Process) The counting process N is a mixed Poisson process if

$$P\left(\bigcap_{i=1}^{n}\{N(t_i)=k_i\}\right)=\int_0^\infty \prod_{i=1}^{n}\exp(-\lambda t_i)\frac{(\lambda t_i)^{k_i}}{k_i!}\,dG(\lambda), \tag{2.66}$$

where G is a distribution of a strictly positive random variable Λ. We have $E[N(t)|\Lambda=\lambda]=\lambda t$ and $\mathrm{Var}[N(t)]=\mathrm{Var}[\Lambda]t^2+E[\Lambda]t$. This implies that this process is always LRcD if G is not trivial. Note that, in contrast to the renewal case, in this process the interpoint distances can have a finite or infinite second moment, depending on the choice of G.

Example 2.9 (Cox Processes) Let $\Lambda(t)$ $(t\geq 0)$ be a stochastic process with absolutely continuous trajectories, that is,

$$\Lambda(t)=\int_0^t \lambda(u)\,du,$$

where $\lambda(u)$ $(u\geq 0)$ is a stationary, nonnegative random process, called *intensity process*. Suppose, that N is a doubly stochastic Poisson (Cox) process driven by $\Lambda(t)$ $(t\geq 0)$. Recall (Daley and Vere-Jones 1988, p. 263) that

$$\mathrm{Var}\big[N(t)\big]=E\big[\Lambda(t)\big]+\mathrm{Var}\big[\Lambda(t)\big].$$

Therefore, N is LRcD if and only if Λ is long-range dependent in the sense of Definition 1.6.

2.2 Physical Models

2.2.1 Temporal Aggregation

Often observed time series are temporal aggregations of (observable or hidden) data generated on a finer time scale. The original time scale may be continuous, and we observe (or decide to look) at discrete time points only. Typical examples are daily average temperatures, monthly average discharge series of a river and so on. In other cases, the original time scale is discrete, but we observe an aggregate at an even lower time resolution. For instance, many economic variables reported on a monthly basis are obtained by suitable averaging of daily data. Now, suppose for instance that the original process X_i $(i\in\mathbb{Z})$ is stationary and ergodic, and we observe instead the aggregate

$$Y_{i,M}=\sum_{j=(i-1)s+1}^{iM} X_j \quad (i=1,2,\ldots,n) \tag{2.67}$$

for some $M \in \mathbb{N}$. Note that, in this notation, a time interval of length k for the aggregated process $Y_{i,M}$ corresponds to a time interval of length kM on the original time axis. Defining $N = nM$ and partial sums

$$S_N(t) = \sum_{j=1}^{[Nt]} X_j,$$

$Y_{i,M}$ can also be written as

$$Y_{i,M} = S_N(t_i) - S_N(t_{i-1})$$

with $t_i = iM/N$. Now, from the previous chapter we know that the only limit the standardized process

$$Z_N(t) = \frac{S_N(t) - E[S_N(t)]}{\sqrt{\text{var}(S_N(1))}}$$

can converge to is a self-similar process $Z(t)$ (Lamperti 1962, 1972). Also, unless X_i is almost surely constant, the only scaling that is possible is $\text{var}(S_N(1)) \sim L(n)n^{2H}$, where $L(n)$ is a slowly varying function and $0 < H < 1$. If the limit is Gaussian, then $Z(t)$ is necessarily a fractional Brownian motion B_H. One may thus say that self-similar processes play the same fundamental role in statistical inference for stochastic processes, as the normal distribution or more generally infinitely divisible distributions play in inference for (marginal) distributions of random variables. In particular, if second moments exist, then the hyperbolic behaviour of the spectral density at the origin (which also includes the possibility of a constant), $f_Y(\lambda) \sim L_f(\lambda)|\lambda|^{-2d}$, can be considered a fundamental phenomenon.

For temporal aggregation, Lamperti's results mean that, for M large enough, the joint distribution of standardized aggregates

$$Y^*_{i,M} = L^{\frac{1}{2}}(N)N^{-H}\big(Y_{i,M} - E(Y_{i,M})\big) \quad (i = 1, 2, \ldots, n),$$

with L and H chosen appropriately, is approximately the same as for increments of a self-similar process, provided that the original process was stationary. In particular, if second moments exist, then the autocovariances of $Y^*_{i,N}$ can be approximated by

$$\gamma_{Y^*}(k) \approx \frac{1}{2}\big(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\big) \sim H(2H-1)|k|^{2H-2} \quad (k \to \infty) \tag{2.68}$$

and the spectral density by

$$f_{Y^*}(\lambda) \approx 2(1 - \cos\lambda) \sum_{k=-\infty}^{\infty} |\lambda + 2k\pi|^{-2H-1} \sim c_f|\lambda|^{1-2H} \quad (\lambda \to 0), \tag{2.69}$$

where $c_f = (2\pi)^{-1}\sin(\pi H)\Gamma(2H+1)$. In other words, in first approximation, we have the same autocovariances and spectral density as a fractional Gaussian

noise. Note, however, that the asymptotic distribution of $Y^*_{i,M}$ need not be Gaussian because there exist non-Gaussian self-similar processes with finite second moments (see Sect. 1.3.5).

In some situations, $Y_{i,M}$ is obtained by aggregating nonstationary observations X. For instance, many aggregated series in finance and economics are based on integrated processes. The limit of $Y^*_{i,M}$ then changes but is not fundamentally different under fairly general conditions. For instance, Tsai and Chan (2005a) consider fractional ARIMA processes that may be nonstationary due to integration. In other words, let $m \in \{0, 1, 2, \ldots\}$ and $-\frac{1}{2} < d < \frac{1}{2}$. Denote by $\varphi(z) = 1 - \sum_{j=1}^{p} \varphi_j z^j$ and $\psi(z) = 1 + \sum_{j=1}^{q} \psi_j z^j$ polynomials with no zeroes for $|z| \le 1$, and let ε_i be i.i.d. zero mean random variables with variance $\sigma_\varepsilon^2 < \infty$. A fractional ARIMA$(p, m+d, q)$ (or FARIMA$(p, m+d, q)$) process X_i is defined as a solution of

$$\varphi(B)(1-B)^{m+d} X_j = \psi(B)\varepsilon_j \tag{2.70}$$

with B denoting the backshift operator (i.e. $B\varepsilon_j = \varepsilon_{j-1}$ etc.). For example, for $m = 0$, we obtain the stationary FARIMA(p, d, q) process introduced in Sect. 2.1.1.4. For $m = 1$, we have a random-walk-type process where the first difference $\Delta X_j = (1-B)X_j$ is a stationary FARIMA(p, d, q) process, and so on. Now, of course, if we first take the mth difference $\Delta^m X_j = (1-B)^m X_j$ and then aggregate, we are in the same situation as before, i.e. we again obtain stationary increments of a self-similar process in the limit. However, often aggregates are calculated first, before making the original observations stationary. As it turns out, this leads to different limits. Thus, consider X_j defined by (2.70) and the aggregated $Y_{i,M}$ defined by (2.67). Moreover,

$$Y^*_{i,M,m} = \frac{\Delta^m Y_{i,M}}{\sqrt{\mathrm{var}(\Delta^m Y_{i,M})}}$$

will denote the standardized differenced series. Then the following result is derived in Tsai and Chan (2005a):

Theorem 2.11 *As $M \to \infty$, the spectral density of $Y^*_{i,M,m}$ converges to*

$$\begin{aligned} f_{m,d}(\lambda) &= C_{m,d}\big[2(1-\cos\lambda)\big]^{m+1} \sum_{k=-\infty}^{\infty} |\lambda + 2\pi k|^{-2H-1-2m} \\ &= C_{m,d} f^*_{m,d}(\lambda), \end{aligned} \tag{2.71}$$

where

$$C_{m,d} = \left(\int_{-\pi}^{\pi} f^*_{m,d}(\lambda)\, d\lambda \right)^{-1}.$$

The same formula was derived in Beran and Ocker (2000) for the case of $m \in \{0, 1\}$ (and $-\frac{1}{2} < d < \frac{1}{2}$). For analogous results in the context of continuous-time processes, in particular continuous-time FARIMA (also called CARFIMA)

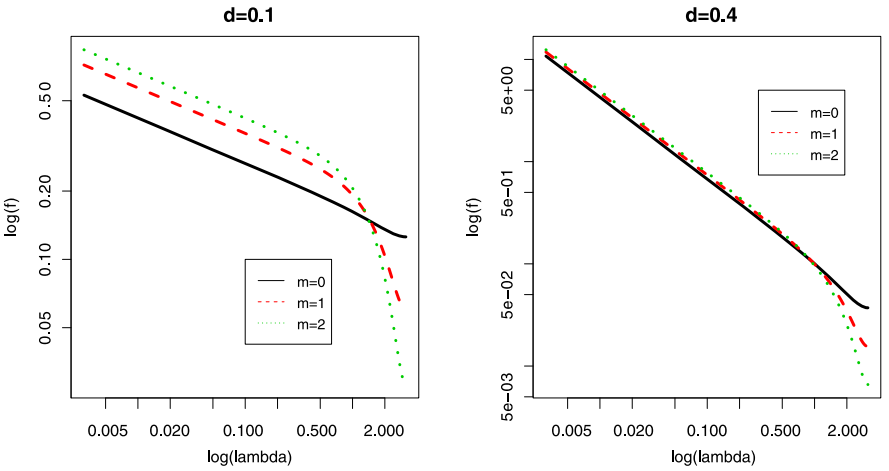

Fig. 2.1 Log-log-plot of the spectral densities $f_{m,d}$ with $m = 0$, 1 and 2 and $d = 0.1$ (*left*) and $d = 0.4$ (*right*) respectively, obtained as limits of temporal aggregation of FARIMA$(p, m + d, q)$ processes

models, see Tsai and Chan (2005b, 2005c). Related papers are also Teles et al. (1999), Hwang (2000), Souza and Smith (2004), Tsai (2006), Paya et al. (2007), Souza (2005, 2007, 2008), Man and Tiao (2006, 2009), Hassler (2011). Moreover, Chambers (1998) showed that (integer) integrated processes keep their order of integrations after aggregation.

For $m = 0$, formula (2.71) is of course the same as (2.69), i.e. $f_{0,d}$ is identical with the spectral density of a fractional Gaussian noise. However, for integrated processes (with $m \geq 1$), the asymptotic dependence structure is different. What remains the same is the preservation of long memory (since it can be shown that, for all m, $f_{m,d}(\lambda) \sim \text{const} \cdot \lambda^{-2d}$ near the origin) and the absence of the initial short-memory parameters φ_j $(j \leq p)$ and ψ_j $(j \leq q)$ in the limit. It is the detailed form of $f_{m,d}$ for nonzero frequencies that depends on m. In particular, for $m \neq 0$, we no longer have increments of a self-similar process. The two plots in Fig. 2.1 show log-log-plots of $f_{m,d}$ for $m = 0, 1, 2$ and $d = 0.1$ (left plot) and $d = 0.4$ (right plot) respectively. For low frequencies up to about $\lambda = 1$, the shapes of $\log f_{m,d}$ $(m = 0, 1, 2)$ are practically the same. The essential difference between the three cases is visible for higher frequencies, however, and is much more pronounced for weaker long memory $(d = 0.1)$.

In summary, one may say that in applications where time series are temporal aggregates, the assumption that the spectral density (of the stationary version) is approximately proportional to λ^{-2d} (for some $-0.5 < d < 0.5$) is a canonical one. In retrospect, it is therefore not surprising that this fact has been noticed empirically by experienced econometricians long before suitable probabilistic models have become available (see e.g. Granger's 1966 Econometrica article entitled "The typical spectral shape of an economic variable").

2.2.2 Cross-Sectional Aggregation

A possible explanation of long memory in observed economic series was suggested by Granger (1980) (also see Robinson 1978 for similar results). He considered independent AR(1) processes $X_{i,t}$ $(i = 1, 2, \ldots)$ with autoregressive parameters φ_i being generated by a distribution G and demonstrated heuristically that the normalized aggregated process $N^{-1/2} \sum_{i=1}^{N} X_{i,t}$ converges to a long-memory process, provided that G puts enough weight near the unit root (but within the stationary range $(-1, 1)$). Many authors took up this topic subsequently, working out a detailed mathematical theory and extending the result to more general processes. References include for instance Goncalves and Gouriéroux (1988), Ding and Granger (1996), Igloi and Terdik (1999), Abadir and Talmain (2002), Leipus and Viano (2002), Kazakevičius et al. (2004), Davidson and Sibbertsen (2005), Leipus et al. (2004), Zaffaroni (2004, 2007a, 2007b), Beran et al. (2010), Giraitis et al. (2010). Also see e.g. references in Baillie (1996).

To be specific, we will (as in Granger 1980) look at aggregation of AR(1) processes. The following definition will be needed.

Definition 2.5 A stochastic process X_t $(t \in \mathbb{N})$ is called strictly asymptotically stationary if the finite-dimensional distributions of $X_{t_1}, \ldots, X_{t_k}$ $(k \in \mathbb{N}, 0 \le t_1 < \cdots < t_k < \infty)$,

$$F_{t_1,\ldots,t_k;n}(x_1, \ldots, x_k) = P(X_{t_1+n} \le x_1, \ldots, X_{t_n+n} \le x_n),$$

converge weakly to the finite-dimensional distributions of a strictly stationary process as $n \to \infty$. The process is called weakly asymptotically stationary if

$$\lim_{n\to\infty} E[X_n] = \mu \in \mathbb{R}$$

and

$$\lim_{t\to\infty} cov(X_n, X_{n+k}) = \gamma(k)$$

for all $k \ge 0$, where $\gamma(k)$ is an even non-negative definite function.

Now, consider a panel of N independent asymptotically stationary normal AR(1) processes, each of length n, where the AR(1) coefficients φ_i of the individual series are i.i.d., and their square is Beta distributed. Thus, we have a sequence of processes $X_{i,t}$ $(i = 1, 2, \ldots)$

$$X_{i,t} = \varphi_i X_{i,t-1} + \varepsilon_{i,t}$$

with $\varepsilon_{i,t}$ i.i.d. standard normal and φ_i i.i.d. with density

$$g_\varphi(x) = \frac{2}{B(\alpha, \beta)} x^{2\alpha-1}\left(1 - x^2\right)^{\beta-1} \quad \left(x \in (0, 1); \alpha, \beta > 1\right)$$

(where $B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function) and independent of all $\varepsilon_{i,t}$. A heuristic argument can now be given as follows. The spectral density of the standardized aggregated

$$X_t^{(N)}=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}X_{i,t}$$

is equal to the average of the individual spectral densities

$$f_i(\lambda)=\frac{\sigma_\varepsilon^2}{2\pi}\left|1-\varphi_i e^{-i\lambda}\right|^{-2}.$$

As $N\to\infty$, we therefore obtain the spectral density

$$f(\lambda)=E\big[f_i(\lambda)\big]=\int_0^1 f_i(\lambda)g_\varphi(\lambda)\,d\lambda,$$

which can be shown to be proportional to λ^{-2d} as $\lambda\to 0$, with $d=1-\beta/2$. The same result also holds conditionally, i.e. when the $AR(1)$ processes are initiated recursively (see Beran et al. 2010, Leipus et al. 2006). Suppose that the initial values $X_{i,0}$ are i.i.d. with $E[X_{i,0}]=0$ and $|X_{i,0}|\le C_0$ for all i and some constant $C_0<\infty$. We also assume that $X_{i,0}$ are independent of all $\varepsilon_{i,t}$ $(i,t\ge 1)$ and of φ_j $(j\ne i)$. Then it follows that for each i, $X_{i,t}$ $(t\in\mathbb{N})$ is a zero-mean strictly and weakly asymptotically stationary process with

$$\operatorname{var}(X_{i,t})\to E\big[\big(1-\varphi^2\big)^{-1}\big]$$

as $t\to\infty$. This expected value is finite since

$$E\big[\big(1-\varphi^2\big)^{-1}\big]=\frac{2}{B(\alpha,\beta)}\int_0^1 x^{2\alpha-1}\big(1-x^2\big)^{\beta-2}\,dx$$

with $\beta>1$. Note, that $X_{i,0}$ is allowed to depend on φ_i. For each i, we may therefore choose the distribution of the initial value $X_{i,0}$ such that, conditionally on φ_i, it is arbitrarily close to a normal distribution with variance $(1-\varphi_i^2)^{-1}$, which one would get under stationarity of $X_{i,t}$ $(t\in\mathbb{N})$.

The covariance function of each process $X_{i,t}$ $(t\in\mathbb{N})$ is given by

$$cov(X_{i,t},X_{i,t+k})=E\left[\varphi^k\sum_{j=1}^{t-1}\varphi^{2j}\right]+E\big[\varphi^{2t+k}X_{1,0}^2\big],$$

which, by dominated convergence, tends to

$$\gamma(k) := E\left[\frac{\varphi^k}{1-\varphi^2}\right] = \int_0^1 \frac{2}{B(\alpha,\beta)} x^{2\alpha-1+k}\left(1-x^2\right)^{\beta-2} dx$$
$$= \frac{B(\alpha+\frac{k}{2},\beta-1)}{B(\alpha,\beta)} \int_0^1 \frac{2}{B(\alpha+\frac{k}{2},\beta-1)} x^{2(\alpha+\frac{k}{2})-1}\left(1-x^2\right)^{(\beta-1)-1} dx$$
$$= \frac{B(\alpha+\frac{k}{2},\beta-1)}{B(\alpha,\beta)} \tag{2.72}$$

as $t \to \infty$. This implies

$$\gamma(k) \sim ck^{1-\beta} \quad (k \to \infty),$$

where $d = 1 - \beta/2$, and the constant $c > 0$ depends on α and β. Hence, *unconditionally*, each process $X_{i,t}$ $(t \in \mathbb{N})$ is a stationary long-memory process if $\beta \in (1, 2)$. However, the long-memory behaviour is not observable if only one of the series is observed. The reason is that the random nature of φ is not visible or, in other words, the data yield only information about the *conditional* distribution of the process (given φ). A single sample path $X_{i,t}(\omega)$ $(t \geq 0)$ is not distinguishable from a path of an AR(1) process with coefficient $\varphi_i(\omega)$. In this sense, $X_{i,t}$ $(t \in \mathbb{N})$ is not an ergodic process (Robinson 1978, Oppenheim and Viano 2004). Ergodicity can be recovered only by observing an increasing number N of replicates and considering the normalized aggregated

$$X_t^{(N)} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_{i,t} \quad (t \in \mathbb{N}).$$

By definition, $X_t^{(N)}$ exhibits the same autocovariance function as each individual series. The difference is, however, that given an individual series, we are only able to estimate the conditional dependence structure given φ_i, whereas the random mechanism generating this coefficient is hidden and cannot be estimated. This is of course different when we observe an increasing number of replicates. For the aggregated process, we obtain in the limit (as $N \to \infty$) an ergodic Gaussian process X_t^{∞} $(t \in \mathbb{N})$ with covariance function (2.72). The existence of the limit is formulated in the following theorem (Beran et al. 2010, also see Oppenheim and Viano 2004). To state the result, we consider convergence of sample paths $X_t^{(N)}$ $(t \in \mathbb{N})$ in the Hilbert space $\mathcal{H}_\varepsilon^2$ (for some $\varepsilon > 0$) of real sequences x_t $(t \in \mathbb{N})$ such that

$$\sum_{t \geq 0} \frac{x_t^2}{(t+1)^{1+\varepsilon}} < \infty$$

with the inner product between two sequences x_t, y_t $(t \in \mathbb{N})$ defined by

$$\langle x_t, y_t \rangle := \sum_{t \geq 0} \frac{x_t y_t}{(t+1)^{1+\varepsilon}}.$$

Proposition 2.1 *As $N \to \infty$, the process $X_t^{(N)}$ $(t \geq 0)$ converges weakly in the space $\mathcal{H}_\varepsilon^2$ to a zero-mean Gaussian stationary process X_t^∞ $(t \geq 0)$ with autocovariance function* (2.72).

Proof First note that the sample paths $X_t^{(N)}$ $(t \geq 0)$ are almost surely in $\mathcal{H}_\varepsilon^2$. This can be seen as follows. Since $E[X_{1,t}^2]$ converges to $E[(1-\varphi_1^2)^{-1}]$ as $t \to \infty$, there is a finite constant c with $E[(X_t^{(N)})^2] \leq c$ for all $t \geq 1$ and $N \geq 1$. This implies for every $N \geq 1$,

$$E\left[\sum_{t=0}^{\infty} \frac{(X_t^{(N)})^2}{(t+1)^{1+\varepsilon}}\right] \leq c^2 \sum_{t\geq 0} (t+1)^{-1-\varepsilon} < \infty.$$

The convergence of the finite-dimensional distributions of $X_t^{(N)}$ $(t \geq 0)$ follows directly from a multivariate central limit theorem (note in particular that the sample paths are already exactly Gaussian) and the convergence of the autocovariances to (2.72). Finally, the most difficult property to check is tightness. Here, one can use sufficient conditions given by Suquet (1996), namely, for every $n \geq 0$,

$$\lim_{a\to\infty} \sup_N P\left(\sum_{t\geq n} \frac{X_t^{(N)}}{(t+1)^{1+\varepsilon}} > a\right) = 0$$

and, for every $a > 0$,

$$\lim_{n\to\infty} \sup_N P\left(\sum_{t\geq n} \frac{X_t^{(N)}}{(t+1)^{1+\varepsilon}} > a\right) = 0$$

(also see Oppenheim and Viano 2004). Indeed, both equations are simple consequences of Chebyshev's inequality

$$\begin{aligned} P\left(\sum_{t\geq n} \frac{X_t^{(N)}}{(t+1)^{1+\varepsilon}} > a\right) &\leq a^{-2} E\left[\sum_{t\geq n} \frac{X_t^{(N)}}{(t+1)^{1+\varepsilon}}\right]^2 \\ &\leq a^{-2} c \sum_{t\geq n} (t+1)^{-1-\varepsilon}. \end{aligned}$$

□

Example 2.10 Figure 2.2 shows a histogram of $N = 1000$ simulated values of φ_i (Fig. 2.2(a)), together with the density function g_φ where $\alpha = 1.1$ and $\beta = 1.2$. Figures 2.2(b) through (e) display (in log-log-coordinates) the corresponding spectral densities f_i (left panel) and their average (right panel), together with a dotted reference line with slope -0.8. In Figs. 2.2(b) and (c) we consider f_i $(i = 1, 2, \ldots, 10)$ only, whereas the plots in (d) and (e) are based on f_i $(i = 1, 2, \ldots, 1000)$. As one can see, aggregating the first 10 processes did not lead to a straight line in log-log-coordinates, whereas for $N = 1000$, one seems to be quite close already to the limiting straight line with slope $-2d = \beta - 2 = -0.8$.

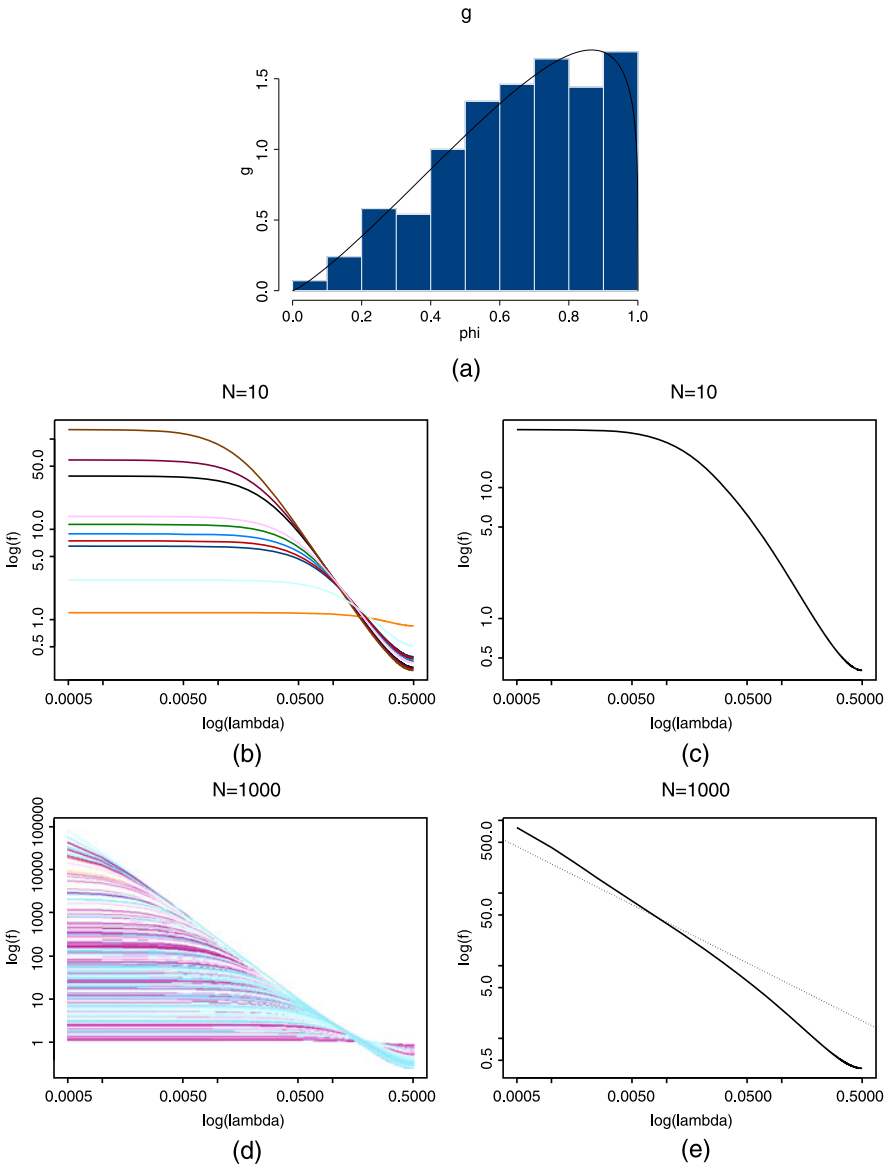

Fig. 2.2 (**a**) Shows a histogram of $N = 1000$ simulated values of φ_i (**a**), together with the density function g_φ. (**b**) through (**e**) display (in log-log-coordinates) the corresponding spectral densities f_i and their average. In (**b**) and (**c**) we consider f_i $(i = 1, 2, \ldots, 10)$ only, whereas the plots in (**d**) and (**e**) are based on f_i $(i = 1, 2, \ldots, 1000)$

2.2.3 Particle Systems, Turbulence, Ecological Systems

Kolmogorov (1940, 1941) introduced fractional Brownian motion, $1/f$ noise and related processes while investigating turbulent flows (see also for instance Batchelor 1953, Cassandro and Jona-Lasinio 1978, Marinari et al. 1983, Eberhard and Horn 1978, Frisch 1995, Barndorff-Nielsen et al. 1998, Anh et al. 1999, Barndorff-Nielsen and Leonenko 2005, Leonenko and Ruiz-Medina 2006). Long memory also plays an important role in explicit models for particle systems, in particular in the context of phase transition (see e.g. Domb and Lebowitz 2001 and all previous and subsequent volumes of this series, Stanley 1971, 1987; Liggett 2004). A typical model used in statistical mechanics is a random field on an m-dimensional lattice $T = \mathbb{Z}^m$. The interpretation is that the values X_t $(t \in T)$ represent the state of a particle at location t. Usually X_t assumes values in a polish space $\mathbb{X}$. Interactions between particles are characterized by a pair potential $\Phi = (\Phi_{i,j})_{i,j \in T}$ where each $\Phi_{i,j}(x, y)$ is a function describing the potential energy of the two interacting particles at locations i and j. Configurations $x = (x_t)_{t \in T} \in \mathbb{X}^T$ are functions on T, and they are assumed to be random, i.e. realizations of a random field $(X_t)_{t \in T} \in \mathbb{X}^T$. For a finite subset $S \subset T$, the energy of a configuration $x_S = (x_t)_{t \in S}$ on S is given by

$$E_{x_S} = \sum_{\{i,j\} \subseteq S} \Phi_{i,j}(x_i, x_j) + \sum_{i \in S, j \notin S} \Phi_{i,j}(x_i, x_j).$$

The distribution of $(X_t)_{t \in T}$ is assumed to be given by a Gibbs measure (associated with the potential Φ) that is absolutely continuous with respect to a measure ν on $\mathbb{X}^T$ (e.g. a Lebesgue or a Bernoulli measure). The Gibbs measure is defined by conditional densities of finite configurations x_S given the remaining configuration x_{S^c} of the form

$$dP(x_S \mid x_{S^c}) = \frac{1}{Z_S(x_{S^c})} \exp(-E_{x_S})\, d\nu(x).$$

Here, Z_S is a normalizing constant so that, up to a proportionality factor, the conditional distribution of x_S is fully described by the potential Φ. Pure phases are characterized by extreme elements in the set of all Gibbs measures, the set itself being convex. For all other phases, the corresponding Gibbs measure can be represented as a mixture of the "pure" measures. For references in this context, see e.g. Kolmogorov (1937), Dobrushin (1968a, 1968b, 1968c, 1969, 1970), Lanford and Ruelle (1968, 1969), Ruelle (1968, 1970), Föllmer (1975), Cassandro and Jona-Lasinio (1978), Kosterlitz and Thouless (1978), Künsch (1980), Sokal (1981), Georgii (1988), Bolthausen et al. (1995), Lavancier (2006). The existence of a Gibbs measure is directly linked to the occurrence of a phase transition. Given Φ and ν, a phase transition occurs if there exists more than one Gibbs measure.

In the simplest case, x_t represents the spin at location t with values in $\mathbb{X} = \{-1, 1\}$, and ν is a point measure with mass $\frac{1}{2}$ at -1 and 1. In the Ising model

(originally introduced to understand fluid dynamics and ferromagnetism), the potential is defined by

$$\Phi_{i,j} = \beta x_i x_j$$

if $\|i - j\| = 1$ and zero otherwise. The constant $\beta > 0$ is inverse temperature. For a one-dimensional lattice, i.e. $T = \mathbb{Z}$, there is a unique Gibbs measure for any β so that no phase transition occurs. In two dimensions, i.e. $T = \mathbb{Z}^2$, more than one Gibbs measure exists, and hence phase transition takes place for the critical value $\beta = \beta_c = \frac{1}{2}\log(1 + \sqrt{2})$ (Onsager 1944). Moreover, the Gibbs measures are stationary. Similarly, for all higher dimensions, phase transition occurs for a dimension-specific critical inverse temperature β_c (Dobrushin 1965). Phase transition is directly linked to long-range dependence as follows (Kaufman and Onsager 1949; Fisher 1964). Using the notation $k = i - j \in \mathbb{Z}^m$, we have $cov(X_i, X_j) = \gamma(\|k\|)$, i.e. the covariance is a function of the Euclidian distance $\|k\|$ only. If $\beta \neq \beta_c$, then $\gamma(\|k\|)$ tends to zero exponentially, whereas a hyperbolic decay with nonsummable covariances is obtained for $\beta = \beta_c$. More specifically, denoting by $\kappa_B \approx 1.38 \times 10^{-23}\ \mathrm{JK}^{-1}$ the Boltzmann constant and by $\mu \in [0, 2]$ a dimension-dependent critical parameter, one obtains $\gamma(\|k\|) \sim \|k\|^{-1} \exp(-\kappa_B \|k\|)$ for $\beta \neq \beta_c$ (as $k \to \infty$) and

$$\gamma\big(\|k\|\big) \sim \|k\|^{2-m-\mu} = \|k\|^{2d-m}$$

with $d = 1 - \frac{1}{2}\mu$. Since $\mu \in [0, 2]$, we have $d > 0$ and, by an m-dimensional Riemann sum approximation, as $\|k\| \to \infty$,

$$V_n = \sum_{\|k\| \le n} \gamma\big(\|k\|\big) \sim C \sum_{0 < \|k\| \le n} \|k\|^{2d-m} \sim C^* n^{2d} \to \infty$$

with nonzero constants C, C^*. For instance, for a two-dimensional lattice ($m = 2$), $\mu = \frac{1}{4}$, so that $\gamma(\|k\|) \sim \|k\|^{-\frac{1}{4}}$, and V_n diverges to infinity at the rate $\sqrt{n}$. In contrast, for $\beta \neq \beta_c$, the exponential decay implies that V_n converges to a finite constant.

Another standard case is $\mathbb{X} = \mathbb{R}$ with

$$\Phi_{i,i}(x_i, x_i) = \beta\left[\frac{1}{2} J(0) x_i^2 + e \cdot x_i\right],$$

$$\Phi_{i,j}(x_i, x_j) = \beta J(i - j) x_i x_j \quad (i \neq j)$$

and ν the Lebesgue measure on $\mathbb{R}^m$. The constants β and e correspond to inverse temperature and an external magnetic field respectively. Moreover, the so-called potential $[J(k)]_{k \in T}$ ($J(k) \in \mathbb{R}$) is positive definite, symmetric and summable. A very elegant result on the existence of Gibbs measures can be derived for the case where $e = 0$, i.e. when there is no external magnetic field (see e.g. Dobrushin 1980, Künsch 1980, Georgii 1988): phase transitions depend on J only, not on the temperature, and the existence of at least one Gibbs measure is equivalent to

$$\int_{[-\pi,\pi]^m} \frac{1}{\hat{J}(\lambda)}\, d\lambda < \infty,$$

where

$$\hat{J}(\lambda) = \sum_{k \in \mathbb{Z}^m} J(k) e^{ik'\lambda}$$

is the Fourier transform of J. The pure phases are Gaussian with autocovariance function

$$\gamma(k) = \int_{[-\pi,\pi]^m} \frac{1}{\hat{J}(\lambda)} e^{ik'\lambda}\, d\lambda = \int_{[-\pi,\pi]^m} \frac{1}{\hat{J}(\lambda)} e^{i(k_1\lambda_1 + \cdots + k_m\lambda_m)}\, d\lambda.$$

Moreover, if only Gibbs measures with existing second moments are considered, then the existence of several Gibbs measures is equivalent to $\hat{J}(\lambda)$ having at least one root in $[-\pi, \pi]^m$. Since $\hat{J}^{-1}(\lambda)$ plays the role of a spectral density for the pure phases, this means that phase transition is equivalent to the spectral density having at least one pole. In this sense, phase transition is linked to long-range dependence. The following example follows from Lavancier (2006).

Example 2.11 Let $m = 2$, and define for $u = (u_1, u_2)' \in \mathbb{Z}^2$ with $u_1 = u_2 = k$,

$$J(u) = \rho_d(k) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(k+1-d)},$$

where $-\frac{1}{2} < d < 0$. Otherwise, for $u_1 \neq u_2$, set $J(u) = 0$. This means that on the diagonal, $J(u)$ is a function of $k = u_1 = u_2$ ($k \in \mathbb{Z}$) and identical with the autocorrelation function of an antipersistent FARIMA$(0, d, 0)$ process. As $k \to \infty$, the correlations are proportional to k^{2d-1}, and the spectral density converges to zero at the origin at the rate $O(\lambda^{2d^*})$ with $0 < d^* = -d < \frac{1}{2}$. For the Fourier transform $\hat{J}$, we have

$$\begin{aligned} \hat{J}(\lambda) &= \sum_{u \in \mathbb{Z}^2} J(u) e^{iu'\lambda} = \sum_{k=-\infty}^{\infty} \rho(k) e^{ik(\lambda_1+\lambda_2)} \\ &= \left|1 - e^{-i(\lambda_1+\lambda_2)}\right|^{2d^*} \cdot \frac{\Gamma^2(1-d)}{\Gamma(1-2d)}. \end{aligned}$$

Thus, $\hat{J}^{-1}(\lambda)$ is integrable, but along the line $\lambda_1 = -\lambda_2$, we have $\hat{J}(\lambda) = 0$. This implies that phase transition occurs. The autocovariances of X_t along the diagonal are of the form

$$\begin{aligned} cov(X_t, X_{t+(k,k)'}) &= c_1 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{ik(\lambda_1+\lambda_2)} \left|1 - e^{-i(\lambda_1+\lambda_2)}\right|^{-2d^*} d\lambda_1\, d\lambda_2 \\ &= c_2 \int_{-\pi}^{\pi} e^{ikx} \left|1 - e^{-ix}\right|^{-2d^*} dx \\ &= c_3 \rho_{d^*}(k) \sim c_4 k^{2d^*-1}, \end{aligned} \tag{2.73}$$

where we substituted $x = \lambda_1 + \lambda_2$, and $c_1, \ldots, c_3$ are suitable constants. In other words, along the diagonal we have the same type of long-range dependence as for a FARIMA$(0, d^*, 0)$ process with $0 < d^* < \frac{1}{2}$. The correlation structure is however not isotropic, since Eq. (2.73) does not apply to off-diagonal directions. For example, for $u = (k, pk)'$ with $p \notin \{0, 1\}$, we have

$$\begin{aligned} cov(X_t, X_{t+u}) &= c_1 \int_{-\pi}^{\pi} e^{i(p-k)\lambda_2} \int_{-\pi}^{\pi} e^{ik(\lambda_1+\lambda_2)} \left|1 - e^{-i(\lambda_1+\lambda_2)}\right|^{-2d^*} d\lambda_1 \, d\lambda_2 \\ &= c_2 \rho_{d^*}(k) \int_{-\pi}^{\pi} e^{i(p-1)k\nu} \, d\nu = c_3 \frac{\sin \pi k(p-1)}{k(p-1)} \rho_{d^*}(k), \end{aligned} \tag{2.74}$$

where c_1, c_2, c_3 are suitable constants, and, for $x = 0$, the value of $x^{-1} \sin x$ is understood as the limit as $x \to 0$. In particular, for $p \in \mathbb{Z} \setminus \{0, 1\}$, the correlation is zero.

For a review of some probabilistic aspects of long memory in the Ising model, see Pipiras and Taqqu (2012).

One of the central questions closely related to phase transition and long-range dependence is percolation (see e.g. Kesten 1982, Stauffer and Aharony 1994, Chakrabarti et al. 2009). Percolation is originally concerned with the movement of fluids in porous material, but applications go far beyond this specific situation (see e.g. Bunde and Havlin 1995, Bak 1996, Vanderzande 1998) including conductivity, sol–gel transition and polymerization, spread of epidemies, ecological systems, computer and social networks. A standard set up in "bond percolation" is a lattice, e.g. $\mathbb{Z}^m$, or network/graph with edges (paths) between vertices that are located on the lattice. The edges are "open" with probability p. (In "site percolation" vertices instead of edges are present or absent with a certain probability.) The events at different edges are assumed to be independent. Considering a finite area $A \subset \mathbb{Z}^m$, one would like to know the probability that there is a connected path from one "end" of A to the other. Practically speaking, this means for instance that a fluid dipped on top of a porous stone flows all the way to the bottom. In percolation theory, one considers the limit $A \to \mathbb{Z}^m$ and thus the question what the probability $\pi(p)$ of an infinite path (or cluster) is. Clearly, $\pi(p)$ is monotonically nondecreasing in p with $\pi(0) = 1 - \pi(1) = 0$. Thus, due to Kolomogorov's 0–1 law, there is a critical probability p_c such that $\pi(0) = 0$ for $p < p_c$ and $\pi(p) = 1$ for $p > p_c$. While p_c generally depends on the local geometry of the graph, various quantities describing the clusters are believed (and partially proved) to be universal. An important part of mathematical percolation theory (see e.g. Kesten 1982, Durrett 1984, Madras and Slade 1996, Grimmett 1999, Járai 2003) is therefore the probabilistic characterization of clusters. For $p = p_c$, clusters have fractal properties, and hyperbolic laws are obtained. For example, in bond percolation, the probability $P_{\text{same}}(r)$ that two sites at locations x and y at distance $r = \|x - y\|$ are in the same connected component is proportional to $r^{2-m-\eta}$, where η is known to be zero for $m \geq 19$ (and believed to be zero for $m \geq 7$). Note that this probability may be considered as a

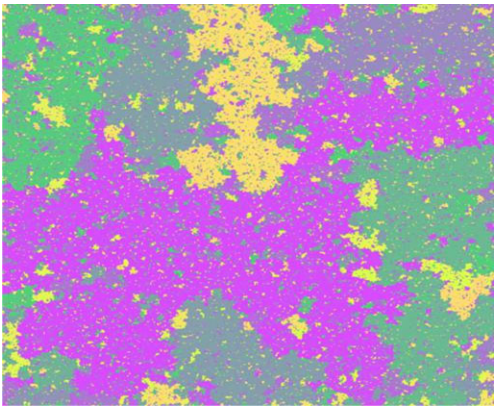

Fig. 2.3 Clusters obtained after percolation at p_c with $P_{\text{same}}(r) \propto r^{-5/24}$. (Figure courtesy of Prof. Hans Jürgen Herrmann, Computational Physics for Engineering Materials, ETH Zurich)

specific measure of dependence. A typical picture with $P_{\text{same}}(r) \propto r^{-5/24}$ is shown in Fig. 2.3.

A more complex version of percolation is so-called "long-range percolation". In contrast to standard percolation with nearest-neighbor connections only, each vertex can be connected to any other arbitrarily remote vertex. This is combined with hyperbolic probabilities (and sometimes also with Ising and related models), with interesting and partially still unexplored connections to long-memory random fields. For instance, the probability that x and y are in the same connected graph is assumed to be proportional to $\|x - y\|^{-\alpha}$ for some $\alpha > 0$. For literature on long-range percolation and related topics, see e.g. Fröhlich and Spencer (1982), Weinrib (1984), Newman and Shulman (1986), Imbrie and Newman (1988), Meester and Steif (1996), Menshikov et al. (2001), Berger (2002), Coppersmith et al. (2002), Abete et al. (2004), de Lima and Sapozhnikov (2008), Trapman (2010), Biskup (2004, 2011), Crawford and Sly (2011), and references therein.

Other results on particle systems and long-range dependence include for instance particle branching systems (Gorostiza and Wakolbinger 1991, Gorostiza et al. 2005, Bojdecki et al. 2007) and random interlacements (see e.g. Sznitman 2010).

Random fields with long memory are also an important part of ecological modelling. One approach is inspired by interacting particle systems in physics similar to the discussion above (see e.g. Bramson et al. 1996, Durrett and Levin 1996). Under certain conditions, Bramson et al. (1996) obtain a hyperbolic dependence between the number of observed species and the area where data are sampled. Another approach that leads to a hyperbolic law is based on latent long-memory fields (Ghosh 2009). The reason for considering latent processes is that observed spatial or space-time data are often regulated or influenced by unobserved processes such as water supply, soil quality, wind etc. A detailed account of the approach in Ghosh (2009) is given in Sect. 9.4. For related applied literature in this context, see also e.g. Scheuring (1991), Harte et al. (1999).

2.2.4 Network Traffic Models

In their pioneering papers Leland et al. (1993a, 1993b) analysed internet traffic data, more precisely time series representing the number of packets sent from a local network. They found that the data exhibits self-similarity over a certain range of scales. Subsequent studies (e.g. Leland et al. 1994, Beran et al. 1995, Paxson and Floyd 1995, Crovella and Bestavros 1997) revealed that classical Poisson modelling fails. Since then, "long-range dependence", "self-similarity" and "high variability" have become important issues in the analysis of network data.

To capture these phenomena, one uses models that can mimic the physical behaviour of a network. For traffic data in telecommunication networks or on the internet, several models play a crucial role:

- Renewal reward process: Levy and Taqqu (1987, 2000), Pipiras and Taqqu (2000b), Pipiras et al. (2004), Hsieh et al. (2007), Taqqu and Levy (1986).
- ON–OFF process: Taqqu et al. (1997), Heath et al. (1998), Greiner et al. (1999), Jelenkvovič and Lazar (1999), Mikosch et al. (2002), Leipus and Surgailis (2007).
- Infinite source Poisson model ($M/G/\infty$): Konstantopoulos and Lin (1998), Resnick and van den Berg (2000), Mikosch et al. (2002), Maulik et al. (2002).
- Error duration process: Parke (1999), Hsieh et al. (2007).

There are a number of modifications of these models, such as Poisson cluster processes or the fractal shot-noise model; see Klüppelberg et al. (2003), Klüppelberg and Kühn (2004), Lowen and Teich (2005), Faÿ et al. (2006), Mikosch and Samorodnitsky (2007), Fasen and Samorodnitsky (2009), Rolls (2010), Dombry and Kaj (2011). We also refer to Taqqu (2002), Willinger et al. (2003), Gaigalas (2004), Deo et al. (2006a) for an overview. Applications of these models go far beyond computer networks. For example, renewal reward and error duration processes have been used for modelling economic data (see e.g. Hsieh et al. 2007, Deo et al. 2009).

We will describe such models under the common umbrella of "shot-noise" processes.

2.2.4.1 Shot-Noise Processes

Consider a stationary point process τ_j ($j \in \mathbb{Z}$) on the real line with rate λ, let N be the associated counting process, and $X_j = \tau_j - \tau_{j-1}$ ($j \in \mathbb{Z}$) be the corresponding stationary sequence of interrenewal times. Recall the convention $\tau_{-1} < 0 \le \tau_0$. Now consider independent copies $Y_j(\cdot)$ ($j \ge 1$) of a stochastic process $Y_1(t)$ ($t \in \mathbb{R}$) and define for $t \ge 0$,

$$W(t) = \sum_{j=-\infty}^{\infty} Y_j(t - \tau_{j-1}). \tag{2.75}$$

If $Y_j(t) = 0$ for $t < 0$, then we can interpret $W(t)$ in the following way. At random times τ_{j-1} we initiate a "shock" (transmission) described by a stochastic process

$Y_j(\cdot)$. There is no specific limit for a duration and "size" of each transmission, unless we impose further conditions on $Y_j(\cdot)$. In particular, if η_j (durations) is an i.i.d. sequence of positive random variables and $Y_j(u) = 1\{0 < u < \eta_j\}$, then

$$W(t) = \sum_{j=-\infty}^{\infty} 1\{\tau_{j-1} \le t < \tau_{j-1} + \eta_j\} \tag{2.76}$$

can be interpreted as the number of active sources at time t. We note that if $\eta_j \le X_j = \tau_j - \tau_{j-1}$, then at time t we can have only one source active, like in a renewal reward or an ON–OFF process considered below. In this case the duration sequence η_j ($j \in \mathbb{Z}$) is *not* independent of the interarrival times X_j. If there is no dependence between the sequences η_j and X_j, then there are possibly many sources active, like in the Infinite Source Poisson model that will be introduced below.

Lemma 2.6 *Assume that* $Y_j(u) = 0$ *for* $u < 0$, $\int_0^\infty E[|Y(u)|]\,du < \infty$ *and* τ_j ($j \in \mathbb{Z}$) *is a stationary renewal process with rate* λ. *Then*

- $W(t)$ ($t > 0$) *is stationary.*
- $E[W(t)] = \lambda \int_0^\infty E[Y(u)]\,du$ *for each* $t \ge 0$.

Proof The stationarity is clear from the stationarity of the underlying point process τ_j. As for the mean,

$$E\big[W(0)\big] = \sum_{j=-\infty}^{0} E\big[Y_j(-\tau_{j-1})\big] = \sum_{j=-\infty}^{-1} \int_{-\infty}^{0} E\big[Y_{j+1}(-u)\big]\,dP_{\tau_j}(u)$$

$$= E\left[\int_{-\infty}^{0} Y_{j+1}(-u) \sum_{j=-\infty}^{-1} dP_{\tau_j}(u)\right],$$

where P_{τ_j} is the distribution of τ_j. Now, recall that $\sum_{j=0}^{\infty} dP_{\tau_j}(u)$ ($u > 0$) is the renewal density function times du and thus equals $\lambda\,du$. Likewise, $\sum_{j=-\infty}^{-1} dP_{\tau_j}(u)$ is the renewal density function for the negative real line. Thus, $E[W(0)] = \int_0^\infty E[Y(u)]\,du$ follows. □

The result of Lemma 2.6 is valid for a very general class of point processes. Moreover, one can also write a formula for the covariance function of $W(t)$ (Leipus and Surgailis 2007). In the case of (2.76) the formula for the covariance function can be obtained, without any assumptions on the independence between the sequences X_j and η_j (Mikosch and Samorodnitsky 2007). This involves however a deeper knowledge of point process theory, and we deliberately choose to work with specific models instead.

Example 2.12 (Renewal Reward Process) Assume that τ_j ($j \in \mathbb{Z}$) is a renewal process with rate λ and interarrival times X_j ($j \in \mathbb{Z}$). Set $Y_j(u) = 1\{0 < u < X_j\}Y_j^*$,

where Y_j^* ($j \in \mathbb{Z}$) is an i.i.d. sequence of random variables with finite mean. Assuming that the sequences X_j and Y_j^* are mutually independent, the resulting process

$$W(t) = \sum_{j=-\infty}^{\infty} Y_j^* 1\{\tau_{j-1} \le t < \tau_{j-1} + X_j\} = \sum_{j=-\infty}^{\infty} Y_j^* 1\{\tau_{j-1} \le t < \tau_j\} \quad (t > 0)$$

is called *a renewal reward process*. Alternatively, if the underlying renewal process τ_j is defined on $(0, \infty)$ only, we may consider

$$W(t) = Y_0^* 1\{0 \le t < \tau_0\} + \sum_{j=1}^{\infty} Y_j^* 1\{\tau_{j-1} \le t < \tau_j\} = Y_{N(t)} \quad (t \ge 0),$$

where τ_0 has the distribution $F^{(0)}(x) = \mu^{-1} \int_0^x \bar{F}(u)\,du$, F is the distribution of X_1, and $\mu = E[X_1]$. In other words, at the renewal time τ_{j-1} we have a shock of size Y_j that lasts for the duration X_j. At a given time t, only one shock contributes to $W(t)$. Now let us look at the autocovariance function. We assume for simplicity that $E[Y_1] = 0$ and $E[Y_1^2] < \infty$. From Lemma 2.6 we have

$$E\big[W(t)\big] = \lambda E\big[Y_1^*\big] \int_0^{\infty} P(X_1 > u)\,du = E\big[Y_1^*\big].$$

Since $W(0) = Y_0^*$ and Y_j^* are independent, we obtain

$$cov\big(W(0), W(t)\big) = E\big[Y_1^{*2}\big] P(\tau_0 > t).$$

If we assume that $P(X_1 > x) = x^{-\alpha} L(x)$, $\alpha \in (1, 2)$, where $L(x)$ is slowly varying at infinity, then

$$P(\tau_0 > t) = \lambda \int_t^{\infty} u^{-\alpha} L(u)\,du \sim \frac{\lambda}{\alpha - 1} t^{1-\alpha} L(t),$$

and thus the covariances are not summable. Furthermore, the relation between $\operatorname{var}(\int_0^t W(u)\,du)$ and $Cov(W(0), W(t))$, together with Lemma 1.1, yields

$$\operatorname{var}\left(\int_0^t W(u)\,du\right) \sim \frac{2E[Y_1^{*2}]}{\mu(\alpha - 1)(2 - \alpha)(3 - \alpha)} t^{3-\alpha} L(t).$$

Thus, in this model, long memory can be generated by heavy-tailed interarrival times. In contrast, if the duration has a finite variance, i.e. $E[X_1^2] < \infty$, then

$$\operatorname{var}\left(\int_0^t W(u)\,du\right) = t E\big[X_1^2\big] E\big[Y_1^{*2}\big].$$

Example 2.13 (ON–OFF Process) Assume that τ_j ($j \in \mathbb{Z}$) is a renewal process with rate λ and interpoint distances X_j ($j \in \mathbb{Z}$). Consider two mutually independent sequences $X_{j,\mathrm{on}}$ and $X_{j,\mathrm{off}}$ ($j \in \mathbb{Z}$) of i.i.d. random variables with common distribution function F_{on}, F_{off} and expected values μ_{on} and μ_{off} respectively. Suppose

that $X_j = X_{j,\text{on}} + X_{j,\text{off}}$ ($j \in \mathbb{Z}$), so that $E[X_1] = \mu = \mu_{\text{on}} + \mu_{\text{off}}$. The first sequence represents ON intervals, during which a source generates traffic (at a fixed rate, say 1). The second sequence represents OFF periods during which the source remains silent. Set $Y_j(u) = 1\{0 < u < X_{j,\text{on}}\}$. The resulting process

$$W(t) = \sum_{j=-\infty}^{\infty} 1\{\tau_{j-1} \le t < \tau_{j-1} + X_{j,\text{on}}\} \quad (t \ge 0)$$

is called *ON–OFF process*. Thus, at the renewal time τ_{j-1} we have a shock of size 1 that lasts for a period of length $X_{j,\text{on}}$. At a given time t, only one shock contributes to $W(t)$. In other words,

$$W(t) = 1 \quad \text{if time } t \text{ is in the ON interval,}$$
$$W(t) = 0 \quad \text{if time t is in the OFF interval.}$$

Application of Lemma 2.6 yields

$$E\big[W(t)\big] = \lambda \int_0^{\infty} P(X_{1,\text{on}} > u)\, du = \frac{\mu_{\text{on}}}{\mu}.$$

Typically in the literature one assumes that the underlying renewal process τ_j is defined on the positive real line. In this case, in order to assure stationarity, the renewal epochs are defined as

$$\tau_0 = \tilde{X}_0,$$
$$\tau_k = \tau_0 + \sum_{j=1}^{k} (X_{j,\text{off}} + X_{j,\text{on}}) = \tau_0 + \sum_{j=1}^{k} X_j \quad (k \ge 1),$$

where the first renewal epoch $\tau_0 = \tilde{X}_0$ is set equal to

$$\tilde{X}_0 = \xi(\tilde{X}_{\text{on}} + \tilde{X}_{\text{off}}) + (1 - \xi)\tilde{X}_{\text{off}},$$

where

$$P(\tilde{X}_{\text{on}} > x) = \frac{1}{\mu_{\text{on}}} \int_x^{\infty} \bar{F}_{\text{on}}(u)\, du =: \bar{F}_{\text{on}}^{(0)}(x),$$

$$P(\tilde{X}_{\text{off}} > x) = \frac{1}{\mu_{\text{off}}} \int_x^{\infty} \bar{F}_{\text{off}}(u)\, du =: \bar{F}_{\text{off}}^{(0)}(x),$$

and ξ is a Bernoulli random variable with $P(\xi = 1) = \mu_{\text{on}}/(\mu_{\text{on}} + \mu_{\text{off}})$. All random variables ξ, $X_{0,\text{on}}$, $X_{0,\text{off}}$ are assumed to be independent. With the notation above, we can write the ON–OFF process as

$$W(t) = \xi 1\{0 \le t < \tilde{X}_{\text{on}}\} + \sum_{j=1}^{\infty} 1\{\tau_{j-1} \le t < \tau_{j-1} + X_{j,\text{on}}\} \quad (t > 0). \tag{2.77}$$

In particular, we have

$$E\big[W(t)\big] = P\big(W(t) = 1\big) = E[\xi]P(\tilde{X}_{\text{on}} > t) + \sum_{j=1}^{\infty} P(\tau_{j-1} \le t < \tau_{j-1} + X_{j,\text{on}}).$$

By conditioning on τ_{j-1} and recalling the definition of the renewal function

$$\tilde{U}(t) = \sum_{j=0}^{\infty} P(\tau_j \le t) = \frac{1}{\mu_{\text{on}} + \mu_{\text{off}}} t,$$

we have

$$\begin{aligned}
&\sum_{j=1}^{\infty} P(\tau_{j-1} \le t < \tau_{j-1} + X_{j,\text{on}}) \\
&\quad = \sum_{j=0}^{\infty} \int_0^t \bar{F}_{\text{on}}(t-u)\, dP_{\tau_j}(u) \\
&\quad = \int_0^t \bar{F}_{\text{on}}(t-u)\, d\tilde{U}(u) = \frac{1}{\mu} \int_0^t \bar{F}_{\text{on}}(t-u)\, du \\
&\quad = \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} \frac{1}{\mu_{\text{on}}} \int_0^t \bar{F}_{\text{on}}(t-u)\, du = \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} P(\tilde{X}_{\text{on}} < t).
\end{aligned}$$

Hence,

$$\begin{aligned}
E\big[W(t)\big] &= \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} P(X_{0,\text{on}} > t) + \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} P(X_{0,\text{on}} < t) \\
&= \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} = \frac{\mu_{\text{on}}}{\mu},
\end{aligned}$$

which means that we obtain the same mean as before. To generate long memory, it is typically assumed that the ON and OFF periods are heavy-tailed, i.e.

$$\bar{F}_{\text{on}}(x) = C_{\text{on}} x^{-\alpha_{\text{on}}}, \quad \alpha_1 \in (1, 2), \tag{2.78}$$

$$\bar{F}_{\text{off}}(x) = C_{\text{off}} x^{-\alpha_{\text{off}}}, \quad \alpha_2 \in (1, 2), \tag{2.79}$$

where $C_{\text{on}}, C_{\text{off}}$ are finite and positive constants. More generally, the constants can be replaced by arbitrary slowly varying functions. Note also that, since $\alpha := \min(\alpha_1, \alpha_2) > 1$, the mean ON and OFF times are finite. The asymptotic decay of the autocovariance function is then as stated in the following lemma.

Lemma 2.7 *Consider the stationary ON–OFF process $W(t)$ $(t \ge 0)$ such that (2.78)–(2.79) hold with $\alpha_{\text{on}} < \alpha_{\text{off}}$. Then, as $u \to \infty$,*

$$\gamma_W(u) = cov\big(W(0), W(u)\big) \sim C_{\text{on}} \frac{\mu_{\text{off}}^2}{(\alpha_{\text{on}} - 1)(\mu_{\text{on}} + \mu_{\text{off}})^3} u^{-(\alpha_{\text{on}} - 1)}.$$

The proof of this result is technical and requires extended knowledge of renewal theory (see Heath et al. 1998 or Taqqu et al. 1997 for the Laplace transform method). It is therefore omitted here.

Example 2.14 (ON–OFF Process, Continued) The lemma implies that, if $\alpha_{\text{on}} \in (1,2)$, then $Cov(W(0), W(u))$ is not integrable and the process $W(t)$ is long-range dependent in the sense of Definition 1.4. For the integrated process $\int_0^t W(u)\,du$, we also have long-range dependence in the sense of Definition 1.6. This can be seen by applying Lemma 1.1 to obtain

$$\begin{aligned}\operatorname{var}\left(\int_0^t W(u)\,du\right) &= \int_0^t \left(\int_0^v \gamma_W(u)\,du\right) dv \\ &\sim C_{\text{on}} \frac{\mu_{\text{off}}^2}{\mu^3(\alpha_{\text{on}}-1)(2-\alpha_{\text{on}})(3-\alpha_{\text{on}})} t^{3-\alpha_{\text{on}}} \\ &=: C_{\text{on}} \sigma_{\text{on-off}}^2 t^{2H}\end{aligned}$$

with $H = (3-\alpha_{\text{on}})/2 > \frac{1}{2}$.

Example 2.15 (Infinite Source Poisson Process) Assume that τ_j ($j \in \mathbb{Z}$) is a Poisson process with rate λ. Set $Y_j(u) = 1\{0 < u < \eta_j\} Y_j^*$, where the random variables η_j, Y_j^*, X_j ($j \in \mathbb{Z}$) are mutually independent, i.i.d. and positive. The resulting process

$$W(t) = \sum_{j=-\infty}^{\infty} Y_j^* 1\{\tau_{j-1} \le t < \tau_{j-1} + \eta_j\} \quad (t > 0)$$

is called an *infinite source Poisson process*. Here, at times τ_{j-1}, shocks of size Y_j^* of a Poisson process occur and last for the duration η_j. At a given time t, all past shocks may contribute to $W(t)$. In queueing theory, one usually sets $Y_j^* = 1$, which leads to the following interpretation. Customers arrive according to a Poisson process τ_j, and each customer requests a service for a time period of length η_{j+1}. Given an infinite number of available servers, the process $W(t)$ describes the number of customers at time t. The model is called $M/G/\infty$. The letter "M" stands for "exponential" arrivals, "G" stands for a general service distribution, and ∞ for the number of servers. If Y_j^* and η_j have a finite mean, then $E[Y(u)] = E[U_1]P(\eta_1 > u)$, so that

$$E\big[W(t)\big] = \lambda E[Y_1] \int_0^\infty P(\eta_1 > u)\,du = \lambda E[Y_1] E[\eta_1].$$

Furthermore,

$$cov\big(W(0), W(t)\big) = E\big[Y_1^2\big]\lambda \int_t^\infty P(\eta_1 > u)\,du.$$

If we assume that $P(\eta_1 > u) = u^{-\alpha}L(u)$, $\alpha \in (1,2)$, then

$$cov\big(W(0), W(t)\big) \sim E\big[Y_1^2\big]\lambda\frac{1}{\alpha-1}t^{1-\alpha}L(t),$$

and hence,

$$\text{var}\big(W(t)\big) \sim 2E\big[Y_1^2\big]\lambda\frac{1}{(\alpha-1)(2-\alpha)(3-\alpha)}t^{3-\alpha}L(t).$$

Example 2.16 (Error Duration Process) Assume that $\tau_j = j$ ($j \in \mathbb{Z}$) is a deterministic sequence. Set $Y_j(u) = 1\{0 < u < \eta_{j+1}\}Y_j^*$, where the random variables η_j, Y_j^* ($j \in \mathbb{Z}$) are mutually independent, i.i.d. and positive. The resulting process

$$W(t) = \sum_{j=-\infty}^{\infty} Y_j^* 1\{j \le t < j + \eta_{j+1}\} \quad (t > 0)$$

is called the *error duration process*. Here, we have, at each deterministic time j, a shock of size Y_j^* that lasts for a period of length η_{j+1}. Although the model is similar to the infinite source Poisson process, due to the lack of a Poisson structure, computations are much more difficult. Hsieh et al. (2007) showed that, if the support of η is a subset of the positive integers, then

$$cov\big(W(0), W(k)\big) = \text{var}(Y_1)\sum_{j=k}^{\infty} p_j,$$

where $p_k = P(\eta_1 \ge k)$. In particular, if $p_j = P(\eta_1 \ge j) = L(j)j^{-\alpha}$ with $\alpha \in (1,2)$, then $W(j)$ ($j \in \mathbb{N}$) has long memory.

Models based on renewal processes are discussed for instance in Levy and Taqqu (1986, 1987, 2000, 2001), Abry and Flandrin (1994), Daley and Vesilo (1997), Resnick (1997), Heath et al. (1998), Daley (1999), Igloi and Terdik (1999), Daley et al. (2000), Pipiras and Taqqu (2000b), Gao and Rubin (2001), Kulik and Szekli (2001), Cappé et al. (2002), Kaj (2002), Maulik et al. (2002), Mikosch et al. (2002), Gaigalas and Kaj (2003), Hernández-Campos et al. (2002), Taqqu and Wolpert (1983), Leipus and Surgailis (2007).

2.2.5 Continuous-Time Models

2.2.5.1 General Remarks

Originally, in the early works of Mandelbrot, long-range dependence has been considered in connection with self-similar processes and thus in continuous time. Such

processes occur naturally as limiting processes (see Sects. 1.3.5 and 2.2.1). However, most statistical models and techniques have been developed for time series in discrete time. One of the reasons is the simplicity of fractional differencing (in discrete time) in general and fractional ARIMA models in particular. Yet, in some applications continuous time is essential. This is for instance the case in finance where stochastic differential equations and the Itô calculus, embedded in the world of Brownian motion, or more generally semimartingales, are key ingredients for pricing formulas. Moreover, the availability of high-frequency data has also increased the demand for fractional time series models in continuous time (see e.g. Bauwens et al. 2008, Bauwens and Hautsch 2009). It is therefore not surprising that, after the initial success of fractional time series models in discrete time, there has been a growing interest in developing suitable long-memory models in continuous time. Not surprisingly, many results are motivated by financial applications. At the same time, there has been an ongoing controversy how empirical findings of long-range dependence should be explained economically and whether long-memory processes make any sense in the financial context. A meanwhile classical controversy is for instance the question of arbitrage when it comes to modelling log-returns. For example, fractional Brownian motion is not a semimartingale so that the standard non-arbitrage arguments cannot be applied (see e.g. Mandelbrot 1971, Rogers 1997). Rogers (1997) shows however that a fractional Brownian motion can be modified so that long-range dependence is untouched and at the same time arbitrage is removed because the modified process is a semimartingale. Some authors suggest instead to link the apparent arbitrage to transaction costs. For example, Guasoni (2006) shows that proportional transaction costs can eliminate arbitrage opportunities from a geometric fractional Brownian motion with an arbitrary continuous deterministic drift. In their review paper, Bender et al. (2007) discuss the contradictory results in the literature on the existence or absence of a riskless gain and point out the importance of the chosen class of admissible trading strategies. Much less controversial is the application of continuous-time long-memory processes when it comes to modelling volatilities. The literature in this respect is fast growing. As an illustration, one particular example will be discussed below.

Many references to continuous-time processes with long-range dependence or antipersistence can be found in Embrechts and Maejima (2002), Bender et al. (2007), Biagini et al. (2008) and Mishura (2008). Further references are for instance Anh et al. (2009), Barndorff-Nielsen and Shephard (2001), Bender (2003a, 2003b), Brockwell and Marquardt (2005), Brody et al. (2002), Chambers (1996), Cheridito et al. (2003), Comte (1996), Comte and Renault (1996), Decreusefond and Üstünel (1999), Duncan et al. (2000), Elliott and van der Hoek (2003), Ercolani (2011), Guasoni (2006), Hu (2005), Hu and Nualart (2010), Hu and Øksendal (2003), Igloi and Terdik (1999), Jasiak (1998), Kleptsyna and Le Breton (2002), Kleptsyna et al. (2000), Le Breton (1998), Leonenko and Taufer (2005), Maejima and Yamamoto (2003), Mandelbrot (1997), Matsui and Shieh (2009), Norros et al. (1999), Pipiras and Taqqu (2000a, 2003), Simos (2008), Tsai (2009), Tsai and Chan (2005a, 2005b, 2005c, 2005d), Viano et al. (1994), Zähle (1998).

2.2.5.2 Volatility Models in Continuous Time

In Sect. 2.1.3 we gave examples of stochastic volatility models in discrete time. When dealing with high-frequency transaction-level data one needs corresponding models in continuous time. We recall that the classical Black–Scholes model assumes that a stock price $S(t)$ behaves like

$$dS(t) = \mu S(t)\,dt + yS(t)\,dB(t),$$

where $y > 0$, $\mu \in \mathbb{R}$, and $B(t)$ is a Brownian motion. The solution is

$$S(t) = S(0)\exp\big(\big(\mu - y^2/2\big)t + yB(t)\big),$$

the so-called geometric Brownian motion. The disadvantage of this model is that the volatility y is constant. The most common solution is to replace y by a (positive) stochastic process $Y(t)$. Barndorff-Nielsen and Shephard (2001) suggest a process defined as a strictly stationary solution of the stochastic differential equation

$$dY(t) = -aY(t)\,dt + \sigma\,dZ(t), \tag{2.80}$$

where $a > 0$, and $Z(\cdot)$ is a Lévy process with finite or infinite variance. The solution is given by

$$Y(t) = e^{-at}Y(0) + \sigma\int_0^t e^{-a(t-u)}\,dZ(u).$$

If Z is a Lévy subordinator (that is, a strictly increasing process), then the process $Y(t)$ is strictly positive and hence may play the role of a volatility. On the other hand, if $Z(t)$ is a standard Brownian motion, then the equation is interpreted as an Itô equation, and we obtain an Ornstein–Uhlenbeck process. If Z is a Lévy process, then Y is called CAR(1), that is, a Continuous Autoregressive Process of order 1. The name can be explained as follows. Let $Z(u) = B(u)$ be a Brownian motion. If we consider the process $Y(t)$ at discrete time points $t \in \mathbb{N}$, then for $X_t = Y(t)$, we have

$$X_{t+1} = e^{-a}X_t + \sigma e^{-a(t+1)}\int_t^{t+1} e^{\mu u}\,dB(u)$$

or

$$X_{t+1} = \phi X_t + R_t,$$

where $\phi = e^{-a} < 1$ and $R_t \sim N(0, \frac{1}{2a}(e^{2a} - 1)\sigma^2)$. Hence, we obtain an AR(1) process. This corresponds to the fact that processes $Y(t)$ obtained as solutions of (2.80) have summable covariances.

One of the possibilities to incorporate long-range dependence is to consider the modified stochastic differential equation

$$dY(t) = -\mu Y(t)\,dt + \sigma\,dB_H(t), \tag{2.81}$$

where $B_H(t)$ is a fractional Brownian motion. The strictly stationary solution has long memory; however, the process $Y(t)$ is not strictly positive. This is a common problem in modelling long memory in volatility. We refer to Maejima and Yamamoto (2003), Cheridito et al. (2003), Buchmann and Klüppelberg (2005, 2006), Brockwell and Marquardt (2005) and Hu and Nualart (2010).

A possible solution to assure positivity is to generate long memory by aggregation of short-memory Ornstein–Uhlenbeck processes, as in the case of AR(1) sequences. We refer to Igloi and Terdik (1999), Oppenheim and Viano (2004), Leonenko and Taufer (2005) and Barndorff-Nielsen and Stelzer (2011a, 2011b).

A different model stems from empirical observations that suggest that durations between trades exhibit long memory, which then propagates to volatility. A possible model that incorporates this behaviour can be described as follows (see Deo et al. 2009). Suppose that the log-price process can be described as

$$P^*(t) = \log P(t) = \log P(0) + \sum_{j=1}^{N(t)} Y_j^*,$$

where $N(t)$ is the number of transactions up to time t, and Y_j^* ($j \in \mathbb{N}$) is an i.i.d. sequence of zero-mean random variables with finite variance. The sequence represents unobservable "shocks" at transaction times. Note that $\sum_{j=1}^{N(t)} Y_j^*$ is almost the renewal reward process studied in Example 2.12. Indeed, in the setting of that example we have

$$\int_0^t W(u)\,du = \sum_{j=0}^{N(t)-1} Y_j^* + (t - \tau_{N(t)-1}) Y_{N(t)}^* \quad (t > \tau_0)$$

and $\int_0^t W(u)\,du = tY_0$ if $0 < t < \tau_0$. Thus, we expect that the process $P^*(t)$ has similar long-memory properties as the integrated renewal reward process. Indeed, we have

$$\begin{aligned}\operatorname{var}\big(P^*(t)\big) &= E\big\{E\big[\big(P^*(t) - E\big(P^*(t)\big)\big)^2 | N\big]\big\} \\ &= E\big[\operatorname{var}\big(P^*(t)|N\big)\big] + \operatorname{var}\big\{E\big[P^*(t)|N\big]\big\},\end{aligned}$$

where conditioning is on the entire counting process $N(t)$ ($t \geq 0$). Since the random variables Y_j^* are i.i.d., we conclude that

$$\begin{aligned}\operatorname{var}\big(P^*(t)\big) &= E\big[N(t)\big]E^2\big[Y_1^2\big] + E\big[Y_1^2\big]\operatorname{var}\big(N(t)\big) \\ &= \lambda t E\big[Y_1^2\big] + E^2[Y_1]\operatorname{var}\big(N(t)\big).\end{aligned}$$

Thus, the variance grows faster than at a linear rate if and only if $E[Y_1] \neq 0$. However, here we assumed that Y_j are centered. Hence, there is no long memory in the log-price process $P^*(t)$. On the other hand, there is long memory in the so-called *realized* volatility. To be more specific, let R_j ($j \in \mathbb{N}$) be log-returns at equally spaced

calendar times,

$$R_j = \log P(j) - \log P(j-1) = \sum_{j=N(j-1)+1}^{N(j)} Y_j \quad (j \in \mathbb{N}).$$

A realized volatility at time i is defined as

$$V_i = \sum_{j=1}^{i} R_j^2.$$

Then

$$\operatorname{var}(V_i) = E\big[\operatorname{var}(V_i|N)\big] + \operatorname{var}\big[E(V_i|N)\big]. \tag{2.82}$$

Noting that

$$E\big[R_j^2|N\big] = E\big[Y_1^2\big]\big(N(j) - N(j-1)\big), \tag{2.83}$$

we obtain

$$\operatorname{var}\big[E(V_i|N)\big] = E\big[Y_1^2\big]\operatorname{var}\left(\sum_{j=1}^{i}\big(N(j) - N(j-1)\big)\right) = E\big[Y_1^2\big]\operatorname{var}\big(N(i)\big). \tag{2.84}$$

(Furthermore, under additional moment assumption, the second term in (2.82) does not contribute asymptotically.) Thus, Eqs. (2.82) and (2.84) imply that the variance of V_i is proportional to the variance of $N(i)$, whereas (2.83) means that the expected value of V_i is proportional to $E[N(i)]$. Consequently, any counting process N that is LRcD implies long memory in the realized volatility.

2.2.6 *Fractals*

Ever since the pioneering work by Benoit Mandelbrot and his highly influential series of books (Mandelbrot 1977, 1983), fractals have become a prime example of mathematical ideas immersing not only all scientific disciplines but also virtually all aspects of daily live. Apart from their usefulness, one key to the popularity of fractals is their beauty (see e.g. Peitgen and Richter 1986). The occurrence of fractal structures in nature is meanwhile widely accepted, and numerous physical explanations have been suggested in the vast literature on the topic (see e.g. Pietronero and Tosatti 1986, Avnir 1989, Becker and Dörfler 1989, Aharoni and Feder 1990, Heck and Pedang 1991, Vicsek 1992, McCauley 1993, Barnsley 1993, Xie 1993, McMullen 1994, Gouyet 1996, Rodriguez-Iturbe and Rinaldo 1997, Turcotte 1997, Meakin 1998, Mandelbrot 1999, 2002). For an excellent elementary but mathematical introduction to fractals, see e.g. Falconer (2003). A very brief introduction will be given in Sect. 3.6.

Many simple fractals are self-similar geometric objects. Loosely speaking, this means that the same geometric shapes can be seen no matter how closely one looks at it. The analogous property for random geometric objects is probabilistic self-similarity. In particular, for stochastic processes X_t ($t \in \mathbb{R}$), we are back to the definition introduced in Sect. 1.3.5. Since stochastic self-similarity combined with stationary increments leads to hyperbolic behaviour of the spectral density near the origin, this leads to the notion of long memory and antipersistence. Thus, the occurrence of fractals in nature (and in the arts) is another "physical" explanation why long memory and antipersistence are fundamentally important for explaining natural phenomena.

Chapter 3
Mathematical Concepts

In this chapter we present some mathematical concepts that are useful when deriving limit theorems for long-memory processes.

We start with a general description of univariate orthogonal polynomials in Sect. 3.1, with particular emphasis on Hermite polynomials in Sect. 3.1.2. Under suitable conditions, a function G can be expanded into a series

$$G(x) = \sum_{j=0}^{\infty} g_j H_j(x)$$

with respect to an orthogonal basis consisting of Hermite polynomials $H_j(\cdot)$ ($j \in \mathbb{N}$). Such expansions are used to study sequences $G(X_t)$ where X_t ($t \in \mathbb{Z}$) is a Gaussian process with long memory (see Sect. 4.2.3). Hermite polynomials can also be extended to the multivariate case. This is discussed in Sect. 3.2.

Subsequently, we discuss the validity of expansions in terms of so-called Appell polynomials. These results are applied in Sect. 4.2.5 to study sequences $G(X_t)$ where X_t is a linear process. The problem becomes very difficult when the assumption of normality is dropped, because of the loss of orthogonality. A detailed theory is presented in Sect. 3.3. This section also includes the notion of Wick products, which are very useful in the context of limit theorems for long-memory processes. The main reason is the so-called diagram formula (see Sect. 3.4.3), which simplifies the calculation of joint cumulants.

An introduction to wavelets is given in Sect. 3.5. Wavelet basis functions are defined via scaling and are therefore natural tools when it comes to stochastic processes with hyperbolic scaling properties. Moreover, they are known to be useful in the context of nonparametric estimation of trends and other functions that may not be continuous or differentiable everywhere. The statistical applications will mainly be discussed in Chaps. 5 and 7. Here, in Sect. 3.5, basic formulas needed in wavelet analysis are introduced.

The chapter concludes with a brief introduction to fractals, including basic definitions and a selection of essential results.

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_3, © Springer-Verlag Berlin Heidelberg 2013

3.1 Orthogonal Polynomials and Expansions

3.1.1 Orthogonal Polynomials—General Introduction

Orthogonal polynomials play an important role in mathematics in general. For a systematic introduction to the topic, see for instance Szegö (1939), Jackson (1941, 2004), Boas and Buck (1964), Chihara (1978) and El Attar (2006) (also Abramowitz and Stegun 1965, Chap. 22). As we will see in Chap. 4, in the context of long-memory processes orthogonal polynomials are very useful for deriving limit theorems and defining certain classes of non-Gaussian and nonlinear processes.

Many important orthogonal polynomials can be defined via a differential equation of the following form. Let $Q(x)$ be a polynomial of degree $p_Q \le 2$, and $L(x)$ a linear polynomial. Then we are looking for a twice differentiable function f such that

$$Q(x)f''(x) + L(x)f'(x) - \lambda f(x) = 0. \tag{3.1}$$

Defining the differential operator $Df = Qf'' + Lf'$, this can be rewritten as

$$Df = \lambda f, \tag{3.2}$$

i.e. f is an eigenfunction, and λ an eigenvalue of D. For suitable choices of Q and L, it turns out that the solutions f are polynomials that are orthogonal to each other with respect to a suitable scalar product and the corresponding eigenvalues depend on the degree of the polynomial Q and L. More specifically, $f \in \{P_0, P_1, \dots\}$, where P_j are polynomials of degree j, and the corresponding eigenvalues are given by

$$\lambda_j = j\left(\frac{j-1}{2}Q'' + L'\right). \tag{3.3}$$

Note that for $Q(x) = a_0 + a_1 x + a_2 x^2$ and $L(x) = b_0 + b_1 x$, $Q'' \equiv 2a_2$ and $L' = b_1$, and the eigenvalues are just constants. Orthogonality is achieved in the following sense:

Lemma 3.1 *Define* $\log R(x)$ *as an antiderivative of* $\frac{L(x)}{Q(x)}$, *i.e.*

$$R(x) = \exp\left(\int \frac{L(y)}{Q(y)}\,dy\right), \qquad w(x) = \frac{R(x)}{Q(x)}$$

and, for real functions on the real line, the scalar product

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)g(x)w(x)\,dx.$$

Furthermore, let

$$S_{jk}(x) = R(x)\big[P_j(x)P_k'(x) - P_j'(x)P_k(x)\big]$$

and assume that, for $j \neq k$, $\lambda_j \neq \lambda_k$ *and*

$$\lim_{x \to \pm\infty} S_{jk}(x) = 0.$$

Then, for all $j \neq k$,

$$\langle P_j, P_k \rangle = 0.$$

Proof By definition, Eq. (3.1) holds for P_j and P_k, that is

$$Q P_j'' + L P_j' - \lambda_j P_j = 0,$$
$$Q P_k'' + L P_k' - \lambda_k P_k = 0.$$

Multiplying these two equations from the left by $RQ^{-1}P_k$ and $RQ^{-1}P_j$, respectively, we obtain

$$R P_j'' P_k + R Q^{-1} L P_j' P_k - \lambda_j R Q^{-1} P_j P_k = 0,$$
$$R P_k'' P_j + R Q^{-1} L P_k' P_j - \lambda_k R Q^{-1} P_k P_j = 0.$$

Subtracting the left-hand terms from each other and using $w = RQ^{-1}$ yields

$$R\big(P_k'' P_j - P_j'' P_k\big) + R Q^{-1} L\big(P_k' P_j - P_j' P_k\big) = (\lambda_k - \lambda_j) P_j P_k w. \tag{3.4}$$

A straightforward computation yields

$$\begin{aligned} \frac{d}{dx} S_{jk} &= \frac{d}{dx}\big\{R\big[P_j P_k' - P_j' P_k\big]\big\} \\ &= R\big(P_k'' P_j - P_j'' P_k\big) + R Q^{-1} L\big(P_k' P_j - P_j' P_k\big), \end{aligned}$$

and we recognize the latter expression as the left-hand side of (3.4). Hence, integrating both sides of (3.4) from $-\infty$ to ∞, we obtain

$$\big[S_{jk}(x)\big]_{-\infty}^{\infty} = (\lambda_j - \lambda_k)\langle P_j, P_k \rangle.$$

Thus, $[S_{jk}(x)]_{-\infty}^{\infty} = 0$ and $\lambda_j \neq \lambda_k$ implies $\langle P_j, P_k \rangle = 0$. □

An explicit formula for P_j can be given as follows (Rodrigues 1816).

Lemma 3.2 (Rodrigues Formula) *Under suitable conditions on* Q *and* L,

$$P_j(x) = c_j \frac{1}{w(x)} \frac{d^j}{dx^j}\big[w(x) Q^j(x)\big],$$

where c_j *are suitable constants.*

In the following sections we will discuss special examples that are of particular interest in the context of long-memory processes.

3.1.2 Hermite Polynomials and Hermite Expansion

In the context of Gaussian long-memory processes, the most useful orthogonal polynomials are Hermite polynomials. They are defined by $Q(x) = 1$, $L(x) = -x$ and $\lambda_j = -j$, or in other words, they fulfill the Hermite differential equation

$$f''(x) - xf'(x) + jf(x) = 0. \tag{3.5}$$

Using the previous notation, we have

$$R(x) = \exp\left(\int \frac{L(y)}{Q(y)}\,dy\right) = \exp\left(-\int y\,dy\right) = e^{-\frac{x^2}{2}} = w(x). \tag{3.6}$$

Rodrigues' formula (Lemma 3.2) then yields

$$P_j(x) = c_j e^{\frac{x^2}{2}} \frac{d^j}{dx^j}\left[e^{-\frac{x^2}{2}}\right].$$

The standard choice of c_j is $(-1)^j$. Thus, we have the following definition.

Definition 3.1 The jth Hermite polynomial $H_j(x)$ $(j = 0, 1, 2, \ldots)$ is equal to

$$P_j(x) := H_j(x) = (-1)^j \exp\left(\frac{x^2}{2}\right)\frac{d^j}{dx^j}\exp\left(-\frac{x^2}{2}\right). \tag{3.7}$$

In particular,

$$\begin{aligned} &H_0(x) = 1, \qquad H_1(x) = x, \qquad H_2(x) = x^2 - 1, \qquad H_3(x) = x^3 - x, \\ &H_4(x) = x^4 - 6x^2 + 3, \qquad H_5(x) = x^5 - 10x^3 + 15x. \end{aligned} \tag{3.8}$$

Lemma 3.1 and direct calculation for $j = k$ imply

$$\langle H_j, H_k\rangle = \int_{-\infty}^{\infty} H_j(x)H_k(x)\varphi(x)\,dx = \delta_{jk}\cdot j!, \tag{3.9}$$

where

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

is the standard normal density function. The L^2-space equipped with this scalar product will be denoted by

$$L^2(\mathbb{R}, \varphi) = \left\{G : \mathbb{R} \to \mathbb{R}, \|G\|^2 = \int G^2(x)\varphi(x)\,dx < \infty\right\}.$$

Since H_j are orthogonal in $L^2(\mathbb{R}, \varphi)$, a natural question is whether they build a basis, i.e. whether any function $G \in L^2(\mathbb{R}, \varphi)$ has a unique representation (in

$L^2(\mathbb{R}, \varphi)$) in terms of Hermite polynomials. The answer is affirmative. Before we state the result and prove it, we note that the Hermite polynomials fulfill the following recursive formulas:

$$H_{j+1}(x) = xH_j(x) - H_j'(x), \tag{3.10}$$

$$H_{j+1}'(x) = (j+1)H_j(x),$$

$$H_{j+1}(x) = xH_j(x) - jH_{j-1}(x). \tag{3.11}$$

Lemma 3.3 *$\{H_j,\ j = 0, 1, 2, \ldots\}$ is an orthogonal basis in $L^2(\mathbb{R}, \varphi)$.*

Proof It only remains to show that the family H_j ($j \in \mathbb{N}$) is complete, i.e. that every function in $L^2(\mathbb{R}, \varphi)$ can be represented by Hermite polynomials. From the recursive formulas (3.10) it follows that $H_0, \ldots, H_k$ span the same space (in $L^2(\mathbb{R}, \varphi)$) as $1, x, x^2, \ldots, x^k$. Thus, it is sufficient to show that

$$\langle x^j, G\rangle = \int_{-\infty}^{\infty} x^j G(x)\varphi(x)\,dx = 0 \quad (j = 0, 1, 2, \ldots)$$

implies $G(x) \equiv 0$ (in $L^2(\mathbb{R}, \varphi)$). Consider the complex function

$$m(z) = \sum_{j=0}^{\infty} \frac{z^j}{j!}\langle x^j, G\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(x)e^{zx - \frac{x^2}{2}}\,dx.$$

Then $m(z)$ is an entire function, and

$$m(it) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(x)e^{-\frac{x^2}{2}} \cdot e^{itx}\,dx$$

is the Fourier transform of $\tilde{G}(x) = G(x)\exp(-x^2/2)/\sqrt{2\pi}$. (Recall that an entire function is infinitely complex differentiable and equal to its Taylor series everywhere.) However, if the Fourier transform is equal to zero for all $t \in \mathbb{R}$, then $\tilde{G} \equiv 0$ and hence $G \equiv 0$ (almost everywhere). □

Why are Hermite polynomials important in the context of random variables and stochastic processes? Suppose that Z is a standard normal random variable and define $Y = G(Z)$. Then Y can be represented uniquely (in $L^2(\Omega)$) in terms of Hermite polynomials $H_j(Z)$:

Lemma 3.4 *Let $Z \sim N(0, 1)$, and let G be such that $E[G(Z)] = 0$ and $E[G^2(Z)] < \infty$. Then $G(Z)$ has the unique representation (in $L^2(\Omega)$)*

$$G(Z) = \sum_{k=1}^{\infty} g_k H_k(Z) = \sum_{k=1}^{\infty} \frac{J(k)}{k!} H_k(Z) \tag{3.12}$$

with (Hermite) coefficients

$$g_k = \frac{J(k)}{\|H_k\|^2} = \frac{J(k)}{\langle H_k, H_k \rangle} = \frac{J(k)}{k!}, \tag{3.13}$$

$$J(k) = \langle G, H_k \rangle = E\big[G(Z)H_k(Z)\big]. \tag{3.14}$$

Sometimes $J(k)$ instead of g_k are called Hermite coefficients. As we will see in Sect. 4.2.3, it is essential to know what the lowest value of k with a nonzero Hermite coefficient is:

Definition 3.2 Let Z be a standard normal random variable, and G be a function such that $E[G(Z)] = 0$ and $E[G^2(Z)] < \infty$. Then the Hermite rank m of G is the smallest integer $k \geq 1$ such that

$$g_k = E\big[G(Z)H_k(Z)\big] \neq 0.$$

Another useful definition is the following.

Definition 3.3 For $x \in \mathbb{R}$, $z \in \mathbb{C}$,

$$M_{\text{Hermite}}(x, z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} H_k(x)$$

is called the generating function.

We claim that

$$M_{\text{Hermite}}(x, z) = \exp\left(xz - \frac{z^2}{2}\right). \tag{3.15}$$

Indeed, $\exp(xz - \frac{z^2}{2})$, as a function of z, can be expanded as

$$\exp\left(xz - \frac{z^2}{2}\right) = \sum_{k=0}^{\infty} \frac{z^k}{k!} \left[\frac{d^k}{dz^k} \exp\left(xz - \frac{z^2}{2}\right)\right]_{z=0}.$$

Now, formula (3.15) follows by noting

$$\left[\frac{d^k}{dz^k} \exp\left(xz - \frac{z^2}{2}\right)\right]_{z=0} = H_k(x).$$

Formula (3.15) implies that if $X \sim N(\mu, 1)$, then $E[H_j(X)] = \mu^j$. Indeed,

$$E\big[M_{\text{Hermite}}(X, z)\big] = E\left[\exp\left(Xz - \frac{z^2}{2}\right)\right] = \exp(\mu z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} \mu^k$$

and

$$E\big[M_{\text{Hermite}}(X,z)\big] = \sum_{k=0}^{\infty} \frac{z^k}{k!} E\big[H_k(X)\big].$$

Thus, the formula for moments comes from comparing coefficients of the both expansions. In particular, for $\mu = 0$, $E[H_j(X)] = 0$ (which also follows by orthogonality and $H_0(x) \equiv 1$).

Furthermore, for real numbers $a_1, \ldots, a_k$ such that $a_1^2 + \cdots + a_k^2 = 1$, we have

$$H_q\left(\sum_{j=1}^{k} a_j x_j\right) = \sum_{q_1+\cdots+q_k=q} \frac{q!}{q_1! \ldots q_k!} \prod_{j=1}^{k} a_j^{q_j} H_{q_j}(x_j). \tag{3.16}$$

This formula is particularly useful to derive the following lemma.

Lemma 3.5 *For a pair of jointly standard normal random variables Z_1, Z_2 with covariance $\rho = cov(Z_1, Z_2)$, we have*

$$cov\big(H_m(Z_1), H_m(Z_2)\big) = m!\rho^m, \tag{3.17}$$

whereas for $j \neq k$,

$$cov\big(H_j(Z_1), H_k(Z_2)\big) = 0. \tag{3.18}$$

Proof Write $Z_2 = \rho Z_1 + \sqrt{1-\rho^2}\xi$, where ξ is independent of Z_1 and standard normal. Then, applying (3.16) and recalling that $E[H_{q_2}(\xi)] = 0$ unless $q_2 = 0$, we have

$$\begin{aligned}
&E\big[H_m(Z_1)H_m(Z_2)\big] \\
&\quad = E\big[H_m(Z_1)H_m\big(\rho Z_1 + \sqrt{1-\rho^2}\xi\big)\big] \\
&\quad = \sum_{q_1+q_2=m} \frac{m!}{q_1!q_2!} \rho^{q_1} \big(\sqrt{1-\rho^2}\big)^{q_2} E\big[H_m(Z_1)H_{q_1}(Z_1)\big] E\big[H_{q_2}(\xi)\big] \\
&\quad = \rho^m \big[H_m^2(Z_1)\big] = \rho^m \langle H_m, H_m\rangle = m!\rho^m.
\end{aligned}$$

In the latter equation we used formula (3.9) for the inner product of Hermite polynomials. This proves (3.17). The second formula (3.18) can be proven analogously. □

Lemma 3.4 implies that the variance of $G(Z)$ can be decomposed into (orthogonal) contributions of the Hermite coefficients,

$$\operatorname{var}\big(G(Z)\big) = \sum_{k=1}^{\infty} g_k^2 k! = \sum_{k=1}^{\infty} \frac{J^2(k)}{k!}. \tag{3.19}$$

Similarly, Lemma 3.5 implies

$$cov\big(G(Z_1), G(Z_2)\big) = \sum_{k=1}^{\infty} \frac{J^2(k)}{k!}\rho^k. \tag{3.20}$$

Example 3.1 Let $G(x) = H_1(x) = x$. Then $J(1) = cov(G(Z)Z) = \text{var}(Z^2) = 1$, so that G has Hermite rank 1. (This can also be seen directly because H_2 is by definition orthogonal to all other Hermite polynomials.) For Z_1, Z_2 standard normal with $\rho = cov(Z_1, Z_2)$, we obviously have $cov(G(Z_1), G(Z_2)) = \rho J(1)/1! = \rho$.

Example 3.2 Let $G(x) = H_2(x) = x^2 - 1$. Then, $J(1) = E(Z^3 - Z) = 0$ and $J(2) = E[(Z^2 - 1)^2] = 2$, so that the Hermite rank is 2. (This can also be seen directly because H_2 is by definition orthogonal to all other Hermite polynomials.) Moreover, $cov(Z_1^2 - 1, Z_2^2 - 1) = \rho^2 J^2(2)/2! = 2\rho^2$.

3.1.3 Laguerre Polynomials

The Laguerre polynomials $P_j(x) = L_j^{(\alpha)}(x)$ are obtained from (3.1) by setting $Q(x) = x$, $L(x) = \alpha + 1 - x$ (with $\alpha > -1$) and $\lambda_j = -j$, and considering $x \geq 0$ only. Thus, $L_j^{(\alpha)}$ are solutions of Laguerre's equation

$$xf''(x) + (\alpha + 1 - x)f'(x) + jf(x) = 0 \quad (x \geq 0). \tag{3.21}$$

This implies

$$\begin{aligned} R(x) &= \exp\left[\int \left(\frac{\alpha+1}{y} - 1\right) dy\right] = x^{\alpha+1}e^{-x}, \\ w(x) &= x^{\alpha}e^{-x}1\{x \geq 0\}, \end{aligned} \tag{3.22}$$

and

$$P_j(x) := L_j^{(\alpha)}(x) = c_j x^{-\alpha} e^x \frac{d^j}{dx^j}\big(x^{j+\alpha}e^{-x}\big).$$

The usual standardization is $c_j = \Gamma(\alpha + 1)/\Gamma(j + \alpha + 1)$, so that we obtain the following definition.

Definition 3.4 The jth generalized or associated Laguerre polynomials $L_j^{(\alpha)}(x)$ $(j \geq 0)$ are defined by

$$L_j^{(\alpha)}(x) = \frac{\Gamma(\alpha + 1)}{\Gamma(j + \alpha + 1)} x^{-\alpha} e^x \frac{d^j}{dx^j}\big(x^{j+\alpha}e^{-x}\big).$$

For $\alpha = 0$, $L_j^{(0)} =: L_j$ are called (simple) Laguerre polynomials.

For probabilistic and statistical applications, the most interesting case is $\alpha = 0$ because it can be associated with the exponential distribution. The first few Laguerre polynomials with $\alpha = 0$ are

$$L_0(x) = 1, \qquad L_1(x) = -x + 1, \qquad L_2(x) = \frac{1}{2}\big(x^2 - 4x + 2\big), \qquad \ldots.$$

By definition (simple) Laguerre polynomials are orthonormal in $L^2(\mathbb{R}_+, \psi)$ where $\psi(x) = \exp(-x)1\{x \geq 0\}$ is the standard exponential density function, i.e.

$$\langle L_j, L_k \rangle = \int_0^\infty L_j(x) L_k(x) e^{-x}\, dx = \delta_{jk}.$$

Similarly to Hermite polynomials, one can show that every function in $L^2(\mathbb{R}_+, \psi)$ can be represented by Laguerre polynomials since $\{L_j, j = 0, 1, 2, \ldots\}$ is an orthonormal basis in $L^2(\mathbb{R}_+, \psi)$.

Thus, for any function $G \in L^2(\mathbb{R}_+, \psi)$, there is a unique representation

$$G(x) = \sum_{k=0}^{\infty} g_k L_k(x),$$

$$g_k = \langle G, L_k \rangle = \int_0^\infty G(x) L_k(x) e^{-x}\, dx.$$

In other words, if Z is a standard exponential random variable, then the transformed random variable $G(Z)$ can be represented as

$$G(Z) = \sum_{k=0}^{\infty} g_k L_k(Z),$$

$$g_0 = E\big[G(Z)\big], \qquad g_k = cov\big(G(Z), L_k(Z)\big) \quad (k \geq 1).$$

In analogy to Hermite polynomials, one can then define the Laguerre rank.

Definition 3.5 Let Z be a standard exponential random variable, and G a function such that $E[G(Z)] = 0$ and $E[G^2(Z)] < \infty$. Then the Laguerre rank m of G is the smallest integer $k \geq 1$ such that

$$g_k = E\big[G(Z) L_k(Z)\big] \neq 0.$$

Further useful properties of Laguerre polynomials are for instance

$$\begin{aligned} (j+1) L_{j+1}(x) &= (2j + 1 - x) L_j(x) - j L_{j-1}(x), \\ x L_j'(x) &= j\big[L_j(x) - L_{j-1}(x)\big]. \end{aligned} \tag{3.23}$$

The importance of Laguerre polynomials for random variables and stochastic processes is due to the importance of the exponential distribution, which is obtained

for instance when considering interarrival times between events of a homogeneous Poisson process. Applications include for example survival analysis (e.g. in medicine or credit risk modelling) and queuing networks (e.g. computer networks).

Example 3.3 Let Z be a standard exponential random variable. Estimates of the survival function $S(z_0) = P(Z > z_0)$ are essentially based on the variable $Y = 1\{Z > z_0\}$. The Laguerre rank of the centred variable $G(Z) = Y - E(Y)$ is equal to 1 because

$$\begin{aligned} g_1 = \langle G, Z\rangle &= -\int_0^\infty G(x) x e^{-x}\, dx \\ &= -\int_{z_0}^\infty x e^{-x}\, dx = -e^{-z_0}(z_0+1) \neq 0 \quad \text{for all } z_0 > 0. \end{aligned}$$

This plays a role when observed life times $Z_1, \ldots, Z_n$ are strongly correlated. For instance Leonenko et al. (2001, 2002) derive the asymptotic distribution of the Kaplan–Meier estimator for censored survival times using a Laguerre polynomial expansion and the notion of a Laguerre rank, analogous to the Hermite polynomial expansion and Hermite rank.

3.1.4 Jacobi Polynomials

Jacobi polynomials $P_j^{(\alpha,\beta)}$ are obtained from (3.1) by setting $Q(x) = 1 - x^2$ and $L(x) = \beta - \alpha - (\alpha + \beta + 2)x$, where $\alpha, \beta > -1$, and x is restricted to $(-1, 1)$. The eigenvalues are $\lambda_j = -j(j + 1 + \alpha + \beta)$. Thus, Jacobi's equation is

$$\left(1-x^2\right) f''(x) + \left[\beta - \alpha - (\alpha+\beta+2)x\right] f'(x) - \lambda f(x) = 0 \quad (-1 < x < 1). \tag{3.24}$$

The solutions are given by

$$P_j^{(\alpha.\beta)}(x) = \frac{\Gamma(\alpha + j + 1)}{j!\Gamma(\alpha + \beta + j + 1)} \sum_{k=0}^{j} \binom{j}{k} \frac{\Gamma(\alpha + \beta + j + k + 1)}{\Gamma(\alpha + \beta + 1)} \left(\frac{x-1}{2}\right)^k, \tag{3.25}$$

where $-1 < x < 1$. Orthogonality is obtained with respect to the weight function (on the interval $(-1, 1)$)

$$w(x) = (1 - x)^\alpha (1 + x)^\beta, \tag{3.26}$$

i.e. for $j \neq k$,

$$\left\langle P_j^{(\alpha,\beta)}, P_k^{(\alpha,\beta)} \right\rangle = \int_{-1}^{1} P_j^{(\alpha,\beta)}(x) P_k^{(\alpha,\beta)}(x)(1 - x)^\alpha (1 + x)^\beta\, dx = 0. \tag{3.27}$$

Why are Jacobi polynomials of interest in the context of random variables and stochastic processes? Special types of Jacobi polynomials, so-called Gegenbauer and Legendre polynomials (see next two sections), are important for modelling seasonal long-range dependence (see Sect. 5.12.2). Also, Jacobi polynomials come up in the context of efficient regression estimation (see Sect. 7.1.2).

3.1.5 Gegenbauer Polynomials

In the following we will use Pochhammer's symbol $(a)_k = \Gamma(a+k)/\Gamma(a)$ and the notation $F_{p,q}$ for the hypergeometric function

$$F_{p,q}(z|a_1,\ldots,a_p;b_1,\ldots,b_q) = \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k \cdots (a_p)_k}{(b_1)_k (b_2)_k \cdots (b_q)_k} z^k.$$

Gegenbauer polynomials are Jacobi polynomials with $\alpha = \beta > -\frac{1}{2}$, i.e. with $Q(x) = 1 - x^2$, $L(x) = -2(\alpha+1)x$ and $\lambda_j = -j(j+1+2\alpha)$. Usually, one uses a new parameter $\kappa = \alpha + \frac{1}{2} > 0$. Then we obtain the definition of Gegenbauer polynomials $C_j^{(\kappa)}$,

$$C_j^{(\kappa)}(x) = \frac{(2\kappa)_j}{(\kappa+\frac{1}{2})_j} P_j^{(\kappa-\frac{1}{2},\kappa-\frac{1}{2})}(x), \tag{3.28}$$

which are solutions of Gegenbauer's equation

$$\left(1-x^2\right) f''(x) - 2\left(\kappa + \frac{1}{2}\right) x f'(x) + j(j+2\kappa) f(x) = 0.$$

(Note that for $\kappa = 0$, one defines $C_j^{(0)}(1) := \frac{2}{j}$.) The jth polynomial can also be written as

$$C_j^{(\kappa)}(x) = F_{p,q}\left(\frac{1-x}{2}\middle| -j, j+2\kappa; \kappa+\frac{1}{2}\right) \cdot \frac{(2\kappa)_j}{j!}.$$

For numeric calculations, the recursion formula

$$C_j^{(\kappa)}(x) = \frac{1}{j}\left[2x(j+\kappa-1)C_{j-1}^{(\kappa)}(x) - (j+2\kappa-2)C_{j-2}^{(\kappa)}(x)\right]$$

is useful, with initiation $C_0^{(\kappa)}(x) = 1$, $C_1^{(\kappa)}(x) = 2\kappa \cdot x$. The Gegenbauer polynomials are orthogonal with respect to the weight function $w(x) = (1-x^2)^{\kappa-\frac{1}{2}}$ on $(-1,1)$. More specifically,

$$\left\langle C_j^{(\kappa)}, C_k^{(\kappa)} \right\rangle = \int_{-1}^{1} C_j^{(\kappa)}(x) C_k^{(\kappa)}(x) \left(1-x^2\right)^{\kappa-\frac{1}{2}} dx = \delta_{jk} \cdot \frac{\pi 2^{1-2\kappa} \Gamma(j+2\kappa)}{j!(j+\kappa)\Gamma^2(\kappa)}.$$

In the context of long-memory processes, Gegenbauer polynomials are easier to understand by looking at their generating function

$$M_G(x, z; \kappa) = \sum_{j=0}^{\infty} z^j C_j^{(\kappa)}(x) = \left(1 - 2xz + z^2\right)^{-\kappa}.$$

The so-called Gegenbauer processes (see Gray et al. 1989, 1994; Giraitis and Leipus 1995; Woodward et al. 1998; Sect. 5.12.2) are defined as stationary solutions of

$$\varphi(B)\left(1 - 2uB + B^2\right)^d X_t = \psi(B)\varepsilon_t$$

where φ, ψ are the usual autoregressive and moving average polynomials, ε_t ($t \in \mathbb{Z}$) are i.i.d. zero mean random variables, and $u \in (-1, 1)$ is a parameter. More explicitly, this may be written as

$$M_G(u, B; -d)X_t = \sum_{j=0}^{\infty} C_j^{(-d)}(u) B^j X_t = \varphi^{-1}(B)\psi(B)\varepsilon_t = Y_t,$$

where Y_t is an ARMA(p, q) process. In particular, for $p = q = 0$, the coefficients π_j in the autoregressive representation $X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + \varepsilon_t$ are Gegenbauer polynomials evaluated at u.

3.1.6 Legendre Polynomials

Legendre polynomials are Gegenbauer polynomials with $\alpha = \beta = 0$, i.e. $\kappa = \frac{1}{2}$. Thus, $Q(x) = 1 - x^2$, $L(x) = -2x$, $\lambda_j = -j(j+1)$, and Legendre's equation is

$$\left(1 - x^2\right) f''(x) - 2x f'(x) + j(j+1) f(x) = 0.$$

More explicitly, Legendre polynomials are given by $P_0 = 1$ and

$$P_j(x) = \frac{2^{-j}}{j!} \frac{d^j}{dx^j}\left[\left(x^2 - 1\right)^j\right].$$

They are orthogonal with respect to the weight function $w(x) = 1\{-1 < x < 1\}$,

$$\langle P_j, P_k \rangle = \int_{-1}^{1} P_j(x) P_k(x)\, dx = \delta_{jk} \cdot \frac{2}{2j+1}.$$

The generating function is

$$M_{\text{Legendre}}(x, z) = \sum_{j=0}^{\infty} z^j P_j(x) = \frac{1}{\sqrt{1 - 2xz + z^2}}.$$

In the long-memory context, extensions of Legendre polynomials to non-integer degrees are useful. The so-called associated Legendre functions P_a^b can be defined either by replacing differentiation f' and f'' by fractional differentiation (see Sect. 7.3) or directly as solutions of the Legendre equation

$$\left(1-x^2\right)f''(x)-2xf'(x)+\left[a(a+1)-\frac{b^2}{1-x^2}\right]f(x)=0,$$

where $a, b \in \mathbb{C}$. The explicit formula is

$$P_a^b(z)=\frac{1}{\Gamma(1-b)}\left(\frac{1+z}{1-z}\right)^{\frac{b}{2}}F_{2,1}\left(\frac{1-z}{2}\middle| -a, a+1; 1-b\right) \quad \left(|1-z|<2\right).$$

These functions are useful for calculating the autocovariance function of Gegenbauer processes (see Chung 1996a).

3.2 Multivariate Hermite Expansions

The notion of Hermite polynomials considered in Sect. 3.1.2 can be extended to the multivariate case. Let $\mathbf{X}=(X_1,\ldots,X_k)^T$ be a k-dimensional Gaussian vector with expected value zero and covariance matrix

$$\Sigma=\left[cov(X_i,X_j)\right]_{i,j=1,\ldots,k},$$

and denote by I_k the $k\times k$ identity matrix. Set $\mathbf{q}=(q_1,\ldots,q_k)^T$, $\mathbf{q}!=q_1!\cdots q_k!$, $|\mathbf{q}|=q_1+\cdots+q_k$, $\mathbf{x}=(x_1,\ldots,x_k)$, $\mathbf{x}^{\mathbf{q}}=x_1^{q_1}\cdots x_k^{q_k}$, $\partial\mathbf{x}^{\mathbf{q}}=\partial x_1^{q_1}\cdots\partial x_k^{q_k}$ and

$$\left(\frac{d}{d\mathbf{x}}\right)^{\mathbf{q}}=\frac{\partial^{|\mathbf{q}|}}{\partial\mathbf{x}^{\mathbf{q}}}=\frac{\partial^{q_1+\cdots+q_k}}{\partial x_1^{q_1}\cdots\partial x_k^{q_k}}.$$

Definition 3.6 The $\mathbf{q}$th Hermite polynomial ($\mathbf{q}\in\mathbb{N}^k$) is equal to

$$H_{\mathbf{q}}(\mathbf{x};\Sigma)=\frac{(-1)^{|\mathbf{q}|}}{\phi_\Sigma(\mathbf{x})}\left(\frac{d}{d\mathbf{x}}\right)^{\mathbf{q}}\phi_\Sigma(\mathbf{x}), \tag{3.29}$$

where

$$\phi_\Sigma(\mathbf{x})=\frac{1}{(2\pi)^{k/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)$$

is the density of $\mathbf{X}$.

Hermite polynomials are orthogonal w.r.t. their dual polynomials,

$$\begin{aligned}\tilde{H}_{\mathbf{q}}(\mathbf{x};\,\Sigma) &= \frac{(-1)^{|\mathbf{q}|}}{\phi_\Sigma(\Sigma\mathbf{y})}\left(\frac{d}{d\mathbf{y}}\right)^{\mathbf{q}}\phi_\Sigma(\Sigma\mathbf{y}) \\ &= \frac{(-1)^{|\mathbf{q}|}}{\phi_\Sigma(\mathbf{x})}\left(\frac{d}{d\mathbf{y}}\right)^{\mathbf{q}}\phi_\Sigma(\Sigma\mathbf{y})\Big|_{\mathbf{y}=\Sigma^{-1}\mathbf{x}},\end{aligned}$$

where $\mathbf{y} = \Sigma^{-1}\mathbf{x}$, in the sense that (cf. (3.9) in the univariate case)

$$\langle \tilde{H}_{\mathbf{q}}, H_{\mathbf{r}}\rangle = \int_{\mathbf{R}^k} \tilde{H}_{\mathbf{q}}(\mathbf{x};\,\Sigma)H_{\mathbf{r}}(\mathbf{x};\,\Sigma)\phi_\Sigma(\mathbf{x})\,d\mathbf{x} = \mathbf{q}!\delta_{\mathbf{qr}}, \tag{3.30}$$

where $\delta_{\mathbf{qr}} = 1$ if $\mathbf{q} = \mathbf{r}$ and zero otherwise.

Definition (3.29) is not particularly useful for constructing multivariate Hermite polynomials. More useful results are obtained via the generating function introduced in the univariate case in Definition 3.3:

$$M_{\text{Hermite}}(x, z) = \exp\left(xz - \frac{z^2}{2}\right) = \sum_{k=0}^{\infty}\frac{z^k}{k!}H_k(x).$$

Note that for $X \sim N(0, 1)$,

$$\sum_{k=0}^{\infty}\frac{z^k}{k!}E\big[(x + iX)^k\big] = E\big[e^{z(x+iX)}\big] = M_{\text{Hermite}}(x, z).$$

Therefore, $H_k(x) = E[(x + iX)^k]$ $(k \geq 0)$. This formula can be extended to the multivariate case (see e.g. Withers 2000; also Barndorff-Nielsen and Pedersen 1979):

Lemma 3.6 *Let* $\mathbf{Y} \sim N(0, \Sigma^{-1})$, $\mathbf{y} = \Sigma^{-1}\mathbf{x}$ *and* $(\mathbf{y} + i\mathbf{Y})^{\mathbf{q}} = \prod_{j=1}^{k}(y_j + iY_j)^{q_j}$. *Then the following formula holds for multivariate Hermite polynomials defined in* (3.29):

$$H_{\mathbf{q}}(\mathbf{x};\,\Sigma) = E\big[(\mathbf{y} + i\mathbf{Y})^{\mathbf{q}}\big].$$

Proof Recall that the characteristic function of $\mathbf{Y} \sim N(0, \Sigma^{-1})$ is given by

$$E\big[\exp\big(i\mathbf{z}'\mathbf{Y}\big)\big] = \exp\left(-\frac{1}{2}\mathbf{z}'\Sigma^{-1}\mathbf{z}\right).$$

Recalling that $\mathbf{y} = \Sigma^{-1}\mathbf{x}$, a Taylor expansion in $\mathbb{R}^k$ leads to

$$\begin{aligned}\sum_{j_1,\ldots,j_k=0}^{\infty}\frac{z_1^{j_1}\cdots z_k^{j_k}}{j_1!\cdots j_k!}E\big[(\mathbf{y} + i\mathbf{Y})^{\mathbf{j}}\big] &= \sum_{\mathbf{j}\in\mathbb{N}_0^k}\frac{\mathbf{z}^{\mathbf{j}}}{\mathbf{j}!}E\big[(\mathbf{y} + i\mathbf{Y})^{\mathbf{j}}\big] \\ &= E\big[\exp\big(\mathbf{z}'(\mathbf{y} + i\mathbf{Y})\big)\big]\end{aligned}$$

$$= \exp\bigl(\mathbf{z}'\mathbf{y}\bigr)\exp\biggl(-\frac{1}{2}\mathbf{z}'\Sigma^{-1}\mathbf{z}\biggr)$$
$$= \exp\bigl(\mathbf{z}'\Sigma^{-1}\mathbf{x}\bigr)\exp\biggl(-\frac{1}{2}\mathbf{z}'\Sigma^{-1}\mathbf{z}\biggr).$$

The last expression equals

$$\frac{\phi_\Sigma(\mathbf{x}-\mathbf{z})}{\phi_\Sigma(\mathbf{x})}.$$

Noting that

$$\biggl[\biggl(\frac{d}{d\mathbf{z}}\biggr)^{\mathbf{j}}\phi_\Sigma(\mathbf{x}-\mathbf{z})\biggr]_{\mathbf{z}=0} = (-1)^{|\mathbf{j}|}\biggl(\frac{d}{d\mathbf{x}}\biggr)^{\mathbf{j}}\phi_\Sigma(\mathbf{x}),$$

the Taylor expansion of $\phi_\Sigma(\mathbf{x}-\mathbf{z})$ (as a function of $\mathbf{z}$) leads to

$$\sum_{\mathbf{j}\in\mathbb{N}_0^k}\frac{\mathbf{z}^{\mathbf{j}}}{\mathbf{j}!}E\bigl[(\mathbf{y}+i\mathbf{Y})^{\mathbf{j}}\bigr] = \frac{1}{\phi_\Sigma(\mathbf{x})}\sum_{\mathbf{j}\in\mathbb{N}_0^k}\frac{\mathbf{z}^{\mathbf{j}}}{\mathbf{j}!}(-1)^{|\mathbf{j}|}\biggl(\frac{d}{d\mathbf{x}}\biggr)^{\mathbf{j}}\phi_\Sigma(\mathbf{x}).$$

□

Example 3.4 Let $\mathbf{q} = (1,1)$. Then

$$H_{\mathbf{q}}(\mathbf{x};\Sigma) = E\bigl[(y_1+iY_1)(y_2+iY_2)\bigr] = y_1y_2 - E[Y_1Y_2].$$

Note that $H_{1,1}(\mathbf{x};\Sigma)$ is expressed here in terms of $\mathbf{y} = \Sigma^{-1}\mathbf{x}$. In particular, if the covariance matrix is $\Sigma = I_2$, then $H_{1,1}(\mathbf{x};\Sigma) = H_1(x_1)H_1(x_2) = x_1x_2$ since $\mathbf{y} = \mathbf{x}$ in this case.

Example 3.5 Let $\mathbf{q} = (1,2)$. Then

$$H_{\mathbf{q}}(\mathbf{x};\Sigma) = H_{1,2}(\mathbf{x};\Sigma) = E\bigl[(y_1+iY_1)(y_2+iY_2)^2\bigr]$$
$$= y_1\bigl(y_2^2 - E\bigl(Y_2^2\bigr)\bigr) - 2y_2E(Y_1Y_2).$$

Again, if $\Sigma = I_2$, then $H_{1,2}(\mathbf{x};\Sigma) = H_1(x_1)H_2(x_2)$.

In general, if $\Sigma = I_k$, then $H_{\mathbf{q}}(\mathbf{x};\Sigma) = \prod_{j=1}^k H_{q_j}(x_j)$. In other words, if the components of the vector $\mathbf{X}$ are independent, then a multivariate Hermite polynomial is a product of univariate ones.

The examples show that multivariate Hermite polynomials have quite a complicated form and may not be suitable in the context of limit theorems. In fact, as we will see below, it is sufficient to consider Gaussian random vectors with i.i.d. $N(0,1)$ components, i.e.

$$\tilde{\mathbf{X}} = (\tilde{X}_1,\dots,\tilde{X}_k)^T \sim N(0,I_k).$$

To see this, define

$$H^*_{\mathbf{q}}(\mathbf{x}) = H^*_{q_1,\ldots,q_k}(x_1,\ldots,x_k) = \prod_{j=1}^{k} H_{q_j}(x_j). \tag{3.31}$$

As indicated in (3.30), $H^*_{\mathbf{q}}$ ($\mathbf{q} \in \mathbb{N}^k$) form an orthonormal basis in $L^2(\mathbb{R}, \phi_{I_k})$. Let $G \in L^2(\mathbb{R}, \phi_{I_k})$ and define

$$J(G, \tilde{\mathbf{X}}, \mathbf{q}) = J(G, I_k, \mathbf{q}) = \langle G, H^*_{\mathbf{q}} \rangle = E\big[G(\tilde{\mathbf{X}}) H^*_{\mathbf{q}}(\tilde{\mathbf{X}})\big].$$

The Hermite rank

$$\tau(G, \tilde{\mathbf{X}}) = \tau(G, I_k)$$

of G with respect to $\tilde{\mathbf{X}}$, or in other words with respect to the distribution $N(0, I_k)$, is the largest integer τ such that

$$J(G, I_k, \mathbf{q}) = 0 \quad \text{for all } 0 < |\mathbf{q}| < \tau, \tag{3.32}$$

where $|\mathbf{q}| = q_1 + \cdots + q_k$. Note that this is the same as the largest integer τ such that

$$\langle G(\tilde{\mathbf{X}}), \tilde{\mathbf{X}}^{\mathbf{q}} \rangle = E\left[G(\tilde{\mathbf{X}}) \prod_{j=1}^{k} \tilde{X}_j^{q_j} \right] = 0 \quad \text{for all } 0 < |\mathbf{q}| < \tau.$$

As in the univariate case, we therefore can write down an orthogonal expansion

$$G(\tilde{X}_1,\ldots,\tilde{X}_k) = E\big[G(\tilde{\mathbf{X}})\big] + \sum_{|\mathbf{q}| \geq \tau(G, I_k)} \frac{J(G, I_k, \mathbf{q})}{q_1! \cdots q_k!} \prod_{j=1}^{k} H_{q_j}(\tilde{X}_j). \tag{3.33}$$

Now consider $\mathbf{X} \sim N(0, \Sigma)$. Then $\mathbf{X}$ is equal in distribution to $U(\tilde{\mathbf{X}}) = \Sigma^{\frac{1}{2}} \tilde{\mathbf{X}}$. Thus, we may apply expansion (3.33) to the function $\tilde{G}(\tilde{\mathbf{X}}) = G \circ U(\tilde{\mathbf{X}})$, which then has the Hermit rank $\tau(\tilde{G}, \tilde{\mathbf{X}}) = \tau(G \circ U, \tilde{\mathbf{X}}) = \tau(G \circ U, I_k)$ with respect to $\tilde{\mathbf{X}}$. We therefore have the expansion

$$G(\mathbf{X}) = \tilde{G}(\tilde{\mathbf{X}}) = E\big[G(\mathbf{X})\big] + \sum_{|\mathbf{q}| = \tau(\tilde{G}, \tilde{\mathbf{X}})}^{\infty} \frac{J(G \circ U, \tilde{\mathbf{X}}, \mathbf{q})}{q_1! \cdots q_k!} \prod_{j=1}^{k} H_{q_j}(\tilde{X}_j).$$

Now let us define ("Hermite") coefficients of G with respect to $\mathbf{X}$ by

$$J(G, \mathbf{X}, \mathbf{q}) = J(G, \Sigma, \mathbf{q}) = E\big[G(\mathbf{X}) H^*_{\mathbf{q}}(\mathbf{X})\big]. \tag{3.34}$$

Here we write "Hermite" in quotation marks because initially $J(G, \mathbf{X}, \mathbf{q})$ does not have the same straightforward interpretation as before. The reason is that the polynomials $H^*_{\mathbf{q}}(\mathbf{X})$ no longer constitute an orthogonal basis in the L^2-space defined

by the distribution of $\mathbf{X}$. Nevertheless, it turns out that the definition of $J(G, \Sigma, \mathbf{q})$ is meaningful when it comes to determining that part of $G(\mathbf{X})$ which is relevant for limit theorems (of sums). The reason is that, as for $\tilde{\mathbf{X}}$, the space spanned by $H^*_{\mathbf{q}}(\mathbf{X})$ ($|\mathbf{q}| \le p$) is the same as the space spanned by all (multivariate) polynomials in $X_1, \ldots, X_k$ up to degree p. If we define the Hermite rank $\tau(G, \mathbf{X}) = \tau(G, \Sigma)$ of G with respect to $\mathbf{X}$ as the largest integer τ such that

$$J(G, \Sigma, \mathbf{q}) = 0 \quad \text{for all } 0 < |\mathbf{q}| < \tau, \tag{3.35}$$

then this is the same as the largest integer such that

$$\left\langle G(\mathbf{X}), \mathbf{X}^{\mathbf{q}} \right\rangle = E\left[G(\mathbf{X}) \prod_{j=1}^{k} X_j^{q_j} \right] = 0 \quad \text{for all } 0 < |\mathbf{q}| < \tau.$$

However, $\mathbf{X}$ is obtained from $\tilde{\mathbf{X}}$ by a one-to-one linear transformation $\mathbf{X} = U(\tilde{\mathbf{X}}) = \Sigma^{\frac{1}{2}}\tilde{\mathbf{X}}$. The space spanned by polynomials in $X_1, \ldots, X_k$ of order $\mathbf{q}$ with $|\mathbf{q}| \le p$ is therefore the same as the one spanned by polynomials in $\tilde{X}_1, \ldots, \tilde{X}_k$ with $|\mathbf{q}| \le p$. Therefore the condition that $E[G(\mathbf{X}) \prod_{j=1}^{k} X_j^{q_j}] = 0$ for all $\mathbf{q}$ with $|\mathbf{q}| \le p$ is the same as the condition that

$$E\left[G(\mathbf{X}) \prod_{j=1}^{k} \tilde{X}_j^{q_j} \right] = E\left[G \circ U(\tilde{\mathbf{X}}) \prod_{j=1}^{k} \tilde{X}_j^{q_j} \right] = 0$$

for all $\mathbf{q}$ with $|\mathbf{q}| \le p$. This implies that the values of $\tau(G, \Sigma)$ and $\tau(G \circ U, I_k)$ are the same. The result can be summarized as follows (see Arcones 1994, p. 2249):

Lemma 3.7 *Let* $\mathbf{X} \sim N(0, \Sigma)$ *and*

$$\tilde{\mathbf{X}} = U(\mathbf{X}) = \Sigma^{-\frac{1}{2}}\mathbf{X} \sim N(0, I_k).$$

Then the Hermite rank $\tau(G, \Sigma)$ *of* G *with respect to* $\mathbf{X}$ *is the same as the Hermite rank* $\tau(G \circ U, I_k)$ *of* $G \circ U$ *with respect to* $\tilde{\mathbf{X}}$, *i.e.*

$$\tau(G, \Sigma) = \tau(G \circ U, I_k). \tag{3.36}$$

Note, however that in general $\tau(G, \Sigma) \ne \tau(G, I_k)$ (see examples below). Moreover, the coefficients $J(G, \Sigma, \mathbf{q})$ and $J(G \circ U, I_k, \mathbf{q})$ are not the same in general. Nevertheless, from a point of view of limit theorems, there is no need to consider the entire class of multivariate Hermite polynomials $H_{\mathbf{q}}$. First of all, due to $\tau(G, \Sigma) = \tau(G \circ U, I_k) =: \tau$, the Hermite rank of $G(\mathbf{X})$ can be determined by calculating either $J(G, \Sigma, \mathbf{q})$ or $J(G \circ U, I_k, \mathbf{q})$ (whatever is easier). To identify the asymptotically relevant part (terms with $|\mathbf{q}| = \tau$) of $G(\mathbf{X})$, one can switch to the representation $G \circ U(\tilde{\mathbf{X}})$. If the limit theorem is for a sum of $G(\mathbf{X}_t)$, then in the

long-memory context the relevant part consists of all contributions with $|\mathbf{q}| = \tau$, i.e.

$$\tilde{G}(\tilde{\mathbf{X}}) = \sum_{\substack{1 \le q_1, \dots, q_k \le \tau \\ |\mathbf{q}| = \tau}} \frac{J(G \circ U, \tilde{\mathbf{X}}, \mathbf{q})}{q_1! \cdots q_k!} H_{\mathbf{q}}^*(\tilde{\mathbf{X}})$$

$$= \sum_{\substack{1 \le q_1, \dots, q_k \le \tau \\ |\mathbf{q}| = \tau}} \frac{J(G \circ U, \tilde{\mathbf{X}}, \mathbf{q})}{q_1! \cdots q_k!} \prod_{j=1}^{k} H_{q_j}(\tilde{X}_j).$$

Note that some but not all of the coefficients in the sum may be zero. Finally, we can write the asymptotically relevant part in terms of the original vector $\mathbf{X}$ by applying the inverse transformation U^{-1},

$$G(\mathbf{X}) = \tilde{G}\big(\Sigma^{\frac{1}{2}} \tilde{\mathbf{X}}\big).$$

Thus, in summary, only the special Hermite polynomials $H_{\mathbf{q}}^*$ (as defined in (3.31)) and the corresponding expansion for $N(0, I_k)$-distributed variables are needed.

Example 3.6 Consider $G(y_1, y_2) = y_1 y_2$, and let X_1 and X_2 be independent $N(0, 1)$. Thus, $\Sigma = I_2$, so that $X_1 = \tilde{X}_1$, $X_2 = \tilde{X}_2$, and we have to consider $J(G, I_2, \mathbf{q})$, where I_2 is the 2×2 identity matrix. From

$$J\big(G, I_2, (1, 0)\big) = J\big(G, I_2, (0, 1)\big) = E[\tilde{X}_1 \tilde{X}_2 \tilde{X}_j] = 0 \quad (j = 1, 2),$$

$$J\big(G, I_2, (2, 0)\big) = J\big(G, I_2, (0, 2)\big) = E\big[\tilde{X}_1 \tilde{X}_2 H_2(\tilde{X}_j)\big] = 0 \quad (j = 1, 2)$$

and

$$J\big(G, I_2, (1, 1)\big) = E\big[\tilde{X}_1 \tilde{X}_2 H_1(\tilde{X}_1) H_1(\tilde{X}_2)\big] = E\big[\tilde{X}_1^2\big] E\big[\tilde{X}_2^2\big] = 1$$

we conclude that the Hermite rank (of G with respect to $\mathbf{X}$) is

$$\tau(G, I_2) = 2$$

and the only nonzero Hermite coefficient with $|\mathbf{q}| = 2$ is obtained for $\mathbf{q} = (1, 1)$.

Example 3.7 As before, we consider $G(y_1, y_2) = y_1 y_2$, but now we assume X_1, X_2 to be correlated $N(0, 1)$ variables. This can be written as

$$\mathbf{X} = (X_1, X_2)^T = \big(\tilde{X}_1, \gamma \tilde{X}_1 + \sqrt{1 - \gamma^2}\, \tilde{X}_2\big)^T$$

with $\gamma = cov(X_1, X_2)$, $0 < |\gamma| < 1$ and $\tilde{X}_1, \tilde{X}_2$ as in the previous example. In other words, $\mathbf{X} = U(\tilde{\mathbf{X}}) = \Sigma^{\frac{1}{2}} \tilde{\mathbf{X}}$ with

$$\Sigma^{\frac{1}{2}} = \begin{pmatrix} 1 & 0 \\ \gamma & \sqrt{1 - \gamma^2} \end{pmatrix}.$$

The rank τ of G with respect to $\mathbf{X}$ is equal to $\tau(G, \Sigma)$ or equivalently $\tau(G \circ U, I_k)$. As an exercise, we calculate τ both ways. For the coefficients $J(G \circ U, I_k, \mathbf{q})$, we obtain

$$J\big(G \circ U, I_2, (1,0)\big) = J\big(G \circ U, I_2, (0,1)\big) = E\big[\tilde{X}_1\big(\gamma \tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)\tilde{X}_1\big] = 0,$$

$$J\big(G \circ U, I_2, (1,1)\big) = E\big[\tilde{X}_1\big(\gamma \tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)\tilde{X}_1\tilde{X}_2\big] = \sqrt{1-\gamma^2},$$

$$J\big(G \circ U, I_2, (0,2)\big) = E\big[\tilde{X}_1\big(\gamma \tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)\big(\tilde{X}_2^2 - 1\big)\big] = 0,$$

and

$$J\big(G \circ U, I_2, (2,0)\big) = E\big[\tilde{X}_1\big(\gamma \tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)\big(\tilde{X}_1^2 - 1\big)\big] = 2\gamma.$$

Thus, $\tau(G \circ U, I_2) = 2$. In fact, $G \circ U$ is exactly equal to the contribution of terms with $|\mathbf{q}| = 2$, namely

$$\begin{aligned} G \circ U(\tilde{X}_1, \tilde{X}_2) - \gamma &= \tilde{X}_1\big(\gamma \tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big) - \gamma \\ &= \frac{2\gamma}{2!} H_2(\tilde{X}_1) + \frac{\sqrt{1-\gamma^2}}{1!} H_1(\tilde{X}_2) H_1(\tilde{X}_2). \end{aligned}$$

For the coefficients $J(G, \Sigma, \mathbf{q})$, we have

$$J\big(G, \Sigma, (1,0)\big) = J\big(G, \Sigma, (0,1)\big) = E[X_1 X_2 X_j] = 0 \quad (j = 1, 2)$$

and using Lemma 3.5,

$$\begin{aligned} J\big(G, \Sigma, (1,1)\big) &= E[X_1 X_2 X_1 X_2] = E\big[H_2(X_1) H_2(X_2)\big] + 1 = 2\gamma^2 + 1, \\ J\big(G, \Sigma, (2,0)\big) &= E\big[X_1 X_2 \big(X_1^2 - 1\big)\big] = E\big[H_3(X_1) H_1(X_2)\big] = 2\gamma \end{aligned}$$

and

$$J\big(G, \Sigma, (0,2)\big) = E\big[X_1 X_2 \big(X_2^2 - 1\big)\big] = E\big[H_1(X_1) H_3(X_2)\big] = 2\gamma.$$

Thus, as it should be according to (3.36), the Hermite rank $\tau(G, \Sigma) = 2$ is the same as $\tau(G \circ U, I_2)$. Note, however, that there are now three nonzero coefficients.

Example 3.8 Let $G(y_1, y_2) = H_2(y_1) H_2(y_2) = (y_1^2 - 1)(y_2^2 - 1)$ and $\mathbf{X} = \tilde{\mathbf{X}}$. Thus we consider $J(G, I_2, \mathbf{q})$. Since only odd powers are involved, we have

$$J\big(G, I_2, (1,0)\big) = J\big(G, I_2, (0,1)\big) = J\big(G, I_2, (1,1)\big) = 0.$$

Also, for $\mathbf{q} = (2,0)$, we obtain

$$J\big(G, I_2, (2,0)\big) = E\big[H_2^2(\tilde{X}_1) H_2(\tilde{X}_2)\big] = E\big[H_2^2(\tilde{X}_1)\big] E\big[H_2(\tilde{X}_2)\big] = 0,$$

and, by symmetry, $J(G, I_2, (0,2)) = J(G, I_2, (2,0)) = 0$. For $|\mathbf{q}| = 3$, only odd powers are involved, so that $J(G, I_2, \mathbf{q}) = 0$. Finally, for $|\mathbf{q}| = 4$, one has for example

$$J\big(G, I_2, (2,2)\big) = E\big[H_2^2(\tilde{X}_1)H_2^2(\tilde{X}_2)\big] = E\big[H_2^2(\tilde{X}_1)\big]E\big[H_2^2(\tilde{X}_2)\big] \neq 0.$$

Thus, the Hermite rank of G with respect to $\mathbf{X} = \tilde{\mathbf{X}}$ is 4.

Example 3.9 Consider the previous function $G(y_1, y_2) = H_2(y_1)H_2(y_2)$, however with

$$\mathbf{X} = (X_1, X_2)^T = \big(\tilde{X}_1, \gamma\tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)^T,$$

where $\gamma = cov(X_1, X_2)$, $0 < |\gamma| < 1$. Then

$$J\big(G \circ U, I_2, (1,0)\big) = J\big(G \circ U, I_2, (0,1)\big) = 0.$$

For

$$J\big(G \circ U, I_2, (1,1)\big) = E\big[H_2(\tilde{X}_1)H_2\big(\gamma\tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)\tilde{X}_1\tilde{X}_2\big],$$

we use Eq. (3.16) to write

$$H_2\big(\gamma\tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big) = \big(1-\gamma^2\big)H_2(\tilde{X}_2) + \gamma^2 H_2(\tilde{X}_1) + 2\gamma\sqrt{1-\gamma^2}\tilde{X}_1\tilde{X}_2.$$

Then

$$J\big(G \circ U, I_2, (1,1)\big) = 2\gamma\sqrt{1-\gamma^2}E\big[H_2(\tilde{X}_1)\tilde{X}_1\tilde{X}_2\tilde{X}_1\tilde{X}_2\big] = 4\gamma\sqrt{1-\gamma^2}.$$

Thus, in contrast to the previous case with independent components, for correlated normal variables X_1, X_2, the Hermite rank $\tau(G, \Sigma) = \tau(G \circ U, I_2)$ of G is 2 instead of 4. This illustrates that a multivariate Hermite rank can be changed just by changing the correlation structure between the components of the normal vector $\mathbf{X}$. (The example also illustrates that for correlated components, it would be wrong to interpret $\tau(G, I_2)$ as the Hermite rank. The correct Hermite rank is obtained only if one calculates $\tau(G, \Sigma)$ or $\tau(G \circ U, I_2)$.)

Example 3.10 Let

$$G(y_1, y_2) = H_2(y_1)y_2$$

and $\mathbf{X} = \tilde{\mathbf{X}} \sim N(0, I_2)$. We have

$$J\big(G, I_2, (1,0)\big) = E\big[\tilde{X}_1 H_2(\tilde{X}_1)\tilde{X}_2\big] = E\big[\tilde{X}_1 H_2(\tilde{X}_1)\big]E[\tilde{X}_2] = 0$$

and

$$J\big(G, I_2, (0,1)\big) = E\big[H_2(\tilde{X}_1)\tilde{X}_2^2\big] = E\big[H_2(\tilde{X}_1)\big]E\big[\tilde{X}_2^2\big] = 0.$$

For $|\mathbf{q}| = 2$, we also similarly have $J(G, I_2, \mathbf{q}) = 0$. For $|\mathbf{q}| = 3$, however we have for example

$$J\big(G, I_2, (2,1)\big) = E\big[H_2^2(\tilde{X}_1)\tilde{X}_2^2\big] = E\big[H_2^2(\tilde{X}_1)\big] = 2.$$

Thus, G has Hermite rank $\tau(G, I_2) = 3$.

Example 3.11 Let

$$G(y_1, y_2) = H_2(y_1)y_2$$

as before, but $cov(X_1, X_2) = \gamma$ (and $E(X_i) = 0$, $\mathrm{var}(X_i) = 1$). Then for instance

$$J\big(G, \Sigma, (1,0)\big) = E\big[X_1 H_2(X_1)X_2\big] = E\big[\big(H_3(X_1) + 2X_1\big)X_2\big] = 2E[X_1X_2] = 2\gamma.$$

Thus, the Hermite rank $\tau(G, \Sigma) = \tau(G \circ U, I_2)$ is now equal to one. Again, this is an example where introducing a correlation between X_1 and X_2 changes the Hermite rank.

Example 3.12 Let $\mathbf{X} = \tilde{\mathbf{X}} \sim N(0, I_2)$, and consider $G(y_1, y_2) = G_1(y_1)G_2(y_2)$, where the two (centred) functions G_i $(i = 1, 2)$ have (univariate) Hermite ranks m_1 and m_2. Then

$$E\big[G(\tilde{\mathbf{X}})H_{q_1}(\tilde{X}_1)H_{q_2}(\tilde{X}_2)\big] = E\big[G_1(\tilde{X}_1)H_{q_1}(\tilde{X}_1)\big]E\big[G_1(\tilde{X}_2)H_{q_2}(\tilde{X}_2)\big].$$

Now, if $E[G_1(\tilde{X}_1)] = E[G_2(\tilde{X}_2)] = 0$, then

$$E\big[G_1(\tilde{X}_1)H_{q_1}(\tilde{X}_1)\big] = 0 \quad (0 \le q_1 \le m_1 - 1),$$
$$E\big[G_2(\tilde{X}_2)H_{q_2}(\tilde{X}_2)\big] = 0 \quad (0 \le q_2 \le m_2 - 1).$$

For $\mathbf{q} = (m_1, m_2)$, on the other hand, both expected values are non-zero. Thus, the Hermite rank is

$$\tau(G, I_2) = m_1 + m_2.$$

If the expected value of one the functions G_1, G_2 is not zero, then the Hermite rank changes. For instance, if $E[G_1(\tilde{X}_1)] \neq 0$, then $E[G_1(\tilde{X}_1)H_0(\tilde{X}_1)] = E[G_1(\tilde{X}_1)] \neq 0$, so that

$$E\big[G(\tilde{\mathbf{X}})H_0(\tilde{X}_1)H_{m_2}(\tilde{X}_2)\big] = E\big[G_1(\tilde{X}_1)H_0(\tilde{X}_1)\big]E\big[G_1(\tilde{X}_2)H_{m_2}(\tilde{X}_2)\big] \neq 0.$$

The same argument applies if $E[G_2(\tilde{X}_2)] \neq 0$. Therefore, in either case G has Hermite rank

$$\tau(G, I_2) = m_1 \wedge m_2 = \min\{m_1, m_2\}.$$

Example 3.13 Let $G_1(x) = \exp(px)$, where p is a an integer. Then the univariate Hermite rank of G_1 is one. Now let $\mathbf{X} = \tilde{\mathbf{X}} \sim N(0, I_2)$ and consider

$$G(y_1, y_2) = G_1(y_1)G_1(y_2) = e^{p(y_1+y_2)}.$$

Then we know from the previous example that the Hermite rank $\tau(G, I_2)$ is one.

Example 3.14 Consider the same function $G(y_1, y_2) = G_1(y_1)G_1(y_2) = \exp(p(y_1 + y_2))$ as in the previous example, however $\mathbf{X} \sim N(0, \Sigma)$ with $N(0, 1)$ marginals and $cov(X_1, X_2) = \gamma$ $(0 < |\gamma| < 1)$. Then

$$\begin{aligned} J\big(G, \Sigma, (1,0)\big) &= E\big[X_1 \exp\big(p(X_1 + X_2)\big)\big] \\ &= E\big[\tilde{X}_1 \exp\big(p\big((1+\gamma)\tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big)\big)\big] \\ &= E\big[\tilde{X}_1 e^{p(1+\gamma)\tilde{X}_1}\big] E\big[e^{p\sqrt{1-\gamma^2}\tilde{X}_2}\big] \\ &= M'\big(p(1+\gamma)\big) M\big(p\sqrt{1-\gamma^2}\big), \end{aligned}$$

where $M(u) = \exp(\frac{1}{2}u^2)$ is the moment generating function of the standard normal distribution. Thus,

$$\tau(G, \Sigma) = \tau(G \circ U, I_2) = 1,$$

i.e. the Hermite rank with respect to the correlated vector $\mathbf{X}$ is also one. This example will be applied in the section on stochastic volatility models.

The next example illustrates that monotonicity of the Hermite ranks for univariate functions may not be transferred to the bivariate case.

Example 3.15 Let $G(y_1, y_2) = G_1(y_1)G_2(y_2)$, where $G_j(x) = H_m(x)$ $(j = 1, 2)$, and consider a dependent vector $\mathbf{X} \sim N(0, \Sigma)$ with $N(0, 1)$ marginals and $cov(X_1, X_2) = \gamma$ $(0 < |\gamma| < 1)$. Due to the recursion

$$H_{m+1}(x) = x H_m(x) - m H_{m-1}(x)$$

and Lemma 3.5,

$$\begin{aligned} J\big(G, \Sigma, (0,1)\big) = J\big(G, \Sigma, (1,0)\big) &= E\big[H_m(X_1)H_m(X_2)X_1\big] \\ &= E\big[H_m(X_1)H_m(X_2)X_1\big] \\ &= E\big[\big(H_{m+1}(X_1) + m H_{m-1}(X_1)\big)H_m(X_2)\big] = 0. \end{aligned}$$

For $\mathbf{q} = (2, 0)$, however we may use formula (3.16) to first obtain

$$\begin{aligned} H_m(X_2) &= H_m\big(\gamma\tilde{X}_1 + \sqrt{1-\gamma^2}\tilde{X}_2\big) \\ &= \sum_{m_1+m_2=m} \frac{m!}{m_1!m_2!} \gamma^{m_1} \big(1-\gamma^2\big)^{\frac{m_2}{2}} H_{m_1}(\tilde{X}_1) H_{m_2}(\tilde{X}_2). \end{aligned}$$

Then

$$J\big(G, \Sigma, (2,0)\big) = E\big[H_m(X_1)H_m(X_2)H_2(X_1)\big] = E\big[H_m(\tilde{X}_1)H_2(\tilde{X}_1)H_m(X_2)\big]$$

can be written as

$$\sum_{m_1+m_2=m} \frac{m!}{m_1!m_2!} \gamma^{m_1} \left(1-\gamma^2\right)^{\frac{m_2}{2}} E\left[H_m(\tilde{X}_1) H_2(\tilde{X}_1) H_{m_1}(\tilde{X}_1) H_{m_2}(\tilde{X}_2)\right].$$

Since $E[H_m(\tilde{X}_1)H_2(\tilde{X}_1)H_{m_1}(\tilde{X}_1)H_{m_2}(\tilde{X}_2)]$ factorizes into $E[H_{m_2}(\tilde{X}_2)]$ times the expected value of the other three factors, all terms in the sum are zero except for $m_2 = 0$, $m_1 = m$, where we have

$$\gamma^m E\left[H_m^2(\tilde{X}_1) H_2(\tilde{X}_1)\right] \neq 0.$$

This means that for any $m \geq 1$, the Hermite rank $\tau(G, \Sigma) = \tau(G \circ U, I_2)$ with respect to the dependent vector $\mathbf{X}$ is equal to 2, no matter which value the univariate Hermite rank $m \geq 1$ the two factors G_i $(i = 1, 2)$ have.

3.3 Appell Polynomials

3.3.1 General Motivation

In the previous section, expansions of transformed random variables $G(X)$ in terms of orthogonal polynomials $P_j(X)$ were obtained. There are however only very few distribution families where this is possible, including the normal and exponential. This is in particular true for stochastic processes where a sufficiently simple closed-form expression for the marginal distribution may not even exist. It is then unclear how to find suitable orthogonal polynomials and whether they exist at all. Most of the polynomials discussed above were indeed originally found in a completely different mathematical context such as differential equations etc., and the application to transformations of random variables and processes came as a byproduct. As it turns out, a general theory of polynomial expansions can be developed for linear processes

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{3.37}$$

with i.i.d. zero-mean innovations ε_t defined in $L^2(\Omega)$. The generality of the approach comes at a price however since the corresponding polynomials are no longer orthogonal except in the Gaussian case. Some care is needed therefore to obtain meaningful expressions and definitions. In the following sections, a brief outline of this approach is given. The original, and rather extended, literature is scattered in various fields of mathematics (see for instance Appell 1880, 1881; Meixner 1934; Boas and Buck 1964; Anderson 1967; Ozhegov 1965, 1967; Kazmin 1969a, 1969b; Bateman and Erdelyi 1974; Szegö 1974; Bourbaki 1976).

The general question is as follows. Let X_t be the stationary linear process defined by (3.37) and denote by $F_X = P(X \leq x)$ the marginal distribution of X_t.

The question is whether it is possible to find polynomials P_j such that, for any function G with $E[G(X)] = 0$ and $E[G^2(X)] < \infty$ there is a unique representation $G(X) = \sum_{j=0}^{\infty} g_j P_j(X)$ with equality defined in $L^2(\Omega)$. The idea is to use so-called Appell polynomials, which are defined in terms of the moment generating function of X. Unfortunately, Appell polynomials are no longer orthogonal (except if ε_t are normally distributed). This can cause problems with respect to the calculation of the coefficients, uniqueness of the representation and the definition of the so-called Appell rank (the analogue to the Hermite rank). The theory of Appell polynomials is therefore quite involved, and, in the context of linear processes, open questions remain.

3.3.2 Definition

Let X be a univariate real-valued random variable with distribution F_X. For simplicity, suppose first that the moment generating function

$$m_X(z) = E\left(e^{zX}\right)$$

is finite in an open neighborhood of zero, $U_r = \{|z| < r\}$, where r is a suitable positive number. Then this implies that $m_X(z)$ is analytic on U_r. Since $m_X(0) = 1 \neq 0$, we can conclude that there is a $\delta \leq r$ such that

$$m_{\text{inv}}(z) = \frac{1}{m_X(z)}$$

is analytic on U_δ, and the same is true for

$$\frac{\exp(z)}{m_X(z)} = \exp(z) m_{\text{inv}}(z).$$

Thus we have the power series representations

$$m_{\text{inv}}(z) = \sum_{j=0}^{\infty} \frac{a_{\text{inv},j}}{j!} z^j, \qquad \exp(z) = \sum_{j=0}^{\infty} \frac{z^j}{j!}$$

and

$$\exp(z) m_{\text{inv}}(z) = \left(\sum_{j=0}^{\infty} \frac{z^j}{j!}\right)\left(\sum_{j=0}^{\infty} \frac{z^j}{j!} a_{\text{inv},j}\right) = \sum_{j=0}^{\infty} \frac{z^j}{j!} b_j$$

with

$$b_j = \sum_{k=0}^{j} \binom{j}{k} a_{\text{inv},k}. \tag{3.38}$$

To define Appell polynomials, we introduce the function

$$M_X(x,z) = \frac{\exp(xz)}{m_X(z)} = \exp(xz) m_{\text{inv}}(z), \tag{3.39}$$

which is called the generating function of Appell polynomials associated with F_X. Then

$$M_X(x,z) = \left(\sum_{j=0}^{\infty} \frac{z^j}{j!} x^j\right)\left(\sum_{j=0}^{\infty} \frac{z^j}{j!} a_{\text{inv},j}\right) = \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x), \tag{3.40}$$

where, as in (3.38),

$$A_j(x) = \sum_{k=0}^{j} \binom{j}{k} a_{\text{inv},k} x^{j-k} = \frac{d^j}{dz^j}\big[M_X(x,z)\big]_{z=0}. \tag{3.41}$$

The coefficients A_j are polynomials in x, of degree j. They were introduced in 1880 by the French mathematician Paul Emile Appell (Appell 1880) and are therefore called Appell polynomials. Thus, we have the following definition.

Definition 3.7 Let $X \sim F_X$. Then the Appell polynomials $A_j(x)$ of order $j = 0, 1, 2, \ldots$ associated with F_X (or X) are defined by

$$M_X(x,z) = \frac{\exp(xz)}{E(e^{zX})} = \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x). \tag{3.42}$$

It should be noted that Appell polynomials are distribution specific. To emphasize this, one should actually use a notation like $A_j^{F_X}$ instead of A_j. However, unless there could be a confusion, we will always write A_j instead.

Definition 3.7 assumes that the moment generating function $m_X(z)$ is finite in an open neighborhood of the origin, which implies, but is stronger than, the assumption that all absolute moments $E[|X|^j]$ are finite. This may not always be the case. Generally we can distinguish three cases: (1) as above, i.e. $m_X(z)$ is finite in U_r for some $r > 0$; (2) $E[|X|^j] < \infty$ for all j, but $\sup_{z \in U_r} |m_X(z)| = \infty$ for any $r > 0$, and (3) there exists a j_0 with $E[|X|^{j_0}] = \infty$. Case (1) is treated above. In case (2), the expansion in (3.42) can be understood formally by matching coefficients. More precisely, this means that the formal series $m_X(z) = \sum_{j=0}^{\infty} \mu_j z^j / j!$ with $\mu_j = E(X^j)$ is understood as a symbolic representation of the sequences $\mu = (\mu_0, \mu_1, \ldots)$. The space of power series is then defined as a space of sequences $\mathscr{P} = \{(a_j)_{j \in \mathbb{N}}, a_j \in \mathbb{R}\}$ endowed with the operations "+" and "·" specified by the usual rules of addition and multiplication for power series (as in Eqs. (3.38) and (3.41)). In case (3), definition (3.39) can be modified to

$$\tilde{M}_X(x,z) = \frac{\exp(xz)}{\tilde{m}_X(z)} = \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x)$$

with

$$\tilde{m}_X(z) = \sum_{j=0}^{j_0-1} \frac{z^j}{j!}\mu_j$$

and j_0 such that all moments up to order $j_0 - 1$ are finite. This way, the Appell polynomials A_0, $A_1, \ldots, A_{j_0-1}$ can be defined, and this definition is compatible with the previous one in cases (1) and (2).

Appell polynomials have the following important properties, which can also be used as an alternative definition:

Lemma 3.8 *For Appell polynomials, we have*

$$\begin{aligned} E\big[A_0(X)\big] &= 1, \\ E\big[A_j(X)\big] &= 0 \quad (j \geq 1) \end{aligned} \tag{3.43}$$

and

$$A_j' = j A_{j-1}. \tag{3.44}$$

Proof For cases (2) and (3), combinatorial proofs of (3.43) can be given. Here, we consider the much easier case (1). Below, it will be shown that, if $m_X(z)$ is finite in U_r, then

$$S_n = \sum_{j=0}^{n} \frac{z^j}{j!} A_j(X) \underset{L^2(\Omega)}{\to} S_\infty = \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(X),$$

where S_∞ is almost surely finite (in $\mathbb{C}$). This implies

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{z^j}{j!} E\big[A_j(X)\big] &= E\left[\sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(X)\right] \\ &= E\left[\frac{\exp(zX)}{m_X(z)}\right] = \frac{m_X(z)}{m_X(z)} = 1, \end{aligned}$$

and hence the result follows by comparing the coefficients of z^j.

For the derivative, we have

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j'(x) &= \frac{d}{dx} \frac{\exp(xz)}{m_X(z)} = \frac{z \exp(xz)}{m_X(z)} \\ &= \sum_{j=0}^{\infty} \frac{z^{j+1}}{(j+1)!} (j+1) A_j(x) = \sum_{j=1}^{\infty} \frac{z^j}{j!} j A_{j-1}(x), \end{aligned}$$

and the result follows again by comparing the coefficients of z^j. □

Example 3.16 For $X \sim N(0,1)$, we have $m_X(z) = \exp(\frac{1}{2}z^2)$, so that from (3.41) and

$$M_X(x,z) = e^{zx-\frac{1}{2}z^2}$$

we obtain

$$\begin{aligned} A_j(x) &= \frac{d^j}{dz^j}\left[\exp\left(zx - \frac{1}{2}z^2\right)\right]_{z=0} \\ &= \frac{d^j}{dz^j}\left[\exp\left(\frac{x^2}{2} - \frac{1}{2}(x-z)^2\right)\right]_{z=0} \\ &= (-1)^j e^{\frac{x^2}{2}} \frac{d^j}{dx^j}\left(e^{-\frac{x^2}{2}}\right) = H_j(x). \end{aligned}$$

Thus, for Gaussian random variables, Appell polynomials coincide with Hermite polynomials.

Example 3.17 Let $X \sim \mathrm{Exp}(\lambda)$, i.e. X is exponentially distributed with cumulative probability distribution $F_X(x) = 1 - e^{-\lambda x}$. Then $m_X(z) = (1 - z/\lambda)^{-1}$ for $|z| < \lambda$, and

$$\begin{aligned} M_X(x,z) &= e^{xz}\left(1 - \frac{z}{\lambda}\right) = \sum_{j=0}^{\infty} \frac{z^j}{j!}x^j - \lambda^{-1}\sum_{j=1}^{\infty} \frac{z^j}{(j-1)!}x^{j-1} \\ &= 1 + \sum_{j=1}^{\infty} \frac{z^j}{j!}\left(x - \frac{j}{\lambda}\right)x^{j-1}, \end{aligned}$$

so that

$$\begin{aligned} A_0(x) &= 1, \\ A_j(x) &= \left(x - \frac{j}{\lambda}\right)x^{j-1} \quad (j \geq 1). \end{aligned} \tag{3.45}$$

Note that in this case Appell polynomials do not coincide with Laguerre polynomials, which were orthogonal w.r.t. the exponential density, see Sect. 3.1.3.

3.3.3 Orthogonality

As we have seen in Example 3.16, for normal random variables, Appell polynomials are identical with Hermite polynomials and hence orthogonal, with a weight function equal (or proportional) to the probability density function. A natural question is then, under which conditions, i.e. for which probability distributions, Appell polynomials are orthogonal. Unfortunately, it turns out that the normal distribution

is the only one where orthogonality is achieved. Before we can see why this is the case, the following two lemmas are needed. The first result provides an expression for powers x^j in terms of Appell polynomials.

Lemma 3.9 *Let $m_X(z)$ be finite in U_r for some $r > 0$. Then*

$$x^j = \sum_{k=0}^{j} \binom{j}{k} \mu_{j-k} A_k(x). \tag{3.46}$$

Proof The assumption implies the power series representation

$$m_X(z) = \sum_{j=0}^{\infty} \frac{z^j}{j!} \mu_j,$$

and hence (see (3.39), (3.40)),

$$\begin{aligned} e^{zx} = m_X(z) M_X(x, z) &= m_X(z) \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x) \\ &= \left(\sum_{j=0}^{\infty} \frac{z^j}{j!} \mu_j \right) \left(\sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x) \right) = \sum_{j=0}^{\infty} \frac{z^j}{j!} b_j \end{aligned}$$

with

$$b_j(x) = \sum_{k=0}^{j} \binom{j}{k} \mu_{j-k} A_k(x).$$

On the other hand,

$$e^{zx} = \sum_{j=0}^{\infty} \frac{z^j}{j!} x^j,$$

so that

$$b_j = b_j(x) = x^j. \qquad \square$$

The second result is a recursion formula. Here, we will need a notation for cumulants. Thus, let

$$\kappa_X(z) = \log m_X(z) = \sum_{j=1}^{\infty} \frac{z^j}{j!} \kappa_j \tag{3.47}$$

be the cumulant generating function. Then

$$\kappa_j = \frac{d^j}{dz^j} \big[\kappa_X(z) \big]_{z=0} \quad (j = 1, 2, \ldots) \tag{3.48}$$

are called cumulants of X. Note that (3.41) can also be written as

$$A_j(x) = \frac{d^j}{dz^j}\left[e^{xz-\kappa_X(z)}\right]_{z=0}. \tag{3.49}$$

Lemma 3.10 *For Appell polynomials, we have*

$$A_{j+1}(x) = xA_j(x) - \sum_{k=0}^{j}\binom{j}{k}\kappa_{j-k+1}A_k(x). \tag{3.50}$$

Proof Using (3.49) and (3.48), we have

$$\begin{aligned}
A_{j+1}(x) &= \frac{d^{j+1}}{dz^{j+1}}\left[e^{xz-\kappa_X(z)}\right]_{z=0} \\
&= \frac{d^j}{dz^j}\left[\left(x-\kappa_X'(z)\right)e^{xz-\kappa_X(z)}\right]_{z=0} \\
&= x\frac{d^j}{dz^j}\left[e^{xz-\kappa_X(z)}\right]_{z=0} - \sum_{k=0}^{j}\binom{j}{k}\frac{d^{j-k}}{dz^{j-k}}\left[\kappa_X'(z)\right]_{z=0}\frac{d^k}{dz^k}\left[e^{xz-\kappa_X(z)}\right] \\
&= xA_j(x) - \sum_{k=0}^{j}\binom{j}{k}\kappa_{j-k+1}A_k(x).
\end{aligned}$$

□

Now we are ready to obtain the result on orthogonality.

Theorem 3.1 *Appell polynomials are orthogonal, i.e.*

$$\langle A_j, A_k\rangle = E\left[A_j(X)A_k(X)\right] = 0 \quad (j \neq k)$$

if and only if $X \sim N(\mu, \sigma^2)$.

Proof From (3.41) we have

$$A_1(X) = \left[\frac{X\exp(Xz)m_X(z) - \exp(xz)m_X'(z)}{m_X^2(z)}\right] = X - \mu_1.$$

Now suppose that for all $j \geq 2$,

$$\langle A_1, A_j\rangle = E\left[(X-\mu_1)A_j(X)\right] = 0.$$

Since $E[A_j(X)] = 0$ $(j \geq 1)$, this implies $E[XA_j(X)] = 0$. Now taking expected values on both sides of (3.50), we obtain

$$\underbrace{E\left[A_{j+1}(X)\right]}_{0} = \underbrace{E\left[XA_j(X)\right]}_{0} - \kappa_{j+1}\underbrace{E\left[A_0(X)\right]}_{1} - \sum_{k=1}^{j}\binom{j}{k}\kappa_{j-k+1}\underbrace{E\left[A_k(X)\right]}_{0},$$

which means that

$$\kappa_{j+1} = 0 \quad (j \geq 2).$$

The only distribution for which all cumulants except κ_1 and κ_2 are zero is the normal distribution (with $\kappa_1 = \mu$ and $\kappa_2 = \sigma^2$). □

In view of the lack of orthogonality, it is not quite easy to answer the following basic questions: (a) which functions $G(x)$ or random variables $G(X)$ have a representation in terms of Appell polynomials?; (b) is the representation unique?; (c) how do we calculate the coefficients? Answers to these questions will be given in the following sections. However, at this point, we may already introduce a definition that will play a central role for limit theorems (see Sect. 4.2.5 on limit theorems for subordinated linear processes):

Definition 3.8 Suppose that a function $G(x)$ has a unique representation

$$G(x) = \sum_{j=m}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(x)$$

with equality defined in an appropriate sense and $a_{\mathrm{app},m} \neq 0$. Then m is called the Appell rank of G.

Example 3.18 Let $X \sim N(0,1)$. The Appell rank of a function G is the same as its Hermite rank, because Appell polynomials coincide with Hermite polynomials.

3.3.4 Completeness and Uniqueness

Although, in general, Appell polynomials are not orthogonal, it is still possible that they build a basis in a suitable space of functions (or random variables). Thus, the next question is which functions $G(x)$ or transformed random variables $G(X)$ may be written as a series expansion in $A_j(x)$ or $A_j(X)$ respectively and in how far this representation is unique. A little step in this direction is Eq. (3.46):

$$x^j = \sum_{k=0}^{j} \binom{j}{k} \mu_{j-k} A_k(x).$$

However, that $G(x) = x^j$ can be represented by $A_0(x), \ldots, A_j(x)$ does not guarantee that this can be carried over for instance to analytic functions. Also, Lemma 3.9 does not say anything about uniqueness. This can be illustrated by the following example.

Example 3.19 Let $X \sim \text{Exp}(1)$. We rewrite (3.45) as

$$A_j(x) = x^j - jx^{j-1} =: A_j^{\exp}(x).$$

We may then write

$$x^j = x^j - jx^{j-1} + j\left(x^{j-1} - (j-1)x^{j-2}\right) + \cdots + j! = \sum_{k=0}^{j} \frac{j!}{k!} A_k^{\exp}(x).$$

(This also follows directly from (3.46) noting that $\mu_{j-k} = (j-k)!$) Consider now the (analytic) function $\psi(x) = \exp(x)$. If we assume that $\psi(x)$ has a series representation in terms of $A_j^{\exp}(x)$, we would have

$$\begin{aligned}\psi(x) &= \sum_{j=0}^{\infty} \frac{x^j}{j!} = \sum_{j=0}^{\infty} \frac{1}{j!} \sum_{k=0}^{j} \frac{j!}{k!} A_k^{\exp}(x) \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{j} \frac{1}{k!} A_k^{\exp}(x) = \sum_{k=0}^{\infty} c_k A_k^{\exp}(x)\end{aligned}$$

with

$$c_k = \sum_{j=k}^{\infty} \frac{1}{k!} = \infty.$$

Obviously, the expansion $\sum_{k=0}^{\infty} c_k A_k^{\exp}(x)$ is not applicable. The problem arises because

$$\exp(xz) = m_X(z) \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x),$$

so that we would have

$$\exp(x) = \exp(xz)|_{z=1} = m_X(1) \sum_{j=0}^{\infty} \frac{1}{j!} A_j(x).$$

But $m_X(z) = (1-z)^{-1}$, so that $m_X(1) = \infty$.

We can conclude that, in spite of Lemma 3.9, not all analytic functions can be represented by Appell polynomials. Instead, one needs to focus on a smaller class of functions. This leads to the following definition.

Definition 3.9 An entire function $\psi : \mathbb{C} \to \mathbb{C}$ is called of exponential order of type τ $(0 < \tau < \infty)$ if there exists a finite number $M > 0$ such that

$$\left|\psi(z)\right| = \left|\psi\left(re^{i\varphi}\right)\right| \leq Me^{\tau r} \tag{3.51}$$

for all $z = re^{i\varphi} \in \mathbb{C}$. The class of functions of type τ is denoted by $\mathbb{E}(\tau)$. Moreover, the exact type of ψ is the smallest τ such that (3.51) holds.

We now can give sufficient conditions in order that $\psi(z)$ can be represented by $A_j(z)$ $(j = 0, 1, 2, \ldots)$ pointwise, i.e. for each fixed z.

Theorem 3.2 *Let $A(z)$ be an analytic function in $U_\tau = \{|z| < \tau\}$, and A_j $(j = 0, 1, 2, \ldots)$ be a sequence of polynomials such that*

$$e^{zw} = \frac{1}{A(w)} \sum_{j=0}^{\infty} \frac{w^j}{j!} A_j(z) \quad (z \in \mathbb{C}).$$

Moreover, let

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j \in \mathbb{E}(\tau_1),$$

where $\tau_1 < \tau$. Then ψ has the representation

$$\psi(z) = \sum_{j=0}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(z) \tag{3.52}$$

with

$$a_{\mathrm{app},j} = \frac{1}{2\pi i} \int_\Gamma w^j \frac{H(w)}{A(w)}\, dw, \tag{3.53}$$

where integration is over the curve $\Gamma = \{|w| = \tau_2\}$ for some $\tau_1 < \tau_2 < \tau$, and $H(w)$ is the Borel transformation

$$H(w) = \sum_{j=0}^{\infty} j! \psi_j z^{-j-1}. \tag{3.54}$$

The convergence in (3.52) is absolute and uniform on compact sets. Moreover,

$$\limsup_{j\to\infty} \sqrt[j]{|a_{\mathrm{app},j}|} \le \tau_1.$$

To prove and understand the theorem, some preliminary results are needed. First, we show that being of exponential order is equivalent to fast convergence of the coefficients in the power series representation.

Lemma 3.11 *We have*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j \in \mathbb{E}(\tau_1)$$

if and only if

$$\limsup_{j\to\infty} \sqrt[j]{j!|\psi_j|} \le \tau_1. \tag{3.55}$$

Proof Suppose first that (3.55) holds. Then, for $\tilde{\tau}_1 > \tau_1$ arbitrarily close to τ,

$$|\psi_j| \le C_1 \frac{\tilde{\tau}_1^j}{j!},$$

and hence,

$$\left|\psi\left(re^{i\varphi}\right)\right| \le \sum_{j=0}^{\infty} |\psi_j| r^j \le C_1 \sum_{j=0}^{\infty} \frac{(\tilde{\tau}_1 r)^j}{j!} = C_1 e^{\tilde{\tau}_1 r}.$$

Since $\tilde{\tau}_1$ is arbitrarily close to τ_1, it then follows that $\psi(z) \in \mathbb{E}(\tau_1)$.

Suppose now that $\psi(z) \in \mathbb{E}(\tau_1)$. Let $r > 0$ and recall Cauchy's inequality

$$|\psi_j| \le r^{-j} \max_{|z|=r} \left|\psi(z)\right|.$$

Then we have

$$|\psi_j| \le r^{-j} M e^{\tau_1 r} = M \tau_1^j e^{\tau_1 r} (\tau_1 r)^{-j}.$$

Now

$$\min_{r>0} \frac{e^{\tau_1 r}}{(\tau_1 r)^j} = \frac{e^j}{j^j},$$

so that we obtain

$$|\psi_j| \le M \frac{(\tau_1 e)^j}{j^j}.$$

By Stirling's formula, $e^j j! j^{-j} \sim \sqrt{2\pi j}$, so that, for j large enough,

$$1 \le \frac{e^j j!}{j^j} < (j+1)e,$$

and hence,

$$\limsup \sqrt[j]{j! e^j j^{-j}} = 1,$$

and

$$\limsup \sqrt[j]{|j!\psi_j|} \le \tau_1 \limsup \sqrt[j]{M} \limsup \sqrt[j]{j! e^j j^{-j}} = \tau_1. \qquad \square$$

Next, we show that the Borel transformation is a convergent Laurent series.

Lemma 3.12 *Let*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j \in \mathbb{E}(\tau_1).$$

Then the Laurent series

$$H(w) = \sum_{j=0}^{\infty} j!\psi_j w^{-j-1} \tag{3.56}$$

is convergent in $\{|w| > \tau'\}$, *where*

$$\tau' = \limsup \sqrt[j]{j!|\psi_j|} \le \tau_1. \tag{3.57}$$

Proof The Laurent series $\sum_{j=-\infty}^{\infty} c_j z^j$ is convergent in $\{r < |z| < R\}$ with

$$r = \limsup_{j\to\infty} \sqrt[j]{|c_{-j}|} = \limsup_{j\to\infty} \sqrt[j]{|(j+1)!\psi_{j+1}|} \le \tau_1$$

by Lemma 3.11, and

$$R = \left(\limsup_{j\to\infty} \sqrt[j]{|c_j|}\right)^{-1} = \infty$$

since $c_j = 0$ for $j \ge 0$. □

Note that usually the Borel transform is denoted by $F(z)$. Here we use the notation $H(z)$ instead to avoid confusion with cumulative distribution functions. Next, it is shown that H can also be written as the Laplace transform of ψ. Indeed, using the power series representation of $\psi(z)$ and partial integration, we obtain

$$\int_0^{\infty} \psi(t)e^{-zt}\,dt = \sum_{j=0}^{\infty} \psi_j \int_0^{\infty} t^j e^{-zt}\,dt = \sum_{j=0}^{\infty} j!\psi_j z^{-j-1} = H(z).$$

Finally, we obtain a representation of ψ as a complex integral over a closed curve around the origin (containing the U_τ neighbourhood of 0). This is also called Borel–Polya representation.

Lemma 3.13 *Let* $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j \in \mathbb{E}(\tau_1)$, *and for an* $\varepsilon > 0$, *define the curve* $\Gamma = \{w \in \mathbb{C} : |w| = \tau_1 + \varepsilon\} = \{w \in \mathbb{C} : w = \gamma(t), t \in [0, 2\pi]\}$, *where* γ *is an injective*, *continuous*, *piecewise differentiable parameterization of the curve. Then*

$$\psi(z) = \frac{1}{2\pi i} \int_{\Gamma} e^{zw} H(w)\,dw = \frac{1}{2\pi i} \int_0^{2\pi} e^{z\gamma(t)} H\big(\gamma(t)\big)\gamma'(t)\,dt. \tag{3.58}$$

Proof Since for $w \in \Gamma$, we have $|w| > \tau_1$, Lemma 3.12 implies that $H(w)$ is convergent on Γ, and since $|w| = \tau_1 + \varepsilon$, convergence is uniform. Also, the power

series representation of $\exp(zw)$ is uniformly convergent on Γ. Therefore, we may exchange summation and integration:

$$\int_\Gamma e^{zw} H(w)\,dw = \int_\Gamma \left(\sum_{j=0}^{\infty} \frac{z^j w^j}{j!} \sum_{j=0}^{\infty} j! \psi_j w^{-j-1} \right) dw$$
$$= \sum_{j=0}^{\infty} j! \psi_j \sum_{k=0}^{\infty} \frac{z^k}{k!} \int_\Gamma w^{k-j-1}\,dw.$$

Now,

$$\gamma(t) = (\tau_1 + \varepsilon) e^{it}, \qquad \gamma'(t) = i(\tau_1 + \varepsilon) e^{it},$$

so that

$$\int_\Gamma w^m\,dw = \int_0^{2\pi} (\tau_1 + \varepsilon)^m e^{imt} \cdot i(\tau_1 + \varepsilon) e^{it}\,dt$$
$$= i(\tau_1 + \varepsilon)^{m+1} \int_0^{2\pi} e^{i(m+1)t}\,dt = 2\pi i \cdot 1\{m = -1\}$$

and

$$\int_\Gamma e^{zw} H(w)\,dw = 2\pi i \sum_{j=0}^{\infty} j! \psi_j \frac{z^j}{j!} = 2\pi i \sum_{j=0}^{\infty} \psi_j z^j = 2\pi i \cdot \psi(z). \qquad \square$$

We now have everything that is needed for the proof of Theorem 3.2:

Proof of Theorem 3.2 The idea is as follows. By assumption

$$e^{zw} = \frac{1}{A(w)} \sum_{j=0}^{\infty} \frac{w^j}{j!} A_j(z). \tag{3.59}$$

On the other hand, we have from (3.58)

$$\psi(z) = \frac{1}{2\pi i} \int_\Gamma e^{zw} H(w)\,dw.$$

The representation of ψ in terms of A_j follows by replacing $\exp(zw)$ in the Borel–Polya representation of ψ by the right-hand side of (3.59),

$$\psi(z) = \frac{1}{2\pi i} \int_\Gamma e^{zw} H(w)\,dw$$
$$= \sum_{j=0}^{\infty} \frac{1}{j!} \underbrace{\left(\frac{1}{2\pi i} \int_\Gamma w^j \frac{H(w)}{A(w)}\,dw \right)}_{a_{\mathrm{app},j}} A_j(z).$$

Here, interchanging integration and summation is possible due to uniform convergence and since $\tau_1 < \tau$ and $A(w)$ is analytic for $|w| < \tau$ (and hence Γ can be chosen such that $A(w)$ is not zero there).

Finally, to give an upper bound for $|a_{\mathrm{app},j}|$, we choose $\tau_2 = \tau_1 + \varepsilon$ such that $\tau_1 < \tau_2 < \tau$, so that

$$\min_{w\in\Gamma}\big|A(w)\big| \geq \min_{|z|<\tau}\big|A(w)\big| = c > 0$$

and by (3.54)

$$\begin{aligned}
|a_{\mathrm{app},j}| &\leq \frac{1}{2\pi}\left|\sum_{k=0}^{\infty} k!\psi_k \int_\Gamma \frac{w^{j-k-1}}{A(w)}\,dw\right| \\
&\leq c^{-1}\frac{1}{2\pi}\sum_{k=0}^{\infty} k!|\psi_k| \int_0^{2\pi} \big|\big(\tau_2 e^{it}\big)^{j-k-1} i\tau_2 e^{it}\big|\,dt \\
&= c^{-1}\frac{1}{2\pi}\left(\sum_{k=0}^{\infty} k!|\psi_k|\cdot 2\pi\,\tau_2^{-k}\right)\tau_2^j.
\end{aligned}$$

Now (3.57) and $\tau_1 < \tau_2$ implies that $k!|\psi_k|\tau_2^{-k} = O(\alpha^k)$ for some $0 < \alpha < 1$, so that

$$|a_{\mathrm{app},j}| \leq \mathrm{const}\cdot\tau_2^j.$$

Since $\tau_2 < \tau$ and τ_2 is arbitrarily close to τ_1, we obtain

$$\limsup \sqrt[j]{|a_{\mathrm{app},j}|} \leq \tau_1. \qquad \square$$

Theorem 3.2 can now be used to obtain an Appell polynomial representation, simply by setting

$$A(z) = \frac{1}{m(z)} = m_{\mathrm{inv}}(z).$$

Theorem 3.3 *Let $m_X(z) = E(e^{zX})$ be defined in $U_\tau = \{|z| < \tau\}$ and denote by A_j $(j = 0, 1, 2, \ldots)$ the Appell polynomials associated with X. Suppose that $\psi \in \mathbb{E}(\tau_1)$ for some $\tau_1 < \tau$. Then ψ has the Appell polynomial representation*

$$\psi(z) = \sum_{j=0}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(z) \quad (z \in \mathbb{C}) \tag{3.60}$$

with

$$a_{\mathrm{app},j} = \frac{1}{2\pi i}\int_\Gamma w^j H(w) m_X(w)\,dw, \tag{3.61}$$

where integration is over the curve $\Gamma = \{|w| = \tau_2\}$ *for some* $\tau_1 < \tau_2 < \tau$, *and* $H(w)$ *is the Borel transform*

$$H(w) = \sum_{j=0}^{\infty} j!\psi_j w^{-j-1}.$$

The convergence in (3.60) *is absolute and uniform on compact sets. Moreover,*

$$\limsup_{j\to\infty} \sqrt[j]{|a_j|} \le \tau_1.$$

Finally, it also follows from the theorem that the coefficients in (3.61) are real numbers:

Corollary 3.1 *Under the assumptions of Theorem* 3.3, *we have* $\psi(x) \in \mathbb{R}$ *for all* x *as well as* $a_{\mathrm{app},j} \in \mathbb{R}$.

Proof Since $m_X(z)$ is finite for $z \in U_\tau$, we have $A(z) \neq 0$ in the same domain. Also, the coefficients of $m_X(z)$ are real numbers, and

$$a_{\mathrm{app},j} = \sum_{k=0}^{\infty} k!\psi_k \frac{1}{2\pi i} \int_\Gamma m_X(w) w^{j-k-1}\, dw.$$

For $j - k - 1 \ge 0$, $m_X(w)w^{j-k-1}$ is an analytic function on U_τ, so that Cauchy's integral theorem implies

$$\frac{1}{2\pi i} \int_\Gamma m_X(w) w^{j-k-1}\, dw = 0.$$

On the other hand, for $j - k - 1 \le -1$, Cauchy's integral formula yields

$$\begin{aligned} \frac{1}{2\pi i} \int_\Gamma m_X(w) w^{j-k-1}\, dw &= \frac{1}{2\pi i} \int_\Gamma m_X(w) w^{-(k-j+1)}\, dw \\ &= \frac{1}{(k-j)!} \frac{d^{k-j}}{dz^{k-j}} \big[m_X(z)\big]_{z=0} \in \mathbb{R}. \end{aligned}$$

□

In general, it may be somewhat tedious to calculate the Appell coefficients via complex integration in (3.61). Fortunately, (3.61) can be replaced by simpler formulas such as the following.

Corollary 3.2 *Let* $m_X(z) = E(e^{zX})$ *be defined in* $U_\tau = \{|z| < \tau\}$. *Denote by* A_j $(j = 0, 1, 2, \ldots)$ *the Appell polynomials associated with* X *and by* $\mu_j = E(X^j)$ *the moments of* X. *Suppose that* $\psi \in \mathbb{E}(\tau_1)$ *for some* $\tau_1 < \tau$. *Then* ψ *has the Appell polynomial representation*

$$\psi(z) = \sum_{j=0}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(z) \quad (z \in \mathbb{C}) \tag{3.62}$$

with

$$a_{\mathrm{app},j} = \sum_{k=j}^{\infty} \frac{k!}{(k-j)!} \psi_k m_X^{(k-j)}(0) = \sum_{k=j}^{\infty} \frac{k!}{(k-j)!} \psi_k \mu_{k-j}. \tag{3.63}$$

In view of these results, we gain a better understanding of Example 3.19:

Example 3.20 Let $X \sim \mathrm{Exp}(1)$. Then $A_j(x) = x^j - jx^{j-1}$ (see (3.45)). Consider now $\psi(z) = z^l$ for some $l \geq 0$. Then clearly $\psi \in \mathbb{E}(\tau_1)$ for all $\tau_1 > 0$. Moreover, $m_X(z) = (1-z)^{-1}$ is defined for $U_1 = \{|z| < 1\}$. Thus, we may choose $\tau_1 < \tau$ so that (3.63) is applicable. This yields

$$\begin{aligned} a_j &= 0 \quad (j \geq l+1), \\ a_j &= \frac{l!}{(l-j)!} \mu_{l-j} = l! \quad (0 \leq j \leq l), \end{aligned}$$

and

$$z^l = \sum_{j=0}^{l} \frac{l!}{j!} A_j(z).$$

Example 3.21 For X as in the previous example, consider now $\psi(z) = \exp(z)$. Then $\psi \in \mathbb{E}(\tau_1)$ with $\tau_1 \geq 1$, but not for $\tau_1 < 1$. In other words, ψ is of the exact type $\mathbb{E}(1)$. However, $m_X(z) = (1-z)^{-1}$ is analytic in $U_1 = \{|z| < \tau\}$ only for $\tau \leq 1$. Thus, there is no $\tau_1 < \tau$ for which ψ would be in $\mathbb{E}(\tau_1)$ and Corollary 3.2 is not applicable. In fact, the explicit calculation in Example 3.19 shows that an expansion of ψ in terms of these Appell polynomials is not possible.

Theorem 3.3 does not necessarily imply that the Appell polynomial representation is unique. A perhaps surprising example can be given as follows:

Example 3.22 We saw that for $X \sim \mathrm{Exp}(1)$ and corresponding Appell polynomials, we have, for all $l \geq 0$,

$$\frac{z^l}{l!} = \sum_{j=0}^{l} \frac{1}{j!} A_j(z).$$

This implies

$$\lim_{l \to \infty} \frac{z^l}{l!} = 0 = \sum_{j=0}^{\infty} \frac{1}{j!} A_j(z). \tag{3.64}$$

In other words, we obtain a nontrivial (i.e. with nonzero coefficients) pointwise representation of the function $\psi(z) \equiv 0$! But, of course, the representation $0 = \sum_{j=0}^{\infty} 0 \cdot A_j(z)$ is also correct. Thus, there are at least two different representations of zero by Appell polynomials.

This example illustrates that pointwise convergence of nonorthogonal representations can lead to rather counterintuitive results. We may ask the question whether this pathology disappears when other types of convergence are considered. Since we are dealing with random variables in $L^2(\Omega)$, we consider $L^2(\Omega)$-convergence. In general, pointwise convergence does not imply $L^2(\Omega)$-convergence. This is sufficient to resolve the problem in Example 3.22:

Example 3.23 For $X \sim \mathrm{Exp}(1)$ as above,

$$\left\| \frac{X^l}{l!} - 0 \right\|_{L^2}^2 = \left(\frac{1}{l!}\right)^2 E\left[X^{2l}\right] = \left(\frac{1}{l!}\right)^2 (2l)! \to \infty$$

as $l \to \infty$. Thus, in the $L^2(\Omega)$-norm, $x^l/l!$ does not converge to zero, so that the series $\sum A_j(x)/l!$ does not represent $\psi(x) \equiv 0$. In contrast, the representation

$$0 = \sum_{j=0}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(x)$$

with all $a_{\mathrm{app},j} = 0$ is of course correct.

In order to achieve not only pointwise but also L^2-convergence, stronger conditions are needed. This is formulated in the following theorem, together with an even simpler formula for the Appell coefficients.

Theorem 3.4 *Let $m_X(z)$ be defined in $U_\tau = \{|z| < \tau\}$, denote by A_j the Appell polynomials associated with X and suppose that $\psi \in \mathbb{E}(\tau_1/2)$ with $\tau_1 < \tau$. Then*

$$\psi(X) \underset{L^2(\Omega)}{=} \sum_{j=0}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(X) \tag{3.65}$$

with

$$a_{\mathrm{app},j} = E\left[\psi^{(j)}(X)\right]. \tag{3.66}$$

Proof First, we show that

$$\psi_n(X) = \sum_{j=0}^{n} \frac{a_{\mathrm{app},j}}{j!} A_j(X)$$

is a Cauchy sequence in $L^2(\Omega)$. From Theorem 3.2 we know that $\limsup \sqrt[j]{|a_{\mathrm{app},j}|} \leq \tau_1/2$, so that, for $\tau_2 \in (\tau_1, \tau)$ and N large enough,

$$|a_{\mathrm{app},j}| \leq \left(\frac{\tau_2}{2}\right)^j \quad (j \geq N).$$

Hence, for $M \geq N$,

$$\left\| \psi_M(X) - \psi_N(X) \right\|^2_{L^2(\Omega)} \leq \int \Delta^2_{N,M}(x)\, dP(x) \tag{3.67}$$

with

$$\Delta_{N,M}(x) = \sum_{j=N+1}^{M} \frac{(\tau_2/2)^j}{j!} \left| A_j(x) \right|.$$

Now,

$$\frac{e^{xz}}{m_X(z)} = \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x)$$

is absolutely convergent for $z = \tau_2/2 < \tau$, so that certainly pointwise $\Delta_{N,M} \to 0$ as $N \to \infty$ (and $M \geq N$). Thus, the left-hand side in (3.67) converges to zero if we can write

$$\lim_{N\to\infty} \int \Delta^2_{N,M}(x)\, dP(x) = \int \lim_{N\to\infty} \Delta^2_{N,M}(x)\, dP(x).$$

To show this, the dominated convergence theorem can be applied as follows. From (3.41) we have

$$A_j(x) = \sum_{k=0}^{j} \binom{j}{k} a_{\mathrm{inv},k} x^{j-k} = \sum_{k=0}^{j} \binom{j}{k} a_{\mathrm{inv},j-k} x^k. \tag{3.68}$$

Setting $t = \tau_2/\tau$, we have

$$\begin{aligned}
\sum_{j=0}^{\infty} \frac{t^j}{j!} \left| A_j(x) \right| &\leq \sum_{j=0}^{\infty} \frac{t^j}{j!} \sum_{k=0}^{j} \binom{j}{k} |a_{\mathrm{inv},j-k}| |x|^k \\
&= \sum_{j=0}^{\infty} \sum_{k=0}^{j} \frac{t^{j-k}}{j!} \frac{j!}{(j-k)!k!} |a_{\mathrm{inv},j-k}| \big(|x|t\big)^k \\
&\leq \left(\sum_{p=0}^{\infty} \frac{(|x|t)^p}{p!} \right) \left(\sum_{q=0}^{\infty} \frac{|a_{\mathrm{inv},q}|}{q!} t^q \right) \\
&= e^{|x|t} \cdot \mathrm{const}\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\int \Delta^2_{N,M}(x)\, dP(x) &\leq \mathrm{const} \cdot \int e^{2|x|t}\, dP(x) \\
&\leq \mathrm{const} \cdot \big[m_X(2t) + m_X(-2t) \big].
\end{aligned}$$

Since $2t = \tau_2 < \tau$, $m_X(\pm 2t)$ are finite, so that the dominated convergence theorem applies.

Next, we need to derive (3.66). Since

$$\psi(x) = \sum_{k=0}^{\infty} \psi_k x^k,$$

we have, for $j \geq 1$,

$$\psi^{(j)}(x) = \sum_{k=j}^{\infty} \psi_k \frac{k!}{(k-j)!} x^{k-j}$$

and

$$E\big[\psi^{(j)}(X)\big] = \sum_{k=j}^{\infty} \psi_k \frac{k!}{(k-j)!} \mu_{k-j},$$

which is however equal to $a_{\mathrm{app},j}$ due to Eq. (3.63). The same argument applies for $j = 0$. □

Example 3.24 Consider $X \sim \mathrm{Exp}(1)$, the corresponding Appell polynomials $A_j(x) = x^j - jx^{j-1}$ (see (3.45)) and $\psi(z) = z^l$. Then $m_X(z) = (1-z)^{-1}$ is defined for $z \in U_1 = \{|z| < 1\}$ and $\psi \in \mathbb{E}(\tau_1/2)$ for all $\tau_1 > 0$. Therefore, we may choose $\tau_1/2 < \tau/2 = 1/2$, and Theorem 3.4 is applicable. The coefficients are therefore obtained by

$$a_{\mathrm{app},j} = E\big[\psi^{(j)}(X)\big] = 0 \quad (j \geq l+1),$$

and for $0 \leq j \leq l$,

$$\begin{aligned} a_{\mathrm{app},j} &= E\big[\psi^{(j)}(X)\big] = E\big[l(l-1)\cdots(l-j+1)X^{l-j}\big] \\ &= l(l-1)\cdots(l-j+1)\mu_{l-j} = l!. \end{aligned}$$

Then

$$X^l \underset{L^2(\Omega)}{=} \sum_{j=0}^{l} \frac{l!}{j!} A_j(X).$$

Example 3.25 Consider again $X \sim \mathrm{Exp}(1)$, but $\psi(z) = e^z$. Then $\psi \in \mathbb{E}(1)$, which is also the exact type, however $m_X(z)$ is analytic for $|z| < 1$ only. Therefore, Theorem 3.4 is not applicable.

Example 3.26 Consider again $X \sim \mathrm{Exp}(1)$, but $\psi(z) \equiv 0$. Obviously, Theorem 3.4 is applicable, and the coefficients are obtained by

$$a_{\mathrm{app},j} = E\big[\psi^{(j)}(X)\big] = E[0] = 0$$

for all $j \geq 0$. In particular, the rather strange nontrivial representation of zero (3.64) is excluded.

An important question has not been answered yet, namely in how far and under which circumstances we can be sure that the Appell polynomial representation of $\psi(z)$ (pointwise) or $\psi(X)$ is unique. With respect to pointwise convergence, a counterexample was the representation of $\psi(z) \equiv 0$. For the $L^2(\Omega)$-representation of $\psi(X)$, relatively simple general conditions for uniqueness can be given as follows (see e.g. Giraitis 1985). We consider the L^2-space (denoted by L^2_X) of random variables measurable with respect to the σ-algebra generated by the random variable X and, without loss of generality, expected value zero. This is a Hilbert space equipped with the scalar product equal to the covariance and hence also a Banach space with norm $\|X\|^2 = \langle X, X\rangle$. The question is now whether $A_j(X)$ $(j = 0, 1, 2, \ldots)$ is a Schauer (i.e. a complete and minimal) basis in this Banach space. To be specific, we recall some standard definitions.

Definition 3.10 Let B be a Banach space over $K = \mathbb{R}$ or $\mathbb{C}$. Then a sequence of elements $v_j \in B$ $(j \in \mathbb{N})$ is called complete if $\overline{\text{span}\{v_j, j \in \mathbb{N}\}}$ (i.e. the closure of all linear combinations of v_js) is equal to B.

Definition 3.11 A system $v_j \in B$ $(j \in \mathbb{N})$ is called minimal (or a minimal system) if for all k, $v_k \notin \overline{\text{span}\{v_j, j \neq k\}}$.

Definition 3.12 A system $v_j \in B$ $(j \in \mathbb{N})$ is called a Schauder basis if it is complete and minimal or, equivalently (see e.g. Banach 1948), if for every $w \in B$, there exists a unique sequence $\alpha_j \in K$ $(j \in \mathbb{N})$ such that

$$w = \sum_{j=0}^{\infty} \alpha_j v_j .$$

A partial answer with respect to uniqueness can be given as follows (see e.g. Giraitis 1985):

Theorem 3.5 *Let $X \sim F_X$ and suppose that $m_X(r) < \infty$ for some $r > 0$. Then $A_j(X)$ $(j \in \mathbb{N})$ is a minimal system in L^2_X if and only if the following three conditions hold*: (i) *F_X has a density $p_X = F_X'$*; (ii) *$p_X \in C^\infty(\mathbb{R})$*; (iii) *$Q_j = p_X^{(j)}/p_X \in L^2(F_X)$ $(j \in \mathbb{N})$.*

The sequence of functions Q_j is also called a biorthogonal system to the sequence A_j in $L^2(F_X)$, in the sense that

$$\begin{aligned} \langle A_j, Q_k\rangle_{L^2(F_X)} &= \int A_j(x) Q_k(x) f_X(x)\, dx \\ &= \int A_j(x) p_X^{(k)}(x)\, dx = \delta_{jk} \cdot j!(-1)^j . \end{aligned}$$

This can also be rephrased as follows. Let $\tilde{Q}_j = Q_j(-1)^j/j!$. Then, given the two Banach spaces, $L^2_X(\Omega)$ and $L^2(F_X)$, and the bilinear form $\langle\cdot,\cdot\rangle : L^2_X(\Omega) \times L^2(F_X) \to \mathbb{R}$, the two sets $\{A_j,\, j \in \mathbb{Q}\} \subset L^2_X(\Omega)$ and $\{\tilde{Q}_j,\, j \in \mathbb{N}\} \subset L^2(F_X)$ are such that $\langle A_j, \tilde{Q}_k\rangle = \delta_{jk}$.

The detailed proof of Theorem 3.5 is rather technical and is therefore omitted here. However, an intuitive explanation why Q_j is a biorthogonal system to A_j can be given as follows. From (3.65), (3.66) we have for $\psi(z) = A_j(z)$,

$$\psi(z) = A_j(z) = \frac{a_{\text{app},j}}{j!} A_j(z)$$

and hence

$$\begin{aligned} a_{\text{app},k} &= j!\delta_{jk} = E\big[\psi^{(k)}(X)\big] \\ &= \int \psi^{(k)}(x) p_X(x)\, dx = \int A_j^{(k)}(x) p_X(x)\, dx. \end{aligned}$$

If partial integration can be applied, then we obtain

$$\begin{aligned} \int A_j^{(k)}(x) p_X(x)\, dx &= (-1)^k \int A_j(x) p_X^{(k)}(x)\, dx \\ &= (-1)^k \int A_j(x) Q_k(x) p_X(x)\, dx = (-1)^k \langle A_j, Q_k\rangle_{L^2(F_X)}. \end{aligned}$$

An interesting, and at first sight surprising, consequence of Theorem 3.5 is the following:

Corollary 3.3 *Let $X \sim F_X$ such that $E(e^{rX}) < \infty$ for some $r > 0$, but suppose that $p_X = F_X'$ does not exist. Then A_j $(j \in \mathbb{N})$ is not a minimal system in $L^2_X(\Omega)$. In particular, there exists at least one function ψ with $\psi(X) \in L^2_X(\Omega)$ for which more than one $L^2_X(\Omega)$-representation by $A_j(X)$ $(j \in \mathbb{N})$ exists and hence the Appell rank is not defined.*

The following example illustrates how it is possible to obtain more than one representation.

Example 3.27 Let $P(X = 0) = P(X = 1) = \frac{1}{2}$. Then, for any two functions G, H,

$$G(x) \underset{L^2_X(\Omega)}{=} H(x)$$

if and only if $G(0) = H(0)$ and $G(1) = H(1)$. The equation defining Appell polynomials,

$$\frac{\exp(xz)}{\frac{1}{2}(1 + \exp(z))} = \sum_{j=0}^{\infty} \frac{z^j}{j!} A_j(x),$$

implies $A_j(0) \neq 0$, $A_j(1) \neq 0$ for all j. Consider now any function ψ. Then, for any $j \neq k$, one can find real coefficients a, b such that

$$\psi(x) \underset{L^2_X(\Omega)}{=} aA_j(x) + bA_k(x).$$

The reason is that all one has to do is to solve the system of two equations,

$$\psi(0) = aA_j(0) + bA_k(0),$$
$$\psi(1) = aA_j(1) + bA_k(1).$$

Since $A_j(x)$ and $A_k(x)$ are not zero for $x = 0, 1$, this is always possible. Thus, we may conclude that for Bernoulli variables, there are infinitely many representations of $\psi(X)$. Obviously, the definition of an Appell rank does not make any sense here.

3.3.5 Extension Beyond Analytic Functions

So far, Appell polynomial expansions have been considered for analytic functions that do not grow too fast. In some applications however one needs to go beyond analytic functions. For instance, the indicator function $G(x) = 1\{X \leq x\}$ used in the empirical distribution function

$$F_n(x) = n^{-1} \sum_{t=1}^{n} 1\{X_t \leq x\}$$

is not analytic. The two main questions are (a) whether a unique Appell polynomial expansion exists for certain classes of functions that are not necessarily analytic, and (b) how to calculate the Appell coefficients.

With respect to the second questions, formula (3.66) is applicable only if ψ is j-times differentiable. Thus, to calculate the coefficients $a_{\text{app},j}$, ψ would have to be infinitely differentiable. The questions is thus what to do if ψ is either not differentiable at all or if it is differentiable almost everywhere (w.r.t. Lebesgue measure), but the derivative is always zero. The latter is for instance the case for the indicator function. Fortunately, formula (3.66) can be rewritten by (formal) partial integration so that derivatives are not required:

Lemma 3.14 *Let $X \sim F_X$ be such that the assumptions of Theorem 3.5 hold, and ψ be such that it has a unique representation*

$$\psi(X) \underset{L^2_X(\Omega)}{=} \sum_{j=0}^{\infty} \frac{a_{\text{app},j}}{j!} A_j(X).$$

Moreover, assume that there is a sequence of analytic functions G_n such the assumptions of Theorem 3.4 hold and G_n converges to ψ "sufficiently fast". Then

$$a_j = (-1)^j \int \psi(x) p_X^{(j)}(x)\, dx. \tag{3.69}$$

A heuristic justification can be as follows. From Theorem 3.4 we have

$$G_n(X) = \sum_{j=0}^{\infty} g_{j,n} X^j \underset{L^2_X(\Omega)}{=} \sum_{j=0}^{\infty} \frac{a_{\text{app},j,n}}{j!} A_j(X)$$

with

$$\begin{aligned} a_{\text{app},j,n} &= E\big[G_n^{(j)}(X)\big] = \int G_n^{(j)}(x) p_X(x)\, dx \\ &= \cdots = (-1)^j \int G_n(x) p_X^{(j)}(x)\, dx. \end{aligned}$$

Thus, if G_n converges to ψ such that the limit in n can be interchanged with integration and summation, then

$$\begin{aligned} \psi(X) &= \lim_{n\to\infty} \sum_{j=0}^{\infty} \frac{a_{\text{app},j,n}}{j!} A_j(X) \\ &= \sum_{j=0}^{\infty} \frac{\lim_{n\to\infty} a_{\text{app},j,n}}{j!} A_j(X) \end{aligned}$$

and

$$\begin{aligned} a_{\text{app},j} &= \lim_{n\to\infty} a_{\text{app},j,n} = (-1)^j \lim_{n\to\infty} \int G_n(x) p_X^{(j)}(x)\, dx \\ &= (-1)^j \int \lim_{n\to\infty} G_n(x) p_X^{(j)}(x)\, dx = (-1)^j \int \psi(x) p_X^{(j)}(x)\, dx. \end{aligned}$$

Now we address question (a), the validity of the Appell polynomial expansion for suitable classes of nonanalytic functions. A partial answer can be found in Schützner (2009), Ramm (1980) and Beran and Schützner (2008). Recall (see Definition 3.9) the notion of the set $\mathbb{E}(\tau)$. The general idea is to define a suitable subclass of $\tilde{\mathbb{E}}(\tau) \subset \mathbb{E}(\tau)$ and approximate ψ by functions from $\tilde{\mathbb{E}}(\tau)$. Specifically, we use the following definition.

Definition 3.13 Let $L^2[-\tau,\tau] = \{g : \int_{-\tau}^{\tau} |g(t)|^2\, dt < \infty\}$. Then

$$\tilde{\mathbb{E}}(\tau) = \left\{ G : G(z) = \int_{-\tau}^{\tau} g(t) e^{itz}\, dt, z \in \mathbb{C}, g \in L^2[-\tau,\tau] \right\}.$$

First, it can be shown that $\tilde{\mathbb{E}}(\tau)$ is indeed a subset of $\mathbb{E}(\tau)$.

Lemma 3.15 $\tilde{\mathbb{E}}(\tau) \subset \mathbb{E}(\tau)$.

Proof Recall that by Morena's theorem, $\int_\Gamma G(z)\,dz = 0$ for all closed curves $\Gamma \subset U = \{z : |z| < \tau\}$ implies that G is analytic on U. For $G \in \tilde{\mathbb{E}}(\tau)$, we have

$$\begin{aligned}\int_\Gamma G(z)\,dz &= \int_\Gamma \left(\int_{-\tau}^{\tau} g(t) e^{itz}\,dt \right) dz \\ &= \int_{-\tau}^{\tau} g(t) \left(\int_\Gamma e^{itz}\,dz \right) dt = 0.\end{aligned}$$

Since this is true for all closed curves, it follows that G is entire. Using the notation $z = x + iy$, an upper bound for G can be given by

$$\left|G(z)\right| \le \int_{-\tau}^{\tau} \left|g(t)\right| e^{-t|y|}\,dt \le e^{\tau|z|} \sqrt{\int_{-\tau}^{\tau} \left|g(t)\right|^2 dt},$$

so that $G \in \mathbb{E}(\tau)$. □

Moreover, $\tilde{\mathbb{E}}(\tau)$ is identical with all functions in $\mathbb{E}(\tau)$ that are square integrable:

Lemma 3.16 *Let $G \in \mathbb{E}(\tau)$. Then $G \in \tilde{\mathbb{E}}(\tau)$ if and only if $\int_{-\infty}^{\infty} |G(x)|^2\,dx < \infty$.*

Proof We refer to classical results from analysis, namely: For "$\Longrightarrow$", note that by Plancherel's theorem we have

$$\int_{-\infty}^{\infty} \left|G(x)\right|^2 dx = 2\pi \int_{-\tau}^{\tau} \left|g(\omega)\right|^2 d\omega < \infty,$$

where $g(\omega)$ is the Fourier transform of G.

The reverse direction "$\Longleftarrow$" is more complicated but is well known as the Paley–Wiener theorem (see e.g. Boas 1954, p. 210). □

Using these results, it then can be shown that the following uniform approximations are possible (see Ramm 1980; Schützner 2006 and Beran and Schützner 2008):

Corollary 3.4 *Let $\psi \in C(\mathbb{R})$. Then, for any $a, \tau > 0$, there exists a sequence of functions $G_n \in \mathbb{E}(\tau)$ such that, as $n \to \infty$,*

$$\sup_{x \in [-a,a]} \left|\psi(x) - G_n(x)\right| \to 0. \tag{3.70}$$

Moreover, if ψ is also bounded, then the sequence may be chosen such that $|G_n(x)| \le C_0 < \infty$ for all n and x, and

$$\sup_{x\in\mathbb{R}} \big|\psi(x) - G_n(x)\big| \to 0. \tag{3.71}$$

It should be noted that condition (3.70) alone is not sufficient to guarantee the existence and uniqueness of an Appell polynomial expansion in general. Depending on additional assumptions, an additional formal proof is required.

3.4 Multivariate Appell Polynomials and Wick Products

3.4.1 Definition

The notion of Appell polynomials can be extended to multivariate random variables. Let $X = (X_1, \dots, X_k)^T \in \mathbb{R}^k$ be a k-dimensional random variable with distribution F_X, and

$$m_X(z) = E\big[\exp\big(z^T X\big)\big] = E\left[\exp\left(\sum_{i=1}^{k} z_i X_i\right)\right] \quad \big(z \in \mathbb{C}^k\big)$$

its moment generating function.

Definition 3.14 The Appell polynomials

$$A^F_{j_1,\dots,j_k}(x_1, \dots, x_k) = A_{j_1,\dots,j_k}(x_1, \dots, x_k)$$

are defined as coefficients in the expansion

$$\frac{\exp(z^T x)}{m_X(z)} = \sum_{j_1,\dots,j_k=0}^{\infty} \frac{z_1^{j_1} \cdots z_k^{j_k}}{j_1! \cdots j_k!} A_{j_1,\dots,j_k}(x_1, \dots, x_k). \tag{3.72}$$

We then write $A_{j_1,\dots,j_k}(X_1, \dots, X_k) = A^F_{j_1,\dots,j_k}(x_1, \dots, x_k)|_{x=X}$.

It is often convenient to write Appell polynomials in terms of so-called Wick products:

Definition 3.15 The Wick product of $X = (X_1, \dots, X_k)^T \sim F$ is defined by

$$:x_1, \dots, x_k:^F = A^F_{1,\dots,1}(x_1, \dots, x_k) = \frac{\partial^k}{\partial z_1 \cdots \partial z_k} \frac{\exp(z' x)}{m_X(z)}.$$

We then write

$$:X_1, \dots, X_k: = :x_1, \dots, x_k:^F\big|_{x=X}. \tag{3.73}$$

Moreover, for the empty set, we define $:\emptyset: = 1$.

Since

$$A_{j_1,\dots,j_k}(x_1,\dots,x_k)=\frac{\partial^{j_1+\cdots+j_k}}{\partial z_1^{j_1}\cdots\partial z_k^{j_k}}\left[\frac{\exp(z'x)}{m_X(z)}\right]_{z=0},$$

we can express Appell polynomials in terms of Wick products by

$$\begin{aligned}A_{j_1,\dots,j_k}(X_1,\dots,X_k)&=:\underbrace{X_1,\dots,X_1}_{j_1},\dots,\underbrace{X_k,\dots,X_k}_{j_k}:\\&=:{X_1'}^{j_1},\dots,{X_k'}^{j_k}:.\end{aligned}$$

In particular, for $k=1$, $X_1=X$ and

$$A_j(X)=:\underbrace{X,\dots,X}_{j}:=:X'^{\,j}:$$

with the obvious notation $:X'^{\,j}:$. The Appell polynomial generating function can then also be written as

$$\frac{\exp(z^Tx)}{m_X(z)}=\sum_{j_1,\dots,j_k=0}^{\infty}\frac{z_1^{j_1}\cdots z_k^{j_k}}{j_1!\cdots j_k!}:{x_1'}^{j_1},\dots,{x_k'}^{j_k}:.$$

As in the univariate case, an equivalent recursive definition of multivariate Appell polynomials and of Wick products can be given as follows.

Definition 3.16 Let $x=(x_1,\dots,x_k)^T=(x_i)_{i=1,\dots,k}\in\mathbb{R}^k$, $j=(j_1,\dots,j_k)^T\in\mathbb{N}^k$, $x^{(i)}=(x_1,\dots,x_{i-1},x_{i+1},\dots,x_k)^T=(x_l)_{l\neq i}$, $j^{(i)}=(j_l)_{l\neq i}$ and $0=(0,\dots,0)^T$. Then the Appell polynomials associated with distribution F are defined by

$$\begin{aligned}A_0&\equiv 1,\\\frac{\partial}{\partial x_i}A_j(x)&=j_iA_{j^{(i)}}\big(x^{(i)}\big),\\E\big[A_j(X)\big]&=0\quad(j\neq 0).\end{aligned}\tag{3.74}$$

Moreover, the corresponding Wick products are defined by

$$\begin{aligned}:\emptyset:&=1,\\\frac{\partial}{\partial x_i}\big[:x:^F\big]&=:x^{(i)}:,\\E[:X:]&=0\quad(k\geq 1).\end{aligned}$$

Example 3.28 Assume that the components of X are independent. Then

$$A_{j_1,\dots,j_k}(x_1,\dots,x_k)=\prod_{l=1}^{k}A_{j_l}(x_l).$$

In particular, if $X_1, \ldots, X_k$ are standard normal, then $A_{j_1,\ldots,j_k}(x_1, \ldots, x_k)$ is the product of univariate Hermite polynomials.

Example 3.29 Bivariate Appell polynomials are defined by

$$\sum_{j_1,j_2=0}^{\infty} \frac{z_1^{j_1}}{j_1!}\frac{z_2^{j_2}}{j_2!} A_{j_1,j_2}(x_1, x_2) = \frac{\exp(z_1x_1 + z_2x_2)}{E[\exp(z_1X_1 + z_2X_2)]}.$$

If $E(X_1) = E(X_2) = 0$, then

$$A_{1,1}(X_1, X_2) = X_1X_2 - E(X_1X_2) = X_1X_2 - cov(X_1, X_2)$$

and

$$A_{1,2}(X_1, X_2) = X_1X_2^2 - X_1E\big(X_2^2\big) - 2X_2^2 cov(X_1, X_2) - cov\big(X_1, X_2^2\big).$$

For instance, if (X_1, X_2) are jointly normal with $X_i \sim N(0, 1)$, $i = 1, 2$, and correlation ρ, then

$$\begin{aligned} A_{1,1}(X_1, X_2) &= X_1X_2 - \rho, \\ A_{1,2}(X_1, X_2) &= X_1X_2^2 - X_1 - 2X_2^2\rho - E\big(X_1X_2^2\big) \\ &= X_1\big(X_2^2 - 1\big) - 2X_2^2\rho - E\big(X_1X_2^2\big). \end{aligned}$$

In particular, if $\rho = 0$, then

$$A_{1,2}(X_1, X_2) = X_1\big(X_2^2 - 1\big) = H_1(X_1)H_2(X_2) = H_{\mathbf{q}}(X),$$

where $\mathbf{q} = (1, 2)$, and $H_{\mathbf{q}}$ is the multivariate Hermite polynomial defined in (3.31).

3.4.2 Connection to Cumulants and Other Important Properties

Wick products are very useful in the context of limit theorems for long-memory processes. The main reason is the so-called diagram formula (or rather formulas), which simplifies the calculation of joint cumulants. First, in this section, some basic properties of cumulants, Wick products and Appell polynomials will be discussed. The diagram formula will be introduced in the next section.

First we recall some standard definitions.

Definition 3.17 The cumulant generating function of a random vector $X = (X_1, \ldots, X_k)^T$ is defined by

$$\kappa_X(t) = \log m_X(t) = \log E\big(e^{t^T X}\big),$$

provided that $m_X(t)$ is well defined in an open neighbourhood of the origin. More generally, without assuming existence of m_X or finite moments, κ_X is defined by

$$\kappa_X(t) = \log \varphi_X(t),$$

where

$$\varphi_X(t) = E\big(e^{it^T X}\big)$$

is the characteristic function of X.

If $m_X(t)$ exists in an open neighbourhood of 0, then $\kappa_X(z)$ can be written as a power series

$$\kappa_X(z) = \kappa_X(z_1, \ldots, z_k) = \sum_{j_1,\ldots,j_k=0}^{\infty} \frac{z_1^{j_1} \cdots z_k^{j_k}}{j_1! \cdots j_k!} \kappa_{j_1,\ldots,j_k},$$

and the coefficients

$$\kappa_{j_1,\ldots,j_k} = \kappa\big(X_1^{j_1}, \ldots, X_k^{j_k}\big) = \frac{\partial^{j_1+\cdots+j_k}}{\partial z^{j_1} \cdots \partial z^{j_k}} \big[\kappa_X(z)\big]_{z=0}$$

are called joint cumulants of $X_1^{j_1}, \ldots, X_k^{j_k}$. Similarly, when using the characteristic function, then

$$\kappa_{j_1,\ldots,j_k} = (-1)^{j_1+\cdots+j_k} \frac{\partial^{j_1+\cdots+j_k}}{\partial z^{j_1} \cdots \partial z^{j_k}} \big[\kappa_X(z)\big]_{z=0}.$$

Cumulants are more useful than moments when dealing with limit theorems. One reason is that, when based on the characteristic function, moments need not exist. A second reason is that κ is multilinear and independence is equivalent to all joint cumulants being zero:

Lemma 3.17 *Denote by π an arbitrary permutation of $1, 2, \ldots, k$. Then*

$$\kappa(X_1, \ldots, X_k) = \kappa(X_{\pi(1)}, \ldots, X_{\pi(k)})$$

and κ is multilinear, i.e. for

$$X_i = \sum_{j=1}^{m} c_{ij} Y_{ij},$$

we have

$$\kappa(X_1, \ldots, X_k) = \sum_{j_1,\ldots,j_k=0}^{m} c_{1j_1} \cdots c_{kj_k} \kappa(Y_{1j_1}, \ldots, Y_{kj_k}).$$

Moreover, if the random variables $\{X_i, i \in W_1\}$ *are independent of the r.v.* $\{X_i, i \in W_2\}$, *where* $W = W_1 \cup W_2 = \{1, 2, \ldots, k\}$, *then*

$$\kappa(X_1, \ldots, X_k) = 0.$$

The importance of Wick products is due to their direct relationship to cumulants:

Theorem 3.6 *Let* $W = \{1, 2, \ldots, k\}$ *and* $X = (X_i)_{i \in W}$. *For* $V = \{i_1, \ldots, i_l\} \subseteq W$ *define*

$$X^V = \prod_{j=1}^{l} X_{i_j}, :X^V: = :X_{i_1}, \ldots, X_{i_l}:$$

and

$$\kappa_V = \kappa\big(X^V\big) = \kappa(X_{i_1}, \ldots, X_{i_l}) = \frac{\partial^l}{\partial z_1 \cdots \partial z_l} \log E\left[\exp\left(\sum_{j=1}^{l} z_j X_{i_j}\right)\right].$$

Then, for any $i \in W$,

$$:X^W: = \big(:X^{W\setminus\{i\}}:\big) \cdot X_i - \sum_{\substack{V \subseteq W \\ V \ni i}} \big(:X^{W\setminus V}:\big) \cdot \kappa\big(X^V\big). \tag{3.75}$$

Proof Without loss of generality, let $i = k$. Then

$$\begin{aligned} :X^W: &= \left\{ \frac{\partial^{k-1}}{\partial z_1 \cdots \partial z_{k-1}} \left[\frac{\partial}{\partial z_k} \frac{\exp(z^T x)}{m_X(z)} \right] \right\}_{z=0} \\ &= \left\{ \frac{\partial^{k-1}}{\partial z_1 \cdots \partial z_{k-1}} \left[\frac{\exp(z^T x)}{m_X(z)} x_k - \frac{\exp(z^T x)}{m_X(z)} \frac{\partial}{\partial z_k} m_X(z) \right] \right\}_{z=0}, \end{aligned}$$

which is equal to

$$(:x_1, \ldots, x_{k-1}:) \cdot x_k - \left\{ \frac{\partial^{k-1}}{\partial z_1 \cdots \partial z_{k-1}} \left[\frac{\exp(z^T x)}{m_X(z)} \frac{\partial}{\partial z_k} m_X(z) \right] \right\}_{z=0}.$$

The result then follows by noting that

$$\left[\frac{\partial^{l+1}}{\partial z_{i_1} \cdots \partial z_{i_l} \partial z_k} \kappa_{X^W}(z) \right]_{z=0} = \kappa_V$$

with $V = \{i_1, \ldots, i_l, k\}$ and applying the product rule. □

Example 3.30 Let $(X_1, X_2) \sim N(0, \Sigma)$ with $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = \rho$. Then

$$m_X(z) = \exp\left[\frac{1}{2}\left(z_1^2 + z_2^2\right) + \rho z_1 z_2\right],$$

$$\frac{\exp(x_1 z_1 + x_2 z_2)}{m_X(z)} = \exp\left[x_1 z_1 + x_2 z_2 - \rho z_1 z_2 - \frac{1}{2}\left(z_1^2 + z_2^2\right)\right],$$

$$A_1(x_j) = :x_j: = \frac{\partial}{\partial z_j}\left[\frac{\exp(x_1 z_1 + x_2 z_2)}{m_X(z)}\right]_{z=0} = x_j,$$

$$A_{1,1}(x) = :x_1 x_2: = \frac{\partial^2}{\partial z_1 \partial z_2}\left[\frac{\exp(x_1 z_1 + x_2 z_2)}{m_X(z)}\right]_{z=0} = x_1 x_2 - \rho,$$

$$A_{1,2}(x) = :x_1\left(x_2'\right)^2: = \frac{\partial^3}{\partial z_1 \partial^2 z_2}\left[\frac{\exp(x_1 z_1 + x_2 z_2)}{m_X(z)}\right]_{z=0}$$

$$= x_1\left(x_2^2 - 1\right) - 2\rho x_2.$$

Now, consider $W = \{1, 2\}$ and $i = 1$. Then $: X^W := X_1 X_2 - \rho$ and $:X^{W\setminus\{i\}}: = :X_2: = X_2$. On the other hand, using formula (3.75), we have $V = \{1\}$ and $\{1, 2\}$ respectively and

$$\sum_{\substack{V \subseteq W \\ V \ni i}} \left(:X^{W\setminus V}:\right) \cdot \kappa\left(X^V\right) = (:X_2:)\kappa(X_1) + (:\emptyset:)\kappa(X_1, X_2)$$

$$= 0 + \rho = \rho.$$

Thus,

$$:X^W: = \left(:X^2:\right) \cdot X_1 - \sum_{\substack{V \subseteq W \\ V \ni i}} \left(:X^{W\setminus V}:\right) \cdot \kappa\left(X^V\right) = X_1 X_2 - \rho.$$

A further important property of Wick products is that they factorize under independence:

Theorem 3.7 *Let $X_1, \ldots, X_k$ be independent and define*

$$X_i'^{\,j} = (\underbrace{X_i, \ldots, X_i}_{j}).$$

Then, for any $j_1, \ldots, j_k$,

$$:X_1'^{\,j_1}, X_2'^{\,j_2}, \ldots, X_k'^{\,j_k}: = \prod_{i=1}^{k} :X_i^{j_i}:.$$

Proof Let $X^{(1)} = (X_i)_{i\in W_1}$ be independent of $X_2 = (X_i)_{i\in W_2}$ and $W = W_1 \cup W_2$. Then

$$\frac{\exp(\sum_{i\in W} x_i z_i)}{m_X(z)} = \frac{\exp(\sum_{i\in W_1} x_i z_i)}{m_{X^{(1)}}(z)} \frac{\exp(\sum_{i\in W_2} x_i z_i)}{m_{X^{(2)}}(z)}.$$

Since each of the terms can be written as a power series in $z_1, \ldots, z_k$, the result then follows by comparing the coefficients of $z_1^{j_1} \cdots z_k^{j_k}/(j_1! \cdots j_2!)$. □

This result simplifies the calculation of Appell polynomials for sums of independent random variables:

Theorem 3.8 *Let X, Y be independent, $X \sim F_X$, $Y \sim F_Y$, $X + Y \sim F_{X+Y} = F_X * F_Y$. Then*

$$A_j^{F_{X+Y}}(X+Y) = \sum_{k=0}^{j} \binom{j}{k} A_k^{F_X}(X) A_{j-k}^{F_Y}(Y). \tag{3.76}$$

Proof The result follows from multilinearity of Wick products,

$$\begin{aligned} A_j^{F_{X+Y}}(X+Y) &= :X+Y, \ldots, X+Y: \\ &= :X, X+Y, \ldots, X+Y: + :Y, X+Y, \ldots, X+Y: \\ &= \cdots = \sum_{k=0}^{j} \binom{j}{k} \left(:X'^k:\right) \cdot \left(:Y'^{(j-k)}:\right). \end{aligned}$$

□

Now we come back to linear processes

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j},$$

where ε_t ($t \in \mathbb{N}$) are i.i.d. zero mean random variables. Since linear processes are sums of independent random variables, Theorem 3.8 can be extended to calculate Appell polynomials for such processes. The following result is due to Avram and Taqqu (1987).

Theorem 3.9 *Let $X_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j}$, $\sum a_j^2 < \infty$, $E(\varepsilon_1) = 0$, $E(\varepsilon_1^{2M}) < \infty$ for some $0 < M < \infty$. Then, for $j \le M$,*

$$A_j^{F_X}(X) = \sum_{p\in I} \frac{j!}{p_1! \cdots p_l!} \sum_{i\in J_l} \prod_{r=1}^{l} \left[a_{i_r}^{p_r} A_{p_r}^{F_\varepsilon}(\varepsilon_{i_r})\right], \tag{3.77}$$

where

$$I = \left\{p = (p_1, \ldots, p_l) \in \mathbb{N}^l : 1 \le l \le j, 1 \le p_1 \le \cdots \le p_l \le j, p_1 + \cdots + p_l = j\right\},$$
$$J_l = \left\{i = (i_1, \ldots, i_l) \in \mathbb{N}^l : i_r \ne i_s \ (r \ne s), i_j < i_{j+1} \textit{ for } p_j = p_{j+1}\right\}.$$

Proof The idea is to apply (3.76) to the truncated sum

$$X_{t,N} = \sum_{j=1}^{N} a_j \varepsilon_{t-j}$$

and to carry it over to the limit (as $N \to \infty$) by using a martingale convergence theorem. For finite N, we have due to multilinearity of the Wick product and independence of ξ_i,

$$\begin{aligned} A_j^{F_N}(X_N) &= :\underbrace{X_N, \ldots, X_N}_{j}: \\ &= \sum_{p \in I} \frac{j!}{p_1! \cdots p_l!} \sum_{i \in J_l} \left(\prod_{r=1}^{l} a_{i_r}^{p_r} \right) :\underbrace{\varepsilon_{i_1}, \ldots, \varepsilon_{i_1}}_{p_1}, \ldots, \underbrace{\varepsilon_{i_l}, \ldots, \varepsilon_{i_l}}_{p_l}: \\ &= \sum_{p \in I} \frac{j!}{p_1! \cdots p_l!} \sum_{i \in J_l} \left(\prod_{r=1}^{l} a_{i_r}^{p_r} \right) \prod_{r=1}^{l} :\underbrace{\varepsilon_{i_r}, \ldots, \varepsilon_{i_r}}_{p_r}:. \end{aligned}$$

To carry this over to $N \to \infty$, one can show that X_N ($N \in \mathbb{N}$) is a martingale and for all even $k \le 2M$, $E(X_N^k) \le \text{const} \cdot E(\varepsilon_1^k)$. This then implies the $L^k(\Omega)$-convergence of X_N and almost sure convergence of $A_j^{F_N}(X_N)$ and the sum above. □

3.4.3 Diagram Formulas

Diagram formulas provide a combinatorial simplification of joint moments and cumulants. This is very useful in the context of limit theorems where one would like to show that certain terms dominate others (see Sect. 4.2.5).

Before writing down diagram formulas, the following definitions and notations are needed. We will denote by W a table with k rows $W_1, \ldots, W_k$ that may be of different length. Thus, denoting just the position in the table, we have $W_j = \{(j,1), \ldots, (j,m_j)\}$ $(1 \le j \le k)$, where m_j is the length of row j. Considered as a set of "positions", the table W can be written as $W = \bigcup_{j=1}^{k} W_j$. The rows $W_1, \ldots, W_k$ define a specific partition of the set W (i.e. a complete decomposition into disjoint sets). More generally, we consider arbitrary partitions $V_1, \ldots, V_r$, $W = \bigcup_{j=1}^{r} V_j$, $V_i \cap V_j = \emptyset$ $(i \ne j)$.

Definition 3.18 Each partition of a table W is called a diagram (or graph) and is denoted by

$$(V)_r = (V_1, \ldots, V_r) = \gamma.$$

Each V_j is called an edge of γ. The set of all diagrams on W is denoted by

$$\Gamma_W = \{\gamma : \gamma = (V)_r = \text{partition of } W\}.$$

An important characteristic of diagrams is whether they can be partitioned according to their association to rows:

Definition 3.19 A diagram $\gamma \in \Gamma_W$ is called connected if there are no two sets $K_1,\ K_2 \neq \emptyset$ such that $K_1 \cup K_2 = \{1, \ldots, k\}$ and for each $V_j \in \gamma$, one has either $V_j \subseteq \bigcup_{i \in K_1} W_i$ or $V_j \subseteq \bigcup_{i \in K_2} W_i$.

In other words, for a connected diagram, it is not possible to separate the rows into two groups such that some of the V_js are in the first set of rows and the other V_js are in the other ones. The set of connected diagrams is denoted by Γ_W^c.

For normal distributions, all cumulants higher than 2 are zero. This is the reason for the following definition.

Definition 3.20 A diagram $\gamma = (V)_r$ with $|V_1| = \cdots = |V_r| = 2$ is called normal or Gaussian. The set of all normal diagrams is denoted by $\Gamma_W^{\mathcal{N}}$.

Furthermore, we distinguish edges that are in one row only:

Definition 3.21 If $V_j \subseteq W_i$ for some i, then V_j is called a flat edge. The set of all graphs with no flat edges is denoted by $\Gamma_W^{\neq}$.

Combining the notations, we also have

$$\Gamma_W^{\neq,c} = \Gamma_W^{\neq} \cap \Gamma_W^c$$

for connected diagrams with no flat edges and

$$\Gamma_W^{\neq,\mathcal{N}} = \Gamma_W^{\neq} \cap \Gamma_W^{\mathcal{N}}$$

for normal diagrams with no flat edges.

Example 3.31 Some examples of diagrams are displayed in Figs. 3.1, 3.2, 3.3 and 3.4.

The following fundamental diagram formulas can be derived by induction (Giraitis and Surgailis 1986; also see Malyshev and Minlos 1991). The detailed proof is rather involved is therefore omitted here. For index sets $A = \{i_1, \ldots, i_l\}$, we use the notation

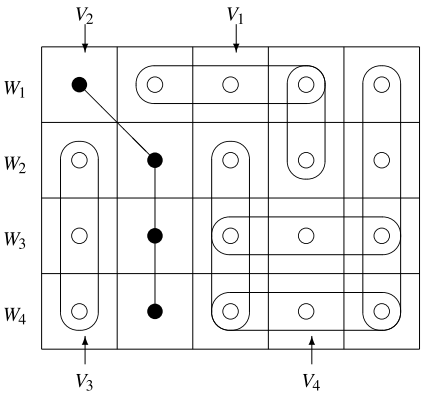

Fig. 3.1 Diagram $\gamma = (V_1, V_2, V_3, V_4) \in \Gamma_W^{\not\perp,c} \setminus \Gamma_W^{\not\sim}$

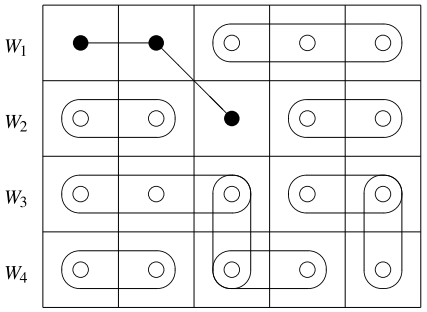

Fig. 3.2 Diagram $\gamma = (V_1, V_2, V_3, V_4, V_5, V_6) \in \Gamma_W \setminus (\Gamma_W^c \cup \Gamma_W^{\not\perp})$

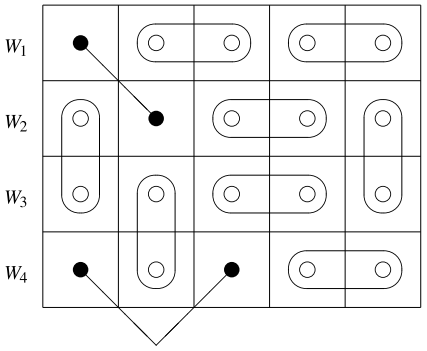

Fig. 3.3 Diagram $\gamma = (V_1, \ldots, V_{10}) \in \Gamma_W^c \cap \Gamma_W^{\not\sim}$

$$X^A = \prod_{i \in A} X_i, :X^A: = :X_{i_1}, \ldots, X_{i_l}:$$

and

$$X'^A = (X_{i_1}, \ldots, X_{i_l}).$$

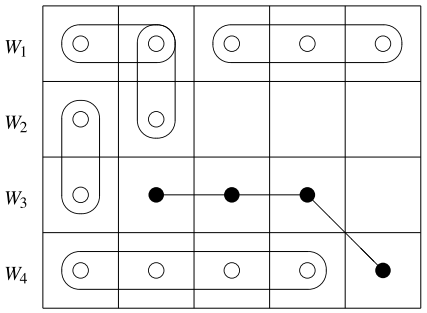

Fig. 3.4 Diagram $\gamma = (V_1, V_2, V_3, V_4, V_5) \in \Gamma_W \setminus (\Gamma_W^{\not-} \cap \Gamma_W^{\not\sim})$

Theorem 3.10 *The following holds for any table* $W = \bigcup_{j=1}^{k} W_j$:

$$E\left(\prod_{j=1}^{k} X^{W_j}\right) = \sum_{\gamma=(V)_r \in \Gamma_W} \kappa\left(X'^{V_1}\right) \cdots \kappa\left(X'^{V_r}\right), \tag{3.78}$$

$$\kappa\left(X^{W_1}, \ldots, X^{W_k}\right) = \sum_{\gamma \in \Gamma_W^c} \kappa\left(X'^{V_1}\right) \cdots \kappa\left(X'^{V_r}\right), \tag{3.79}$$

$$E\left(\prod_{j=1}^{k} :X^{W_j}:\right) = \sum_{\gamma=(V)_r \in \Gamma_W^{\not-}} \kappa\left(X'^{V_1}\right) \cdots \kappa\left(X'^{V_r}\right), \tag{3.80}$$

$$\kappa\left(:X^{W_1}:, \ldots, :X^{W_k}:\right) = \sum_{\gamma \in \Gamma_W^{\not-,c}} \kappa\left(X'^{V_1}\right) \cdots \kappa\left(X'^{V_r}\right). \tag{3.81}$$

In particular, remarkable is that for the Wick product, flat edges are removed completely. For the joint cumulant, the sum even reduces to connected graphs without flat edges. As a side product, we obtain another proof that Wick products have zero expectation: Defining a table with one row, $W = W_1 = \{1, \ldots, k\} \neq \emptyset$, no non-flat edges exist, so that (3.80) implies $E(:X^W:) = 0$.

The following examples illustrate some direct applications of Theorem 3.10.

Example 3.32 Consider $W = W_1 \cup W_2$ with $W_1 = \{1, 2\}$, $W_2 = \{1\}$. We associate the positions in the table with random variables, namely $(1, 1) \longleftrightarrow X_1$, $(1, 2) \longleftrightarrow X_2$ and $(2, 1) \longleftrightarrow X_3$. For simplicity of notation, we will write the random variables instead of the positions. Then the set of connected graphs is $\Gamma_W^c = \{\gamma_1, \gamma_2, \gamma_3\}$ with $\gamma_1 = (V)_1 = \{\{X_1, X_2, X_3\}\}$, $\gamma_2 = (V)_2 = \{\{X_1, X_3\}, \{X_2\}\}$ and $\gamma_3 = (V)_2 = \{\{X_1\}, \{X_2, X_3\}\}$. Therefore,

$$\begin{aligned}\kappa\left(X^{W_1}, X^{W_2}\right) &= \kappa(X_1 \cdot X_2, X_3) \\ &= \sum_{\gamma \in \Gamma_W^c} \kappa\left(X'^{V_1}\right) \cdots \kappa\left(X'^{V_r}\right)\end{aligned}$$

$$= \kappa(X_1, X_2, X_3) + \kappa(X_1, X_3)\kappa(X_2) + \kappa(X_1)\kappa(X_2, X_3).$$

Similarly, $\Gamma_W^{\not=} = \{\gamma_1\}$ with $\gamma_1 = (V)_1 = \{V_1\}$ and $\Gamma_W^{\not=} = \Gamma_W^{\not=,c}$. Therefore,

$$E\left[\prod_{j=1}^{k} :X^{W_j}:\right] = E\big[(:X_1, X_2:)(:X_3:)\big]$$

$$= \sum_{\gamma \in \Gamma_W^{\not=,c}} \kappa\big(X'^{V_1}\big) \cdots \kappa\big(X'^{V_r}\big) = \kappa\big(X'^{V_1}\big) = \kappa(X_1, X_2, X_3).$$

Example 3.33 Let X_1, X_2, X_3 be jointly normal with $E(X_i) = \mu_i$, $\mathrm{var}(X_i) = 1$ and $cov(X_i, X_j) = \rho_{ij} = \rho_{ji}$. Then $\kappa(X_1, X_2, X_3) = 0$, $\kappa(X_i, X_j) = \rho_{ij}$ and $\kappa(X_i) = \mu_i$. From the previous example we then conclude that

$$\kappa(X_1 X_2, X_3) = \mu_2 \rho_{13} + \mu_1 \rho_{23}$$

and

$$E\big[(:X_1, X_2:)(:X_3:)\big] = \kappa(X_1, X_2, X_3) = 0.$$

Note also that $:X_1X_2: = X_1X_2 - \rho_{12}$ and $:X_3: = X_3$, so that this means

$$E\big[(X_1X_2 - \rho_{12})X_3\big] = 0.$$

For the ordinary product, we obtain

$$E[X_1X_2X_3] = \sum_{\gamma \in \Gamma_W^c} + \sum_{\gamma \in \Gamma_W \setminus \Gamma_W^c} = \sum_{\gamma \in \Gamma_W \setminus \Gamma_W^c}$$

$$= \kappa(X_1)\kappa(X_2)\kappa(X_3) + \kappa(X_1, X_2)\kappa(X_3) + \kappa(X_1, X_3)\kappa(X_2)$$

$$= \mu_1\mu_2\mu_3 + \mu_3\rho_{12} + \mu_2\rho_{13}.$$

Example 3.34 Let X be Poisson distributed with $E(X) = \lambda = 1$. We would like to calculate the variance of the corresponding Appell polynomials, $\mathrm{var}(A_j^F(X))$. Since $m_X(z) = \exp(e^z - 1)$, we have

$$\kappa_X(z) = e^z - 1 = \sum_{j=1}^{\infty} \frac{z^j}{j!},$$

so that

$$\kappa_j = \kappa(\underbrace{X, \ldots, X}_{j}) = 1 \quad (j \geq 1).$$

Now define the table $W = W_1 \cup W_2$ with $W_1 = \{1, \ldots, j\}$, $W_2 = \{j+1, \ldots, 2j\}$ and associate each position in the table with X. Then (3.80) implies

$$E\big(A_j^2(X)\big) = E\big[(:X^{W_1}:)(:X^{W_2}:)\big] = \sum_{\gamma \in \Gamma_W^{\neq}} \underbrace{\kappa\big(X'^{V_1}\big)\cdots\kappa\big(X'^{V_r}\big)}_{=1} = N_W^{\neq},$$

where $N_W^{\neq}$ is the number of diagrams without flat edges. Thus, the task of calculating the variance of A_j is reduced to the combinatorial question of counting the number of elements in $|\Gamma_W^{\neq}|$.

For Gaussian random variables, Theorem 3.10 leads to simplified formulas for joint moments and cumulants of Hermite polynomials where only correlations occur:

Corollary 3.5 *Let $X_1, \ldots, X_k$ be jointly normal with $E(X_i) = 0$, $\mathrm{var}(X_i) = 1$ and $\rho_{ij} = E(X_i X_j)$. For given integers $m_j \geq 1$, define a table W with k rows W_j of length m_j and the positions in row j associated with X_j (i.e. W_j corresponding to $(X_j, \ldots, X_j)$). Then*

$$E\left[\prod_{j=1}^{k} H_{m_j}(X_j)\right] = \sum_{\gamma \in \Gamma_W^{\neq,\mathcal{N}}} \prod_{1 \leq i < j \leq k} \rho_{ij}^{l_{ij}}, \tag{3.82}$$

and

$$\kappa\big(H_{m_1}(X_1), \ldots, H_{m_k}(X_k)\big) = \sum_{\gamma \in \Gamma_W^{\neq,c,\mathcal{N}}} \prod_{1 \leq i < j \leq k} \rho_{ij}^{l_{ij}}, \tag{3.83}$$

where, for each γ, $l_{ij} = l_{ij}(\gamma)$ is the number of edges between rows W_i and W_j.

Proof Theorem 3.10 implies

$$E\left[\prod_{j=1}^{k} H_{m_j}(Y_j)\right] = E\left[\prod_{j=1}^{k} :\underbrace{Y_j, \ldots, Y_j}_{m_j}:\right] = \sum_{\gamma \in \Gamma_W^{\neq}} \kappa\big(X'^{V_1}\big)\cdots\kappa\big(X'^{V_r}\big).$$

Since for jointly Gaussian variables all higher-order cumulants are zero, the sum reduces to

$$\sum_{\gamma \in \Gamma_W^{\neq,\mathcal{N}}} \kappa\big(X'^{V_1}\big)\cdots\kappa\big(X'^{V_r}\big).$$

Since for each pair X_i, X_j $(i \neq j)$, $\kappa(X_i, X_j) = \rho_{ij}$, the result follows by counting the number of pairs connecting each pair of rows. □

This result can be used to derive covariances for Hermite polynomial transformations of for stationary processes:

Corollary 3.6 *Let X_t $(t \in \mathbb{Z})$ be a stationary Gaussian process with $E(X_t) = 0$, $\mathrm{var}(X_t) = 1$ and $\rho(k) = corr(X_t, X_{t+k})$. Then*

$$cov\big(H_j(X_t), H_l(X_{t+k})\big) = j!\rho^j(k) \cdot \delta_{j,l},$$

and

$$\kappa\big(H_{n_1}(X_1), \ldots, H_{n_k}(X_k)\big) = \sum_{\gamma \in \Gamma_W^{\neq, c, \mathcal{N}}} \prod_{1 \leq i < j \leq k} \rho_{ij}^{l_{ij}}.$$

Proof Suppose that $j \neq l$. Then, for each $\gamma \in \Gamma_W^{\neq}$, there exists an edge V_i with more than two elements. Therefore, $\Gamma_W^{\neq, \mathcal{N}} = \emptyset$, and the covariance is zero due to formula (3.82). For $j = l$, the result is obtained from (3.82), $\mathrm{var}(H_j) = j!$ and the fact that, when W consists of two rows of length j, then the number of elements in a diagram $\gamma \in \Gamma_W^{\mathcal{N}}$ is equal to j. □

3.5 Wavelets

3.5.1 The Continuous Wavelet Transform (CWT)

In this section we discuss basic properties of wavelet functions. First, theoretical results on the so-called continuous wavelet decomposition appeared in the early 1980s (Morlet et al. 1982; Grossmann and Morlet 1985). A classical monograph on the topic is Daubechies (1992). Statistical applications of wavelets were mainly initiated by a series of papers by Donoho, Johnstone and others (Donoho and Johnstone 1994, 1995, 1997; Donoho et al. 1995). Also see Brillinger (1994, 1996), Hall and Patil (1996a, 1996b), Johnstone (1999), Johnstone and Silverman (1997) and Abramovich et al. (1998). In time series analysis, the main applications include nonparametric trend estimation (see Sect. 7.5), spectral estimation and estimation of the long-memory parameter d (see Sect. 5.7). A more detailed discussion of mathematical properties of wavelets can be found for instance in Mallat (1989), Strang (1989), Daubechies (1992), Antoniadis and Oppenheim (1995), Cohen and Ryan (1995), Neumann and von Sachs (1995), Härdle et al. (1998), Steeb (1998), Percival and Walden (2000), Pinsky (2002), Vidakovic (1999) and references therein. Much earlier references are also for instance Haar (1910) and Gabor (1946).

Loosely speaking, a wavelet is a square-integrable function $\psi : \mathbb{R} \to \mathbb{R}$ such that

$$\int_{-\infty}^{\infty} \psi(x)\,dx = 0. \tag{3.84}$$

This condition means that ψ must have an oscillatory behaviour. This, together with rescaling (see below) and the fact that often ψ is also assumed to have a compact support, justifies the name *wavelet*. It is also convenient to assume (without loss of generality) that

$$\|\psi\|^2 = \int_{-\infty}^{\infty} \big|\psi(x)\big|^2\,dx = 1. \tag{3.85}$$

An important property of wavelets is the number of vanishing moments. If

$$\int x^k \psi(x)\,dx = 0 \quad (0 \le k \le M-1), \qquad \int x^M \psi(x)\,dx \neq 0, \tag{3.86}$$

then we say that ψ has M vanishing moments. If we denote by

$$\hat{\psi}(\lambda) = \int_{-\infty}^{\infty} \psi(x) e^{-i\lambda x}\,dx \tag{3.87}$$

the Fourier transform of ψ, then

$$\int x^k \psi(x)\,dx = i^k \frac{d^k \hat{\psi}(\lambda)}{d\lambda^k}\bigg|_{\lambda=0}. \tag{3.88}$$

Therefore, if ψ has M vanishing moments, then $\hat{\psi}^{(k)}(0) = 0$ for $k = 0, \dots, M-1$ and $\hat{\psi}^{(M)}(0) \neq 0$. Hence, the Taylor expansion at the origin yields, as $\lambda \to 0$,

$$\big|\hat{\psi}(\lambda)\big| = \frac{|\hat{\psi}^{(M)}(0)|}{M!}|\lambda|^M + o\big(|\lambda|^M\big).$$

Example 3.35 The most basic example of a wavelet is the a Haar wavelet

$$\psi(x) = 1\left\{x \in \left(0, \frac{1}{2}\right]\right\} - 1\left\{x \in \left(\frac{1}{2}, 1\right)\right\}.$$

It has one vanishing moment since $\int \psi(x)\,dx = \int_0^{\frac{1}{2}} dx - \int_{\frac{1}{2}}^1 dx = 0$. Note also that $\int \psi^2(x)\,dx = 1$.

Example 3.36 A typical wavelet with infinite support is the Mexican hat wavelet defined as the second derivative of the standard normal density,

$$\psi(x) = \frac{d^2}{dx^2}\frac{1}{\sqrt{2\pi}}\exp\big(-x^2/2\big).$$

It has two vanishing moments.

The idea of wavelets is obtain a representation of square-integrable functions $g \in L^2(\mathbb{R})$ in terms of local functions ("wavelets", little waves) with different frequencies. This is similar to a Fourier series representation. However, in contrast to the Fourier series representation with fixed global sine and cosine functions, we not only have a decomposition in terms of frequencies, but also a local representation ("localization") that highlights local features of the function. Thus, consider the Hilbert space $L^2(\mathbb{R})$ of measurable complex-valued functions on $\mathbb{R}$ equipped with the scalar product

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)\overline{g(x)}\,dx$$

(for $f, g \in L^2(\mathbb{R})$) and the corresponding norm $\|g\| = \sqrt{\langle g, g\rangle}$. The wavelet function has to satisfy the admissibility condition

$$0 < C_\psi = 2\pi \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\lambda)|^2}{|\lambda|}\,d\lambda < \infty. \tag{3.89}$$

Note that (3.89) is stronger than condition (3.84) because a necessary condition for (3.84) is $\hat{\psi}(0) = \int \psi(\lambda)\,d\lambda = 0$. To obtain a decomposition of functions $g \in L^2(\mathbb{R})$ into "wavelets", one defines an infinite number of shifted and rescaled versions of ψ by

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right) \quad (a, b \in \mathbb{R},\ a \neq 0).$$

The scaling factor a is called dilation parameter, and b is the translation parameter. Note that $\psi_{1,0} = \psi$. Due to the factor $|a|^{-\frac{1}{2}}$ the L^2-norm of all wavelets $\psi_{a,b}(x)$ is the same. Usually one takes a ψ-function with $\|\psi\|^2 = \int |\psi(x)|^2\,dx = 1$, so that we have, for all a, b,

$$\|\psi_{a,b}\|^2 = \int \left|\psi_{a,b}(x)\right|^2 dx = 1.$$

Now we would like to express a function $g \in L^2(\mathbb{R})$ in terms of the wavelet functions $\psi_{a,b}(x)$ $(a, b \in \mathbb{R},\ a \neq 0)$. This leads to the definition of the continuous wavelet transform (CWT)

$$g \to T_g(a, b) \quad (a, b \in \mathbb{R}, a \neq 0),$$

where

$$T_g(a, b) = \int_{-\infty}^{\infty} g(x)\overline{\psi_{a,b}(x)}\,dx \tag{3.90}$$

(i.e. for a given function g, $T_g : (\mathbb{R} \setminus \{0\}) \times \mathbb{R} \to \mathbb{C}$ is a function of a, b). Note that

$$T_{f-g}(a, b) = T_f(a, b) - T_g(a, b). \tag{3.91}$$

It can be shown that (see Daubechies 1992, Proposition 2.4.1, p. 24)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a^{-2}T_g(a,b)\overline{T_f(a,b)}\,da\,db = C_\psi\langle g, f\rangle.$$

In particular,

$$C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a^{-2}\big|T_g(a,b)\big|^2\,da\,db = \|g\|^2,$$

so that, together with (3.91), we have

$$T_f(a,b) = T_g(a,b) \quad\Longleftrightarrow\quad f = g \quad \big(\text{in}L^2(\mathbb{R})\big).$$

This implies that g can be reconstructed perfectly from T_g. Note that heuristically, for $x \neq y$,

$$\begin{aligned}
\breve{\psi}(x-y) &= C_\psi^{-1}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} a^{-2}\psi_{a,b}(x)\psi_{a,b}(y)\,da\right)db\\
&= C_\psi^{-1}\int a^{-2}\big\langle\psi(\cdot),\psi\big(\cdot+(y-x)/a\big)\big\rangle\,da\\
&= -C_\psi^{-1}\int\big\langle\psi(\cdot),\psi\big(\cdot+(y-x)v\big)\big\rangle\,dv\\
&= -C_\psi^{-1}\left\langle\psi(\cdot),\int\psi\big(\cdot+(y-x)v\big)\,dv\right\rangle = 0,
\end{aligned}$$

where the last equality follows from (3.84). For $x = y$, the integral $C_\psi^{-1}\int a^{-2}\,da = \infty$, but $\breve{\psi}(z)$ $(z \in \mathbb{R})$ can be understood as a generalized function that is identical with the Dirac function $\delta(z)$, i.e. for well-behaved functions, we have

$$\int f(v)\breve{\psi}(v)\,dv = \int f(v)\delta(v)\,dv = f(0).$$

A concrete formula for the reconstruction of g is therefore obtained by

$$\begin{aligned}
&C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a^{-2}T_g(a,b)\psi_{a,b}(x)\,da\,db\\
&\quad= \int_{-\infty}^{\infty} g(y)\left[C_\psi^{-1}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} a^{-2}\psi_{a,b}(y)\psi_{a,b}(x)\,da\right)db\right]dy\\
&\quad= \int_{-\infty}^{\infty} g(y)\delta(y-x)\,dy = g(x).
\end{aligned}$$

3.5.2 The Discrete Wavelet Transform (DWT)

The one-to-one mapping $g \to T_g(a,b)$ transforms functions of one variable to functions of two variables, but it is obviously not a parsimonious representation of g. It is in fact possible to reduce a and b to a countable set. The coarsest discretization that is possible without loss of information is a diadic one, i.e.

$$a \in \left\{2^{-j}, j \in \mathbb{Z}\right\}, \qquad b \in \left\{k2^{-j}, j,k \in \mathbb{Z}\right\}. \tag{3.92}$$

For other admissible choices, see e.g. Daubechies (1992), Heil and Walnut (1989).

The next aim is to obtain a nice orthonormal countable basis, with diadic dilation and translation parameters. An elegant approach is the so-called multiresolution analysis initiated by Mallat. One starts with a function $\phi \in L^2(\mathbb{R})$ such that the set of dilated functions $\{\phi_{0k}, k \in \mathbb{Z} : \phi_{0k}(x) = \sqrt{N}\phi(Nx-k)\}$ (with N a positive integer) is an orthonormal system. Usually, one standardizes ϕ so that

$$\int \phi(x)\,dx = 1, \qquad \int \phi^2(x)\,dx = 1. \tag{3.93}$$

For simplicity of presentation, we will use the most frequently used value of $N = 1$ in the following. Denote by V_0 all functions in $L^2(\mathbb{R})$ that can be represented as a linear combination of ϕ_{0k} $(k \in \mathbb{Z})$. Since ϕ_{0k} are orthonormal, each function $g \in V_0$ has a unique representation $g(x) = \sum_{k=-\infty}^{\infty} \alpha_k \phi_{0k}(x)$ with $\alpha_k = \langle g, \phi_{0k}\rangle$ and $\|g\|^2 = \sum \alpha_k^2$. To obtain a basis in $L^2(\mathbb{R})$, one then defines

$$\phi_{jk}(x) = 2^{\frac{j}{2}}\phi\left(2^j x - k\right) \quad (j,k \in \mathbb{Z}).$$

Note that it is sufficient to keep the translation parameter at the same scale (here with steps of size one for example). For each j, we obtain a different subspace V_j generated by (possibly infinite) linear combinations of $\phi_{jk}(x)$ $(k \in \mathbb{Z})$. It can be written as

$$\begin{aligned} V_j &= \operatorname{span}\left\{2^{j/2}\phi\left(2^j \cdot -k\right), k \in \mathbb{Z}\right\} \\ &= \operatorname{span}\left\{\phi_{jk}(\cdot), k \in \mathbb{Z}\right\} \\ &= \left\{g \in L^2(\mathbb{R}) : g(x) = h\left(2^j\right), h \in V_0\right\}. \end{aligned}$$

In each V_j the functions ϕ_{jk} $(k \in \mathbb{Z})$ build an orthonormal basis. In order that we can represent all functions in $L^2(\mathbb{R})$, we need to make sure that the L^2-closure of the union of all these sets is equal to $L^2(\mathbb{R})$. In other words, ϕ has to be such that

$$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}). \tag{3.94}$$

Furthermore, in order that V_j at different dilation levels (resolution levels) are sufficiently different, one likes to have

$$\bigcap_{j\in\mathbb{Z}} V_j = \{0\}. \tag{3.95}$$

This leads to the following definition.

Definition 3.22 Let ϕ be such that $\cdots V_{-2} \subseteq V_{-1} \subseteq V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots$, and (3.94) and (3.95) hold. Then $\{V_j, j \in \mathbb{Z}\}$ is called a multiresolution analysis (MRA) of $L^2(\mathbb{R})$. The function ϕ is called a scaling function or father wavelet.

Example 3.37 The Haar scaling (or father) function is given by $\phi(x) = 1\{x \in [0,1]\}$. Then

$$\phi_{jk}(x) = 2^{j/2}\phi\big(2^j x - k\big) = 2^{j/2} 1\big\{x \in \big[2^{-j}k, 2^{-j}k + 2^{-j}\big]\big\}.$$

Thus, we are approximating L^2-functions by step functions. It is well known that step functions are dense in $L^2(\mathbb{R})$, so that (3.94) holds. Condition (3.95) is obvious. This means that ϕ is indeed a father wavelet. Note also that for $k \neq k'$, $\langle \phi_{jk}, \phi_{jk'} \rangle = 0$. However, not all functions in the system are orthogonal. For instance, $\langle \phi_{0k}, \phi_{1k} \rangle = \frac{1}{2}\sqrt{2}$.

As illustrated in the example, in general, the system of functions ϕ_{jk} $(j, k \in \mathbb{Z})$ is a basis in $L^2(\mathbb{R})$, but not an *orthogonal* one. In fact, since $V_0 \subseteq V_1$ and $\phi = \phi_{0,0} \in V_0$, we can write

$$\phi(x) = \phi_{00}(x) = \sum_k u_k \phi_{1k}(x) = 2^{\frac{1}{2}} \sum_k u_k \phi(2x - k) \tag{3.96}$$

with

$$u_k = \langle \phi_{00}, \phi_{1k} \rangle = \int_{\mathbb{R}} \phi(x)\phi_{1k}(x)\,dx = 2^{\frac{1}{2}} \int \phi(x)\phi(2x - k)\,dx.$$

The family u_k $(k \in \mathbb{Z})$ is called a low-pass filter.

To obtain an orthogonal basis, let W_j be the orthogonal complement of V_j in V_{j+1}, i.e.

$$W_j = V_{j+1} \ominus V_j = V_{j+1} \cap V_j^{\perp}.$$

Since, the sequence of sets V_j is nested, we can choose an arbitrary initial "resolution level" $J \in \mathbb{Z}$ and obtain the orthogonal decomposition

$$\bigcup_{j\in\mathbb{Z}} V_j = V_J \oplus W_J \oplus W_{J+1} \oplus \cdots = V_J \oplus \bigcup_{j=J}^{\infty} W_j.$$

For $L^2(\mathbb{R})$, we then have

$$L^2(\mathbb{R}) = \overline{V_J \oplus \bigcup_{j=J}^{\infty} W_j}.$$

Now we can choose an orthonormal basis in V_J given by ϕ_{Jk} $(k \in \mathbb{Z})$. Then we define for each W_j $(j \geq J)$ a corresponding basis consisting of functions ψ_{jk} $(k \in \mathbb{Z})$. This can be done as follows. For illustration, suppose for instance that $J = 0$, and let $\psi = \psi_{0,0}$ be a function such that $\psi_{0,k}(\cdot) := \psi(\cdot - k)$ $(k \in \mathbb{Z})$ is an orthonormal system in W_0. Since W_0 is orthogonal on V_0, ψ_{0k} $(k \in \mathbb{Z})$ and ϕ are orthogonal. Again, as in (3.96), since $V_1 = V_0 \cup W_0$, we can write $\psi(\cdot)$ as a linear combination of the base system from V_1,

$$\psi(x) = \sum_{k=-\infty}^{\infty} v_k \phi_{1,k}(x) = 2^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} v_k \phi(2x - k), \tag{3.97}$$

where

$$v_k = \langle \psi, \phi_{1k} \rangle = \int_{\mathbb{R}} \psi(x)\phi_{1k}(x)\, dx = 2^{\frac{1}{2}} \int \psi(x)\phi(2x - k)\, dx.$$

The family v_k $(k \in \mathbb{Z})$ is called a high-pass filter because we reach W_0 that is the higher-resolution part of $V_1 = V_0 \oplus W_0$, instead of the lower-resolution part V_0 where we would arrive via the coefficients u_k. The low- and high-pass filters are related by (see Vidakovic 1999, (3.34)):

$$v_k = (-1)^k u_{1-k}. \tag{3.98}$$

Since for each j, the functions

$$\psi_{j,k}(\cdot) = 2^{\frac{j}{2}} \psi\left(2^j \cdot - k\right) \quad (k \in \mathbb{Z})$$

form an orthonormal basis in W_j, where $V_{j+1} = V_j \oplus W_j$, we end up with an orthonormal basis in $L^2(\mathbb{R})$ consisting of ϕ_{0k} $(k \in \mathbb{Z})$ and ψ_{jk} $(j \geq 0, k \in \mathbb{Z})$. More generally, we may start with orthonormal basis functions $\phi_{J,k}$ $(k \in \mathbb{Z})$ in V_J (for any fixed integer J) and complete the basis by corresponding orthonormal basis functions ψ_{jk} $(j \geq J, k \in \mathbb{Z})$ in $W_J, W_{J+1}, \ldots$. For any $J \in \mathbb{Z}$, the system

$$\phi_{Jk}(k \in \mathbb{Z}), \qquad \psi_{jk}(j \geq J, k \in \mathbb{Z})$$

is an orthonormal basis in $L^2(\mathbb{R})$. The coarsest resolution level J is also called the decomposition level. Since a countable set of dilation and translation parameters is used, the mapping of g to these coefficients (or the coefficients themselves) are also called Discrete Wavelet Transform (DWT)—in contrast to the CWT defined in (3.90). Note that the distinction between DWT and CWT has nothing to do with x being continuous or not. Both methods are originally devised for functions $g(x)$ of

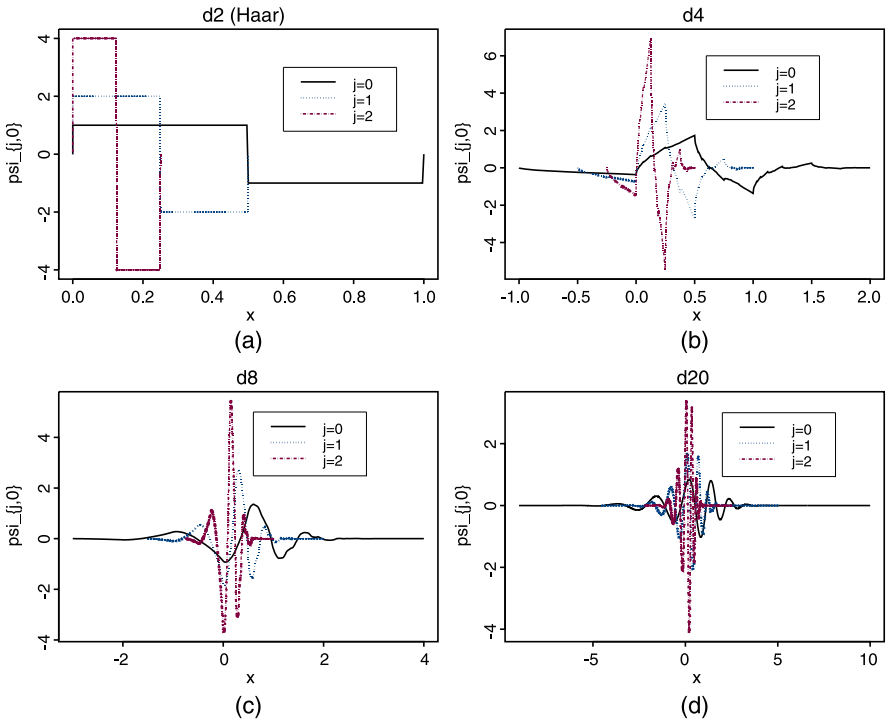

Fig. 3.5 Different mother wavelet functions at resolution levels $j = 0$, 1 and 2: (**a**) $d2$ (Haar); (**b**) $d4$; (**c**) $d8$; (**d**) $d20$

a continuous argument $x \in \mathbb{R}$. For further details, such as conditions on ϕ to achieve certain properties of ψ, see for instance the books listed at the beginning of this section.

Example 3.38 Haar wavelets are generated by $\phi(x) = 1\{x \in [0, 1]\}$. Equation (3.96) is easily verified since

$$\begin{aligned}\frac{1}{\sqrt{2}}\phi_{10}(x) + \frac{1}{\sqrt{2}}\phi_{11}(x) &= \phi(2x) + \phi(2x-1) \\ &= 1\left\{x \in \left[0, \frac{1}{2}\right]\right\} + 1\left\{x \in \left[\frac{1}{2}, 1\right]\right\} \\ &= 1\{x \in [0,1]\} = \phi(x).\end{aligned}$$

The mother wavelet function is equal to

$$\psi(x) = \frac{1}{\sqrt{2}}\phi_{10}(x) - \frac{1}{\sqrt{2}}\phi_{11}(x) = 1\left\{x \in \left[0, \frac{1}{2}\right]\right\} - 1\left\{x \in \left[\frac{1}{2}, 1\right]\right\}.$$

Figure 3.5(a) shows ψ_{0j} for $j = 0, 1, 2$.

Table 3.1 Properties of Daubechies wavelets

Wavelet	N	No. vanishing moments	No. derivatives	α
$d2$ (Haar)	1	1	0	0
$d4$	3	2	0	0.55
$d6$	5	3	1	1.09
$d8$	7	4	1	1.69
$d10$	9	5	1	1.97
$d12$	11	6	2	2.19
$d14$	13	7	2	2.46
$d16$	15	8	2	2.76
$d18$	17	9	3	3.07
$d20$	19	10	3	3.38

Example 3.39 Daubechies (1992) constructed compactly supported wavelets with a given degree of smoothness and a given number of vanishing moments M (see e.g. Sect. 3.4.5 of Vidakovic 1999 for a brief introduction and Daubechies 1992 for a full treatment). Several wavelets from the Daubechies family ($d2$, $d4$, $d8$ and $d20$; note that $d2$ is the Haar wavelet) are plotted in Figs. 3.5(a) through (d), at resolution levels $j = 0, 1, 2$ and dilation $k = 0$. Table 3.1 gives an overview of the first few Daubechies wavelets and their properties. Smoothness is characterized by the Hölder exponent α where $|\psi(x) - \psi(y)| \leq c|x - y|^{\alpha}$; the support of ϕ and ψ is given by $[0, N]$.

Example 3.40 Figure 3.6 shows the approximation of a function $g(t)$ (top in Fig. 3.6(a)) by $\phi_{0,k}$ and $\psi_{j,k}$ ($j = 0, 1, 2, 3$) using the Daubechies wavelet $d4$. Approximations, starting with the coarsest level $j = 0$ (curve at the bottom), and successively adding more levels up to $j = 3$, are shown in Fig. 3.6(a). The actual function is plotted on top. The corresponding coefficients are displayed in Fig. 3.6(b). As one can see, the father wavelet (denoted by "$s2$" in the figure) (at the coarsest level) captures the main long-term tendency of the function. The mother wavelets $\psi_{j,k}$ then add more details that are due to departures from the main (locally averaged) level of g. Using only four resolution levels ($j = 0, 1, 2, 3$) leads already to a reasonably good approximation of g.

In summary, since the $L^2(\mathbb{R})$ space is equal to the closure of

$$\bigcup_{j=-\infty}^{\infty} W_j = \bigoplus_{j=-\infty}^{\infty} W_j,$$

every function $g \in L_2(\mathbb{R})$ can be decomposed into orthogonal components by

$$g(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} b_{j,k} \psi_{jk}(x),$$

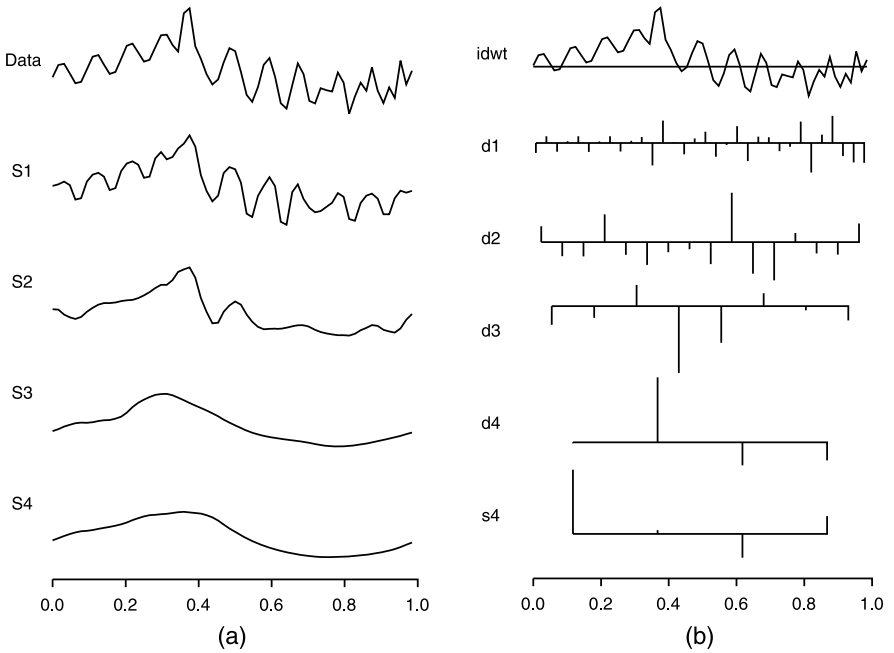

Fig. 3.6 Approximation of a function $g(t)$ (*curve at the top* of (**a**) and (**b**)) by $\phi_{0,k}$ and $\psi_{j,k}$ ($j = 0, 1, 2, 3$) using Daubechies' wavelet $d4$. (**a**) Shows approximations, starting with the coarsest level $j = 0$ (*curve at the bottom*) and successively adding more levels up to $j = 3$. The actual function is plotted on top. The corresponding coefficients are displayed in (**b**). The picture was generated with the S-Plus wavelets module. (The notation in (**b**) is somewhat confusing: "*d1*", "*d2*", "*d4*", "*s4*" have nothing to do with the type of the wavelet but rather denote in reversed order, and shifted by 1, the resolution level j)

where $b_{jk} = \langle g, \psi_{jk} \rangle = \int g(x)\psi_{jk}(x)\,dx$. At the same time, for any $J \in \mathbb{Z}$, the $L^2(\mathbb{R})$ can be written as a closure of $V_J \oplus \bigcup_{j=J}^{\infty} W_j$, so that another orthogonal decomposition is given by

$$g(x) = \sum_{k=-\infty}^{\infty} a_{Jk}\phi_{Jk}(x) + \sum_{j \geq J} \sum_{k=-\infty}^{\infty} b_{jk}\psi_{jk}(x), \tag{3.99}$$

where b_{Jk} is as before, and $a_{jk} = \langle g, \psi_{jk} \rangle = \int g(x)\psi_{jk}(x)\,dx$. The lowest index J is called decomposition level (sometimes also resolution level) because the father wavelets ϕ_{Jk} based on ϕ (with $\int \phi(x)\,dx = 1$ and support $[0, 1]$) approximate the function by pasting the functions $a_{Jk}\phi_{Jk}(x)$ ($k \in \mathbb{Z}$) next to each other, on an equidistant grid of x-values with step size 2^{-J}. For instance, for Haar wavelets, we obtain an approximation of g by step functions that are constant on adjacent intervals of length 2^{-J}. The second term in (3.99) captures deviations of g from the simple form given by the father wavelets approximation, starting with piecewise approximations by $b_{Jk}\psi_{Jk}$ ($k \in \mathbb{Z}$) on intervals of length 2^{-J} and continuing to

additional improvements of the approximation by $b_{jk}\psi_{jk}$ ($k \in \mathbb{Z}$) at arbitrarily fine resolution levels with 2^{-j} tending to zero. For this reason, the first term in (3.99) is sometimes called the approximation part, and the second term the details part. This terminology may not be ideal because it suggests that one would always approximate g by father wavelets only. This is of course not the case. In general, unless J is relatively large, a good approximation of g includes several levels of the mother wavelets as well. Note also the contrast between the conditions $\int \phi(x)\,dx = 1$ and $\int \psi(x)\,dx = 0$. This reflects the feature that ϕ_{Jk} estimates the local level of the function g (at the lowest resolution level J) and the mother wavelets ψ_{jk} capture the remaining "oscillations" of g around its local "average" (which is represented by $a_{Jk}\phi_{Jk}$). The quantities a_{Jk} and b_{jk} are therefore also called *scaling* and *wavelet* coefficients respectively (although the word "scaling" may be somewhat ambiguous in this context).

An important property of (regular) wavelets is its effect on polynomials. Let

$$g(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p$$

be a polynomial of order p and suppose that ψ has $M \geq p+1$ vanishing moments as defined in (3.86). Then, for all j, k,

$$\int g(x)\psi_{jk}(x)\,dx = \sum_{i=0}^{p} \beta_i \int x^i \psi_{jk}(x)\,dx = 0.$$

This will be crucial when studying long-memory processes later and when looking at nonparametric trend estimation based on the wavelet decomposition.

3.5.3 Computational Aspects and the Transition from Discrete to Continuous Time

A recursive algorithm for calculating the coefficients (i.e. the DWT) can be obtained as follows. If $g \in V_0$, then we have another expansion in terms of father wavelets only,

$$g(x) = \sum_{k\in\mathbb{Z}} a_{0,k}\phi(x-k),$$

where $a_{0,k} = \int g(x)\phi(x-k)\,dx$. On the other hand, $\phi \in V_0 \subseteq V_1$, so that (cf. (3.96)) $\phi(x) = 2^{\frac{1}{2}} \sum_{m\in\mathbb{Z}} u_m \phi(2x-m)$ and

$$\begin{aligned} a_{0,k} &= \int g(x)\left\{2^{\frac{1}{2}} \sum_{m=-\infty}^{\infty} u_m \phi\big(2(x-k)-m\big)\right\} dx \\ &= \sum_{m=-\infty}^{\infty} u_m \int g(x) 2^{\frac{1}{2}} \phi\big(2x-(2k+m)\big)\,dx = \sum_{m=-\infty}^{\infty} u_m a_{1,2k+m}. \end{aligned}$$

In general, for arbitrary $j \in \mathbb{Z}$, we have

$$a_{j,k} = \sum_{m=-\infty}^{\infty} u_m a_{j+1,2k+m} \tag{3.100}$$

and, by (3.98),

$$b_{j,k} = \sum_{m=-\infty}^{\infty} v_m a_{j+1,2k+m} = \sum_m (-1)^m u_{1-m} a_{j+1,2k+m}. \tag{3.101}$$

In other words, the coefficients associated with the projection of g on a smaller space V_j can be computed in terms of the coefficients associated with the larger space V_{j+1}. Hence, starting with most detailed description of g, we go down to a coarser and coarser level. Equations (3.100) and (3.101) are also called cascade algorithm for the DWT.

Now we turn to an issue that is of particular interest in time series analysis. Equation (3.99) describes the wavelet expansion of a continuously observed function $g(x)$ ($x \in \mathbb{R}$). Now assume that we observe $y_0, \ldots, y_{n-1}$ associated with the x-values $x = 0, 1, \ldots, n-1$. For illustration, we will focus on a zero-mean stationary time series $Y_t(\omega)$ with values y_i observed at time points $t \in \{0, 1, \ldots, n-1\}$ (i.e. $Y_t = y_t$ for $t = i$). We will use the notation $f_Y(\lambda)$ ($\lambda \in [-\pi, \pi]$) for the spectral density of Y_t and $y = (y_0, \ldots, y_{n-1})^T$ for the vector of observed values. For simplicity, we assume that $n = 2^J$ for some $J \geq 1$. One way of adopting the techniques above is to artificially create a function (sample path) $\tilde{Y}(t) = g(t; \omega)$ in continuous time ($t \in \mathbb{R}$). There is not one unique way of doing this, and the choice of the method may depend on the purpose. For instance, Veitch et al. (2000) suggest that, if the focus is on second-order properties of $Y_t(\omega)$ ($t \in \mathbb{Z}$), then the synthetic continuous-time version should be such that these properties are preserved. This means that the process $\tilde{Y}(t; \omega)$ ($t \in \mathbb{R}$) has expected value zero and spectral density

$$\begin{aligned} f_{\tilde{Y}}(\lambda) &= f_Y(\lambda) \quad \big(\lambda \in [-\pi, \pi]\big), \\ f_{\tilde{Y}}(\lambda) &= f_Y(\lambda) \quad \big(|\lambda| > \pi\big). \end{aligned}$$

One of several possible solutions is

$$\tilde{Y}_t = g(t; \omega) = \sum_{j=-\infty}^{\infty} \frac{\sin \pi(t-j)}{\pi(t-j)} Y_j, \tag{3.102}$$

where the equality is in $L^2(\Omega)$. Now $\tilde{Y}_t = g(t; \omega)$ is defined for all $t \in \mathbb{R}$, and, given ω (i.e. an observed path of $\tilde{Y}_t$), we can proceed with the DWT as discussed above. Note that for practical applications with a finite number of observed values Y_t ($t = 0, \ldots, n-1$), one has to truncate the sum, i.e. set unobserved values equal to zero.

Another continuous-time version of Y_t that is often used is the step function

$$\tilde{Y}_t(\omega) = g(t;\omega) = \sum_{i=0}^{n-1} y_i 1\{t \in [i, i+1)\}.$$

If we use a father wavelet function ϕ with support $[0,1]$, $\int \phi(x)\,dx = 1$ and $\int \phi^2(x)\,dx = 1$, then the projection of g on V_0 (which is generated by the orthonormal basis $\phi_{0,k}$) is equal to

$$\tilde{Y}_t^*(\omega) = g^*(t;\omega) = \sum_{k=0}^{n-1} y_k \phi_{0,k}(t) = \sum_{k=0}^{n-1} y_k \phi(t-k).$$

Therefore the observed values y_t $(t = 0, 1, \ldots, n-1)$ are interpreted as coefficients in the approximation of $\tilde{Y}_t(\omega) = g(t;\omega)$ using resolution levels $j \leq 0$. Now we can decompose this function into lower-resolution components down to a certain coarsest level $J \leq 0$. The cascade algorithm defined in (3.100) and (3.101) can be used to obtain the corresponding coefficients a_{Jk} $(k \in \mathbb{Z})$ and b_{jk} $(J \leq j \leq 0, k \in \mathbb{Z})$. Since $g^*(t;\omega) = 0$ for $t < 0$ and $t > k$, one actually only has to calculate a finite number of coefficients.

3.6 Fractals

In the context of time series analysis, fractal behaviour is often mentioned as synonym for long-range dependence. Though there are strong connections between the two notions, they are also in some sense completely different. To see the connections and differences, it is necessary to understand some of the basic definitions in fractal geometry. Ever since the pioneering books by Mandelbrot (1977, 1983) and a sequence of papers in applied journals (e.g. Mandelbrot and van Ness 1968; Mandelbrot and Wallis 1968a, 1968b, 1969a, 1969b, 1969c), the theory of fractals and their applications have developed at an enormous speed. Here we can only give a tiny glimpse of a few basic concepts. A beautiful concise introduction to some of the mathematical principles is for instance Falconer (2003).

There is no "official" consensus on the definition of a fractal. However, what is generally agreed on is that the Hausdorff measure and Hausdorff dimension play a key role. One possible definition of a fractal is then for example that it is a set $A \subseteq \mathbb{R}^k$ whose Hausdorff dimension $\dim_H A$ is not an integer.

The Hausdorff measure and dimension in $\mathbb{R}^k$ are defined as follows. The general idea comes from measuring the length, area, volume etc. of geometric objects using approximations by a union of increasingly small simple geometric shapes such as straight lines, circles, balls etc. Consider a set $A \subseteq \mathbb{R}^k$ and a cover $\mathscr{U} = \{U_i, i = 1, 2, \ldots\}$ of A by a countable number of open sets U_i, i.e. $A \subseteq \bigcup_{i=1}^{\infty} U_i$. For any $\delta > 0$, we say that $\mathscr{U}$ is a δ-cover if the diameter of each U_i in $\mathscr{U}$ is at most δ, i.e.

$$\|U_i\| = \sup_{x,y \in \mathscr{U}} \|x - y\| \leq \delta,$$

where $\|x - y\|$ is the Euclidian distance between x, y. Denote by $\mathscr{C}_\delta(A)$ the set of all δ-covers of A and define, for each real number $s > 0$,

$$\mathscr{H}^s_\delta(A) = \inf_{\mathscr{U} \in \mathscr{C}_\delta(A)} \sum_{i=1}^{\infty} \|U_i\|^s.$$

Note that, for $s = 1$, this corresponds to an approximation of the length of A, for $s = 2$, it approximates the area, and so on. By definition, $0 \le \mathscr{H}^s_\delta(A) \le \infty$, and $\mathscr{H}^s_\delta(A)$ is monotonically nondecreasing in δ. Therefore the infimum over δ exists (even though it may be infinite). Therefore we may define the so-called s-dimensional Hausdorff measure of A by

$$\mathscr{H}^s(A) = \lim_{\delta \to 0} \mathscr{H}^s_\delta(A)$$

(it can indeed be shown that $\mathscr{H}^s$ is a measure). Note that for $s = 1, 2, 3, \ldots$, this corresponds to the usual definitions of length, area, volume etc. In fact, if A is a Borel set, and $s = 1, 2, 3, \ldots$ is an integer, then $\mathscr{H}^s(A)$ is equal to a constant times the usual measure of length, area, volume etc. Since ultimately diameters δ smaller than one are used, $\mathscr{H}^s$ is monotonically nonincreasing in s. For $s = 0$, $\mathscr{H}^s(A)$ is equal to the number of points in A. Thus, if the number of points in A is infinite, then $\mathscr{H}^s(A) = \infty$. As s increases, $\mathscr{H}^s(A)$ remains infinite until a certain value s_0 where $\mathscr{H}^s(A) = 0$ for all $s > s_0$. For s_0 itself, $\mathscr{H}^s(A)$ may take any value between (and including) zero and infinity. The Hausdorff dimension (or Hausdorff–Besicovitch dimension; Hausdorff 1918; Besicovitch 1928) is then defined as this exponent s_0 where the value of $\mathscr{H}^s(A)$ flips, i.e.

$$\dim_H A = \inf\bigl\{s \ge 0 : \mathscr{H}^s(A) = 0\bigr\} = \sup\bigl\{s \ge 0 : \mathscr{H}^s(A) = \infty\bigr\}.$$

For simple geometric objects, $\dim_H A$ is an integer because $\mathscr{H}^s$ with $s \in \mathbb{N}$ is proportional to the usual Lebesgue measure. There are however many interesting sets where $\mathscr{H}^s$ is not an integer. A very intuitive way of constructing such sets is by iterative application of a set of functions $f_1, \ldots, f_p$, also called iterated function system (IFS). First of all, a function $f : \mathbb{R}^k \to \mathbb{R}^k$ is called a similarity of scale $c > 0$ if $\|f(x) - f(y)\| = c\|x - y\|$ for all points x, y. If $c < 1$, then f is a contracting similarity. Given contracting similarities $f_1, \ldots, f_p$ with scales $c_1, \ldots, c_p$ one defines for every set $E \subseteq \mathbb{R}^k$, the transformation

$$f(E) = \bigcup_{j=1}^{p} f_j(E).$$

Then it can be shown that there is a unique set A, also called the attractor of the IFS, such that for any $i \ge 1$,

$$f^i(A) = A.$$

(Here f^i means that we apply the transformation i times.) Moreover, this set can be obtained by infinite iteration starting with an arbitrary set E for which $f_j(E) \subseteq E$ $(1 \le j \le k)$, namely

$$A = \bigcap_{i=1}^{\infty} f^i(E).$$

By definition, the attractor A is self-similar in the sense that it is a union of copies of itself at different scales (see e.g. Falconer 2003, Chap. 9). From the definition of similarities it immediately follows that

$$\mathscr{H}^s\big(f_j(A)\big) = c_j^s \mathscr{H}^s(A).$$

Under a so-called open set condition which essentially implies that $f_1(A), \ldots, f_p(A)$ are "almost" disjoint (see e.g. Falconer 2003, Sect. 9.2), we then have

$$\begin{aligned}\mathscr{H}^s(A) &= \mathscr{H}^s\left(\bigcup_{j=1}^{p} f_j(A)\right) = \sum_{j=1}^{p} \mathscr{H}^s\big(f_j(A)\big) \\ &= \mathscr{H}^s(A) \sum_{j=1}^{p} c_j^s.\end{aligned}$$

Thus, in cases where for $s_0 = \dim_H A$ we have $0 < \mu_{s_0}(A) < \infty$ (which in fact can be shown under the given assumptions), we obtain the condition

$$\sum_{j=1}^{p} c_j^{s_0} = 1.$$

In other words, the Hausdorff dimension of A is equal to the solution s_0 of this equation.

Example 3.41 One of the most famous examples of a fractal is the Cantor set. The construction starts with the interval $[0, 1]$. In a first step one removes the middle third of the interval to obtain the two sets $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$. In a second step, one again removes the middle thirds from each of these sets, and so on. The Cantor set is then the intersection of all sets obtained during the infinite iteration process. This can also be described as the attractor of the IFS (in $\mathbb{R}$) with $f_1(x) = \frac{1}{3}x$ and $f_2(x) = \frac{1}{3}x + \frac{2}{3}$. Since $c_1 = c_2 = \frac{1}{3}$, we have the equation $2(\frac{1}{3})^{s_0} = 1$ so that

$$s_0 = \dim_H A = \log 2 / \log 3 \approx 0.6309.$$

Example 3.42 Another classical fractal is the Sierpiński triangle (or Sierpiński gasket). It is constructed starting with a filled equilateral triangle which is divided into four smaller equilateral triangles, with the midpoints of each of the three sides of the original triangle as the new vertices. The triangle in the middle is then

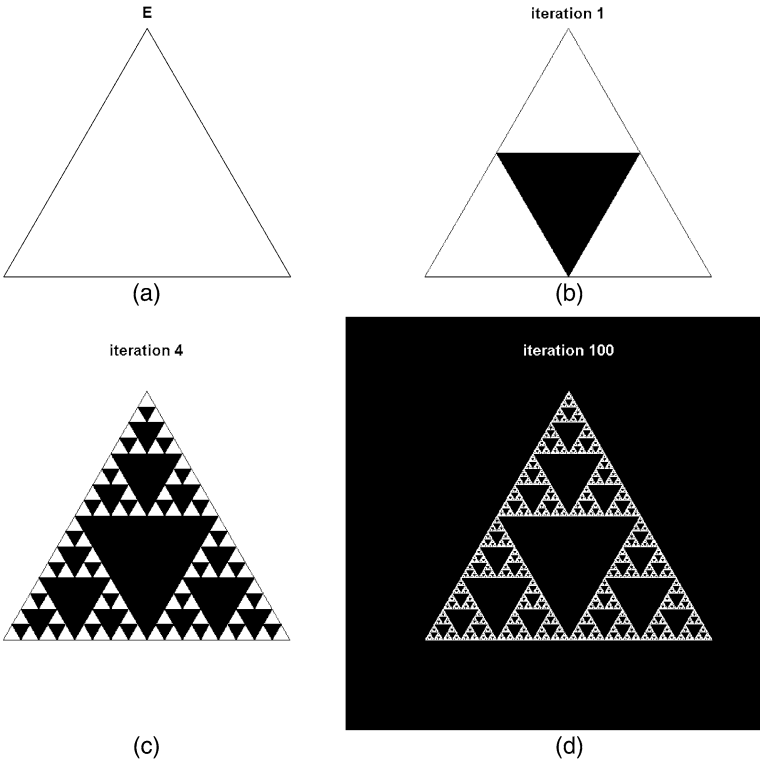

Fig. 3.7 Recursive construction of the Sierpiński triangle: initial set E (**a**) and $A_k = \bigcap_{i=1}^{k} f^i(E)$ for $k = 1$, 4 and 100 respectively (**b**), (**c**), (**d**). Note that the *white area* within the boundary of the initial triangle represents A_k. The figures were created using the R-function *spt* (programmer: Bin Wang)

removed, and the whole procedure is repeated for each of the remaining triangles, and so on. This can also be described as the attractor of the IFS given by an initial equilateral triangle with side length 1 and left lower corner at the origin, and the functions (in $\mathbb{R}^2$) $f_1(x_1, x_2) = \frac{1}{2}(x_1, x_2)$, $f_2(x_1, x_2) = \frac{1}{2}(x_1 + \frac{1}{2}, x_2)$ and $f_3(x_1, x_2) = \frac{1}{2}(x_1 + \frac{1}{4}, x_2 + \frac{\sqrt{3}}{4})$. Thus we have $c_j = \frac{1}{2}$ $(j = 1, 2, 3)$ so that $\sum c_j^s = 3(\frac{1}{2})^s = 1$ leads to the Hausdorff dimension $s_0 = \log 3/\log 2 \approx 1.585$. Figure 3.7 shows different steps in the iteration converging to the attractor.

The practical application of the Hausdorff dimension is quite difficult, in particular when dealing with observed data. Various alternative definitions have therefore been suggested in the literature. The best known is the so-called box-counting dimension. Denote by N_δ the minimal number of sets U_i needed for a δ-cover of A. As $\delta \to 0$, one usually has $N_\delta \sim c\delta^{-s}$ for some $0 < s < \infty$. If that is so, then

$$s = -\lim_{\delta \to 0} \log N_\delta / \log \delta =: \dim_B A$$

is called box-counting dimension of A. More generally, even if it is not clear whether this limit exists, one can at least define the lower and upper box-counting dimension, $\underline{\dim}_B A$ and $\overline{\dim}_B A$, by replacing lim by lim inf and lim sup respectively. This definition is very convenient for applications, in particular since it is possible to replace general open sets U_i by more specific ones, such as closed balls, cubes etc. Since one uses special coverings for the box-counting dimensions, one has $\mathscr{H}^s_\delta(A) \leq N_\delta \delta^s$, and hence,

$$\dim_H A \leq \underline{\dim}_B A \leq \overline{\dim}_B A.$$

Thus, the box-counting dimension is useful for obtaining upper bounds for the Hausdorff dimension. This is in particular interesting if the value is not an integer.

The construction of self-similar sets is only one of many possibilities for obtaining fractals (sets whose Hausdorff dimension is not an integer). The notion of (exact, deterministic) self-similarity is too rigid for general applications. When dealing with random objects, one likes to replace it by *stochastic* self-similarity as defined before. Thus, recall that for instance a stochastic process X_t ($t \in \mathbb{R}$) is called self-similar with self-similarity parameter H if for any $c > 0$, the rescaled process $c^{-H} X_{ct}$ has the same distribution as X_t. The same definition applies to random fields X_t with $t \in \mathbb{R}^m$ for some $m \geq 1$. More generally, one looks at processes in the domain of attraction of a self-similar process.

The obvious question now is whether there is a universal connection between the *stochastic* self-similarity parameter H and the Hausdorff dimension of sample paths. In general, this questions can be asked for processes in the domain of attraction of a self-similar process. To be specific, we consider the Hausdorff dimension of random graphs

$$A_{X,\text{graph}}(\omega) = \left\{ \big(t, X(t,\omega)\big) : t \in [0,1]^m \subset \mathbb{R}^m \right\} \subseteq \mathbb{R}^{m+1}.$$

(Note that this is a different question than finding the Hausdorff dimension of the one-dimensional set $\tilde{A} = \{x \in \mathbb{R} : x = X(t), t \in [0,1]^m\}$.) Examining the meaning of H on one hand and $\dim_H A_{X,\text{graph}}$ on the other hand, it becomes quite obvious that there is no universal formula that would link H with the Hausdorff dimension. For instance, if we consider not necessarily self-similar processes with existing second moments, then the parameter H only determines the long-term behaviour of autocorrelations. The detailed autocorrelations and the marginal distribution can be chosen quite freely. The Hausdorff measure on the other hand characterizes very local geometric properties. To establish a relationship between H and $\dim_H A_{X,\text{graph}}(\omega)$, one therefore needs to add more detailed specifications, such as self-similarity, symmetry, marginal distribution etc. For instance, for general Lévy processes, the relationship is quite complex (see e.g. Nolan 1988; Manstavičius 2007 and references therein). Also see Hall and Roy (1994) for results in the context of stationary processes, Kôno (1986), and Talagrand (1995) and Xiao (1997a, 1997b) for self-similar processes in general and fractional Brownian motion in particular.

A simple connection between autocovariances and Hausdorff dimension can be established for certain self-similar processes. Consider first a Gaussian self-similar process, i.e. fractional Brownian motion B_H. Then the following holds.

Theorem 3.11 *Let* $A_{X,\text{graph}}(\omega) = \{(t, B_H(t,\omega)) : t \in [0,1] \subset \mathbb{R}\} \subseteq \mathbb{R}$. *Then, with probability* 1,

$$\dim_H A_{X,\text{graph}}(\omega) = \dim_B A_{X,\text{graph}}(\omega) = 2 - H. \tag{3.103}$$

Proof (Sketch) In a first step one shows that almost surely we have Hölder continuity in the following sense. Let $0 < \beta < H$. Then there exist constants c and h_0 such that, with probability 1,

$$\left|B_H(t+h) - B_H(t)\right| \le c|h|^\beta$$

for $|h| \le h_0$. This in turn can be used to show that $\overline{\dim}_B A_{X,\text{graph}}(\omega) \le 2 - \beta$ for all $\beta < H$ (see Corollary 11.2 in Falconer 2003), and hence,

$$\dim_H A(\omega) \le \underline{\dim}_B A_{X,\text{graph}}(\omega) \le \overline{\dim}_B A_{X,\text{graph}}(\omega) \le 2 - H.$$

To obtain a lower bound for $\dim_H A_{X,\text{graph}}(\omega)$, one makes use of the particular Gaussian distribution of B_H and potential theory (see e.g. Theorem 16.7 in Falconer 2003 for details). □

Note in particular that, for $H \to 1$, the Hausdorff dimension of the graph tends to one which is the smallest possible dimension for a graph (in $\mathbb{R}^2$). This reflects the increase of dependence between increments of $B_H(t)$. On the other hand, for $H \to 0$, the dimension approaches the maximal dimension 2. This means that the stronger antipersistence is, the more the space is filled out. The same result also holds for symmetric α-stable Lévy processes (see e.g. Falconer 2003, Theorem 16.8, for a sketched proof).

Theorem 3.12 *Let* $X(t) = X(t,\omega)$ *be a symmetric* α*-stable Lévy process with* $0 < \alpha \le 2$ (*i.e.* $X(t)$ *has independent stationary increments with characteristic function* $E[\exp(i(X(t+u) - X(t)))] = c|u|^\alpha$). *Then, with probability* 1,

$$\dim_H A_{X,\text{graph}}(\omega) = \dim_B A_{X,\text{graph}}(\omega) = \max\{1, 2 - 1/\alpha\}.$$

Note that the self-similarity parameter is $H = 1/\alpha$, so that the formula is actually the same as for fractional Brownian motion. For $\alpha = 2$, we obtain Brownian motion with $H = \frac{1}{2}$ and thus indeed a special case of the previous theorem. For $1 < \alpha < 2$, second moments do not exist, but first moments are finite. For $\alpha < 1$, even the first moment does not exist, and the Hausdorff dimension is always 1. In this case, sample paths consist of infinitely many jumps in any time interval.

Theorems 3.11 and 3.12 are obtained under self-similarity and additional assumptions on the marginal distribution. If we remove the assumption of self-similarity, then H and the Hausdorff dimension are completely unrelated in general, even when considering Gaussian processes only. The reason is that now H no longer determines the fully autocorrelation structure, but instead its long-term behaviour only. In contrast, the Hausdorff dimension of Gaussian processes is determined by the behaviour of the autocorrelation function at small distances because this determines the local behaviour of sample paths. To be specific, let $X(t)$ $(t \in \mathbb{R})$ be a stationary Gaussian process with long-memory parameter $d = H - \frac{1}{2} \in (0, \frac{1}{2})$ in the sense that for the autocovariance function, we have

$$\gamma_X(u) = cov\big(X(t), X(t+u)\big) \sim c_\gamma u^{2d-1} = c_\gamma u^{2H-2}$$

as $u \to \infty$, or equivalently, for the spectral density,

$$f_X(\lambda) \sim c_f |\lambda|^{-2d} = c_f |\lambda|^{1-2H}$$

as $\lambda \to 0$. Then the behaviour of $\gamma_X(u)$ as $u \to 0$ is unspecified. The Hausdorff dimension of (the graph of) sample paths is however determined by the behaviour of γ_X at the origin. More specifically, if, as $u \to 0$,

$$\gamma_X(u) = \sigma_X^2\big(1 - c_0|u|^{2\beta} + o\big(|u|^{2\beta}\big)\big)$$

for some $0 < c_0 < \infty$ and $0 < \beta \leq 1$, then

$$\dim_H A_{X,\text{graph}} = 2 - \beta$$

(see e.g. Adler 1981). This means that even when we stay within the realm of stationary Gaussian processes, there is no general relationship between H and the fractal dimension of sample paths. In fact, examples can be constructed where both aspects are completely unrelated, and, on the other hand, there are cases with a one-to-one relationship between H and $\dim_H A_{X,\text{graph}}$. This is illustrated by the following examples.

Example 3.43 Let $X(t) = B_H(t) - B_H(t-1)$ $(t \in \mathbb{R}_+, 0 < H < 1)$ be fractional Gaussian noise. Then the autocovariance function and thus the complete distribution of $X(t)$ is fully specified, namely

$$\begin{aligned}\gamma_X(u) - \sigma_X^2 &= \frac{\sigma_X^2}{2}\big(|u+1|^{2H} - 2|u|^{2H} + |u-1|^{2H}\big) - \sigma_X^2 \\ &\sim \text{const} \cdot |u|^{2H} \quad (\text{as } u \to 0).\end{aligned}$$

Thus, not only for the self-similar process B_H but also for its increments (as defined above), we obtain relationship (3.103), i.e. $\dim_B A_{X,\text{graph}}(\omega) = 2 - H$. Gaussianity, together with self-similarity, carries this relationship through to the increments.

Example 3.44 Let $X(t)$ be a stationary Gaussian process belonging to the so-called Cauchy class (Wackernagel 1998; Gneiting and Schlather 2004 and references therein), which means that the autocovariance function is given by

$$\gamma_X(u) = \left(1 + |u|^{2\beta}\right)^{-\frac{\kappa}{\beta}}$$

for some $0 < \beta \leq 1$ and $\kappa > 0$. Then, as $u \to 0$,

$$\gamma_X(u) = \sigma_X^2\left(1 - c_0|u|^{2\beta} + o\left(|u|^{2\beta}\right)\right),$$

whereas, as $u \to \infty$,

$$\gamma_X(u) \sim c_\gamma u^{-2\kappa}.$$

This means that the Hurst or long-memory parameter $H = 1 - \kappa$ (and $d = H - \frac{1}{2}$) is completely unrelated to the Hausdorff dimension

$$\dim_H A_{X,\text{graph}} = 2 - \beta.$$

Within the Cauchy class it is possible to combine any degree of long memory with any Hausdorff dimension below 2. (This can be generalized to Cauchy classes with an m-dimensional index $t \in \mathbb{R}^m$ to obtain $\dim_H A_{X,\text{graph}} = m + 1 - \beta$.) Figures 3.8(a) through (d) and Figs. 3.9(a) and (b) show simulated sample paths (all with the same random seed) for different values of β and κ. In the first two figures (Fig. 3.8) the long-memory parameter κ is fixed, so that one can see the influence of β on the local structure of the sample paths. As expected, higher values of β lead to smoother paths. This is reflected in a lower Hausdorff dimension. In Figs. 3.9(a) and (b), β is fixed, and one can see that changing κ does not have any influence on the local impression of the graph. Finally, the four Figs. 3.10(a) through (d) show image plots of a Cauchy class random field (with $t \in \mathbb{R}^2$). Again one can see that increasing β leads to a smoother surface.

Since the relationship between H and the Hausdorff dimension depends on specific circumstances, various statistical methods have been suggested for estimating the fractal (Hausdorff, box-counting or related) dimension *directly* (instead of indirect inference via H). Some references are for instance Taylor and Taylor (1991), Smith (1992), Feuerverger et al. (1994), Constantine and Hall (1994), Hall (1995), Hall et al. (1996), Chan and Wood (1997, 2004), Istas and Lang (1997), Kent and Wood (1997), Davies and Hall (1999), Blanke (2004). Most methods are designed for the box-counting dimension. One should bear in mind, however, that in general the box-counting dimension need not coincide with the Hausdorff dimension, though it at least provides an upper limit.

In summary, many interesting fractals can be generated as attractors of iterated function systems (based on similarities f_j). These sets are exactly self-similar, and the Hausdorff dimension follows directly from the scaling factors c_j of the involved functions f_j. Stochastic fractal structures can be obtained by relaxing the assumption of *exact* self-similarity and replacing it by *stochastic* self-similarity with a self-similarity parameter H. This leads to self-similar processes (and processes in their

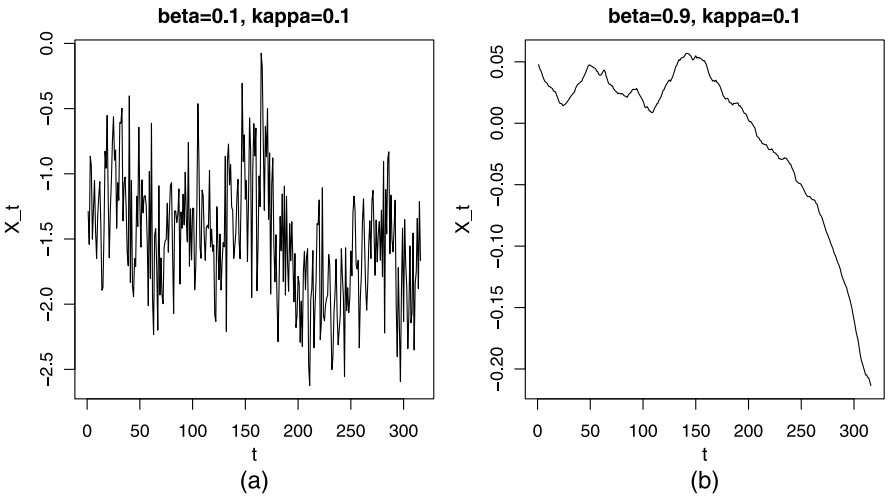

Fig. 3.8 Simulated sample paths of Cauchy class processes with $\kappa = 0.1$, and $\beta = 0.1$ and 0.9 respectively

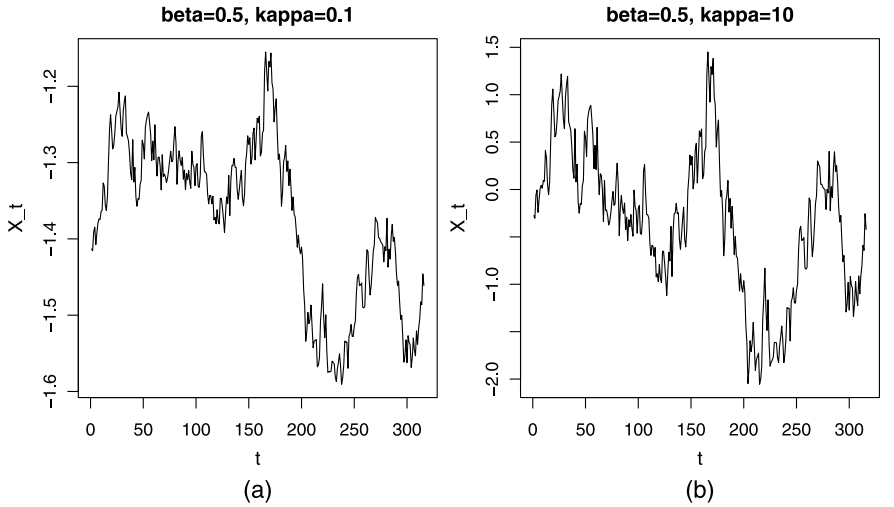

Fig. 3.9 Simulated sample paths of Cauchy class processes with $\beta = 0.5$, and $\kappa = 0.1$ and 10 respectively

domain of attraction), with sample paths encompassing a much larger variety of geometric structures. However, at the same time, the direct connection between self-similarity (specified by the scaling factors c_j in the deterministic and by H in the stochastic case) and the Hausdorff dimension is lost. A one-to-one relationship can be recovered only if suitable additional specifications such as continuity, finite-dimensional distributions etc. are imposed. Some caution is therefore needed

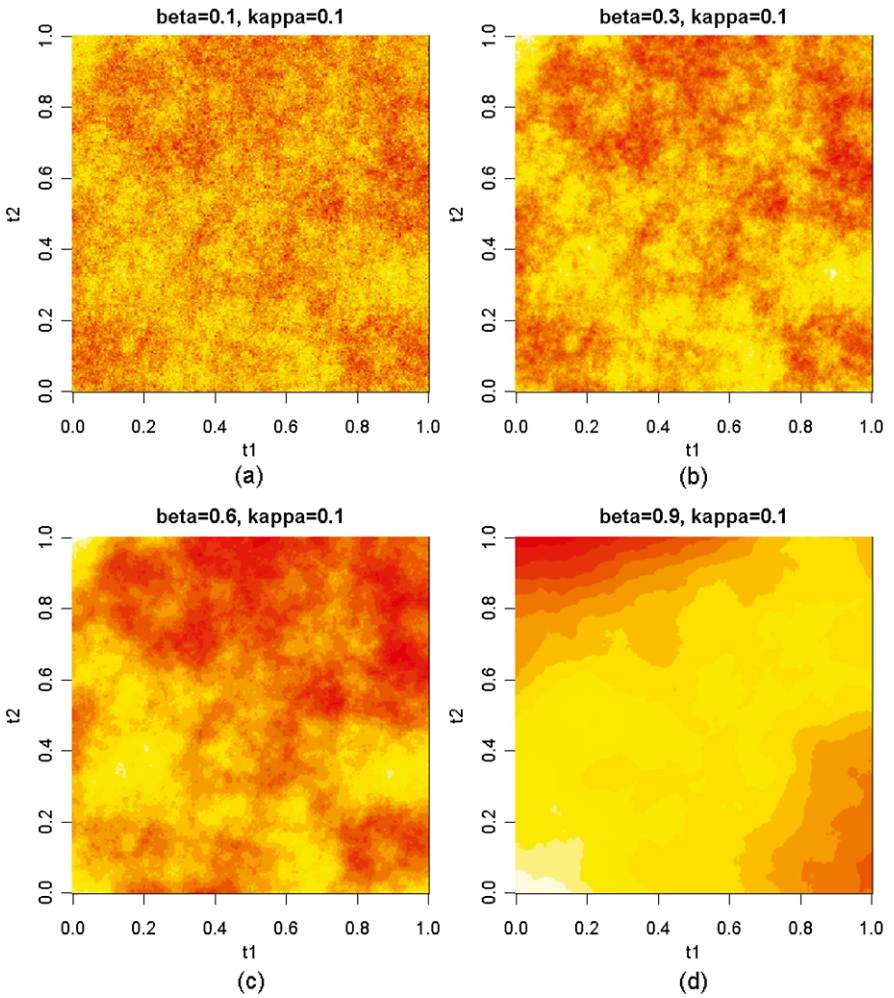

Fig. 3.10 Image plots of simulated spatial processes from the Cauchy class with $\kappa = 0.1$, and $\beta = 0.1, 0.3, 0.6$ and 0.9 respectively

when interpreting estimated values of H as indicators of a certain fractal dimension. The situation changes however when we apply aggregation. For instance, for time series, temporal aggregation ultimately leads back to fractional Gaussian noise (see Chap. 4 and Sects. 2.2.1 and 5.4.6) since local effects are eliminated in the limit. It should also be noted that in most situations one observes data at discrete time points (or more generally discrete values of $t \in \mathbb{R}^m$), whereas strictly speaking, the notions of self-similarity and Hausdorff dimension of sample paths (in the sense of graphs) lead to interesting results in continuous time (or space etc.) only. In this sense general conclusions on the fractal nature of observed data may often be considered as convenient approximations rather than a complete description of the

phenomenon. Further caution is required due to possible discretization effects and noise that may blur the underlying fractal structure.

3.7 Fractional and Stable Processes

In this section we present some theory on two classes of processes that appear as limits in the case of long-memory sequences with finite and infinite variance.

We start with integral representations of fractional Brownian motion (fBm) and Hermite–Rosenblatt processes in Sects. 3.7.1.1 and 3.7.1.2. Both processes are represented as a (multiple) Wiener–Itô integral with respect to a Brownian motion. Such representations date back to Mandelbrot and van Ness (1968). Then, we link the time-domain representation with the spectral representation, which is crucial in the understanding of limit theorems for nonlinear functionals of Gaussian processes. Further material can be found in Taqqu (1978, 2003) and Pipiras and Taqqu (2000a, 2003). Furthermore, Meyer et al. (1999) and Pipiras et al. (2004) discuss wavelet expansions of fractional Brownian motion and the Hermite–Rosenblatt process (this material is not discussed here).

Next, the integral representation of fractional Brownian motion is extended to a construction of Linear Fractional Stable Motion (LFSM) in Sect. 3.7.2. We first recall the point process representation of a Poisson process. This is followed by a brief summary of stable random variables, stable Lévy processes and stable random measures. This material allows us to define an LFSM "replacing" Brownian motion by a Lévy process. A more detailed discussion can be found for instance in Samorodnitsky and Taqqu (1994).

We conclude with a section on fractional calculus.

3.7.1 Fractional Brownian Motion and Hermite–Rosenblatt Processes

3.7.1.1 Integral Representation of fBm

Let M be a real-valued Gaussian process on $[-\pi, \pi]$ with zero mean, (almost surely) right-continuous sample paths and uncorrelated (and hence independent) increments such that

$$\begin{aligned} cov\big(dM(\lambda), dM(\nu)\big) &= E\big[dM(\lambda)\, dM(\nu)\big] = 0 \quad (\lambda \neq \nu) \\ \operatorname{var}\big(dM(\lambda)\big) &= c_M\, d\lambda, \end{aligned}$$

where c_M is a constant. Without loss of generality, we can assume that $c_M = 1$. The process M can also be interpreted as a Gaussian random measure on $[-\pi, \pi]$. For

disjoint sets A and B, we have $E[M(A)M(B)] = 0$. Furthermore, for any $c > 0$ and any sets $A_1, \ldots, A_m$,

$$\big(M(cA_1), \ldots, M(cA_m)\big) \stackrel{\mathrm{d}}{=} c^{1/2}\big(M(A_1), \ldots, M(A_m)\big). \tag{3.104}$$

This equation establishes self-similarity of the Gaussian random measure M with self-similarity parameter $H = \frac{1}{2}$.

Let g be a square-integrable function (with respect to the Lebesgue measure on $[-\pi, \pi]$). Then

$$I(g) = \int g(\lambda)\, dM(\lambda) \tag{3.105}$$

is well defined and is called the Wiener–Itô integral. We can associate

$$M\big([0, x]\big) = \int_0^x dM(\lambda) = B(x), \tag{3.106}$$

where $B(x)$ is a Brownian motion (which is usually defined as a Gaussian process with independent, stationary increments such that its variance is proportional to x). The Gaussian random measure M is also used to construct a fractional Brownian motion (fBm). We start with a commonly used definition of fBm:

Definition 3.23 A Gaussian stochastic process $B_H(u)$ $(u \in \mathbb{R})$ with mean zero is called a fractional Brownian motion with self-similarity (or Hurst) parameter $H \in (0, 1)$ if its covariance function is given by

$$\gamma_H(t, s) = cov\big(B_H(u), B_H(v)\big) = \frac{\sigma^2}{2}\big[|u|^{2H} + v^{2H} - |u - v|^{2H}\big] \quad (u, v \in \mathbb{R}).$$

In order to proceed with the construction of fBm, we note that (3.106) allows us to rewrite the integral (3.105) as

$$I(g) = \int g(x)\, dB(x).$$

The next lemma establishes two basic properties of the Wiener–Itô integral.

Lemma 3.18 *Assume that g, g_1, g_2 are square-integrable functions with respect to the Lebesgue measure. Then*

$$Cov\big(I(g_1), I(g_2)\big) = \int g_1(x) g_2(x)\, dx.$$

In particular, $I(g_1)$ and $I(g_2)$ are independent if and only if $\int g_1(x) g_2(x)\, dx = 0$. Furthermore, $I(g)$ has a normal distribution with mean 0 and variance $\int g^2(x)\, dx$.

Define now

$$s_+ = \begin{cases} s & \text{if } s > 0, \\ 0 & \text{if } s \le 0, \end{cases} \qquad s_- = \begin{cases} -s & \text{if } s < 0, \\ 0 & \text{if } s \ge 0, \end{cases}$$

and consider the kernel

$$\begin{aligned} Q_{u,1}(x; H) &= c_1\big[(u-x)_+^{H-1/2} - (-x)_+^{H-1/2}\big] + c_2\big[(u-x)_-^{H-1/2} - (-x)_-^{H-1/2}\big] \\ &=: c_1 Q_{u,1}^+(x; H) + c_2 Q_{u,1}^-(x; H), \end{aligned} \tag{3.107}$$

where c_1 and c_2 are deterministic constants. We note that the kernel $Q_{u,1}(\cdot; H)$ is square integrable. Indeed, for example the first integrand $(u-x)_+^{H-1/2} - (-x)_+^{H-1/2}$ behaves like $(H-1/2)(-x)^{H-3/2}$ as $x \to -\infty$ and like $(u-x)_+^{H-1/2}$ as $x \to u$. A function $(-x)^{-(3/2-H)}$ $(x \to -\infty)$ is square integrable if $2(3/2-H) > 1$, that is, $H < 1$. Likewise, the function $y^{-(1/2-H)}$ $(y \to 0)$ is square integrable if $2(1/2-H) < 1$, which means $H > 0$.

One can verify that the kernel $Q_{u,1}(\cdot, H)$ has the following properties: for all $0 \le v < u$,

$$Q_{u,1}(x; H) - Q_{v,1}(x; H) = Q_{u-v,1}(x - v; H), \tag{3.108}$$

$$Q_{cu,1}(cx; H) = c^{H-1/2} Q_{u,1}(x; H). \tag{3.109}$$

In particular, the first property reflects the stationarity of increments of a process defined in terms of the kernel $Q_{u,1}(\cdot; H)$. The second property leads to self-similarity with self-similarity parameter H.

Now, we have all tools to represent fBm in terms of a Brownian motion.

Lemma 3.19 *Let $B(u)$ $(u \in \mathbb{R})$ be a standard Brownian motion on $\mathbb{R}$. Define*

$$B_H(u) = \int_{-\infty}^{\infty} Q_{u,1}(x; H)\, dB(x). \tag{3.110}$$

Then $B_H(u)$ $(u \in \mathbb{R})$ is a fractional Brownian motion.

Proof On account of Lemma 3.18, the stochastic integral $\int_{-\infty}^{\infty} Q_{u,1}(x; H)\, dB(x)$ is normal with mean zero. Furthermore, the vector $(B_H(u_1), \ldots, B_H(u_q))$ is multivariate normal for any $u_1 < \cdots < u_q$.

Next, using properties (3.108) and (3.109) of the kernel $Q_{u,1}(\cdot; H)$, the process $B_H(\cdot)$ defined in (3.110) is H-self similar with stationary increments. Therefore, $B_H(0) = 0$ almost surely, and for $u < v$, the covariance function can be expressed as

$$E\big[B_H(u) B_H(v)\big] = \frac{1}{2}\big\{E\big[B_H^2(u)\big] + E\big[B_H^2(v)\big] - E\big[\big(B_H(v-u) - B_H(0)\big)^2\big]\big\}.$$

We now have to evaluate the covariance function of $B_H(u)$. For $u > 0$, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big(Q_{u,1}^{+}(x;H)\big)^2\,dx \\
&\quad= \int_{-\infty}^{\infty} \big[(u-x)_+^{H-1/2} - (-x)_+^{H-1/2}\big]^2\,dx \\
&\quad= \int_0^u (u-x)^{2H-1}\,dx + \int_{-\infty}^0 \big[(u-x)^{H-1/2} - (-x)^{H-1/2}\big]^2\,dx \\
&\quad= \frac{1}{2H}u^{2H} + u^{2H-1}\int_{-\infty}^0 \big[(1-x/u)^{H-1/2} - (-x/u)^{H-1/2}\big]^2\,dx.
\end{aligned}$$

The substitution $v = x/u$ yields

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big(Q_{u,1}^{+}(x;H)\big)^2\,dx \\
&\quad= \frac{1}{2H}u^{2H} + u^{2H}\int_{-\infty}^0 \big[(1-v)^{H-1/2} - (-v)^{H-1/2}\big]^2\,dv \\
&\quad= u^{2H}\left\{\frac{1}{2H} + \int_0^{\infty} \big[(1+v)^{H-1/2} - v^{H-1/2}\big]^2\,dv\right\} =: u^{2H}C_1^2(H).
\end{aligned}$$

Likewise,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big(Q_{u,1}^{-}(x;H)\big)^2\,dx \\
&\quad= \int_{-\infty}^{\infty} \big[(u-x)_-^{H-1/2} - (-x)_-^{H-1/2}\big]^2\,dx \\
&\quad= \int_u^{\infty} \big[(x-u)^{H-1/2} - x^{H-1/2}\big]^2\,dx - \int_0^u x^{2H-1}\,dx \\
&\quad= u^{2H}\left\{\int_1^{\infty} \big[(v-1)^{H-1/2} - v^{H-1/2}\big]^2\,dv - \frac{1}{2H}\right\} =: u^{2H}C_2^2(H).
\end{aligned}$$

Furthermore, for

$$\int_{-\infty}^{\infty} Q_{u,1}^{+}(x;H)\,Q_{u,1}^{-}(x;H)\,dx,$$

only integration over $0 < x < u$ contributes and yields

$$\begin{aligned}
-\int_0^u (u-x)^{H-1/2}x^{H-1/2}\,dx &= -u^{2H}\int_0^1 (1-v)^{H-1/2}v^{H-1/2}\,dv \\
&=: -u^{2H}C_3(H).
\end{aligned}$$

A similar computation holds for $u < 0$. Therefore,

$$\operatorname{var}\big(B_H(u)\big) = u^{2H}\big(c_1^2 C_1^2(H) + c_2^2 C_2^2(H) - 2c_1c_2C_3(H)\big) =: s^{2H} C_4(H),$$

and

$$E\big[B_H(u)B_H(v)\big] = \frac{1}{2}C_4(H)\big(|u|^{2H} + |v|^{2H} - |u-v|^{2H}\big),$$

as required in Definition 3.23. The constant $C_4(H)$ is equal to $\operatorname{var}(B_H(1))$. □

Example 3.45 If we set $c_1 = c_2 = 1$ in the kernel $Q_{u,1}(\cdot; H)$ (cf. (3.107)), then we obtain

$$B_H(u) = \int_{-\infty}^{\infty} \big(|u-x|^{H-1/2} - |x|^{H-1/2}\big)\, dB(x).$$

This is the so-called well-balanced representation of fBm.

If we set

$$c_1 = \frac{1}{C_1(H)} = \left\{\frac{1}{2H} + \int_0^{\infty} \big[(1+v)^{H-1/2} - v^{H-1/2}\big]^2\, dv\right\}^{-1/2}$$

and $c_2 = 0$, then the integral

$$B_H(u) = \frac{1}{C_1(H)} \int_{-\infty}^{\infty} \big[(u-x)_+^{H-1/2} - (-x)_+^{H-1/2}\big]\, dB(x) \tag{3.111}$$

defines a standard fractional Brownian motion with $\operatorname{var}(B_H(1)) = 1$. This representation was used in Mandelbrot and van Ness (1968).

Another representation of fBm is given in Lévy (1953):

$$B_H(u) = \frac{1}{\Gamma(H+\frac{1}{2})} \int_0^u (u-x)^{H-\frac{1}{2}}\, dB(x).$$

This is not a standard fractional Brownian motion since

$$E\big[B_H^2(1)\big] = \frac{1}{2H\Gamma^2(H+\frac{1}{2})}.$$

However, this type of representation is connected to fractional integration, see Sect. 3.7.3.

3.7.1.2 Integral Representation of the Hermite–Rosenblatt Process

To define the Hermite–Rosenblatt process, we have to extend the definition (3.105) of a stochastic integral to the multivariate case. For simplicity, let I be a compact interval and assume that it can be partitioned such that $I = \bigcup I_l$, where $I_l, l = 1, \ldots, k$ are disjoint subintervals. Recall that a function g defined on I is called simple if

$g(x) = \sum_{l=1}^{k} a_l 1\{x \in I_l\}$, where a_l $(l = 1, \ldots, k)$ are real numbers. A function g defined on I^2 is called simple if

$$g(x_1, x_2) = \begin{cases} a_{jl} & \text{if } (x_1, x_2) \in I_j \times I_l, \ j \neq l, \\ 0 & \text{if } (x_1, x_2) \in I_l \times I_l. \end{cases}$$

In particular, the function g vanishes on the "diagonal" (I_l, I_l), $l = 1, \ldots, k$. More generally, a simple function $g : I^m \to \mathbb{R}$ vanishes whenever there are two indices $1 \le i_1 < i_2 \le m$ such that $(x_{i_1}, x_{i_2}) \in I_l \times I_l$, $l = 1, \ldots, k$.

For simple functions $g : I^2 \to \mathbb{R}$, the multiple integral is defined as

$$\sum_{j,l} a_{jl} M(I_j) M(I_l),$$

where M is the Gaussian random measure as in (3.105). This integral extends to a bivariate Wiener–Itô integral

$$I_2(g) = \int_{\mathbb{R}^2} g(x_1, x_2)\, dM(x_1)\, dM(x_2).$$

Since the simple functions used in the construction of the integral vanish on the diagonal, the above integration is in fact defined on $\mathbb{R}^2$ with removed hyperplane $x_1 = x_2$. Furthermore, the integral is well defined if

$$\int_{\mathbb{R}^2} g^2(x_1, x_2)\, dx_1\, dx_2 < \infty. \tag{3.112}$$

Recall (see Lemma 3.18) that the integral $I(g)$ in (3.105) has mean zero. Also, since $E[M(A)M(B)] = 0$ for any disjoint sets A, B, we have $E[I_2(g)] = 0$. We also see why we have to remove the diagonal: if $x_1 = x_2$, then, informally, $E[dM(x_1)\, dM(x_1)] = dx_1$, so that the diagonal would contribute $\int_{\mathbb{R}} g(x_1, x_1)\, dx_1$ yielding a non-zero mean in general.

Lemma 3.20 *Assume that $g, g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}$ are square integrable as in (3.112). Then $E[I_2(g) = 0]$, and*

$$cov\big(I_2(g_1), I_2(g_2)\big) = \int_{\mathbb{R}^2} g^2(x_1, x_2)\, dx_1\, dx_2.$$

Proof We explained why the multiple integral has zero mean. It remains to verify the covariance formula. We have

$$\begin{aligned} &cov\big(I_2(g_1), I_2(g_2)\big) \\ &\quad = E\left[\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} g(x_1, x_2) g(y_1, y_2)\, dM(x_1)\, dM(x_2)\, dM(y_1)\, dM(y_2)\right]. \end{aligned}$$

Since the integration excludes the hyperplanes $x_1 = x_2$ and $y_1 = y_2$ and since the random measure M has the property $E[M(A)M(B)] = 0$ for disjoint sets A and B,

the only contribution comes from integrating over $x_1 = y_1$ and $x_2 = y_2$. Recalling that $\mathrm{var}(dM(x)) = dx$, we obtain the formula for the covariance. □

More generally, for square-integrable functions $g(x_1, \ldots, x_m)$, we consider

$$I_m(g) = \int_{\mathbb{R}^m} g(x_1, \ldots, x_m)\, dM(x_1) \cdots dM(x_m), \tag{3.113}$$

where $\int_{\mathbb{R}^m}$ is the multiple Wiener–Itô integral over $\mathbb{R}^m$, disregarding hyperplanes $x_i = x_j$, $i \neq j$, $i, j = 1, \ldots, m$. Again, by the association $M([0, x]) = B(x)$ we can rewrite the integral as

$$I_m(g) = \int_{\mathbb{R}^m} g(x_1, \ldots, x_m)\, dB(x_1) \cdots dB(x_m). \tag{3.114}$$

We are now ready to define a Hermite–Rosenblatt process.

Definition 3.24 Let $B(\cdot)$ denote a standard Brownian motion on $\mathbb{R}$ and assume that $1 - \frac{1}{2m} < H_0 < 1$. Define, for $u \geq 0$,

$$Z_{m,H_0}(u) = K(m, H_0) \times \left\{ \int_{-\infty}^{\infty} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_{m-1}} \left(\int_0^u \prod_{i=1}^m (s - x_i)_+^{H_0 - \frac{3}{2}}\, ds \right) dB(x_m) \cdots dB(x_1) \right\},$$

where

$$K^2(m, H_0) = \frac{m!(2m(H_0 - 1) + 1)(m(H_0 - 1) + 1)}{(\int_0^\infty (x + x^2)^{H_0 - \frac{3}{2}}\, dx)^m}.$$

The process $Z_{m,H_0}(u)$, $(u \geq 0)$, is called a Hermite–Rosenblatt process. The choice of the constant assures that $E[Z^2_{m,H_0}(1)] = 1$; see the computation below.

The integral above is to be interpreted as iteration of the univariate Wiener–Itô integrals.

We note that the function

$$(x_1, \ldots, x_m) \to \int_0^u \prod_{i=1}^m (s - x_i)_+^{H_0 - \frac{3}{2}}\, ds$$

is symmetric. We therefore may write alternatively

$$Z_{m,H_0}(u) = \frac{K(m, H_0)}{m!} \left\{ \int_{\mathbb{R}^m} \left(\int_0^u \prod_{i=1}^m (s - x_i)_+^{H_0 - \frac{3}{2}}\, ds \right) dB(x_m) \cdots dB(x_1) \right\}. \tag{3.115}$$

Here, the integral is understood as the multiple Wiener–Itô integral defined in (3.113).

As we will show below, the process $Z_{m,H_0}(u)$ is self-similar with self-similarity parameter $H = m(H_0 - 1) + 1$. In particular, for $m = 1$ and $H_0 = H$, we obtain for $u \geq 0$,

$$\begin{aligned} Z_{1,H}(u) = Z_{1,H}(u) &= K(1,H)\left\{\int_{-\infty}^{\infty}\left(\int_0^u (s-x)_+^{H-\frac{3}{2}}\,ds\right)dB(x)\right\} \\ &= K(1,H)\left\{\int_{-\infty}^{0}\left(\int_0^u (s-x)^{H-\frac{3}{2}}\,ds\right)dB(x)\right. \\ &\quad \left. + \int_0^u\left(\int_x^u (s-x)^{H-\frac{3}{2}}\,ds\right)dB(x)\right\}. \end{aligned}$$

Since

$$\int_0^u (s-x)^{H-\frac{3}{2}}\,ds = \frac{1}{H-\frac{1}{2}}\left\{(u-x)^{H-\frac{1}{2}} - (-x)^{H-\frac{1}{2}}\right\}$$

and

$$\int_x^u (s-x)^{H-\frac{3}{2}}\,ds = \frac{1}{H-\frac{1}{2}}(u-x)^{H-\frac{1}{2}},$$

we conclude

$$\begin{aligned} Z_{1,H}(u) &= \frac{K(1,H)}{H-\frac{1}{2}}\left\{\int_{-\infty}^{0}\left((u-x)^{H-\frac{1}{2}} - (-x)^{H-\frac{1}{2}}\right)dB(x)\right. \\ &\quad \left. + \int_0^u (u-x)^{H-\frac{1}{2}}\,dB(x)\right\} \\ &= \frac{K(1,H)}{H-\frac{1}{2}}\int_{-\infty}^{\infty}\left((u-x)_+^{H-\frac{1}{2}} - (-x)_+^{H-\frac{1}{2}}\right)dB(x). \end{aligned}$$

We recognize the Mandelbrot–van Ness representation given in (3.111).

Now, we will establish some properties of the process $Z_{m,H_0}(\cdot)$. First, we will verify that it is self-similar. Next, we will identify its covariance structure. Finally, we will justify that $E[Z_{m,H_0}^2(1)] = 1$.

Lemma 3.21 *The process $Z_{m,H_0}(u)$ $(u \geq 0)$ is H-self-similar with*

$$H = m(H_0 - 1) + 1.$$

Proof We conduct the proof just for $m = 2$, but the general case is analogous. First, we write $(K(m, H_0))^{-1} Z_{m,H_0}(cu)$ as

$$\begin{aligned} &\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{x_1}\left(\int_0^{cu}\prod_{i=1}^{2}(s-x_i)_+^{H_0-\frac{3}{2}}\,ds\right)dB(x_2)\,dB(x_1)\right\} \\ &= c^{2(H_0-\frac{3}{2})}\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{x_1}\left(\int_0^{cu}\prod_{i=1}^{2}\left(\frac{s}{c}-\frac{x_i}{c}\right)_+^{H_0-\frac{3}{2}}\,ds\right)dB(x_2)\,dB(x_1)\right\} \end{aligned}$$

$$= c^{2(H_0-\frac{3}{2})+1}\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{x_1}\left(\int_0^u \prod_{i=1}^{2}\left(s-\frac{x_i}{c}\right)_+^{H_0-\frac{3}{2}} ds\right) dB(x_2)\, dB(x_1)\right\}$$

$$= c^{2(H_0-\frac{3}{2})+1}\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{x_1}\left(\int_0^u \prod_{i=1}^{2}(s-x_i)_+^{H_0-\frac{3}{2}} ds\right) dB(cx_2)\, dB(cx_1)\right\}$$

$$\stackrel{\mathrm{d}}{=} c^{2(H_0-\frac{3}{2})+\frac{2}{2}} Z_{2,H_0}(u) = c^{m(H_0-\frac{3}{2})+\frac{m}{2}} Z_{2,H_0}(u) = c^H Z_{2,H_0}(u),$$

where in the second last equality we used the self-similarity property of Brownian motion, that is $B(c\cdot)$ has the same distribution as $c^{1/2}B(\cdot)$. □

To study further properties of the Hermite–Rosenblatt process, let us consider a process

$$Z_m(u) = \int_{\mathbb{R}^m}\left(\int_0^u \prod_{i=1}^{m} g(s,x_i)\, ds\right) dB(x_m)\cdots dB(x_1). \tag{3.116}$$

Again, we will do computations for $m=2$, but we keep a general m in the notation. It is assumed that the real-valued function $g:\mathbb{R}\times\mathbb{R}$ is such that

$$\int_{-\infty}^{\infty} g^2(s,x)\, dx < \infty.$$

This implies that

$$r(s_1,s_2) := \int_{-\infty}^{\infty} g(s_1,x)g(s_2,x)\, dx < \infty$$

and the process $Z_m(\cdot)$ is well defined. Using the function $r(\cdot,\cdot)$, McKean (1973) gives the following representation for $Z_m(\cdot)$:

$$Z_m(u) = \int_0^u r^{m/2}(s,s) H_m\big(X(s)\big)\, ds,$$

where

$$X(s) = \frac{1}{r^{1/2}(s,s)}\int_{-\infty}^{\infty} g(s,x)\, dB(x),$$

and H_m is the mth Hermite polynomial. In other words, in the definition of $Z_m(\cdot)$, the multiple Wiener–Itô integral is replaced by the standard Itô integral. We note that

$$E\big[X^2(s)\big] = \frac{1}{r(s,s)}\int_{-\infty}^{\infty} g^2(s,x)\, dx = 1$$

and

$$E\big[X(s_1)X(s_2)\big] = \frac{r(s_1,s_2)}{r^{1/2}(s_1,s_1)r^{1/2}(s_2,s_2)}.$$

Thus, using the formula for the covariance of Hermite polynomials (see Lemma 3.5), we have

$$E\big[H_m\big(X(s_1)\big)H_m\big(X(s_2)\big)\big] = m!\left(\frac{r(s_1,s_2)}{r^{1/2}(s_1,s_1)r^{1/2}(s_2,s_2)}\right)^m.$$

Consequently, the covariance structure of the process $Z_m(\cdot)$ is given by

$$E\big[Z_m(u_1)Z_m(u_2)\big] = m!\int_0^{u_1}\int_0^{u_2} r^m(s_1,s_2)\,ds_2\,ds_1.$$

Now, we would like to apply these computations to

$$g(s,x) = (s-x)_+^{H_0-\frac{3}{2}},$$

so that Z_m in (3.116) becomes (up to the constant $K(m,H_0)/m!$) Z_{m,H_0} in representation (3.115). The problem is that the function $x \to g(s,x) = c(s-x)_+^{H_0-\frac{3}{2}}$ is not square integrable w.r.t. Lebesgue measure. However, the functions $g_\varepsilon(s,\cdot) := g(s+\varepsilon,\cdot)$ $(\varepsilon > 0)$ are square integrable and tend monotonically to $g(s,\cdot)$ as $\varepsilon \to 0$. This approach guarantees that $g(s_1,\cdot)g(s_2,\cdot)$ is integrable. This in turn implies the existence of Z_m (see Lemma 2.3 in Taqqu 1978). Consequently, for our choice of $g(s,x)$, we have

$$\begin{aligned} r(s_1,s_2) &= \int_{-\infty}^{\min(s_1,s_2)} (s_1-x)^{H_0-\frac{3}{2}}(s_2-x)^{H_0-\frac{3}{2}}\,dx \\ &= \int_0^\infty x^{H_0-\frac{3}{2}}\big(|s_2-s_1|+x\big)^{H_0-\frac{3}{2}}\,dx \\ &= \underbrace{\int_0^\infty \big(x+x^2\big)^{H_0-\frac{3}{2}}\,dx}_{=:C}\,|s_2-s_1|^{-2(1-H_0)}. \end{aligned}$$

Thus, the covariance structure of the process $Z_m(\cdot)$ is given by

$$E\big[Z_m(u_1)Z_m(u_2)\big] = m!C^m\int_0^{u_1}\int_0^{u_2} |s_2-s_1|^{-2m(1-H_0)}\,ds_2\,ds_1.$$

In particular, for $u_1 = u_2 = 1$, we have

$$E\big[Z_{m,}^2(1)\big] = m!C^m\frac{1}{(2m(H_0-1)+1)(m(H_0-1)+1)}.$$

Now, $Z_{m,H_0}(u)$ in (3.115) equals $\frac{K(m,H_0)}{m!}Z_m(u)$. It is straightforward to verify that $E[Z_{m,H_0}^2(1)] = 1$.

3.7.1.3 Spectral Representation of fBm and Hermite–Rosenblatt Processes

Let M_1 and M_2 be two independent real-valued Gaussian measures as in (3.104). Define by

$$\tilde{M}(A) = \frac{1}{\sqrt{2}}\big(M_1(A) + iM_2(A)\big) \tag{3.117}$$

a complex-valued Gaussian random measure. In particular, for a set A, we have $E[\tilde{M}(A)] = 0$ and $E[|\tilde{M}(A)|^2] = |A|$, where $|\cdot|$ stands for the Lebesgue measure. The goal of this section is to find a spectral representation of fBm and the Hermite–Rosenblatt process.

We start by arguing that a standard Brownian motion $B(u)$ $(u \geq 0)$ can be written as

$$B(u) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{e^{i\lambda u} - 1}{i\lambda}\, d\tilde{M}(\lambda) = \int_{\mathbb{R}} \tilde{h}_u(\lambda)\, d\tilde{M}(\lambda), \tag{3.118}$$

where $\tilde{h}_u(\lambda)$ is the Fourier transform of the function $h_u(s) = 1\{0 \leq u \leq s\}$:

$$\tilde{h}_u(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\lambda s} 1\{0 \leq u \leq s\}\, ds.$$

Indeed, Lemma 3.18 implies that a random variable $B(u)$ defined in (3.118) is Gaussian. The same applies to any vector $(B(u_1), \ldots, B(u_q))$. Furthermore, we have by Parseval's identity

$$\begin{aligned} E\big[B(u)B(v)\big] &= \int_{\mathbb{R}} \tilde{h}_u(\lambda)\overline{\tilde{h}_v(\lambda)}\, d\lambda \\ &= \int_{\mathbb{R}} h_u(x)h_v(x)\, dx = \int_0^{\min(v,u)} dx = \min\{u, v\}. \end{aligned}$$

We recognize the covariance function of a standard Brownian motion. Consequently, the process $B(u)$ $(u \geq 0)$ is indeed a standard Brownian motion.

Now, we recall that $g : \mathbb{R}^m \to \mathbb{C}$ is symmetric if it is invariant under permutation of its indices. Furthermore, it is even if $g(x_1, \ldots, x_m) = \overline{g(-x_1, \ldots, -x_m)}$. Similarly to (3.113), for each symmetric, even, complex-valued function $g : \mathbb{R}^m \to \mathbb{C}$, one can define

$$\int_{\mathbb{R}^m} g(\lambda_1, \ldots, \lambda_m)\, d\tilde{M}(\lambda_1) \cdots d\tilde{M}(\lambda_m). \tag{3.119}$$

The integration in (3.119) disregards hyperplanes $|\lambda_i| = |\lambda_j|$, $i \neq j$.

Now, we are ready to establish the following representation of Hermite–Rosenblatt processes introduced in Definition 3.24.

Proposition 3.1 *Assume that* $1 - \frac{1}{2m} < H_0 < 1$. *The process* $Z_{m,H_0}(u)$ *has the representation*

$$Z_{m,H_0}(u) = K_1(m, H_0) \int_{\mathbb{R}^m} \frac{e^{i(\lambda_1+\cdots+\lambda_m)u-1}}{i(\lambda_1 + \cdots + \lambda_m)} \prod_{j=1}^{m} \frac{1}{|\lambda_j|^{H_0-\frac{1}{2}}} d\tilde{M}(\lambda_1) \cdots d\tilde{M}(\lambda_m),$$

where

$$K_1^2(m, H_0) = \frac{(m(H_0-1)+1)(2m(H_0-1)+1)}{m!\{2\Gamma(2-2H_0)\sin\pi(H_0-\frac{1}{2})\}^m}.$$

To justify this formula, let $\tilde{g}$ be the Fourier transform of g, i.e.

$$\tilde{g}(\lambda_1, \ldots, \lambda_m) = \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} \exp\left(i \sum_{j=1}^{m} \lambda_j x_j\right) g(x_1, \ldots, x_m)\, dx_1 \cdots dx_m.$$

We have the following relation between the multiple Wiener–Itô integral defined in (3.113) and the integral in (3.119). This result was proven in Taqqu (1978).

Lemma 3.22 *Assume that* $g : \mathbb{R}^m \to \mathbb{R}$ *is a symmetric, even and square-integrable real-valued function. Then*

$$\int_{\mathbb{R}^m} g(x_1, \ldots, x_m)\, dB(x_1) \cdots dB(x_m) = \int_{\mathbb{R}^m} \tilde{g}(\lambda_1, \ldots, \lambda_m)\, d\tilde{M}(\lambda_1) \cdots d\tilde{M}(\lambda_m),$$

in the sense of equality of finite-dimensional distributions.

Proof We conduct a proof for $m = 1$ only. If $g \in L^2(\mathbb{R}, \text{Leb})$, then

$$g(x) = \sum_{k=0}^{\infty} c_k \psi_k(x),$$

where ψ_k $(k \geq 0)$ is a complete orthonormal basis in $L^2(\mathbb{R}, \text{Leb})$. Therefore, we can write

$$\int g(x)\, dB(x) = \sum_{k=0}^{\infty} c_k \int \psi_k(x)\, dB(x).$$

On the other hand, $\tilde{\psi}_k(\lambda) = (2\pi)^{-1/2} \int e^{i\lambda x} \psi(x)\, dx$ $(k \geq 0)$ is an orthonormal basis in the set of symmetric, even, complex-valued functions. Furthermore, applying the Fourier transform to g, we obtain

$$\tilde{g}(\lambda) = \sum_{k=0}^{\infty} c_k \tilde{\psi}_k(\lambda)$$

and thus

$$\int \tilde{g}(\lambda)\, d\tilde{M}(\lambda) = \sum_{k=0}^{\infty} c_k \int \tilde{\psi}_k(\lambda)\, d\tilde{M}(\lambda).$$

Now, the random variables $Y_k := \int \psi_k(x)\, dB(x)$ $(k \geq 0)$ are centred Gaussian and $E[Y_k Y_l] = \int \psi_k(x)\psi_l(x)\, dx$. Furthermore, the random variables $\tilde{Y}_k = \int \tilde{\psi}_k(\lambda)\, d\tilde{M}(d\lambda)$ $(k \geq 0)$ are also centred Gaussian, and by Parseval's identity we have

$$E[\tilde{Y}_k \overline{\tilde{Y}_l}] = \int \tilde{\psi}_k(\lambda)\overline{\tilde{\psi}_l(\lambda)}\, d\lambda = \int \psi_k(\lambda)\psi_l(\lambda)\, d\lambda.$$

Therefore, the two sequences Y_k and $\tilde{Y}_k$ $(k \geq 0)$ have the same distribution. It follows that the integrals $\int g(x)\, dB(x)$ and $\int \tilde{g}(\lambda)\, d\tilde{M}(\lambda)$ also have the same distribution. □

Now, we would like to apply this lemma to

$$g(x_1, \ldots, x_m; u) = \int_0^u \prod_{i=1}^m (s - x_i)_+^{H_0 - \frac{3}{2}}\, ds = \int_{\mathbb{R}} 1_{\{0<s<u\}} \prod_{i=1}^m (s - x_i)_+^{H_0 - \frac{3}{2}} ds. \tag{3.120}$$

However, since $\frac{1}{2} < H_0 < 1$, the function $u^{H_0 - \frac{3}{2}} 1\{u > 0\}$ is not in $L^2(\mathbb{R}, \text{Leb})$, nor it is in $L^1(\mathbb{R}, \text{Leb})$. We will overcome this problem by applying a truncation argument.

Proof of Proposition 3.1 In the proof we will consider the case $m = 1$ only. Let $g_T(x; u) = g(x; u)1\{|x| < T\}$, where $g(x; u)$ is defined in (3.120) with $m = 1$. Its Fourier transform is given by

$$\tilde{g}_T(\lambda; u) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\lambda x} \left(\int_{\mathbb{R}} 1\{0 < s < u\}(s - x)_+^{H_0 - \frac{3}{2}}\, ds \right) 1\{|x| < T\}\, dx.$$

Since $g_T(x; u) \to g(x; u)$ as $T \to \infty$ pointwise, we also have $\tilde{g}_T(\lambda; u) \to \tilde{g}(\lambda; u)$. Substituting $x \to s - x =: v$, we obtain

$$\begin{aligned} \tilde{g}_T(\lambda; u) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\lambda(s-v)} \left(\int_{\mathbb{R}} 1\{0 < s < u\}(v)_+^{H_0 - \frac{3}{2}}\, ds \right) 1\big\{|s - v| < T\big\}\, dv \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-i\lambda v}(v)_+^{H_0 - \frac{3}{2}} \left(\int_{\mathbb{R}} e^{i\lambda s} 1\{0 < s < u\}\, ds \right) 1\big\{|s - v| < T\big\}\, dv. \end{aligned}$$

Letting $T \to \infty$, we obtain

$$\tilde{g}(\lambda; u) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-i\lambda v}(v)_+^{H_0 - \frac{3}{2}} \left(\int_{\mathbb{R}} e^{i\lambda s} 1\{0 < s < u\}\, ds \right) dv.$$

This argument requires exchange of integration with $\lim_{T\to\infty}$. This is allowed since $g_T(\lambda; u)$ is uniformly bounded in T (see Taqqu 1979 for details). Thus, recalling

that the Fourier transform of $h_u(s) = 1\{0 < s < u\}$ is

$$\tilde{h}_u(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\lambda s} 1\{0 \le u \le s\}\, ds = \frac{1}{\sqrt{2\pi}} \frac{e^{i\lambda u} - 1}{i\lambda},$$

we have

$$\tilde{g}(\lambda; u) = \tilde{h}_u(\lambda) \left(\int_0^{\infty} e^{-i\lambda v} v^{H_0 - \frac{3}{2}}\, dv \right) = \frac{e^{i\lambda u - 1}}{i\lambda} |\lambda|^{\frac{1}{2} - H_0} \frac{\Gamma(H_0 - \frac{1}{2})}{\sqrt{2\pi}}.$$

By Lemma 3.22,

$$\int_{\mathbb{R}} g(x; u)\, dB(x) = \frac{\Gamma(H_0 - \frac{1}{2})}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{e^{i\lambda u - 1}}{i\lambda} \frac{1}{|\lambda|^{H_0 - \frac{1}{2}}}\, d\tilde{M}(\lambda).$$

Therefore, $Z_{1,H_0}(u)$ defined in Definition 3.24 or in (3.115) equals in the sense of finite-dimensional distributions to

$$Z_{1,H_0}(u) = \frac{\Gamma(H_0 - \frac{1}{2})}{\sqrt{2\pi}} \frac{K(1, H_0)}{1!} \int_{\mathbb{R}} \frac{e^{i\lambda u - 1}}{i\lambda} \frac{1}{|\lambda|^{H_0 - \frac{1}{2}}}\, \tilde{W}(d\lambda).$$

The constant can be further simplified further to

$$\frac{\Gamma(H_0 - \frac{1}{2})}{\sqrt{2\pi}} \frac{K(1, H_0)}{1!} = \left\{ \frac{\Gamma^2(H_0 - \frac{1}{2})(2H_0 - 1)H_0}{(2\pi)^{2/2} (\int_0^{\infty} (x + x^2)^{H_0 - \frac{3}{2}}\, dx)^1} \right\}^{1/2} = K_1(1, H_0)$$

by Eq. (D.5) (see Appendix D).

For a general m,

$$\int_{\mathbb{R}^m} g(x_1, \ldots, x_m; u)\, dB(x_1) \cdots dB(x_m)$$
$$= \left(\frac{\Gamma(H_0 - \frac{1}{2})}{\sqrt{2\pi}} \right)^m \int_{\mathbb{R}^m} \frac{e^{i(\lambda_1 + \cdots + \lambda_m)u - 1}}{i(\lambda_1 + \cdots + \lambda_m)} \prod_{j=1}^{m} \frac{1}{|\lambda_j|^{H_0 - \frac{1}{2}}}\, d\tilde{M}(\lambda_1) \cdots d\tilde{M}(\lambda_m).$$

From this the constant $K_1(m, H_0)$ can be easily computed. □

3.7.2 *Linear Fractional Stable Motion*

3.7.2.1 Poisson Processes

Let E_j be a sequence of independent exponential random variables with mean 1. Define $\Gamma_n = \sum_{j=1}^{n} E_j$ and

$$N = \sum_{k=1}^{\infty} \delta_{\Gamma_k},$$

where $\delta_x(\cdot)$ is the Dirac measure, that is $\delta_x(A) = 1$ if x belongs to the set A and zero otherwise. In other words, $N(A)$ counts the number of points Γ_j that fall in A. This formula defines a Poisson process on $[0, \infty)$. Its mean (intensity) measure is given by $d\lambda(x) = \lambda\, dx$, where λ is a constant. In other words, $E[N(A)] = \lambda \cdot |A|$, where $|A|$ is the Lebesgue measure of A. Such a Poisson process is called homogeneous.

Consider now a transformation $T : (0, \infty) \to (0, \infty)$. Then $\sum_{k=1}^{\infty} \delta_{T(\Gamma_k)}$ is still a Poisson process, but with an intensity measure given by $\lambda \circ T^{-1}$.

Example 3.46 Assume that $T(u) = u^{-1/\alpha}$. Then

$$\sum_{k=1}^{\infty} \delta_{\Gamma_k^{-1/\alpha}} =: \sum_{k=1}^{\infty} \delta_{T_k},$$

is the representation of a Poisson process on $(0, \infty)$ with intensity measure

$$d\tilde{\lambda}(x) = \alpha x^{-(\alpha+1)}, \quad x > 0.$$

Furthermore, if U_n, $n \geq 1$, is a sequence of i.i.d. standard uniform random variables, then

$$N = \sum_{k=1}^{\infty} \delta_{(U_k, T_k)} \tag{3.121}$$

is the representation of a Poisson processs on $[0, 1] \times (0, \infty)$ with intensity measure Leb $\times\, \tilde{\lambda}$.

3.7.2.2 Stable Random Variables

There are several equivalent definitions of stable random variables.

Definition 3.25 A random variable X is stable if for any $n \geq 2$, there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that

$$X_1 + \cdots + X_n \stackrel{d}{=} c_n X + d_n,$$

where $X_1, X_2, \ldots$ are independent copies of X. Necessarily, $c_n = n^{1/\alpha}$, where $\alpha \in (0, 2]$.

Equivalently, stable random variables are characterized in terms of *domains of attraction*:

Definition 3.26 A random variable X is stable if there exist an i.i.d. sequence Y_t, $t \in \mathbb{N}$ and constants $c_n > 0$, $d_n \in \mathbb{R}$ such that

$$\frac{Y_1 + \cdots + Y_n}{c_n} - d_n \stackrel{d}{\to} X.$$

The characteristic function of a stable random variable X is given by

$$Ee^{i\theta X} = \begin{cases} \exp(-\eta^\alpha |\theta|^\alpha (1 - i\beta \text{sign}(\theta) \tan \frac{\pi\alpha}{2})) & \text{if } \alpha \neq 1, \\ \exp(-\eta|\theta|(1 + i\beta \frac{2}{\pi} \text{sign}(\theta) \ln(\theta)) + i\mu\theta) & \text{if } \alpha = 1. \end{cases}$$

Here, $0 < \alpha \leq 2$, $\eta > 0$ is the scale parameter, $-1 \leq \beta \leq 1$ is skewness, and $\mu \in \mathbb{R}$ is a shift parameter. We write $X \sim S_\alpha(\eta, \beta, \mu)$. In particular, X is symmetric α-stable (written as $X \sim S\alpha S$) if $X \sim S_\alpha(\eta, 0, 0)$.

If $\alpha \in (0, 2)$, then stable random variables are heavy tailed. Indeed, if $X \sim S_\alpha(\eta, \beta, \mu)$, then

$$\lim_{x\to\infty} x^\alpha P(X > x) = C_\alpha \frac{1+\beta}{2} \eta^\alpha, \qquad \lim_{x\to\infty} x^\alpha P(X < -x) = C_\alpha \frac{1-\beta}{2} \eta^\alpha,$$

where

$$C_\alpha = \left(\int_0^\infty x^{-\alpha} \sin x \right)^{-1}.$$

Stable Stochastic Process A stochastic process is stable if all linear combinations are stable.

Lévy Measure Let $\|\cdot\|$ be the Euclidean norm. A measure λ on $R^d \setminus \{0\}$ is called a Lévy measure if

$$\int_{0 \leq \|x\| < c} \|x\|^2 d\lambda(x) < \infty$$

for all $c \in (0, \infty)$.

Example 3.47 The measure

$$d\lambda(x) = \alpha \left[\frac{1+\beta}{2} x^{-(\alpha+1)} 1\{0 < x < \infty\} + \frac{1-\beta}{2} (-x)^{-(\alpha+1)} 1\{-\infty < x < 0\} \right] dx$$

is a Lévy measure. Of course, if $\beta = 1$, then we obtain the measure $\tilde{\lambda}$ in Example 3.46.

3.7.2.3 Lévy Processes

A stochastic process $Z(u)$ is called a Lévy process if $Z(0) = 0$, sample paths are in $D[0, 1]$ and Z has stationary and independent increments.

Itô Representation Recall the representation $N = \sum_{k=1}^{\infty} \delta_{(U_k, T_k)}$ (see (3.121)). This Poisson process has the mean measure Leb $\times \tilde{\lambda}$. Its natural extension is a Poisson process with the Lévy measure Leb $\times \lambda$ (see Example 3.47). The simplest α-stable Lévy motion on $[0, 1]$ can be constructed as

$$Z(u) = \sum_{U_k \le u} T_k.$$

In general, an α-stable Lévy motion on $[0, 1]$ can be represented as

$$Z(u) = \lim_{\varepsilon \to 0} \int_0^u \int_{\varepsilon}^{\infty} u\big(N(ds, du) - ds\lambda(du)\big).$$

α-Stable Lévy Motion Based on this representation, we say that $Z(\cdot)$ is a Lévy process with Lévy measure λ. Such a process is called an α-stable Lévy motion, denoted in this book by $\tilde{Z}_\alpha(\cdot)$. If $\beta = 0$ in the definition of the measure λ, then the process is called a symmetric α-stable Lévy motion (denoted as $S\alpha S$).

Using the language of stable random variables, a process $Z_\alpha(s)$ is called an α-stable Lévy motion if it has independent increments, $Z_\alpha(0) = 0$ and $Z_\alpha(s') - Z_\alpha(s) \sim S_\alpha((s' - s)^{1/\alpha}, \beta, 0)$. If $\beta = 0$, then the process is $S\alpha S$. Note also that α-stable Lévy motions are $1/\alpha$-self-similar.

3.7.2.4 Stable Random Measures and Stochastic Integrals

Recall Eq. (3.104). It defines a stable random measure that is self-similar with parameter $1/2$.

Let m be a measure on a space E, and let $\beta : E \to [-1, 1]$. Let M be a random measure defined on sets A such that $m(A) < \infty$. In other words:

- $(M(A_1), \ldots, M(A_k))$ is a random vector;
- If A_j are disjoint, then $M(\bigcup A_j) = \sum_j M(A_j)$.

We say that M is an independently scattered random measure if for disjoint sets $A_1, \ldots, A_k$, the random variables $M(A_1), \ldots, M(A_k)$ are independent. In particular, the Gaussian random measure defined in (3.104) is independently scattered.

Assume that $\alpha \in (0, 2)$. We say that M is an independently scattered α-stable random measure with control measure m if

$$M(A) \sim S_\alpha\left(\big(m(A)\big)^{1/\alpha}, \frac{\int \beta(x) m(dx)}{m(A)}, 0\right).$$

Recall that a Brownian motion can be defined in terms of a Gaussian random measure, cf. (3.106). An analogous result holds for Lévy processes.

Example 3.48 Let M be an α-stable random measure with $m = \text{Leb}$ and $\beta(x) = \beta$. Then $Z_\alpha(s) = M([0, s])$, $s \ge 0$, defines α-stable Lévy motion.

However, random measures are also building blocks for other processes, e.g. via stochastic integrals. We noted this in the construction of a fractional Brownian motion. To extend this to stable processes, let $L^\alpha(E, m)$ be a class of functions such that

$$\int \left| f(x) \right|^\alpha dm(x) < \infty.$$

Then

$$I(f) = \int f(x)\, dM(x) \tag{3.122}$$

is well defined. In particular, if $f(x) = \sum_{j=1}^k f_j 1_{\{x \in A_j\}}$, then

$$I(f) = \sum_{j=1}^k f_j M(A_j). \tag{3.123}$$

The integral $I(f)$ is still a stable random variable. In particular, $I(f) \sim S_\alpha(\eta_f, \beta_f, 0)$, where

$$\eta_f = \left(\left| f(x) \right|^\alpha m(dx) \right)^{1/\alpha}.$$

Example 3.49 In this example, we assume that M is an independently scattered α-stable random measure. Let $f \in L^\alpha(\mathbb{R}, m)$. Then

$$Z(u) = \int_{-\infty}^{\infty} f(u - x)\, dM(x)$$

is called an $S\alpha S$ moving average. The process $Z(s)$ is stationary.

Furthermore, let

$$g_t(x) = f(t - x) - f(t - 1 - x).$$

Then

$$X_t = Z(t) - Z(t - 1) = \int_{-\infty}^{\infty} g_t(x)\, dM(x)$$

defines an α-stable stationary process.

3.7.2.5 Linear Fractional Stable Motion (LFSM)

Fractional Brownian motion has been represented as an integral with respect to a Brownian motion. The integrand was defined in terms of the kernel that appeared in (3.107).

Let M be an independently scattered α-stable measure with the Lebesgue measure as the control measure. We define similarly an LFSM as

$$Z_{H,\alpha}(u) = \int_{-\infty}^{\infty} Q_{u,1}(x; H, \alpha)\, dM(x)$$

or equivalently as

$$Z_{H,\alpha}(u) = \int_{-\infty}^{\infty} Q_{u,1}(x; H, \alpha)\, dZ_{\alpha}(x),$$

where $Z_\alpha(\cdot)$ is an α-stable Lévy motion, and

$$Q_{u,1}(x; H, \alpha) = c_1\big[(u-x)_+^{H-1/\alpha} - (-x)_+^{H-1/\alpha}\big] + c_2\big[(u-x)_-^{H-1/\alpha} - (-x)_-^{H-1/\alpha}\big].$$

The integral is well defined as long as $H > 1/\alpha$. Indeed, for example, the first integrand $(u-x)_+^{H-1/\alpha} - (-x)_+^{H-1/\alpha}$ behaves like $(H-1/\alpha)(-x)^{H-1/\alpha-1}$ as $x \to -\infty$. A function $(-x)^{-(1+1/\alpha-H)}$ $(x \to -\infty)$ is integrable with power α if $\alpha(1+1/\alpha-H) > 1$, that is $H < 1$. The function $(u-x)_+^{H-1/\alpha}$ behaves like $x^{H-1/\alpha}$ as $x \to u$, and then is clearly integrable if $H > 1/\alpha$.

The process is self-similar with stationary and dependent increments. As in the case of fractional Brownian motion, this representation is not unique. For example,

$$Z(u) = \int_{-\infty}^{\infty} \big(|u-x|^{H-1/\alpha} - |x|^{H-1/\alpha}\big)\, dM(x)$$

is called a well-balanced $S\alpha S$ linear fractional stable motion.

3.7.3 Fractional Calculus

Fractional calculus is a useful tool when dealing with limit theorems for long-memory processes. In particular, it provides an elegant way of understanding integral representations of fractional Brownian motion and asymptotic results in the context of piecewise polynomial or spline regression (see Sect. 7.3). Here, we summarize some basics of fractional calculus (see e.g., Samko et al. 1987; for a nice summary ofsome essentials, also see the papers by Pipiras and Taqqu 2000a, 2003). One possible elementary motivation for introducing fractional integrals is the observation that for a real-valued function φ on an interval $[a, b]$, we have

$$\int_a^{t_n} \int_a^{t_{n-1}} \cdots \left(\int_a^{t_1} \varphi(u)\, du \right) dt_1\, dt_2 \cdots dt_{n-1} = \frac{1}{\Gamma(n)} \int_a^{t_n} \varphi(u)(t_n - u)^{n-1}\, du.$$

Replacing n by a positive real value d, we obtain the definition of fractional integral operators as follows:

Definition 3.27 The Riemann–Liouville fractional integrals of order $d > 0$ are defined by

$$\big(I_+^d f\big)(s) = \frac{1}{\Gamma(d)} \int_{-\infty}^{s} f(u)(s-u)^{d-1}\, du = \frac{1}{\Gamma(d)} \int_{-\infty}^{\infty} f(u)(s-u)_+^{d-1}\, du$$

and

$$\big(I_-^d f\big)(s) = \frac{1}{\Gamma(d)} \int_{s}^{\infty} f(u)(u-s)^{d-1}\, du = \frac{1}{\Gamma(d)} \int_{-\infty}^{\infty} f(u)(u-s)_+^{d-1}\, du.$$

Note that $I_+^d f$ and $I_-^d f$ can be understood as operators mapping a function $f(\cdot)$ to the functions $(I_+^d f)(\cdot)$ and $(I_-^d f)(\cdot)$ respectively. These integrals are well defined if $f \in L^p(\mathbb{R})$ with $1 \le p < d^{-1}$ (see Samko et al. 1987, p. 94) in the sense that $|(I_\pm^d f)(s)| < \infty$ for almost all s. A natural extension to $d = 0$ is $I_\pm^0 f := f$, i.e. $I_\pm^0$ is the identity operator.

A slightly more difficult concept is the fractional derivative. A natural approach to defining a fractional derivative of order d is via the inverse operator of $I_\pm^d$ (where $d > 0$). In view of the semigroup property $I_\pm^{d_1} I_\pm^{d_2} = I_\pm^{d_1+d_2}$, one may be tempted to use the integral $\int_s^\infty f(u)(u-s)^{-d-1}\, du$. However, for many functions f, this integral does not exist or is infinite. One way of avoiding this problem is to integrate $(u-s)^{-d}$ first and take the derivative with respect to u afterwards (i.e. take the derivative outside the integral). This leads to the definition of the Riemann–Liouville fractional derivative:

Definition 3.28 For $0 < d < 1$, the Riemann–Liouville derivatives of order d are defined by

$$\big(D_+^d f\big)(u) = \frac{1}{\Gamma(1-d)} \frac{d}{du} \int_{-\infty}^{\infty} f(s)(u-s)_+^{-d}\, ds$$

and

$$\big(D_-^d f\big)(u) = -\frac{1}{\Gamma(1-d)} \frac{d}{du} \int_{-\infty}^{\infty} f(s)(s-u)_+^{-d}\, ds.$$

The reason why these are suitable definitions of fractional derivatives can be seen from the proof of the following lemma.

Lemma 3.23 *Let $0 < \alpha < 1$, and let f be a given function for which $(D_\pm^d f)(u)$ as defined above exists. Then Abel's equation*

$$\big(I_+^d \varphi\big)(s) = f(s)$$

(with unknown φ) has the solution

$$\varphi(u) = \big(D_+^d f\big)(u).$$

The analogous result holds for I_+^d and D_+^d.

Proof The result is obtained by first integrating both sides of Abel's equation multiplied by $(u-s)_+^{-d}$. For the right hand side, we obtain

$$\int_{-\infty}^{\infty} f(s)(u-s)_+^{-d}\,ds,$$

whereas for the left-hand side, we have

$$\begin{aligned}\int_{-\infty}^{\infty} \big(I_+^d\varphi\big)(s)(u-s)_+^{-d}\,ds &= \frac{1}{\Gamma(d)}\int_{-\infty}^{u}\left[\int_{-\infty}^{s}\varphi(v)(s-v)^{d-1}\,dv\right](u-s)^{-d}\,ds\\ &= \frac{1}{\Gamma(d)}\int_{-\infty}^{u}\varphi(v)B(d,1-d)\,dv\\ &= \Gamma(1-d)\int_{-\infty}^{u}\varphi(v)\,dv.\end{aligned}$$

Thus,

$$\int_{-\infty}^{\infty} f(s)(u-s)_+^{-d}\,ds = \Gamma(1-d)\int_{-\infty}^{u}\varphi(v)\,dv,$$

so that

$$\varphi(u) = \frac{1}{\Gamma(1-d)}\frac{d}{du}\int_{-\infty}^{\infty} f(s)(u-s)_+^{-d}\,ds,$$

which is the definition of $(D_+^d f)(u)$. □

More specifically, $(D_+^d f)(u)$ exists and is the left- and right-hand inverse of I_+^d if $0<d<1$ and $f = I_+^d\varphi$ for a function $\varphi \in L^1(\mathbb{R})$. For applications to stochastic integration with respect to fractional Brownian motion, the restriction to L^1-functions is not general enough. What one needs to be able to use are L^p-functions for a p at least equal to 2. This motivates a slightly more complicated definition of the fractional derivative.

Definition 3.29 The Marchaud derivative of order d $(0<d<1)$ is defined by

$$\mathbf{D}_\pm^d f := \lim_{\varepsilon\to 0}\mathbf{D}_{\pm,\varepsilon}^d f$$

with

$$\big(\mathbf{D}_{\pm,\varepsilon}^d f\big)(s) = \frac{d}{\Gamma(1-d)}\int_{\varepsilon}^{\infty}\frac{f(s)-f(s\mp u)}{u^{1+d}}\,du.$$

It can be shown (see Theorem 6.1 in Samko et al. 1987) that if $f = I_\pm^d\phi$ for some $\phi\in L^p(\mathbb{R})$, $d>0$ and $1\le p<d^{-1}$, then $\lim_{\varepsilon\to 0}\mathbf{D}_{\pm,\varepsilon}^d f$ exists and is equal to ϕ in $L^p(\mathbb{R})$ and almost surely. In other words, the Marchaud derivative inverts the fractional integral. For $d<0$, we may thus set $I_\pm^d := (I_\pm^{|d|})^{-1} = \mathbf{D}_\pm^{|d|}$.

Chapter 4
Limit Theorems

4.1 Tools

4.1.1 Introduction

Most statistical procedures in time series analysis (and in fact statistical inference in general) are based on asymptotic results. Limit theorems are therefore a fundamental part of statistical inference. Here we first review very briefly a few of the basic principles and results needed for deriving limit theorems in the context of long-memory and related processes.

4.1.2 How to Derive Limit Theorems?

To prove the convergence of an appropriately normalized process $S_n(\cdot)$, one has to verify the convergence of finite-dimensional distributions and tightness. With respect to the first issue, we usually prove just one-dimensional convergence because in most situations extensions to the multivariate case are straightforward. The tools we describe here are applicable to many statistics, not only partial sums. On the other hand, most of the statistics we will consider are just partial sums.

4.1.2.1 How to Verify Finite-Dimensional Convergence?

Suppose that X_t ($t \in \mathbb{N}$) is a stationary process. One of the common methods for deriving limit theorems is to evaluate its characteristic function. This is however rarely successful in a long-memory setting. An alternative method for partial sums of long-memory sequences is to study the asymptotic behaviour of cumulants. Recall that for a given random variable X, its cumulants are the coefficients in the power series

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_4, © Springer-Verlag Berlin Heidelberg 2013

expansion of $\kappa_X(z) = \log E(e^{zX})$, i.e. $\kappa_j = \kappa_j(X)$ in

$$\kappa_X(z) = \sum_{j=0}^{\infty} \frac{z^j}{j!} \kappa_j.$$

In particular, $\kappa_1 = \mu_X = E(X)$, $\kappa_2 = \sigma_X^2 = \mathrm{var}(X)$. If $E(X) = 0$, then $\kappa_4 = E(X^4) - 3E^2(X^2)$. One of the useful properties of cumulants is that for a normal random variable X, we have $\kappa_j = 0$ for all $j \geq 3$, and this is only the case for the normal distribution. Moreover, a normal distribution is uniquely determined by its moments.

The justification for the approach based on cumulants is the following well-known result (see e.g. Rao 1965):

Theorem 4.1 *Let S_n $(n \in \mathbb{N})$ be a sequence of random variables such that $E[|S_n|^j] < \infty$ for all j, and let Y be a random variable whose distribution is uniquely determined by its moments $\mu_j = E(Y^j)$ $(j \in \mathbb{N})$. Then the convergence of all cumulants $\kappa_j(S_n)$ of S_n $(j \in \mathbb{N})$ to the cumulants $\kappa_j(Y)$ of Y implies that S_n converges to Y in distribution.*

Cumulants are useful if all moments exist. An approach that does not require finiteness of higher-order moments is referred to as a K-dependent approximation method and is adapted from Billingsley (1968, Theorem 4.2).

Proposition 4.1 *Let X_t $(t \in \mathbb{N})$ be a stationary sequence, c_n a sequence of constants, and $X_{t,K}$ $(t \in \mathbb{N})$ a sequence of K-dependent random variables. Define $S_n = \sum_{t=1}^{n} X_t$ and $S_{n,K} = \sum_{t=1}^{n} X_{t,K}$, and suppose that the following holds*:

(a) $c_n^{-1} S_{n,K} \xrightarrow{d} S_K$ *as* $n \to \infty$;
(b) $S_K \xrightarrow{P} S$ *as* $K \to \infty$;
(c)

$$\lim_{K \to \infty} \limsup_{n \to \infty} P\left(c_n^{-1} |S_{n,K} - S_n| > \gamma\right) = 0$$

for each $\gamma > 0$.

Then, as $n \to \infty$,

$$c_n^{-1} S_n \xrightarrow{d} S.$$

To apply this theorem, we mention that if $v_K^2 \to v^2$ as $K \to \infty$, then $N(0, v_K^2) \xrightarrow{d} N(0, v^2)$. Furthermore, this approach requires the following result for K-dependent sequences.

Lemma 4.1 *Let* $X_{t,K}$ $(t \in \mathbb{N})$ *be a stationary sequence of* K*-dependent random variables with* $\mathrm{var}(X_{0,K}) < \infty$*, and define* $S_{n,K} = \sum_{t=1}^{n} X_{t,K}$*. Then*

$$n^{-\frac{1}{2}} S_{n,K} \xrightarrow{d} \sigma_K N(0,1),$$

where $\sigma_K^2 = \mathrm{var}(X_{0,K}) + 2\sum_{j=1}^{K} cov(X_{0,K}, X_{j,K})$.

Another useful result is the following martingale central limit theorem.

Lemma 4.2 *Let* $(X_{t,n}, \mathcal{F}_t)$ $(t \in \mathbb{N},\ n \geq 1)$ *be a martingale difference array, and define* $\tilde{X}_{t,n} = X_{t,n} - E(X_{t,n}|\mathcal{F}_{t-1})$*. Furthermore, assume that the following conditions hold*:

(a) *for each* $\delta > 0$,

$$\sum_{t=1}^{n} E\big(\tilde{X}_{t,n}^2 1\{|\tilde{X}_{t,n}| > \delta\}\big) \to 0,$$

(b)

$$\sum_{t=1}^{n} E\big(\tilde{X}_{t,n}^2 \big| \mathcal{F}_{t-1}\big) \xrightarrow{p} 1.$$

Then

$$\sum_{t=1}^{n} X_{t,n} \xrightarrow{d} N(0,1).$$

4.1.2.2 How to Verify Tightness?

There are several ways to prove tightness. A particularly useful result given in Theorem 15.6 of Billingsley (1968) provides sufficient conditions for tightness in D (the space of right-continuous functions with left limits):

Lemma 4.3 *A stochastic process* $Y_n(u)$ $(u \in [0,1])$ *is tight if there exist* $\eta > 1$, $a > 0$ *and a nondecreasing function* g *such that for all* $v_1 < u < v_2 \in [0,1]$,

$$E\big[\big|Y_n(v_2) - Y_n(u)\big|^a \big|Y_n(u) - Y_n(v_1)\big|^a\big] \leq \big(g(v_2) - g(v_1)\big)^{\eta}.$$

In particular, assume that X_t $(t \in \mathbb{N})$ is a stationary sequence of random variables and G is a function such that $E[G(X_t)] = 0$. Consider the partial sum process

$$S_n(u) = \sum_{t=1}^{[nu]} G(X_t) \quad \big(u \in [0,1]\big). \tag{4.1}$$

Applying Lemma 4.3 to the partial sum process $Y_n(u) = d_n^{-1} S_n(u)$ yields the following result (see Theorem 2.1 in Taqqu 1975).

Lemma 4.4 *Assume that*

(a) $E[G(X_1)] = 0$ *and* $E[G^2(X_1)] < \infty$.
(b) $d_n^2 \sim n^{2d+1} L_S(n)$ *with* $-\frac{1}{2} \le d < \frac{1}{2}$ *and a slowly varying function* L_S.
(c) $E[S_n^2(1)] = O(d_n^2)$.
(d) *There exists* $a > (2d+1)^{-1}$ *such that* $E(|S_n(1)|^{2a}) = O((E[S_n^2(1)])^a)$.

Then $d_n^{-1} S_n(\cdot)$ *is tight.*

Proof Assume for simplicity that $L_S \equiv 1$. We note that the process $S_n(u)$, $u \in [0,1]$, has stationary increments. In particular, for $0 \le u \le v \le 1$, $S_n(v) - S_n(u) \stackrel{d}{=} S_n(v-u)$. Thus, applying the Cauchy–Schwarz inequality and stationarity of increments, we have for $v_1 < u < v_2$, and a suitable constant $0 < C < \infty$,

$$\begin{aligned}
&d_n^{-2a} E\big[\big|S_n(v_2) - S_n(u)\big|^a \big|S_n(u) - S_n(v_1)\big|^a\big] \\
&\quad \le d_n^{-2a} \big(E\big[\big|S_n(v_2 - u)\big|^{2a}\big]\big)^{1/2} \big(E\big[\big|S_n(u - v_1)\big|^{2a}\big]\big)^{1/2} \\
&\quad \le d_n^{-2a} d_n^{2a} \big\{(v_2 - u)^{2d+1} (u - v_1)^{2d+1}\big\}^{a/2} C \le \big\{(v_2 - u)(u - v_1)\big\}^{(d+\frac{1}{2})a} C \\
&\quad \le (v_2 - v_1)^{(2d+1)a} C.
\end{aligned}$$

Since $(2d+1)a > 1$, Billingsley's criterium is fulfilled, and the process is tight. □

If we restrict ourselves to $d > 0$, then Lemma 4.3 leads to a particularly useful criterion in the long-memory case because it amounts to finding a bound on $E[(Y_n(v_2) - Y_n(v_1))^2]$ only.

Lemma 4.5 *Assume that* $Y_n(u)$ $(u \in [0,1])$ *is a stochastic process with stationary increments. If*

$$E\big[\big|Y_n(v_2) - Y_n(v_1)\big|^2\big] \le (v_2 - v_1)^{2d+1}, \tag{4.2}$$

$d > 0$, *then the process is tight.*

Indeed, if we consider again $Y_n(u) = d_n^{-1} S_n(u)$, then

$$\begin{aligned}
&d_n^{-2} E\big[\big|S_n(v_2) - S_n(u)\big|\big|S_n(u) - S_n(v_1)\big|\big] \\
&\quad \le d_n^{-2} \big(E\big[\big|S_n(v_2 - u)\big|^2\big]\big)^{1/2} \big(E\big[\big|S_n(u - v_1)\big|^2\big]\big)^{1/2} \\
&\quad \le d_n^{-2} d_n^2 \big\{(v_2 - u)^{2d+1} (u - v_1)^{2d+1}\big\}^{1/2} C \le \big\{(v_2 - u)(u - v_1)\big\}^{(d+\frac{1}{2})} C \\
&\quad \le (v_2 - v_1)^{(2d+1)} C,
\end{aligned}$$

and the exponent exceeds one since $d > 0$. We note that this approach does not work when $d \le 0$. Hence, in a sense, showing tightness in a long-memory case is easier than in a weakly dependent and antipersistent situation. We note further that

condition (4.2) is almost the same as a moment condition for tightness of processes in C; see Theorem 12.3 in Billingsley (1968).

4.1.2.3 Functional Central Limit Theorem for Processes

The following result is used to establish a functional limit theorem for a sum of independent stochastic processes; see e.g. p. 226 of Whitt (2002).

Lemma 4.6 *Let $X_t(u)$ $(u \in [0,\infty), t \in \mathbb{N})$ be an i.i.d. sequence of processes viewed as random elements in $D[0,\infty)$. If $E(X_1(u)) = 0$, $E(X_1^2(u)) < \infty$ for each $u \in [0,\infty)$ and there exist continuous nondecreasing functions f, g and numbers $a > 1/2$, $b > 1$ such that*

$$E\big[\big(X_1(v) - X_1(u)\big)^2\big] \le \big(g(v) - g(u)\big)^a,$$

$$E\big[\big(X_1(v_2) - X_1(u)\big)^2\big(X_1(u) - X_1(v_1)\big)^2\big] \le \big(g(v_2) - g(v_1)\big)^b,$$

for all $0 \le u < v \le \infty$, $0 \le v_1 < u < v_2 < \infty$, then

$$n^{-1/2}\sum_{t=1}^{n} X_t(u) \Rightarrow G(u),$$

where G is a zero-mean Gaussian process with continuous sample paths, $cov(G(0), G(u)) = cov(X_1(0), X_1(u))$, and $\Rightarrow$ denotes weak convergence in $D[0,\infty)$.

4.1.2.4 Functional Central Limit Theorem for Inverses

The following result, known as Vervaat's lemma (see Vervaat 1972 or De Haan and Ferreira 2006), plays a crucial role in deriving limit theorems for appropriately scaled and normalized quantile processes (as inverses of empirical processes; see Sect. 4.8.2), or counting processes (as inverses of partial sum processes; see Sect. 4.9).

Lemma 4.7 (FCLT for Inverse Functions) *Denote by $D_0([0,\infty))$ the subset of $D[0,\infty)$ consisting of non-decreasing, non-negative, unbounded functions. Let $y_n(\cdot)$ $(n \ge 1)$ be a sequence of elements of $D_0([0,\infty))$. Moreover, let $y(\cdot)$ be a continuous function on $[0,\infty)$, and c_n $(n \ge 1)$ a sequence of positive numbers such that $c_n \to 0$. If*

$$\frac{y_n(u) - u}{c_n} \to y(u)$$

uniformly on compact sets in $[0,\infty)$, then

$$\frac{y_n^{-1}(u) - u}{c_n} \to -y(u)$$

uniformly on compact sets in $[0,\infty)$, *where* $y_n^{-1}(u) := \inf\{v : y_n(v) > u\}$ *is the generalized inverse of* $y_n(\cdot)$.

It is important to mention that the continuity assumption on $y(\cdot)$ cannot be relaxed. If the limiting function has jumps, then the uniform convergence of the inverse processes does not follow necessarily. In particular, this theorem will be applicable to situations where we have weak convergence in $D[0,1]$ equipped with the standard J_1-topology, to a continuous process, and from that we will conclude weak convergence in that topology for the inverse processes. If the limiting process has jumps, we may not be able to conclude weak convergence of the inverse processes in the same topology, even though we may have weak convergence of the original processes. Nevertheless, at least finite-dimensional convergence follows. We refer to Whitt (2002, Chap. 13) for more details.

It is also important to see that in this lemma we assume the identity function to be the correct quantity to subtract. Thus, for instance, when dealing with the empirical distribution function $F_n(x) = n^{-1}\sum_{t=1}^n 1\{X_t \le x\}$ (where $X \sim F_X$), the result actually refers to $\tilde{F}_n(x) = n^{-1}\sum_{t=1}^n 1\{F_X(X_t) \le x\}$ and the corresponding inverse. The reason is that $F_X(X)$ is uniformly distributed, so that we are in the situation described in Vervaat's lemma. The result for F_n (and F_n^{-1}) then follows by the continuous mapping theorem.

4.1.3 Spectral Representation of Stationary Sequences

In this section we collect several standard results on spectral theory for stationary processes. Some of these properties have been used in the preliminary discussion on long memory, see Chap. 1. We state these results without a reference since they can be found in standard textbooks on time series such as Brockwell and Davis (1991).

Recall that for a zero-mean second-order stationary process X_t ($t \in \mathbb{Z}$) with autocovariances $\gamma_X(k)$, there is a spectral distribution function F such that

$$\gamma_X(k) = \int_{-\pi}^{\pi} e^{ik\lambda}\, dF(\lambda).$$

Moreover, X_t has a spectral representation of the form

$$X_t(\omega) = \int_{-\pi}^{\pi} e^{it\lambda}\, dM(\lambda;\omega),$$

where $M(\cdot;\omega)$ is a spectral measure (for simplicity, we will often write $M(\lambda)$ instead of $M(\lambda;\omega)$). The spectral measure is a complex-valued zero mean stochastic process on $[-\pi,\pi]$ with (a.s.) right-continuous sample paths and *uncorrelated* (but not necessarily independent) increments with a variance that is directly related to F. More specifically, we have

$$\begin{aligned} cov\big(dM(\lambda), dM(\nu)\big) &= E\big[dM(\lambda)\,\overline{dM(\nu)}\big] = 0 \quad (\lambda \neq \nu), \\ \operatorname{var}\big(dM(\lambda)\big) &= E\big[\big|dM(\lambda)\big|^2\big] = dF(\lambda). \end{aligned}$$

In particular, if the spectral density exists, then we may write the infinitesimal equation $\operatorname{var}(dM(\lambda)) = E[|dM(\lambda)|^2] = f(\lambda)\,d\lambda$.

It is important to distinguish between the role of the spectral distribution F and the spectral measure M. The spectral distribution determines the autocovariance structure, i.e. linear dependence, of the process only. In contrast, the spectral measure fully specifies the process (in the sense of the probability distribution of sample paths). In the special case where $M = M_\varepsilon$ with $E[|dM_\varepsilon(\lambda)|^2] = \sigma_\varepsilon^2/(2\pi)\cdot d\lambda$ we obtain a white noise process with variance σ_ε^2 where "white noise" stands for uncorrelated observations. This follows directly from the spectral representation

$$\varepsilon_t = \int_{-\pi}^{\pi} e^{it\lambda}\,dM_\varepsilon(\lambda) \quad (t \in \mathbb{Z}) \tag{4.3}$$

since

$$\begin{aligned} E[\varepsilon_t \varepsilon_s] &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{i(t\lambda - s\nu)} E\big[dM_\varepsilon(\lambda)\,\overline{dM_\varepsilon(\nu)}\big] \\ &= \frac{\sigma_\varepsilon^2}{2\pi}\int_{-\pi}^{\pi} e^{i(t-s)\lambda}\,d\lambda = \sigma_\varepsilon^2 \delta_{ts}. \end{aligned}$$

The spectral density of ε_t is $f_\varepsilon(\lambda) = \sigma_\varepsilon^2/(2\pi)$. One should bear in mind that, in general, this does not imply the independence of ε_t ($t \in \mathbb{Z}$). Such a direct conclusion can only be made if $M(\lambda;\omega)$ is a Gaussian process.

A zero mean, purely nondeterministic second-order stationary process always has a Wold decomposition

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = A(B)\varepsilon_t \quad (t \in \mathbb{Z})$$

with uncorrelated (i.e. "white noise") innovations ε_t and $A(z) = \sum a_j z^j$ such that $\sum_{j=0}^{\infty} a_j^2 < \infty$. Therefore, the spectral measure and spectral distribution have a simple form, namely (with equality in the $L^2(\Omega)$ sense)

$$X_t = \int_{-\pi}^{\pi} e^{it\lambda}\,dM_X(\lambda) = \int_{-\pi}^{\pi} e^{it\lambda} A\big(e^{-i\lambda}\big)\,dM_\varepsilon(\lambda) \quad (t \in \mathbb{Z}). \tag{4.4}$$

In other words,

$$dM_X(\lambda) = \left(\sum_{j=0}^{\infty} a_j e^{-ij\lambda}\right) dM_\varepsilon(\lambda) = A\big(e^{-i\lambda}\big)\,dM_\varepsilon(\lambda).$$

The spectral density

$$f_X(\lambda) = \frac{1}{2\pi}\sum_{k=\infty}^{\infty} \gamma_X(k)\exp(-i\lambda k)$$

is then given by

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\sum_{j=0}^{\infty} a_j e^{-ij\lambda}\right|^2 = \frac{\sigma_\varepsilon^2}{2\pi}\left|A\big(e^{-i\lambda}\big)\right|^2.$$

These formulas are valid generally. More specifically, if we consider linear processes only, the ε_ts in the Wold representation are not only uncorrelated but even *independent*. This means that the increments of M_ε are independent (instead of being just uncorrelated). Even more specifically, a *Gaussian* process is a linear process that has normally distributed ε_ts, namely $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. This means that we are in the following situation. The measure M_ε is a Gaussian spectral measure such that for all sets A, $E[M_\varepsilon(A)] = 0$, $E_\varepsilon[M(A \cap B)] = 0$ for all disjoint sets A and B, and $E[M_\varepsilon(A)\overline{M_\varepsilon(A)}] = \sigma_\varepsilon^2|A|/(2\pi)$, where $|\cdot|$ denotes the Lebesgue measure. Moreover, for all $\lambda_1 \le \lambda_2 < \lambda_3 \le \lambda_4$, the increments $M_\varepsilon(\lambda_4) - M_\varepsilon(\lambda_3)$ and $M_\varepsilon(\lambda_2) - M_\varepsilon(\lambda_1)$ are independent. (For simplicity of notation, we will mostly assume that $\sigma_\varepsilon^2 = 1$, which means that $M_\varepsilon(\cdot)$ is a spectral measure of an i.i.d. $N(0, 1)$ sequence.) The Gaussian process X_t is then given by

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_X(\lambda) \quad (t \in \mathbb{N}), \tag{4.5}$$

where M_X is the Gaussian spectral measure defined by

$$dM_X(\lambda) = \left(\sum_{j=0}^{\infty} a_j e^{-ij\lambda}\right) dM_\varepsilon(\lambda) = A\big(e^{-i\lambda}\big)\, dM_\varepsilon(\lambda) =: \sqrt{2\pi}\, a(\lambda)\, dM_\varepsilon(\lambda).$$

Note that in the notation with $a(\lambda)$, the spectral density can be written as

$$f_X(\lambda) = \sigma_\varepsilon^2 \big|a(\lambda)\big|^2.$$

Thus, for $\sigma_\varepsilon^2 = 1$, we have the identity $f_X(\lambda) = |a(\lambda)|^2$.

Another result that is very useful in many situations, such as prediction or (Gaussian) maximum likelihood estimation, is the following factorization of the spectral density. Let us write $\log f_X$ as a Fourier series

$$\log f_X(\lambda) = \sum_{j=-\infty}^{\infty} \alpha_j e^{-ij\lambda}$$

with coefficients

$$\alpha_j = \alpha_{-j} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ij\lambda} \log f_X(\lambda)\, d\lambda. \tag{4.6}$$

Then we obtain the factorization

$$f_X(\lambda) = \exp(\alpha_0)\big|A\big(e^{-i\lambda}\big)\big|^2 = \frac{\sigma_\varepsilon^2}{2\pi}\big|A\big(e^{-i\lambda}\big)\big|^2 =: \frac{\sigma_\varepsilon^2}{2\pi} h_X(\lambda), \tag{4.7}$$

where

$$A(z) = \sum_{j=0}^{\infty} a_j z^j = \exp\left(\sum_{j=1}^{\infty} \alpha_j z^j\right)$$

and

$$\frac{\sigma_\varepsilon^2}{2\pi} = \exp(\alpha_0).$$

The last equation, together with (4.6), implies

$$\alpha_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log f_X(\lambda)\, d\lambda = \log \sigma_\varepsilon^2 - \log 2\pi.$$

For the function $h_X(\cdot)$ defined in (4.7), we therefore obtain

$$\int_{-\pi}^{\pi} \log h_X(\lambda)\, d\lambda = 0. \tag{4.8}$$

This property is particularly useful for the asymptotic theory of (Gaussian) quasi-maximum likelihood estimation.

Finally, the following lemma is useful in spectral analysis of stationary sequences (see Lemma 2 in Moulines et al. 2007a). Consider the spectral radius $Sp(A)$ of an $n \times n$ matrix A, defined as the maximal absolute eigenvalue, or

$$Sp(A) = \sup_{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|\le 1} \mathbf{x}^T A \mathbf{x}.$$

Now let $A = \Sigma_n = [\gamma_X(i-j)]_{i,j=1,\ldots,n}$ be the covariance matrix of $X = (X_1, \ldots, X_n)^T$, where X_t is a zero-mean stationary process with spectral density f_X. Then

$$\begin{aligned}
\mathbf{x}^T A \mathbf{x} &= \sum_{j,l=1}^{n} \gamma_X(j-l) x_j x_l \\
&= \int_{-\pi}^{\pi} f_X(\lambda) \left| \sum_{j=1}^{n} x_j \exp(-ij\lambda) \right|^2 d\lambda \\
&\le \sup_{\lambda\in[-\pi,\pi]} \left| f_X(\lambda) \right| \int_{-\pi}^{\pi} \left| \sum_{j=1}^{n} x_j \exp(-ij\lambda) \right|^2 d\lambda = 2\pi |\mathbf{x}|^2 \sup_{\lambda\in[-\pi,\pi]} \left| f_X(\lambda) \right|,
\end{aligned}$$

where the last expression follows from the Parseval identity. Hence, we have the following result.

Lemma 4.8 *Assume that X_t $(t \in \mathbb{Z})$ is a stationary process with the spectral density f_X. Assume that Σ_n is the covariance matrix of $X_1, \ldots, X_n$. Then*

$$Sp(\Sigma_n) \le 2\pi \sup_{\lambda \in [-\pi,\pi]} \left| f_X(\lambda) \right|.$$

4.2 Limit Theorems for Sums with Finite Moments

4.2.1 Introduction

Let X_t $(t \in \mathbb{N})$ be a stationary process. The asymptotic behaviour of partial sums

$$S_n(u) = S_{n,G}(u) = \sum_{t=1}^{[nu]} G(X_t) \tag{4.9}$$

is at the core of probability theory. In this section we present limit theorems for partial sums associated with long-memory or antipersistent processes. Two types of distinctions have to be made. One is between linear and nonlinear processes. The other is between processes with finite and infinite variance. The case of infinite variance is studied in Sect. 4.3. Depending on which of these cases is considered, different results and mathematical techniques are required.

In this section we discuss finite-variance processes only. We will begin our exposition by assuming that X_t $(t \in \mathbb{N})$ is a Gaussian process, since computations and proofs are technically less challenging than for instance for general Appell polynomials. The limiting phenomena related to partial sums of subordinated Gaussian sequences were observed first by Rosenblatt (1961) and then developed independently by Taqqu (1975, 1977, 1979), Dobrushin (1980) and Dobrushin and Major (1979). Further developments can be found in Breuer and Major (1983), Giraitis and Surgailis (1985), Ho and Sun (1987, 1990), Dehling and Taqqu (1989a, 1989b) and Arcones (1994). Although the original technique in Taqqu (1975) to show convergence to the so-called Hermite–Rosenblatt distribution was based on characteristic functions, the common method to obtain a non-central limit theorem is based on (multiple) Wiener–Itô integrals, together with the diagram formula. For long-memory linear processes, the first result was obtained in Davydov (1970a, 1970b); see also Gorodetskii (1977), Lang and Soulier (2000), Wang et al. (2003).

As for subordinated linear processes, there are two common approaches: Appell polynomials (Surgailis 1981, 1982; Giraitis 1985; Giraitis and Surgailis 1986, 1989; Avram and Taqqu 1987; Surgailis and Vaičiulis 1999; Surgailis 2000; also see Surgailis 2003 for a review) and a martingale decomposition (Ho and Hsing 1996, 1997; Giraitis and Surgailis 1999; Wu 2003; see also Hsing 2000 for a review).

The theory for nonlinear models with long memory is less well developed. EGARCH-type models were considered in Surgailis and Viano (2002), whereas results for LARCH(∞) processes can be found for instance in Giraitis et al. (2000c), Giraitis and Surgailis (2002), Berkes and Horváth (2003), Beran (2006).

4.2.2 Normalizing Constants for Stationary Processes

Before getting into the details of limiting distributions, a first question can be answered relatively easily, namely which normalizing sequences should be used to obtain nondegenerate limits. Let $S_n = \sum_{t=1}^{n} X_t$, where X_t ($t \in \mathbb{N}$) is a stationary sequence with appropriate moment conditions. We consider the asymptotic behaviour of $\operatorname{var}(S_n)$ in three cases: long memory, short memory and antipersistence.

Lemma 4.9 (Long Memory) *Let X_t ($t \in \mathbb{N}$) be a stationary sequence with $\gamma_X(k) \sim L_\gamma(k)k^{2d-1}$ ($k \to \infty$) for some $0 < d < \frac{1}{2}$, where L_γ is slowly varying at infinity. Then, as $n \to \infty$,*

$$\operatorname{var}(S_n) \sim L_S(n)n^{2d+1} \tag{4.10}$$

with

$$L_S(n) = L_1(n) = C_1 L_\gamma(n) = \frac{1}{d\,(2d+1)} L_\gamma(n). \tag{4.11}$$

Proof We have

$$\begin{aligned}\operatorname{var}(S_n) &= n \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right)\gamma_X(k) \\ &\sim n \sum_{\substack{k=-(n-1)\\ k\neq 0}}^{n-1} L_\gamma(k)|k|^{2d-1} - \sum_{\substack{k=-(n-1)\\ k\neq 0}}^{n-1} L_\gamma(k)|k|^{2d}.\end{aligned}$$

The last expression can be written as

$$\begin{aligned}&L_\gamma(n)n^{2d+1}\left[\sum_{\substack{k=-(n-1)\\ k\neq 0}}^{n-1} \frac{L_\gamma(k)}{L_\gamma(n)}\left(\frac{|k|}{p}\right)^{2d-1} n^{-1}\right. \\ &\qquad \left. - \sum_{\substack{k=-(n-1)\\ k\neq 0}}^{n-1} \frac{L_\gamma(k)}{L_\gamma(n)}\left(\frac{|k|}{n}\right)^{2d} n^{-1}\right] \\ &\sim 2L_\gamma(n)n^{2d+1}\left[\int_0^1 u^{2d-1}\,du - \int_0^1 u^{2d}\,du\right] \\ &= 2L_\gamma(n)n^{2d+1}\left(\frac{1}{2d} - \frac{1}{2d+1}\right) = \frac{L_\gamma(n)}{d(2d+1)}n^{2d+1}.\end{aligned}$$

□

Lemma 4.10 (Short Memory) *Let X_t ($t \in \mathbb{N}$) be a stationary sequence with $\sum_{k=-\infty}^{\infty} \gamma_X(k) > 0$ and $\sum_{k=-\infty}^{\infty} |\gamma_X(k)| < \infty$. Then, as $n \to \infty$,*

$$\operatorname{var}(S_n) \sim c_S n \tag{4.12}$$

with

$$c_S = \sum_{k=-\infty}^{\infty} \gamma_X(k). \tag{4.13}$$

Proof Cesaro summability implies

$$\sum_{k=-(n-1)}^{n-1} \frac{k}{n} \gamma_X(k) \to 0,$$

so that

$$\operatorname{var}(S_n) \sim n \sum_{k=-(n-1)}^{n-1} \gamma_X(k) \sim c_S n. \qquad \square$$

Lemma 4.11 (Antipersistence) *Let* X_t $(t \in \mathbb{N})$ *be a stationary sequence with* $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$ $(k \to \infty)$ *for some* $-\frac{1}{2} < d < 0$, *where* L_γ *is slowly varying at infinity, and*

$$\sum_{k=-\infty}^{\infty} \gamma_X(k) = 0.$$

Then, as $n \to \infty$,

$$\operatorname{var}(S_n) \sim L_S(n) n^{2d+1} \tag{4.14}$$

with

$$L_S(n) = \frac{1}{d(2d+1)} L_\gamma(n). \tag{4.15}$$

Proof

$$\begin{aligned} \sum_{k=-(n-1)}^{n-1} \gamma_X(k) &= -2 \sum_{k=n}^{\infty} \gamma_X(k) \sim -2 L_\gamma(n) \sum_{k=n}^{\infty} k^{2d-1} \\ &\sim -2 L_\gamma(n) n^{2d} \int_1^{\infty} u^{2d-1}\, du = \frac{2 L_\gamma(n)}{2d} n^{2d}. \end{aligned}$$

Then the result follows by the same arguments as in the long-memory case. $\square$

Note that in the proof of Lemma 4.11, the Riemann approximation could not be applied to $\sum_{k=-(n-1)}^{n-1} \gamma_X(k)$ directly because u^{2d-1} is not integrable at the origin for $d < 0$. Note also that in the antipersistent case, $L_\gamma(k) < 0$ for k large enough. However, since $L_\gamma(k)$ is multiplied by d^{-1}, the slowly varying function $L_S(n)$ is positive asymptotically.

Taking into account Theorem 1.3, a unified formula including (4.10), (4.12) and (4.14) can be written in terms of the spectral density. Using the notation

$$L_f(\lambda) = L_\gamma\big(\lambda^{-1}\big)\pi^{-1}\Gamma(2d)\sin\left(\frac{\pi}{2} - \pi d\right)$$

and

$$\begin{aligned} \nu(d) &= \frac{2\sin\pi d}{d(2d+1)}\Gamma(1-2d) \quad (d \neq 0), \\ \nu(0) &= \lim_{d\to 0}\nu(d) = 2\pi, \end{aligned} \tag{4.16}$$

we have

$$\operatorname{var}(S_n) \sim \nu(d)L_f\big(n^{-1}\big)n^{2d+1} \sim \nu(d)f_X\big(n^{-1}\big)n.$$

4.2.3 Subordinated Gaussian Processes

We begin our exposition by assuming that X_t ($t \in \mathbb{N}$) are normal random variables because computations and proofs are technically less challenging than in the case of Appell polynomials, for instance. The limiting phenomena related to partial sums of subordinated Gaussian sequences were first observed by Rosenblatt (1961) and then developed independently by Taqqu (1975, 1977, 1979) Dobrushin (1980) and Dobrushin and Major (1979). Further developments can be found in Breuer and Major (1983), Giraitis and Surgailis (1985), Ho and Sun (1987, 1990) and Arcones (1994). Although the original technique in Taqqu (1975) to show convergence to the so-called Hermite–Rosenblatt distribution was based on characteristic functions, the common method to obtain non-central limit theorems is based on (multiple) Wiener–Itô integrals, together with the diagram formula.

4.2.3.1 Moment Bounds and Normalizing Constants

Recall from Sect. 3.1.2 that each function $G(\cdot)$ in $L^2(\mathbb{R}, \phi)$ with $\phi(x) = (2\pi)^{-1/2} \times \exp(-x^2/2)$ can be expanded as

$$G(X) = E\big[G(X)\big] + \sum_{l=1}^{\infty}\frac{J(l)}{l!}H_l(X) = E\big[G(X)\big] + \sum_{l=m}^{\infty}\frac{J(l)}{l!}H_l(X),$$

where $J(l) = E[G(X)H_l(X)]$, X is a standard Gaussian random variable, and m is the Hermite rank of G (i.e. the smallest $m \geq 1$ such that $J(m) \neq 0$). Moreover, recall the formula (3.16) for $H_m(\sum_{j=1}^{l} a_j x_j)$,

$$H_m\left(\sum_{j=1}^{l} a_j x_j\right) = \sum_{m_1+\cdots+m_k=m}\frac{m!}{m_1!\ldots m_k!}\prod_{j=1}^{l} a_j^{m_j}H_{m_j}(x_j). \tag{4.17}$$

This was used for deriving the formula for covariances of Hermite polynomials given in Lemma 3.5. For convenience, we repeat the result here:

Lemma 4.12 *Let X_1, X_2 be a pair of jointly standard normal random variables with covariance $\gamma = cov(X_1, X_2)$. Then*

$$cov\big(H_l(X_1), H_l(X_2)\big) = l!\gamma^l, \tag{4.18}$$

whereas for $j \neq l$,

$$cov\big(H_j(X_1), H_l(X_2)\big) = 0. \tag{4.19}$$

In particular, assume now that

$$\gamma_X(k) \sim L_\gamma(k)k^{2d-1}$$

with $d \in (0, 1/2)$, and consider the sum of $H_m(X_t)$. From Lemma 4.12 we see that if $d > 1 - \frac{1}{2}m^{-1}$, the autocovariance $\gamma_{H_m}(k) = cov(H_m(X_t), H_m(X_{t+k}))$ of the transformed process $H_m(X_t)$ is not summable because it is (up to the slowly varying function) of the order $k^{m(2d-1)}$ with $m(2d-1) > -1$. Using the same argument as in the proof of Lemma 4.9, we then obtain

$$\operatorname{var}\left(\sum_{t=1}^{n} H_m(X_t)\right) = m!\sum_{k=1}^{n}\sum_{j=1}^{n}\gamma_X^m(j-k) \sim L_m(n)n^{(2d-1)m+2}, \tag{4.20}$$

where

$$L_m(n) = m!C_m L_\gamma^m(n) \tag{4.21}$$

and

$$C_m = \frac{2}{[(2d-1)m+1][(2d-1)m+2]}. \tag{4.22}$$

Furthermore, if G has the Hermite rank m, then the variance of $G(X)$ can be decomposed into (orthogonal) contributions of the Hermite coefficients,

$$\operatorname{var}\big(G(X)\big) = \sum_{l=1}^{\infty}\left(\frac{J(l)}{l!}\right)^2 l! = \sum_{l=m}^{\infty}\frac{J^2(l)}{l!}. \tag{4.23}$$

Similarly, if X_1 and X_2 are as in Lemma 4.12,

$$cov\big(G(X_1), G(X_2)\big) = \sum_{l=m}^{\infty}\frac{J^2(l)}{l!}\gamma^l. \tag{4.24}$$

Consequently, applying this to the stationary Gaussian sequence X_t ($t \in \mathbb{N}$), we obtain

$$\gamma_G(k) = cov\big(G(X_t), G(X_{t+k})\big) = \sum_{l=m}^{\infty}\frac{J^2(l)}{l!}\gamma_X^l(k). \tag{4.25}$$

Thus, as $k \to \infty$, the asymptotic behaviour of $cov(G(X_t), G(X_{t+k}))$ is determined by the leading term $(J^2(m)/m!)\gamma_X^m(k)$. From (4.25) we therefore conclude that for a function G with the Hermite rank m, the asymptotic behaviour of the autocovariance is given by

$$\gamma_G(k) \sim \frac{J^2(m)}{m!} L_\gamma^m(k) k^{m(2d-1)} \quad (k \to \infty).$$

Therefore, if $m(1-2d) < 1$, then by the same argument as in (4.20),

$$\operatorname{var}\left(\sum_{t=1}^{n} G(X_t)\right) \sim \frac{J^2(m)}{m!} C_m L_\gamma^m(n) n^{(2d-1)m+2} = \left(\frac{J(m)}{m!}\right)^2 L_m(n) n^{(2d-1)m+2}, \tag{4.26}$$

where C_m is the constant in (4.22), and $L_m(\cdot)$ is the slowly varying function defined in (4.21). Otherwise, if $m(1-2d) > 1$, then

$$\sum_{k=1}^{\infty} \left| cov\big(G(X_t), G(X_{t+k})\big) \right| < \infty.$$

Therefore, one can expect two different types of convergence: either a long-memory type where the normalization for partial sums is

$$n^{-((d-\frac{1}{2})m+1)} L_m^{-\frac{1}{2}}(n) = n^{-\frac{1}{2}-((m-1)/2-d)} L_m^{-\frac{1}{2}}(n) \tag{4.27}$$

or a weakly-dependent type with the usual normalization $n^{-1/2}$.

We conclude the discussion of normalizing constants by mentioning two useful bounds derived by Arcones (1994):

- If $m(1-2d) < 1$, then there is a constant C such that for any function G with Hermite rank m,

$$\operatorname{var}\left(n^{-1} \sum_{t=1}^{n} G(X_t)\right) \le C \gamma_X^m(n) \operatorname{var}\big(G(X_1)\big).$$

- If $m(1-2d) > 1$, then there is a constant C such that for any function G with Hermite rank m,

$$\operatorname{var}\left(n^{-1} \sum_{t=1}^{n} G(X_t)\right) \le C n^{-1} \operatorname{var}\big(G(X_1)\big).$$

The first inequality looks very similar to (4.26). However, the important difference is that the constant C depends on the Gaussian process X_t only and not on the function G.

4.2.3.2 Limiting Distribution

The Hermite rank of $G(x) = x$ is one. Furthermore, $\sum_{t=1}^{[nu]} X_t$ is normally distributed for all n and $u \in [0, 1]$. Therefore, in view of (4.27), the following result is obvious. Note that it is valid for all values of $d \in (-\frac{1}{2}, \frac{1}{2})$, i.e. for long memory ($d \in (0, \frac{1}{2})$), short memory ($d = \frac{1}{2}$) and antipersistence ($d \in (-\frac{1}{2}, 0)$). The limiting process is Gaussian. The dependence structure of the increments depends on d.

Theorem 4.2 *Assume that X_t ($t \in \mathbb{N}$) is a stationary sequence of standard normal random variables such that $f_X(\lambda) = L_f(\lambda)|\lambda|^{-2d}$ with $d \in (-1/2, 1/2)$ and the assumptions of Lemma 4.9 (for $d > 0$, Lemma 4.10) (for $d = 0$) or Lemma 4.11 (for $d < 0$) hold respectively. Let $S_n(u) = \sum_{t=1}^{[nu]} X_t$. Then*

$$n^{-(d+\frac{1}{2})} L_1^{-\frac{1}{2}}(n) S_n(u) \Rightarrow B_H(u) \quad \big(u \in [0, 1]\big),$$

where $B_H(\cdot)$ is a standard fractional Brownian motion with Hurst parameter $H = d + \frac{1}{2}$, "$\Rightarrow$" denotes weak convergence in $D[0, 1]$, and $L_1(n) = L_f(n^{-1})v(d)$ with $v(d)$ defined in (4.16).

Proof As mentioned in the introduction to this chapter, we prove finite-dimensional convergence just in the one-dimensional case. Clearly, $S_n(u)$ is normal, and $r_n^2 = \mathrm{var}(S_n(1))/(n^{2d+1}L_1(n)) \to 1$. Thus, with $d_n^2 = n^{2d+1}L_1(n)$,

$$E\big(e^{i\theta d_n^{-1} S_n(1)}\big) = \exp\left(-\frac{1}{2}\theta^2 r_n^2\right) \to \exp\big(-\theta^2/2\big).$$

Thus, one-dimensional distributions of $S_n(u)$ converge to the standard normal distribution.

For tightness, note that $S_n(1)$ is normal, so that $E[S_n^{2l}(1)]$ ($l \in \mathbb{N}$) is proportional to $(E[S_n^2(1)])^l$. Therefore, the conditions of Lemma 4.4 are fulfilled, and tightness follows. □

We will now present another proof of this theorem. The reason is that it will be easily extendable to more complicated cases of general Hermite polynomials and non-normal random variables. Recall some notions on the spectral representation of stationary time series from Sect. 4.1.3. Let ε_t ($t \in \mathbb{Z}$) be a centred, finite-variance i.i.d. sequence. Then ε_t can be represented in terms of a Gaussian spectral measure with uncorrelated increments,

$$\varepsilon_t = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_\varepsilon(\lambda) \quad (t \in \mathbb{Z}).$$

Recall also that

$$E\big[\big|dM_\varepsilon(\lambda)\big|^2\big] = \frac{\sigma_\varepsilon^2}{2\pi}\, d\lambda = f_\varepsilon(\lambda)\, d\lambda,$$

where $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t)$. Without loss of generality, we will assume that $\sigma_\varepsilon^2 = 1$ in the following. Moreover it will be convenient to use instead of M_ε the spectral measure

$$M_0(A) = \sqrt{2\pi}\, M_\varepsilon(A),$$

so that

$$\varepsilon_t = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{it\lambda}\, dM_0(\lambda)$$

and $E[|dM_0(\lambda)|^2] = d\lambda$. For a linear process $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ $(t \in \mathbb{Z})$ with $\sum_{j=0}^{\infty} a_j^2 < \infty$ (and $\sigma_\varepsilon^2 = 1$), one then has the spectral representation

$$X_t = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_X(\lambda) \quad (t \in \mathbb{Z}) \tag{4.28}$$

with

$$dM_X(\lambda) = \left(\sum_{j=0}^{\infty} a_j e^{-ij\lambda} \right) dM_\varepsilon(\lambda) = A\left(e^{-i\lambda}\right) dM_\varepsilon(\lambda)$$

$$= \frac{1}{\sqrt{2\pi}} A\left(e^{-i\lambda}\right) dM_0(\lambda) =: a(\lambda)\, dM_0(\lambda).$$

The spectral density of X_t is

$$f_X(\lambda) = \frac{1}{2\pi} \left| A\left(e^{-i\lambda}\right) \right|^2 = |a(\lambda)|^2.$$

Assume that $f_X(\lambda) = L_f(\lambda)|\lambda|^{-2d}$ as $\lambda \to 0$ or $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$ as $k \to \infty$. Recall that, under suitable conditions, these assumptions are equivalent to

$$L_f(\lambda) = L_\gamma\left(\lambda^{-1}\right) \pi^{-1} \Gamma(2d) \sin\left(\frac{\pi}{2} - \pi d \right)$$

and

$$L_\gamma(k) = 2 L_f\left(k^{-1}\right) \Gamma(1 - 2d) \sin(\pi d). \tag{4.29}$$

Then $|a(\lambda)| = L_f^{1/2}(\lambda)|\lambda|^{-d}$. Now, we are ready to present an alternative proof of Theorem 4.2. This type of approach was initiated in Dobrushin (1980), Dobrushin and Major (1979); also see Arcones (1994) and Lang and Soulier (2000). We will use a representation of a fractional Brownian motion that appears in Sect. 3.7.1.

Alternative proof of Theorem 4.2 Let $S_n = S_n(1) = \sum_{t=0}^{n-1} X_t$ (note that we take summation from $t = 0$ to $n - 1$) and write the spectral representation

$$
\begin{aligned}
S_n &= \sum_{t=0}^{n-1} \int_{-\pi}^{\pi} e^{it\lambda} dM_X(\lambda) \\
&= \sum_{t=0}^{n-1} \int_{-\pi}^{\pi} e^{it\lambda} a(\lambda)\, dM_0(\lambda) = \int_{-\pi}^{\pi} \left(\sum_{t=0}^{n-1} e^{it\lambda} \right) a(\lambda)\, dM_0(\lambda) \\
&= \int_{-\pi}^{\pi} \frac{e^{i\lambda n} - 1}{e^{i\lambda} - 1} a(\lambda)\, dM_0(\lambda) \\
&= n^{1/2} \int_{-n\pi}^{n\pi} D_n(\lambda/n) a\left(\frac{\lambda}{n}\right) n^{1/2}\, dM_0\big(n^{-1}\lambda\big),
\end{aligned}
$$

where

$$
D_n(\lambda) = \frac{e^{i\lambda n} - 1}{n(e^{i\lambda} - 1)} 1\big\{|\lambda| \le \pi n\big\}. \tag{4.30}
$$

Since $\lim_{u\to 0}(e^{\lambda u} - 1)/u = \lambda$, we conclude that

$$
\lim_{n\to\infty} D_n(\lambda/n) \to \frac{e^{i\lambda} - 1}{i\lambda} =: D(\lambda). \tag{4.31}
$$

Now, $E(|dM_0(n^{-1}\lambda)|^2) = n^{-1} d\lambda$. Hence, $n^{1/2} M_0(n^{-1}A)$ and $M_0(A)$ have the same distribution (as stochastic processes indexed by A), and we can write

$$
S_n \overset{\mathrm{d}}{=} n^{1/2} \int_{-n\pi}^{n\pi} D_n(\lambda/n) a\left(\frac{\lambda}{n}\right) dM_0(\lambda) \approx n^{1/2} \int_{-\infty}^{\infty} D_n(\lambda/n) a\left(\frac{\lambda}{n}\right) dM_0(\lambda).
$$

Consequently, we have two possible scenarios:

- $\lim_{\lambda\to 0} a(\lambda) = a(0) = \sqrt{f_X(0)} \neq 0$. Then we expect

$$
n^{-1/2} S_n \xrightarrow{\mathrm{d}} a(0) \int_{-\infty}^{\infty} \frac{e^{i\lambda} - 1}{i\lambda}\, dM_0(\lambda).
$$

- $a(\lambda) = L_f^{1/2}(\lambda)|\lambda|^{-d}$, $d \in (-1/2, 0) \cup (0, 1/2)$. Then we expect

$$
n^{-(1/2+d)} L_f^{-1/2}\big(n^{-1}\big) S_n \xrightarrow{\mathrm{d}} \int_{-\infty}^{\infty} D(\lambda) \frac{1}{|\lambda|^d}\, dM_0(\lambda). \tag{4.32}
$$

In the latter case, applying (4.21) and (4.22) with $m = 1$ and (4.29), we obtain

$$
L_1(n) = \frac{2\Gamma(1-2d)\sin \pi d}{d(2d+1)} L_f\big(n^{-1}\big) =: K_1^{-2}(1,d) L_f\big(n^{-1}\big).
$$

Thus,

$$
n^{-(1/2+d)} L_1^{-1/2}(n) S_n = K_1(1,d) \int_{-\infty}^{\infty} |\lambda|^{-d} \frac{e^{i\lambda} - 1}{i\lambda}\, dM_0(\lambda).
$$

Recall Proposition 3.1. We can verify that $K_1(1,d)$ agrees with $K_1(1,H)$ there by setting $H = d + \frac{1}{2}$, so that the limiting random variable is $B_H(1)$.

To make the argument (4.32) precise, we note that for $|\lambda| < \pi n$,

$$\left|D_n(\lambda/n) - D(\lambda)\right| = \left|\frac{e^{i\lambda}-1}{n(e^{i\lambda/n}-1)} - \frac{e^{i\lambda}-1}{i\lambda}\right| = O\left(n^{-1}\right)$$

uniformly w.r.t. λ (the bound does not depend on λ). Thus,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left|D_n(\lambda/n) - D(\lambda)\right|^2 d\lambda \\
&\quad = \int_{-n\pi}^{n\pi} \left|D_n(\lambda/n) - D(\lambda)\right|^2 d\lambda \\
&\qquad + \int_{|\lambda|>n\pi} \left|D(\lambda)\right|^2 d\lambda \le O\left(n^{-1}\right) + 2\int_{|\lambda|>n\pi} \frac{1}{|\lambda|^2}\, d\lambda = O\left(n^{-1}\right).
\end{aligned}$$

We conclude that $D_n(\lambda/n)$ converges to $D(\lambda)$ in $L^2(\mathbb{R}, d\lambda)$ (here "$d\lambda$" stands for the Lebesgue measure). Also,

$$n^{-d} L_f^{-1/2}\left(n^{-1}\right) D_n(\lambda/n) a\left(\frac{\lambda}{n}\right)$$

converges in $L^2(\mathbb{R}, d\lambda)$ to $D(\lambda)|\lambda|^{-d}$. Since

$$\begin{aligned}
&E\left[\left(\int_{-\infty}^{\infty}\left(n^{-d} L_f^{-1/2}\left(n^{-1}\right) D_n(\lambda/n) a\left(\frac{\lambda}{n}\right) - D(\lambda)|\lambda|^{-d}\right) dM_0(\lambda)\right)^2\right] \\
&\quad = \int_{-\infty}^{\infty}\left(n^{-d} L_f^{-1/2}\left(n^{-1}\right) D_n(\lambda/n) a\left(\frac{\lambda}{n}\right) - D(\lambda)|\lambda|^{-d}\right)^2 d\lambda \to 0,
\end{aligned}$$

we conclude the convergence in L^2. Thus, the result of Proposition 4.2 follows. □

The limiting distribution in formula (4.32) can be also written as

$$n^{-(1/2+d)} L_f^{-1/2}\left(n^{-1}\right) S_n(1) \xrightarrow{\mathrm{d}} \int_{-\infty}^{\infty} D(\lambda)\, dW_X(\lambda), \tag{4.33}$$

where

$$dW_X(\lambda) = \frac{1}{|\lambda|^d}\, dM_0(\lambda). \tag{4.34}$$

The measure W_X is called the limiting spectral measure that depends (via the parameter d) on the sequence X_t. This representation will be essential in Sect. 4.4.

The longish version of the proof of Theorem 4.2 will allow us to obtain the limiting behaviour of subordinated Gaussian sequences. First, we extend the theorem to partial sum processes $S_{n,H_m}(u) := \sum_{t=1}^{[nu]} H_m(X_t)$, where H_m is the mth Hermite

polynomial. Remarkably, the limit is no longer an fBm process, provided that long memory is strong enough and $m \geq 2$. This was first observed in Rosenblatt (1961), also see Taqqu (1975). Note that their method of proof is based on characteristic functions and is different from the one used in the alternative proof of Theorem 4.2.

Theorem 4.3 *Assume that X_t $(t \in \mathbb{N})$ is a stationary sequence of standard normal random variables such that $\gamma_X(k) \sim L_\gamma(k)k^{2d-1}$ with $d \in (0, 1/2)$. Let $S_{n,H_m}(u) = \sum_{t=1}^{[nu]} H_m(X_t)$. If $m(1-2d) < 1$, then*

$$n^{-(1-m(\frac{1}{2}-d))}L_m^{-1/2}(n)S_{n,H_m}(u) \Rightarrow Z_{m,H}(u) \quad \big(u \in [0,1]\big),$$

where $Z_{m,H}(\cdot)$ is a Hermite–Rosenblatt process with $H = d + \frac{1}{2}$, $\Rightarrow$ denotes weak convergence in $D[0,1]$, and $L_m(n) = m!C_m L_\gamma^m(n)$, see (4.21) and (4.22).

Note that this type of convergence requires long memory to be strong enough. In particular, if $m = 2$, we require $d \in (1/4, 1/2)$. If this is not the case, then the partial sum process has weak dependence properties.

Example 4.1 Assume that $m = 2$. If $d \in (1/4, 1/2)$, then

$$n^{-2d}L_2^{-1/2}(n)\sum_{t=1}^{[nu]}\big(X_t^2 - 1\big) \Rightarrow Z_{2,H}(u),$$

where

$$L_2(n) = 2C_2L_\gamma^2(n),$$

$$C_2 = \frac{1}{(2(2d-1)+1)(2d+1)}.$$

For each fixed $u \in [0,1]$, the limit is non-normal. This will be illustrated by simulations in computer Example 4.3 later in this section.

Proof of Theorem 4.3 The proof is almost a copy of the alternative proof of Theorem 4.2. We replace (4.28) by

$$H_m(X_t) = \int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} e^{it(\lambda_1+\cdots+\lambda_m)}\,dM_X(\lambda_1)\cdots dM_X(\lambda_m)$$

(we refer to Sect. 3.7.1.3 for the formula and the meaning of this integral). Recalling

$$dM_X(\lambda) = \sqrt{2\pi}a(\lambda)\,dM_\varepsilon(\lambda) = a(\lambda)\,dM_0(\lambda),$$

we have

$$S_{n,H_m}(1) = \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \frac{e^{in(\lambda_1+\cdots+\lambda_m)} - 1}{e^{i(\lambda_1+\cdots+\lambda_m)} - 1} \prod_{r=1}^{m} a(\lambda_r)\, dM_0(\lambda_1) \cdots dM_0(\lambda_m)$$
$$= \frac{n}{n^{m/2}} \int \cdots \int D_n\left(\frac{\lambda_1 + \cdots + \lambda_m}{n}\right)$$
$$\times \prod_{r=1}^{m} a\left(\frac{\lambda_r}{n}\right) n^{1/2}\, dM_0\big(n^{-1}\lambda_1\big) \cdots n^{1/2}\, dM_0\big(n^{-1}\lambda_m\big),$$

where the integration is over $[-n\pi, n\pi]^m$. Therefore, if $a(\lambda) = L_f^{1/2}(\lambda)|\lambda|^{-d}$, $d \in (0, 1/2)$, then we expect

$$n^{-(1-m(\frac{1}{2}-d))} L_f^{-m/2}\big(n^{-1}\big) S_{n,H_m}(1)$$
$$\xrightarrow{\mathrm{d}} \int_{\mathbb{R}^m} D(\lambda_1 + \cdots + \lambda_m) \prod_{r=1}^{m} \frac{1}{|\lambda_r|^d}\, dM_0(\lambda_1) \cdots dM_0(\lambda_m), \tag{4.35}$$

cf. (4.31). Again, we identify

$$L_m(n) = m! C_m \big(2\Gamma(1-2d)\sin \pi d\big)^m L_f^m\big(n^{-1}\big) = K_1^{-2}(m,d) L_f^m\big(n^{-1}\big),$$

and from Proposition 3.1 we recognize the representation of the Hermite–Rosenblatt process.

A precise argument for (4.35) is the same as in the case $m = 1$; see the proof of Proposition 4.2. Furthermore, we do not verify tightness here since it will be done in the next theorem. □

Finally, convergence of partial sums $S_{n,G}(u) = \sum_{t=1}^{[nu]} G(X_t)$ is just a consequence of Theorem 4.3, using the so-called reduction principle, proven originally in Taqqu (1975).

Theorem 4.4 *Assume that X_t $(t \in \mathbb{N})$ is a stationary sequence of standard normal random variables such that $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$ $(d \in (0, 1/2))$. Let $S_{n,G}(u) = \sum_{t=1}^{[nu]} G(X_t)$, where G is a function such that $E[G(X_1)] = 0$, $E[G^2(X_1)] < \infty$. If m is the Hermite rank of G and $m(1-2d) < 1$, then*

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) S_{n,G}(u) \Rightarrow \frac{J(m)}{m!} Z_{m,H}(u) \quad \big(u \in [0,1]\big),$$

where $Z_{m,H}(\cdot)$ is a Hermite–Rosenblatt process, $H = d + \frac{1}{2}$, $\Rightarrow$ denotes weak convergence in $D[0,1]$, and L_m is given in (4.21):

$$L_m(n) = m! C_m L_\gamma^m(n).$$

Proof Decompose

$$G(x) = \frac{J(m)}{m!} H_m(x) + \sum_{l=m+1}^{\infty} \frac{J(l)}{l!} H_l(x) =: \frac{J(m)}{m!} H_m(x) + G^*(x).$$

Using (4.18) and (4.25), we have

$$cov\left[\frac{J(m)}{m!} H_m(X_0), \frac{J(m)}{m!} H_m(X_k) \right] = \frac{J^2(m)}{m!} \gamma_X^m(k)$$

and

$$cov\big[G^*(X_0), G^*(X_k)\big] = \sum_{l=m+1}^{\infty} \frac{J^2(l)}{l!} \gamma_X^l(k).$$

Furthermore, for any t, s, the random variables $G^*(X_t)$ and $H_m(X_s)$ are uncorrelated. Therefore,

$$\begin{aligned} \operatorname{var}\left(\sum_{t=1}^{n} G(X_t) \right) &= \sum_{t=1}^{n} \sum_{s=1}^{n} E\big[G^*(X_t) G^*(X_s)\big] + \frac{J^2(q)}{m!} \sum_{t=1}^{n} \sum_{s=1}^{n} \gamma_X^m\big(|t-s|\big) \\ &= \sum_{t=1}^{n} \sum_{s=1}^{n} E\big[G^*(X_t) G^*(X_s)\big] + \left(\frac{J(m)}{m!} \right)^2 \operatorname{var}\left(\sum_{t=1}^{n} H_m(X_t) \right). \end{aligned} \tag{4.36}$$

The Hermite rank of the function G^* is at least $m+1$. Consequently, we have two scenarios. Either $\sum_k \gamma_X^m(k) < \infty$, and then both terms in (4.36) are of the order $O(n)$, or $\sum_k \gamma_X^m(k) = +\infty$, and then the second term dominates the first one. The latter happens if $m(1-2d) < 1$, and in this case the asymptotic behaviour of $\sum_{t=1}^{n} G(X_t)$ is the same as that of $(J(m)/m!) \sum_{t=1}^{n} H_m(X_t)$.

A proof of tightness is immediate. If we set

$$S'_{n,G}(u) := n^{-(m(d-1/2)+1)} L_m^{-m/2}(n) S_n(u),$$

we have

$$E\big[\big(S'_{n,G}(u) - S'_{n,G}(v)\big)^2\big] \sim |u-v|^{m(2d-1)+2}.$$

Since $m(1-2d) < 1$, the exponent is greater than one, and tightness follows from Lemma 4.3. □

In contrast, if the Hermite rank is large enough such that $m(1-2d) > 1$, then we have a weakly dependent-type behaviour of partial sums. The statement and proof of this result is postponed to the section on limit theorems for Appell polynomials.

Example 4.2 We illustrate the theoretical findings by a simulation example. First, we generate $n = 1000$ i.i.d. standard normal random variables X_t and plot the partial

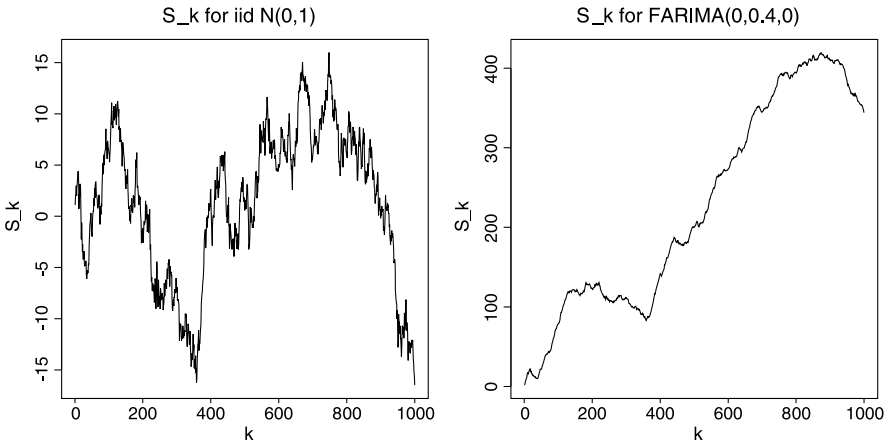

Fig. 4.1 Partial sum sequence $S_k = \sum_{t=1}^{k} X_t$ $(k = 1, \ldots, n)$ with X_t i.i.d. $N(0, 1)$ (*left*) and X_t generated by a FARIMA$(0, 0.4, 0)$ process (*right*)

sum sequence $S_k = \sum_{t=1}^{k} X_t$, $k = 1, \ldots, n$. This procedure is repeated for a Gaussian fractional ARIMA$(0, d, 0)$ process with parameter $d = 0.4$. The corresponding partial sum processes are plotted in Fig. 4.1. They can be considered approximations of a Brownian motion and a fractional Brownian motion with $H = 0.9$ respectively. Note that the path of the fractional Brownian motion is much smoother than the one of Brownian motion. This is due to long memory, which acts like a smoothing filter.

Example 4.3 In this example we generate $n = 1000$ random variables X_t from a Gaussian fractional ARIMA$(0, d, 0)$ process with parameter $d = 0.4$ and compute their sum. This procedure is repeated $N = 1000$ times. A normal probability plot of the $N = 1000$ sums $\sum_{t=1}^{n} X_t$ is displayed in the left panel of Fig. 4.2. The right panel shows a normal probability plot for the sums $\sum_{t=1}^{n} X_t^2$. The non-normal behaviour is clearly visible.

4.2.4 Linear Processes

In this section we consider a causal linear process

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \quad (t \in \mathbb{N}), \tag{4.37}$$

where, without loss of generality, $\sum_{j=0}^{\infty} a_j^2 = 1$, and ε_t $(t \in \mathbb{Z})$ are i.i.d. zero mean random variables with $\mathrm{var}(\varepsilon_1) = \sigma_\varepsilon^2 < \infty$. Thus, $\mathrm{var}(X_1) = \sigma_X^2 = \sigma_\varepsilon^2$. Note that Gaussian processes are included in this definition, but the class is much more general. Three different assumptions on the coefficients will be considered as $j \to \infty$ and with L_a denoting a slowly varying function at infinity:

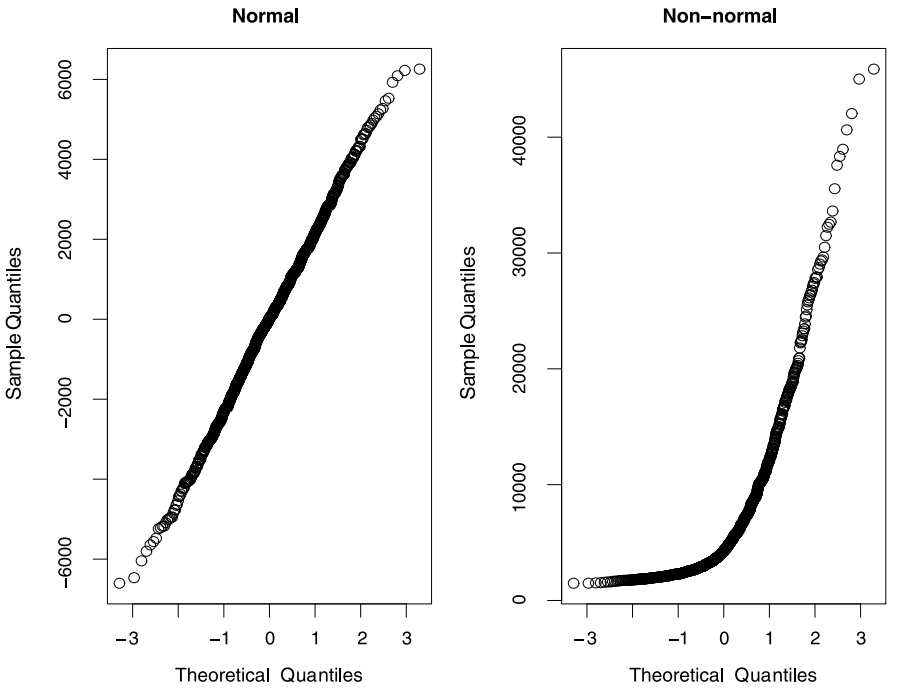

Fig. 4.2 Illustration of Theorem 4.3: normal probability plots of partial sums $\sum_{t=1}^{k} X_t$ (*left*) and $\sum_{t=1}^{k} X_t^2$, where X_t is generated by a FARIMA$(0, 0.4, 0)$ process

- (B1) long memory:

$$a_j \sim L_a(j) j^{d-1} \quad \left(0 < d < \frac{1}{2}\right);$$

- (B2) short memory:

$$\sum_{j=0}^{\infty} |a_j| < \infty, \qquad \sum_{j=0}^{\infty} a_j \neq 0.$$

- (B3) antipersistence:

$$a_j \sim L_a(j) j^{d-1}$$

with $-\frac{1}{2} < d < 0$, and

$$\sum_{j=0}^{\infty} a_j = 0.$$

Under the short-memory assumption (B2), limiting behaviour is classical (see Theorem 4.5); see Brockwell and Davis (1991). Under long memory (B1), the first

result was obtained in Davydov (1970a, 1970b); see also Gorodetskii (1977), Lang and Soulier (2000), Wang et al. (2003).

4.2.4.1 Asymptotic Covariances and Normalizing Constants

The behaviour of the autocovariance function γ_X and the spectral density f_X for the three cases can be characterized as follows. Combining Lemmas 4.13–4.15 with Lemmas 4.9–4.11, respectively, yields the asymptotic behaviour of $\operatorname{var}(S_n)$ (where $S_n(u) = \sum_{t=1}^{[nu]} X_t$, $S_n = S_n(1)$).

Lemma 4.13 *Under assumption* (B1), *we have, as* $\lambda \to 0$ *and* $k \to \infty$ *respectively,*

$$\begin{aligned} f_X(\lambda) &\sim L_f(\lambda)|\lambda|^{-2d}, \\ \gamma_X(k) &\sim L_\gamma(k)k^{2d-1}, \end{aligned} \tag{4.38}$$

where

$$L_\gamma(k) = L_a^2(k) \cdot \sigma_\varepsilon^2 \int_0^\infty v^{d-1}(1+v)^{d-1}\,dv = \sigma_\varepsilon^2 L_a^2(k) B(1-2d, d), \tag{4.39}$$

$B(x, y)$ *denotes the Beta function, and* L_f *is obtained from* L_γ *by* (*cf.* (1.1))

$$L_f(\lambda) = L_\gamma\big(\lambda^{-1}\big)\pi^{-1}\Gamma(2d)\sin\left(\frac{\pi}{2} - \pi d\right). \tag{4.40}$$

Hence, via Lemma 4.9,

$$\operatorname{var}(S_n) \sim L_S(n)n^{2d+1} = \frac{1}{d(2d+1)} L_\gamma(n)n^{2d+1}. \tag{4.41}$$

Proof We have

$$\gamma_X(k) \sim \sigma_\varepsilon^2 \sum_{j=1}^{\infty} L_a(j)L_a(j+k)j^{d-1}(j+k)^{d-1} = \sigma_\varepsilon^2 S_{\infty,k} \cdot k^{2d-1},$$

where

$$S_{\infty,k} = \lim_{n\to\infty} S_{n,k}$$

and

$$\begin{aligned} S_{n,k} &= \sum_{j=1}^{nk} L_a(j)L_a(j+k)\left(\frac{j}{k}\right)^{d-1}\left(\frac{j}{k}+1\right)^{d-1} n^{-1} \\ &= L_a^2(k)\sum_{j=1}^{nk} \frac{L_a(j)}{L_a(k)}\frac{L_a(j+k)}{L_a(k)}\left(\frac{j}{k}\right)^{d-1}\left(\frac{j}{k}+1\right)^{d-1} n^{-1} \\ &\underset{k\to\infty}{\sim} L_a^2(k)\int_0^n v^{d-1}(v+1)^{d-1}\,dv, \end{aligned}$$

where the last approximation is uniform in n. The approximation formula for f_X follows from Theorem 1.3. □

Example 4.4 (ARFIMA Model) Consider an ARFIMA$(0, d, 0)$ model, $d \in (0, 1/2)$. This process has the linear representation $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$, where

$$a_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} \sim \frac{1}{\Gamma(d)} j^{d-1} \quad (j \to \infty).$$

Thus, $L_a \sim 1/\Gamma(d)$, so that

$$\gamma_X(k) \sim c_\gamma k^{2d-1}$$

with

$$\begin{aligned} c_\gamma &= \sigma_\varepsilon^2 \Gamma^{-2}(d) \int_0^\infty v^{d-1}(1+v)^{d-1}\, dv \\ &= \sigma_\varepsilon^2 \Gamma^{-2}(d) B(1-2d, d) = \sigma_\varepsilon^2 \frac{\Gamma(1-2d)\Gamma(d)}{\Gamma^2(d)\Gamma(1-d)} \\ &= \sigma_\varepsilon^2 \frac{\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)} = \frac{\sigma_\varepsilon^2}{\pi} \Gamma(1-2d)\sin(\pi d). \end{aligned}$$

The last equality follows from $\Gamma(d)\Gamma(1-d) = \pi/\sin \pi d$. Moreover,

$$\begin{aligned} L_f(\lambda) &= \frac{\sigma_\varepsilon^2}{\pi} \Gamma(1-2d)\sin(\pi d)\pi^{-1}\Gamma(2d)\sin\left(\frac{\pi}{2} - \pi d\right) \\ &= \frac{\sigma_\varepsilon^2}{\pi} \frac{\sin(\pi d)\sin(\frac{\pi}{2} - \pi d)}{\sin(2\pi d)} = \frac{\sigma_\varepsilon^2}{\pi} \frac{\sin(\pi d)\cos(\pi d)}{\sin(2\pi d)} \\ &= \frac{\sigma_\varepsilon^2}{\pi} \frac{\sin(\pi d)\cos(\pi d)}{2\sin(\pi d)\cos(\pi d)} = \frac{\sigma_\varepsilon^2}{2\pi}, \end{aligned}$$

so that

$$f_X(\lambda) \sim \frac{\sigma_\varepsilon^2}{2\pi} |\lambda|^{-2d}.$$

Lemma 4.14 *Under assumption* (B2), *we have*

$$\sum_{k=-\infty}^{\infty} \left|\gamma_X(k)\right| < \infty, \qquad \sum_{k=-\infty}^{\infty} \gamma_X(k) > 0.$$

If, in addition, $\sum_{j=0}^{\infty} j|a_j| < \infty$, *then* $f_X(\lambda)$ *is continuous on* $[-\pi, \pi]$.

Proof We have

$$\sum_{k=-\infty}^{\infty} \left|\gamma_X(k)\right| = \sigma_\varepsilon^2 \sum_{k=-\infty}^{\infty} \left| \sum_{j=0}^{\infty} a_j a_{j+|k|} \right| \le 2\sigma_\varepsilon^2 \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} |a_j||a_{j+|k|}|$$

$$= 2\sigma_\varepsilon^2 \left(\sum_{j=0}^{\infty} |a_j| \right)^2 < \infty.$$

Furthermore,

$$\sum_{k=-\infty}^{\infty} \gamma_X(k) = 2\pi f_X(0) = 2\pi \frac{\sigma_\varepsilon^2}{2\pi} \left| \sum_{j=0}^{\infty} a_j \right|^2 > 0.$$

To show that f_X is continuous, consider

$$\tilde{a}(\lambda) = \sum_{j=0}^{\infty} a_j e^{-ij\lambda}.$$

Since, as $x \to 0$, $\sin x \sim x$ and $\cos x - 1 \sim x^2/2$, we obtain for $\varepsilon < 1$,

$$\left|\tilde{a}(\lambda+\varepsilon) - \tilde{a}(\lambda)\right| \le \sum_{j=0}^{\infty} |a_j| \left| e^{-ij\lambda} \left(e^{-ij\varepsilon} - 1 \right) \right|$$

$$\le 2\varepsilon \sum_{j=0}^{\infty} j|a_j|,$$

so that $\tilde{a}(\cdot)$ is continuous, and hence so is $f_X(\lambda) = \sigma_\varepsilon^2/(2\pi)\,|\tilde{a}(\lambda)|^2$. □

Lemma 4.15 *Under assumption* (B3), *we have, as* $\lambda \to 0$ *and* $k \to \infty$ *respectively,*

$$f_X(\lambda) \sim L_f(\lambda)|\lambda|^{-2d}, \tag{4.42}$$

$$\gamma_X(k) \sim L_\gamma(k) k^{2d-1}, \qquad \sum_{k=-\infty}^{\infty} \gamma_X(k) = 0, \tag{4.43}$$

where

$$L_\gamma(k) = L_a^2(k) \cdot \sigma_\varepsilon^2 \int_0^\infty v^{d-1} \left[1 - (v+1)^{d-1} \right] du$$

$$= \sigma_\varepsilon^2 L_a^2(k) B(1-2d, d),$$

and L_f *is obtained from* L_γ *by* (4.40).

Proof Similarly to the proof of Lemma 4.13,

$$\gamma_X(k) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} a_j a_{j+k} = \sigma_\varepsilon^2 S_{\infty,k} \cdot k^{2d-1}$$

with $S_{\infty,k} = \lim_{n\to\infty} S_{n,k}$,

$$S_{n,k} = k^{1-2d} \sum_{j=0}^{nk} a_j a_{j+k} = S_{n,k}(1) + S_{n,k}(2)$$

and

$$S_{n,k}(1) = k^{1-2d} \sum_{j=0}^{nk} a_j (a_{j+k} - a_k) \sim L_a^2(n) \int_0^n v^{d-1} \big[(v+1)^d - 1\big]\, du,$$

$$S_{n,k}(2) = k^{1-2d} a_k \sum_{j=0}^{nk} a_j = -k^{1-2d} a_k \sum_{j=nk+1}^{\infty} a_j \sim L_a^2(n) \int_n^\infty v^{d-1}\, dv = o(n),$$

where the approximations are uniform in n. Moreover,

$$\sum_{k=-\infty}^{\infty} \gamma_X(k) = 2\pi f_X(0) = 2\pi \frac{\sigma_\varepsilon^2}{2\pi} \left| \sum_{j=0}^{\infty} a_j \right|^2 = 0.$$

The approximation of f_X for $\lambda \to 0$ follows from Theorem 1.3. □

4.2.4.2 Asymptotic Distribution

Proofs of the next results illustrate different techniques that are applicable in various situations:

- Under short memory (B2), we apply the K-dependent approximation method, i.e. a combination of Proposition 4.1 and Lemma 4.1. This is easier than the cumulant method and does not require restrictive moment assumptions. It is particularly suited for linear processes (see Brockwell and Davis 1991).
- Under long memory (B1), we apply the method based on random spectral measures, as outlined in the alternative proof of Theorem 4.2; see Lang and Soulier (2000).

Theorem 4.5 *Assume that* X_t $(t \in \mathbb{N})$ *is a stationary linear process* (4.37) *such that* (B2) *holds. Then*

$$n^{-1/2} S_n = n^{-1/2} \sum_{t=1}^{n} X_t \to N\big(0, v^2\big),$$

where the variance $v^2 = \sigma_X^2 + 2\sum_{k=1}^{\infty} \gamma_X(k)$.

This theorem can be formulated in terms of functional convergence to Brownian motion.

Proof Let $X_{t,K} = \sum_{j=0}^{K} a_j \varepsilon_{t-j}$. Since the sequence $X_{t,K}$ $(t \in \mathbb{N})$ is K-dependent, an application of Lemma 4.1 yields

$$n^{-1/2} S_{n,K} = n^{-1/2} \sum_{t=1}^{n} X_{t,K} \xrightarrow{d} N\big(0, \nu_K^2\big)$$

with $\nu_K^2 = \operatorname{var}(X_{0,K}) + 2\sum_{k=0}^{K} \gamma_{X_K}(k)$, where

$$\gamma_{X_K}(k) = E[X_{t,K} X_{t+k,K}] = \sigma_\varepsilon^2 \sum_{j=0}^{K} a_j a_{j+k}.$$

Since $\nu_K \to \nu$ as $K \to \infty$, we conclude $N(0, \nu_K^2) \xrightarrow{d} N(0, \nu^2)$. It suffices to prove that for all $\delta > 0$,

$$\lim_{K\to\infty} \limsup_{n\to\infty} P\big(n^{-1/2} |S_n - S_{n,K}| > \delta\big) = 0.$$

The result of our theorem will then follow by Proposition 4.1. By Markov's inequality, it is sufficient to verify that

$$\lim_{K\to\infty} \lim_{n\to\infty} n^{-1} \operatorname{var}(S_n - S_{n,K}) = 0.$$

Let $\bar{X}_{t,K} = X_t - X_{t,K}$. Then

$$\begin{aligned}\lim_{n\to\infty} n^{-1} \operatorname{var}(S_n - S_{n,K}) &= \lim_{n\to\infty} \sigma_\varepsilon^2 \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right) \sum_{j=K+1}^{\infty} a_j a_{j+k} \\ &= \sigma_\varepsilon^2 \sum_{k=-\infty}^{\infty} \sum_{j=K+1}^{\infty} a_j a_{j+k} = \sigma_\varepsilon^2 \sum_{j=K+1}^{\infty} a_j \sum_{k=-\infty}^{\infty} a_{j+k}.\end{aligned}$$

The $\lim_{n\to\infty}$ behaviour above is obtained by applying the dominated convergence theorem. For this, we need $\sum_k \sum_j |a_j a_{j+k}| < \infty$. This is true under the summability condition $\sum_{j=0}^{\infty} |a_j| < \infty$. Under this condition, we can also exchange the summations $\sum_k$ and $\sum_j$. Finally,

$$\lim_{K\to\infty} \lim_{n\to\infty} n^{-1} \operatorname{var}(S_n - S_{n,K}) \le \sum_{k=-\infty}^{\infty} |a_k| \lim_{m\to\infty} \sum_{j=K+1}^{\infty} |a_j| = 0. \qquad \square$$

Under (B1), the asymptotic behaviour of partial sums changes. This result was proven first in Davydov (1970a, 1970b). The method below is adapted from Lang and Soulier (2000), where the reader is referred to for details.

Theorem 4.6 *Assume that X_t $(t \in \mathbb{N})$ is a stationary linear process (4.37) such that the long-memory condition* (B1) *holds, i.e. $a_j \sim L_a(j)\, j^{d-1}$, $d \in (0, \frac{1}{2})$. Then*

$$n^{-(d+\frac{1}{2})} L_S^{-1/2}(n) S_n(u) = n^{-(d+\frac{1}{2})} L_S^{-1/2}(n) \sum_{t=1}^{[nu]} X_t \Rightarrow B_H(u) \quad \big(u \in [0,1]\big),$$

where $B_H(u)$ is a standard fractional Brownian motion, $H = d + \frac{1}{2}$, $\Rightarrow$ denotes weak convergence in $D[0,1]$, and

$$L_S(n) = \frac{1}{d(2d+1)} L_\gamma(n)$$

with L_γ defined in (4.39):

$$\begin{aligned} L_\gamma(k) &= L_a^2(k)\sigma_\varepsilon^2 \int_0^\infty v^{d-1}(v+1)^{d-1}\, dv \\ &= L_a^2(k)\sigma_\varepsilon^2 B(1-2d, d). \end{aligned}$$

Proof We use the spectral method, as in the alternative proof of Theorem 4.2. Recall that any stationary sequence with finite variance can be written as

$$\varepsilon_t = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{it\lambda} M_0(d\lambda), \quad t \in \mathbb{Z}.$$

The only difference between the spectral measure M_0 here and M_0 in the proof of Theorem 4.2 is that the measure here is not necessarily Gaussian. In particular, there is no guarantee that $n^{1/2} M_0(n^{-1}\cdot)$ and $M_0(\cdot)$ have the same distribution. Nevertheless, the same argument can be applied (see Lang and Soulier 2000). □

Example 4.5 (ARFIMA) Assume that X_t $(t \in \mathbb{N})$ is a FARIMA$(0, d, 0)$ model as in Example 4.4. Then

$$\begin{aligned} \gamma_X(k) &\sim c_\gamma k^{2d-1}, \\ c_\gamma &= \frac{\sigma_\varepsilon^2}{\pi} \Gamma(1-2d) \sin(\pi d). \end{aligned}$$

Hence,

$$n^{-(d+\frac{1}{2})} L_S^{-1/2}(n) \sum_{t=1}^{[nu]} X_t \Rightarrow B_H(u)$$

and

$$L_S(n) = c_\gamma \frac{1}{d(2d+1)}.$$

Note that the innovations ε_t do not need to be Gaussian.

4.2.5 Subordinated Linear Processes

Next we consider the case where instead of the linear process X_t $(t \in \mathbb{N})$ a subordinated process, i.e. a transformation $Y_t = G(X_t)$ $(t \in \mathbb{N})$, is observed. Recall that in the Gaussian case asymptotic properties of partial sums of X_t and $H_m(X_t)$ (and, via the reduction principle of Theorem 4.4, of general functionals) can be studied using the spectral method. For linear processes, we applied again the spectral method in Theorem 4.6. However, this extension is not feasible for subordinated linear processes. In this setup, there are two common approaches: Appell polynomials (Surgailis 1982; Giraitis 1985; Giraitis and Surgailis 1986, 1989; Avram and Taqqu 1987; Surgailis and Vaičiulis 1999; Surgailis 2000; see also Surgailis 2003, for overview) and a martingale decomposition (Ho and Hsing 1996, 1997; Wu 2003; see also Hsing 2000 for an overview).

4.2.5.1 Normalizing Constants: Simple Example

Before we develop a general formula, let us consider the simple case of $G(X_t) = X_t^2$.

Example 4.6 Let X_t $(t \in \mathbb{N})$ be a linear process defined by (4.37). Assume that $E[\varepsilon_1^4] < \infty$ and that the long-memory condition (B1) holds. Using formula (4.38) for the covariance of X_t $(t \in \mathbb{N})$, we have

$$\gamma_X^2(k) \sim L_\gamma^2(k) k^{2(2d-1)}.$$

On the other hand,

$$\gamma_X^2(k) = cov^2(X_t, X_{t+k}) = \left(\sum_{j=0}^{\infty} a_j a_{j+k} \right)^2$$

$$= \sum_{j=0}^{\infty} a_j^2 a_{j+k}^2 + \sum_{j,l=0;\ j \neq l}^{\infty} a_j a_l a_{j+k} a_{l+k}.$$

Note that under (B1) the limiting behaviour of $\gamma_X^2(k)$ is determined by the second term. Now,

$$X_0^2 = \sum_{j=0}^{\infty} a_j^2 \varepsilon_{0-j}^2 + \sum_{j,l=0;\ j \neq l}^{\infty} a_j a_l \varepsilon_{0-j} \varepsilon_{0-l} =: X_{0,1} + X_{0,2}.$$

Analogously, we define $X_k^2 := X_{k,1} + X_{k,2}$. Note that $X_{0,1}$ and $X_{k,2}$ are uncorrelated. The same holds for $X_{0,2}$ and $X_{k,1}$. Furthermore,

$$cov(X_{0,1}, X_{k,1}) = E\left[\varepsilon_1^4\right] \sum_{j=0}^{\infty} a_j^2 a_{j+k}^2$$

and

$$cov(X_{0,2}, X_{k,2}) = 2 \sum_{j,l=0;\ j\neq l}^{\infty} a_j a_l a_{j+k} a_{l+k}.$$

Recalling that the second covariance is of a larger order than the first one, we conclude

$$\gamma_{X^2}(k) \sim 2 \sum_{j,l=0;\ j\neq l}^{\infty} a_j a_l a_{j+k} a_{l+k} \sim 2\gamma_X^2(k) \sim 2L_\gamma^2(k) k^{2(2d-1)}.$$

4.2.5.2 Normalizing Constants: Appell Polynomials

Now, we turn our attention to general nonlinear functionals. For a general non-normal distribution, in view of Sect. 3.3, a natural approach is to start with the Wick product $Y_t = A_m(X_t) = :X_t, \ldots, X_t:$ where A_m is the mth Appell polynomial associated with the marginal distribution of X_t. Suppose that $\gamma_X(k)$ is known, either exactly or its asymptotic behaviour. Can we give a simple formula for $\gamma_Y(k)$? In principle, the diagram formulas given in Theorem 3.10 provide an answer because

$$\kappa(Y_t, Y_{t+k}) = \left[\frac{\partial^2}{\partial z_1 \partial z_2} \log E\big[\exp(z_1 Y_t + z_2 Y_{t+k})\big] \right]_{z=0} = \gamma_Y(k).$$

To apply the diagram formula, consider a table W with two rows W_1, W_2 of length m. The positions in W_1 are associated with X_t and those in W_2 with X_{t+k}, i.e. we may write $W_1 = \{\tilde{X}_{(1,1)}, \ldots, \tilde{X}_{(1,m)}\}$ with $\tilde{X}_{(1,t)} = X_t$ and $W_2 = \{\tilde{X}_{(2,1)}, \ldots, \tilde{X}_{(2,m)}\}$ with $\tilde{X}_{(2,j)} = X_{j+k}$. Using the same notation as in Theorem 3.10, we obtain from (3.81)

$$\gamma_Y(k) = \kappa\big(:X^{W_1}:, :X^{W_2}:\big) = \sum_{\gamma \in \Gamma_W^{\neq,c}} \kappa\big(X'^{V_1}\big) \cdots \kappa\big(X'^{V_r}\big). \tag{4.44}$$

Unfortunately, this is a rather complicated expression because in general $\kappa(X'^V)$ may not be zero for any subset V. There is one exception where (4.44) simplifies considerably, namely if X_t ($t \in \mathbb{N}$) is a Gaussian process. In this case, all cumulants $\kappa(X'^V)$ are zero except for normal edges, i.e. $\kappa(X'^V) = 0$ if $|V| \neq 2$, so that the sum in (4.44) is over $\Gamma_W^{\neq,c,\mathcal{N}}$, and, up to a constant, we obtain a sum of correlations to the power m, see Corollary 3.5.

Although (4.44) is complicated, it is possible to give simple asymptotic formulas for $\gamma_Y(k)$ and, consequently, the variance of $S_{n,A_m} = \sum_{t=1}^n A_m(X_t)$. A first simplification can be obtained in the representation of Appell polynomials of linear processes:

Lemma 4.16 *Let X_t $(t \in \mathbb{N})$ be a linear process (4.37) such that the Appell polynomials of its marginal distribution A_m $(m \in \mathbb{N})$ exist. Then*

$$A_m(X_t) = \sum_{k_1,\ldots,k_m=0}^{\infty} a_{k_1} \cdots a_{k_m} (:\varepsilon_{t-k_1} \cdots \varepsilon_{t-k_m}:). \tag{4.45}$$

Proof The result follows from

$$A_m(X_t) = :\underbrace{X_t, \ldots, X_t}_{m}:$$

and multilinearity of the Wick product. □

A direct consequence of this result is a simplified expression for S_n:

Corollary 4.1 *Let X_t $(t \in \mathbb{N})$ be a linear process defined by (4.37) such that the Appell polynomials of its marginal distribution A_m $(m \in \mathbb{N})$ exist. Let*

$$S_{n,A_m} = \sum_{t=1}^{n} A_m(X_t).$$

Then

$$S_{n,A_m} = \sum_{k_1,\ldots,k_m=0}^{\infty} a_{k_1} \cdots a_{k_m} \sum_{t=1}^{n} (:\varepsilon_{t-k_1} \cdots \varepsilon_{t-k_m}:)$$

with $a_k = 0$ for $k < 0$.

Furthermore, the diagram formula can be used to obtain an expression for the asymptotic autocovariance function of the subordinated sequence Y_t $(t \in \mathbb{N})$ under long memory:

Corollary 4.2 *Let X_t $(t \in \mathbb{N})$ be a linear process defined by (4.37) such that the Appell polynomials of its marginal distribution A_m $(m \in \mathbb{N})$ exist and the long-memory assumption* (B1) *holds. Then $Y_t = A_m(X_t)$ has an autocovariance function $\gamma_Y(k)$ with*

$$\begin{aligned} \gamma_Y(k) &\sim m! \gamma_X^m(k) \\ &\sim m! \left(L_a^2(k) \sigma_\varepsilon^2 \int_0^\infty v^{d-1} (v+1)^{d-1} \, dv \right)^m \cdot k^{(2d-1)m} \\ &= m! L_\gamma^m(k) k^{(2d-1)m} \end{aligned} \tag{4.46}$$

as $k \to \infty$, cf. (4.39).

Proof Here, only an outline of the extended proof in Giraitis and Surgailis (1989) and Surgailis and Vaičiulis (1999) is given. Lemma 4.16 and the multilinearity of cumulants imply

$$\begin{aligned}
&cov\big(A_m(X_t), A_m(X_{t+k})\big) \\
&\quad = \kappa\big(A_m(X_t), A_m(X_{t+k})\big) \\
&\quad = \kappa\Bigg(\sum_{j_1,\dots,j_m=0}^{\infty} a_{j_1}\cdots a_{j_m}(:\varepsilon_{t-j_1}\cdots\varepsilon_{t-j_m}:), \\
&\qquad\qquad \sum_{j_1,\dots,j_m=0}^{\infty} a_{j_1}\cdots a_{j_m}(:\varepsilon_{t+k-j_1}\cdots\varepsilon_{t+k-j_m}:)\Bigg) \\
&\quad = \sum_{\substack{j_1,\dots,j_m=0,\\ j'_1,\dots,j'_m=0}}^{\infty} a_{j_1}\cdots a_{j_m}a_{j'_1}\cdots a_{j'_m}\kappa(:\varepsilon_{t-j_1}\cdots\varepsilon_{t-j_m}:,:\varepsilon_{t+k-j'_1}\cdots\varepsilon_{t+k-j'_m}:).
\end{aligned}$$

Now consider a table W with two rows $W_i = \{\varepsilon_{(i,1)},\dots,\varepsilon_{(i,m)}\}$ $(i = 1, 2)$ with $\varepsilon_{(1,s)} = \varepsilon_{t_s}$ and $\varepsilon_{(2,s)} = \varepsilon_{t'_s}$. The diagram formula for cumulants of Wick products implies

$$\kappa(:\varepsilon_{t-j_1},\dots,\varepsilon_{t-j_m}:,:\varepsilon_{t+k-j'_1},\dots,\varepsilon_{t+k-j'_m}:) = \sum_{\gamma\in\Gamma_W^{\neq,c}} \kappa\big(\varepsilon'^{V_1}\big)\cdots\kappa\big(\varepsilon'^{V_r}\big).$$

Using this equation, we have

$$\kappa\big(A_m(X_t), A_m(X_{t+k})\big) = r_{\text{main}} + r_k,$$

where

$$r_{\text{main}} = \sum_{\gamma\in\Gamma_W^{\neq,c,\mathcal{N}}} \sum_{\substack{j_1,\dots,j_m=0\\ j'_1,\dots,j'_m=0}} \Bigg(\prod_{i=1}^{m} a_{j_i}a_{j'_i}\Bigg)\kappa\big(\varepsilon'^{V_1}\big)\cdots\kappa\big(\varepsilon'^{V_r}\big)$$

and

$$r_k = \sum_{\gamma\in\Gamma_W^{\neq,c}\setminus\Gamma_W^{\neq,c,\mathcal{N}}} \sum_{\substack{j_1,\dots,j_m=0\\ j'_1,\dots,j'_m=0}} \Bigg(\prod_{i=1}^{m} a_{j_i}a_{j'_i}\Bigg)\kappa\big(\varepsilon'^{V_1}\big)\cdots\kappa\big(\varepsilon'^{V_r}\big).$$

It can be shown that, as $k\to\infty$, $r_k = o(k^{(2d-1)m})$, so that only diagrams in $\Gamma_W^{\neq,c,\mathcal{N}}$ matter asymptotically. For instance, for $\gamma = \bigcup_{i=1}^{m-1} V_i$ with $V_i = \{(1,i),(2,i)\}$ $(i = 1,\dots,m-2)$ and $V_{m-1} = \{(1,m-1),(2,m-1),(1,m),(2,m)\}$, we have, because

of independence of the random variables ε_i,

$$\kappa\left(\varepsilon'^{V_1}\right)\cdots\kappa\left(\varepsilon'^{V_{m-1}}\right)=0,$$

unless $j'_1=j_1+k,\ldots,j'_{m-1}=j_{m-1}+k$ and $j_{m-1}=j_m$, $j'_{m-1}=j'_m=j_{m-1}+k$. Thus, the contribution of γ to r_m is

$$\sigma_\varepsilon^2\left(\sum_{j=0}^{\infty}a_ja_{j+k}\right)^{m-2}\sum_{j=0}^{\infty}a_j^2a_{j+k}^2\sim\gamma_X^{m-2}(k)L(k)k^{4d-3}=o\left(k^{(2d-1)m}\right).$$

For κ_{main}, the calculation simplifies considerably because each $\gamma\in\Gamma_W^{\neq,c,\mathcal{N}}$ consists of edges $V_j=\{(1,j),(1,\pi(j))\}$ $(j=1,2,\ldots,m)$ where π is a permutation of $\{1,2,\ldots,m\}$. Thus, the number of diagrams in $\Gamma_W^{\neq,c,\mathcal{N}}$ is $|\Gamma_W^{\neq,c,\mathcal{N}}|=m!$. Moreover, for each permutation π,

$$\sum_{\substack{j_1,\ldots,j_m=0\\ j'_1,\ldots,j'_m=0}}\left(\prod_{i=1}^{m}a_{j_i}a_{j'_i}\right)\kappa\left(\varepsilon'^{V_1}\right)\cdots\kappa\left(\varepsilon'^{V_r}\right)=\sigma_\varepsilon^{2m}\left(\sum_{j=0}^{\infty}a_ja_{j+k}\right)^m=\gamma_X^m(k).$$

Thus, taking the sum over all $m!$ permutations, we have

$$r_{\text{main}}=m!\gamma_X^m(k).$$ □

Note that, if X_t $(t\in\mathbb{N})$ is a Gaussian process, then we have the exact relationship $\gamma_{A_m}(k)=m!\gamma_X^m(k)$ for any finite k because all cumulants above order 2 are zero, so that all contributions except those from $\Gamma_W^{\neq,c,\mathcal{N}}$ are zero. (cf. Sect. 4.2.3).

The combination of Lemma 4.9 and formula (4.38) yields an asymptotic formula for the variance of $S_{A_m,n}=\sum_{t=1}^{n}A_m(X_t)$ under the assumption of long memory (see Giraitis and Surgailis 1989; Surgailis and Vaičiulis 1999):

Theorem 4.7 *Let X_t $(t\in\mathbb{N})$ be a linear process defined by (4.37) such that the Appell polynomials A_m $(m\in\mathbb{N})$ of its marginal distribution exist and the long-memory assumption* (B1) *holds. Assume further that $m(1-2d)<1$. Then, as $n\to\infty$,*

$$\operatorname{var}(S_{n,A_m})=\operatorname{var}\left(\sum_{t=1}^{n}A_m(X_t)\right)\sim L_m(n)n^{(2d-1)m+2}$$

with

$$\begin{aligned}L_m(n)&=m!C_mL_\gamma^m(n),\\ C_m&=\frac{2}{((2d-1)m+1)((2d-1)m+2)}\end{aligned}\tag{4.47}$$

and L_γ given by (4.39). On the other hand, if $m(1-2d)>1$, then

$$\operatorname{var}(S_{n,A_m})=O(n).$$

We recognize the same formula as in the Gaussian case, see (4.20). Furthermore, note that, in general, antipersistence is not inherited because the condition that autocovariances add up to zero is destroyed much more easily than nonsummability.

4.2.5.3 Asymptotic Distributions: Appell Polynomials

In the previous sections we obtained asymptotic expressions for the autocovariance function $\gamma_{A_m}(k) = cov(A_m(X_t), A_m(X_{t+k}))$ and the variance $v_n^2 := \mathrm{var}(S_{n,A_m})$. The remaining question is which processes one obtains as limits of $S_{n,A_m}(t)/v_n$. It turns out that, under suitable moment conditions, the only possible limiting processes are Hermite–Rosenblatt processes. In fact this question has been answered in the Gaussian case, see Theorem 4.4.

Theorem 4.8 *Let X_t $(t \in \mathbb{N})$ be a linear process defined by (4.37) such that the Appell polynomials A_m $(m \in \mathbb{N})$ of its marginal distribution exist and the long-memory assumption* (B1) *holds, i.e. $a_j \sim L_a(j)j^{d-1}$, $d \in (0, 1/2)$. Let*

$$S_{n,A_m}(u) = \sum_{t=1}^{[nu]} A_m(X_t) \quad \big(u \in [0,1]\big)$$

and assume that $E(\varepsilon_1^{2j}) < \infty$ for all j. Then, if $m(1-2d) < 1$,

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) S_{n,A_m}(u) \Rightarrow Z_{m,H}(u) \quad \big(u \in [0,1]\big), \tag{4.48}$$

where $Z_{m,H}(\cdot)$ is the Hermite–Rosenblatt process with $H = d + \frac{1}{2}$, $\Rightarrow$ denotes weak convergence in $D[0,1]$, and L_m is given in (4.47):

$$L_m(n) = m! C_m L_\gamma^m(n),$$

$$C_m = \frac{2}{((2d-1)m+1)((2d-1)m+2)},$$

with L_γ given by (4.39):

$$L_\gamma(k) = L_a^2(k) \cdot \sigma_\varepsilon^2 \int_0^\infty v^{d-1}(v+1)^{d-1}\, dv.$$

On the other hand, if $m(1-2d) > 1$, then $\mathrm{var}(S_{n,A_m}) \sim \sigma_S n$ for some $\sigma_S > 0$, and

$$n^{-\frac{1}{2}} S_{n,A_m}(u) \Rightarrow \sigma_S B(u) \quad \big(u \in [0,1]\big), \tag{4.49}$$

where $B(\cdot)$ is a standard Brownian motion, and $\Rightarrow$ denotes weak convergence in $D[0,1]$.

In other words, the asymptotic distribution is the same as in case of Hermite polynomials. Moreover, L_m agrees with L_m in Theorem 4.3.

Proof At first consider the case with $m(1-2d) > 1$. The proof is rather long, so that only a sketch is given here (for details, see e.g. Surgailis 2003). To prove the convergence of finite-dimensional distributions, we use the cumulant method (cf. Theorem 4.1). Recall that for the normal distribution, all cumulants of order $j \geq 3$ equal zero, and there is no other distribution with this property. It is therefore sufficient to show that for $j \geq 3$,

$$\lim_{n\to\infty} \kappa_j\big(n^{-\frac{1}{2}} S_{n,A_m}(t)\big) = n^{-\frac{j}{2}} \lim_{n\to\infty} \kappa\big(\underbrace{S_{n,A_m}(t), \ldots, S_{n,A_m}(t)}_{j}\big) = 0.$$

Without loss of generality, we may fix t at $t = 1$, and we write $S_{n,A_m} = S_{n,A_m}(1)$. Now for $s_1, \ldots, s_j \in \mathbb{N}$, consider a table W with rows

$$W_r = \{X_{(r,1)} = X_{s_r}, \ldots, X_{(r,j)} = X_{s_r}\} \quad (1 \leq r \leq j).$$

Then, because of multilinearity of κ,

$$\begin{aligned}\kappa(S_{n,A_m}, \ldots, S_{n,A_m}) &= \sum_{s_1,\ldots,s_j=1}^{n} \kappa\big(A_m(X_{s_1}), \ldots, A_m(X_{s_j})\big) \\ &= \sum_{s_1,\ldots,s_j=1}^{n} \kappa\big(:X^{W_1}:, \ldots, :X^{W_j}:\big).\end{aligned}$$

The diagram formula implies

$$\kappa\big(:X^{W_1}:, \ldots, :X^{W_j}:\big) = \sum_{\gamma\in\Gamma_W^{\neq,c}} \kappa\big(X'^{V_1}\big)\cdots\kappa\big(X'^{V_r}\big),$$

and hence,

$$\begin{aligned}\kappa_j\big(n^{-\frac{1}{2}} S_{n,A_m}(t)\big) &= \sum_{\gamma\in\Gamma_W^{\neq,c}} n^{-\frac{j}{2}} \sum_{s_1,\ldots,s_j=1}^{n} \kappa\big(X'^{V_1}\big)\cdots\kappa\big(X'^{V_r}\big) \\ &= \sum_{\gamma\in\Gamma_W^{\neq,c}} n^{-\frac{j}{2}} J_{n,\gamma}.\end{aligned}$$

Since the number of diagrams in $\Gamma_W^{\neq,c}$ is finite and does not depend on n, it is sufficient to show that $n^{-\frac{j}{2}} J_{n,\gamma}$ converges to zero. Note first that, for any $s_1, \ldots, s_j$ and $V \subseteq W$,

$$\kappa\big(X'^{V}\big) = \kappa\big(\underbrace{X_{s_1}, \ldots, X_{s_1}}_{|V\cap W_1|\text{-times}}, \ldots, \underbrace{X_{s_j}, \ldots, X_{s_j}}_{|V\cap W_j|\text{-times}}\big).$$

Since X_t ($t \in \mathbb{N}$) is a linear process with i.i.d. innovations ε_j ($t \in \mathbb{Z}$), this can be written as

$$\kappa\left(X'^V\right) = \text{const} \cdot B_{V,s_1,\ldots,s_j},$$

where

$$B_{V,s_1,\ldots,s_j} = \sum_{i=-\infty}^{\infty} a_{i+s_1}^{|V\cap W_1|} \cdots a_{i+s_j}^{|V\cap W_j|}.$$

Hence,

$$\kappa\left(X'^{V_1}\right) \cdots \kappa\left(X'^{V_r}\right) = \text{const} \cdot \prod_{u=1}^{r} B_{V_u,s_1,\ldots,s_j},$$

so that it is sufficient to show that each $n^{-\frac{j}{2}} B_{V_u,s_1,\ldots,s_j}$ converges to zero. This requires a rather laborious detailed argument. However, the essential idea used in Surgailis (2003, Lemma 6.1) is to show this first for a finite moving average process $X_{t,K} = \sum_{j=0}^{K} a_j \varepsilon_{t-j}$ (actually Surgailis allows for a two-sided moving average) and then give an upper bound for the difference between the approximation $J_{n,\gamma}^K$ and $J_{n,\gamma}$ that converges to zero as K tends to infinity. Note that a similar approximation argument was used to establish convergence of partial sums of weakly dependent linear processes, see Theorem 4.5.

Tightness is easier than fidi-convergence but is omitted here; we refer the reader to Giraitis (1985).

Next, consider the case $m(1-2d) < 1$. This case has been considered for instance in Surgailis (1981, 1982), Giraitis and Surgailis (1986, 1989) and Avram and Taqqu (1987); see also Surgailis (2003) for an overview.

Recall from Corollary 4.1 that

$$S_{n,A_m} = \sum_{t=1}^{n} \sum_{j_1,\ldots,j_m=0}^{\infty} a_{j_1} \cdots a_{j_m} (:\varepsilon_{t-j_1} \cdots \varepsilon_{t-j_m}:).$$

Consider

$$U_{n,m} := m! \sum_{t=1}^{n} \sum_{0=j_1<j_2<\cdots<j_m}^{\infty} a_{j_1} \cdots a_{j_m} (:\varepsilon_{t-j_1} \cdots \varepsilon_{t-j_m}:). \tag{4.50}$$

Since the random variables $\varepsilon_{j_1} \cdots \varepsilon_{j_m}$ in this expression are independent, we have

$$:\varepsilon_{j_1} \cdots \varepsilon_{j_m}: = A_1(\varepsilon_{j_1}) \cdots A_1(\varepsilon_{j_m}) = \varepsilon_{j_1} \cdots \varepsilon_{j_m}.$$

Therefore, we may write

$$U_{n,m} = m! \sum_{t=1}^{n} \sum_{0=j_1<j_2<\cdots<j_m}^{\infty} \prod_{s=1}^{m} a_{j_s} \varepsilon_{t-j_s} =: m! \sum_{t=1}^{n} V_{t,m}. \tag{4.51}$$

If we recall now (cf. proof of Theorem 4.6) that

$$\varepsilon_t = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{it\lambda} M_0(d\lambda),$$

where M_0 is a spectral measure with independent increments, then combining argument from the proof of Theorem 4.3 with the proof of Theorem 4.6, we expect that

$$n^{-(1-m(\frac{1}{2}-d))} L_f^{-m/2}\left(n^{-1}\right) U_{n,m} \xrightarrow{d} m! \int_{\lambda_1<\cdots<\lambda_m} D(\lambda_1+\cdots+\lambda_m) dW_X(\lambda_1)\cdots dW_X(\lambda_m), \tag{4.52}$$

where $dW_X(\lambda) = |\lambda|^{-d}\, dM_0(\lambda)$ is the limiting spectral measure defined in (4.34). The spectral-domain function L_f is replaced by the time-domain slowly varying function L_m using the same argument as in the proof of Theorem 4.3:

$$L_m(n) = m! C_m \left(2\Gamma(1-2d)\sin(\pi d)\right)^m L_f^m\left(n^{-1}\right).$$

Then,

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) U_{n,m} \xrightarrow{\mathrm{d}} Z_{m,H}(1). \tag{4.53}$$

Finally,

$$S_{n,A_m} = U_{n,m} + r_{n,m},$$

where the remainder $r_{n,m}$ involves summation over $j_1, \ldots, j_m$ such that at least two indices agree. The remainder is of a smaller order (see Avram and Taqqu 1987 for details).

Tightness is very easy. We use the same argument as in the proof of Theorem 4.4, together with the variance estimates in Theorem 4.7. □

As noted in the proof, in the case with $m(1-2d) < 1$, the convergence of S_{n,A_m} is determined by the term $U_{n,m}$ defined in (4.51). In fact, the convergence equation (4.52) will play a crucial role in some of the results following below.

The assumptions of the theorem can be relaxed in various ways. For instance, in order to obtain the usual central limit theorem in (4.49), only $\sum |\gamma_X(k)|^m < \infty$ is required instead of the specific decay of γ_X (see Surgailis 2003). Moreover, the result can be extended to

$$S_{n,G}(u) = \sum_{t=1}^{[nu]} G(X_t)$$

with

$$G(x) = \sum_{j=m}^{\infty} \frac{a_{\mathrm{app},j}}{j!} A_j(x).$$

Assuming that $a_{\text{app},m} \neq 0$ (i.e. G has Appell rank m), the contribution of $a_{\text{app},m} \times A_m(X_t)/m!$ dominates, provided that $m(1-2d) < 1$. For example, Surgailis (2000) considers arbitrary polynomials G. Furthermore, Surgailis and Vaičiulis (1999) replace independent ε_t ($t \in \mathbb{Z}$) by martingale differences, and Surgailis (2000) considers $\tilde{X}_t = X_t + V_t$ where V_t ($t \in \mathbb{N}$) is a stationary short-memory process.

In view of the fact that for each distribution different Appell polynomials are obtained, and in general they are not orthogonal, it is quite remarkable that the same asymptotic limit is obtained as under Gaussian subordination and Hermite polynomials. Moreover, it is worth noting that, for fixed m, the condition $m(1-2d) < 1$ means that $d > \frac{1}{2}(1-m^{-1})$. Thus, a nonstandard limiting behaviour (which is also called noncentral limit theorem) is achieved for sufficiently strong long-range dependence. The higher the degree m of the Appell polynomial, the stronger dependence has to be to satisfy the condition. This is essentially due to (4.46). Since at the same time d does not exceed $\frac{1}{2}$, there is no such d for $m = 1$. In other words, for X_t ($t \in \mathbb{N}$), a noncentral limit theorem holds for all $0 < d < \frac{1}{2}$.

4.2.5.4 Asymptotic Distributions: Martingale Approach and Power Ranks

Recall now that the jth Appell coefficient can be obtained either by

$$a_{\text{app},j} = E\big[G^{(j)}(X)\big] \tag{4.54}$$

if the jth derivative of G exists and its expected value is not zero (see (3.66)) or by

$$a_{\text{app},j} = (-1)^j \int G(x) p_X^{(j)}(x)\, dx \tag{4.55}$$

(see (3.69)), where $p_X = F_X'$ is the density of X. Note that due to (4.54), a similar definition of Appell rank that has been proposed in the literature is the so-called power rank.

Definition 4.1 Let X be a random variable. The power rank of a function G (with respect to X) is the smallest integer $m \geq 1$ such that $G_\infty^{(m)}(x) \neq 0$, where $G_\infty(x) = E[G(X+x)]$.

Example 4.7 Let F_X be the distribution of a random variable X with $E(X) = 0$. If $G(x) = x^2 - E(X^2)$, then $G_\infty^{(1)}(0) = 2\int u\, dF_X(u) = 2E(X) = 0$. Furthermore, $G_\infty^{(2)}(0) = 2\int dF_X(u) = 2$. This implies that for a centred linear process $X_t = \sum a_j \varepsilon_{t-j}$, the power rank of the quadratic function is always 2, regardless of the distribution of ε_t (and the marginal distribution of X_t).

Using the power rank, Ho and Hsing (1996, 1997) developed a different approach to studying limit theorems for functionals of linear processes. To describe the idea,

let us again consider

$$X_{t,K} = \sum_{j=0}^{K} a_j \varepsilon_{t-j},$$

$$\tilde{X}_{t,K} = X_t - X_{t,K} = \sum_{j=K+1}^{\infty} a_j \varepsilon_{t-j}$$

and

$$G_K(y) := E\big[G(X_{t,K} + y)\big] \quad (K \geq 0), \qquad G_\infty(y) = E\big[G(X_t + y)\big]. \tag{4.56}$$

We also use the convention $G_{-1} = G$ and $\tilde{X}_{0,-1} = X_0$. Note now, that if $\mathcal{F}$ is a sigma field, ξ_A is a random variable that is $\mathcal{F}$-measurable and ξ_B is a random variable that is independent of $\mathcal{F}$ and has distribution F_B, then

$$E\big[G(\xi_A + \xi_B + y)|\mathcal{F}\big] = \int G(\xi_A + v + y)\, dF_B(v) =: G_{B,*}(\xi_A + y) \tag{4.57}$$

and

$$G_*(y) := E\big[G(\xi_A + \xi_B + y)\big] = E\big[G_{B,*}(\xi_A + y)\big]. \tag{4.58}$$

Now let $\mathcal{F}_K = \sigma(\varepsilon_j, -\infty < j \leq K)$ $(K \in \mathbb{Z})$. We apply (4.57) and (4.58) with $(\xi_A, \xi_B, \mathcal{F}) = (\tilde{X}_{t,K-1}, X_{t,K-1}, \mathcal{F}_{t-K})$ and $(\xi_A, \xi_B, \mathcal{F}) = (\tilde{X}_{t,K}, X_{t,K}, \mathcal{F}_{t-(K+1)})$ respectively. We obtain

$$\begin{aligned}
&\sum_{t=1}^{n} \big\{G(X_t) - E\big[G(X_1)\big]\big\} \\
&\quad = \sum_{t=1}^{n} \sum_{K=0}^{\infty} \big\{E\big[G(X_t)|\mathcal{F}_{t-K}\big] - E\big[G(X_t)|\mathcal{F}_{t-(K+1)}\big]\big\} \\
&\quad = \sum_{t=1}^{n} \sum_{K=0}^{\infty} \big(G_{K-1}(\tilde{X}_{t,K-1}) - G_K(\tilde{X}_{t,K})\big) \\
&\quad \approx \sum_{t=1}^{n} \sum_{K=0}^{\infty} \big(G_K(\tilde{X}_{t,K-1}) - G_K(\tilde{X}_{t,K})\big) \\
&\quad \approx \sum_{t=1}^{n} \sum_{K=0}^{\infty} a_t \varepsilon_{t-K} G_K^{(1)}(\tilde{X}_{t,K}) \qquad (4.59) \\
&\quad \approx G_\infty^{(1)}(0) \sum_{t=1}^{n} X_t + \sum_{t=1}^{n} \sum_{K=0}^{\infty} a_K \varepsilon_{t-K} \big(G_K^{(1)}(\tilde{X}_{t,K}) - G_\infty^{(1)}(0)\big). \qquad (4.60)
\end{aligned}$$

The point of this approximation is that the first term in the last expression is just the partial sum of the linear sequence, multiplied by a constant. The first term is of a larger order than the second term. Consequently, using Theorem 4.6, we expect

$$n^{-(d+\frac{1}{2})}L_S^{-1/2}(n)\sum_{t=1}^{n}\{G(X_t)-E[G(X_1)]\}\xrightarrow{d}G_\infty^{(1)}(0)B_H(1).$$

This is useful, of course, only if $G_\infty^{(1)}(0)$, the first power rank of G, does not vanish. If $G_\infty^{(1)}(0)=0$, then the expansion is continued until we obtain a non-vanishing quantity $G_\infty^{(m)}(0)$. In that case we say that the power rank of G is m. If for example the power rank is 2, the expansion reads further

$$\begin{aligned}
&\sum_{j=1}^{n}\{G(X_t)-E[G(X)]\}\\
&\quad=\sum_{t=1}^{n}\sum_{K=0}^{\infty}\{E[G(X_t)|\mathcal{F}_{t-K}]-E[G(X_t)|\mathcal{F}_{t-(K+1)}]\}\\
&\quad\approx G_\infty^{(2)}(0)\sum_{t=1}^{n}\sum_{j_1=0}^{\infty}\sum_{j_2=j_1+1}^{\infty}a_{j_1}a_{j_2}\varepsilon_{t-j_1}\varepsilon_{t-j_2}\\
&\qquad+\sum_{t=1}^{n}\sum_{j_1=0}^{\infty}\sum_{j_2=j_1+1}^{\infty}a_{j_1}a_{j_2}\varepsilon_{t-j_1}\varepsilon_{t-j_2}\big(G_{j_2}^{(2)}(\tilde{X}_{t,j_2})-G_\infty^{(2)}(0)\big).
\end{aligned}$$

As before, the second term in the last expression is of a smaller order than the first one. We recognize the first term as $G_\infty^{(2)}(0)U_{n,2}/2!$ (cf. (4.51)). Therefore, using the convergence result (4.52), we have

$$n^{-2d}L_2^{-1/2}(n)\sum_{j=1}^{n}\{G(X_t)-E[G(X_1)]\}\Rightarrow G_\infty^{(2)}(0)Z_{2,H}(1)/2!.$$

This can be generalized to arbitrary power ranks. There are a lot of technical details missing in the heuristic explanation above. We make it more precise, using a modified version of Ho and Hsing's approach (see Wu 2003). In order to do this, let G be a function, and $p\in\mathbb{N}$. Define (cf. (4.51))

$$T_n(G;p)=\sum_{t=1}^{n}\left\{G(X_t)-E[G(X_1)]-\sum_{r=1}^{p}G_\infty^{(r)}(0)V_{t,r}\right\},$$

where

$$V_{t,r}=\sum_{0\le j_1<\cdots<j_r}\prod_{s=1}^{r}a_{j_s}\varepsilon_{t-j_s}.$$

In particular,

$$T_n(G;1) = \sum_{j=1}^{n} \{G(X_t) - E[G(X_1)] - G_\infty^{(1)}(0)X_t\}.$$

For any random variable Y, let $\|Y\|_r = E^{1/r}[Y^r]$. The following theorem establishes a reduction principle for $T_n(G;p)$ that can be viewed as a counterpart to the Gaussian case (see the proof of Theorem 4.4). We state the result assuming that the slowly varying function L_a in (B1) is constant. The statement can be modified appropriately to incorporate a general slowly varying function $L_a(j)$.

Theorem 4.9 *Let X_t $(t \in \mathbb{N})$ be a linear process defined by (4.37) with coefficients satisfying assumption* (B1) *with $L_a(j) \equiv 1$. Assume that $E[|\varepsilon|^{4+\gamma}] < \infty$ for some $\gamma > 0$ and*

$$\max_{r=1,2,\dots,p+1} \sup_y |G_\infty^{(r)}(y)| < \infty, \tag{4.61}$$

where G_∞ is defined in (4.56).

- *If $(p+1)(1-2d) > 1$, then $\|T_n(G;p)\|_2^2 = O(n)$.*
- *If $(p+1)(1-2d) < 1$, then*

$$\|T_n(G;p)\|_2^2 = O\big(n^{2-(p+1)(1-2d)}\big). \tag{4.62}$$

The proof of this result is postponed to the end of this section. At this moment, let us discuss its consequences and technical assumptions. Assumption (4.61) is in the spirit of Ho and Hsing (1997). Another assumption was considered in Wu (2003). Similarly to definition (4.56), one can argue that

$$G_K^{(r)}(y) := \frac{d}{dy^r} E[G(X_{0,K} + y)] = E[G^{(r)}(X_{0,K} + y)] \quad (K \ge 0),$$
$$G_\infty^{(r)}(y) = E[G^{(r)}(X + y)].$$

For example,

$$\begin{aligned}
&\frac{E[G(X+y+\delta)] - G(X+y)}{\delta} - E[G^{(1)}(X+y)] \\
&= \int \left\{ \frac{G(x+y+\delta) - G(x+y)}{\delta} - G^{(1)}(x+y) \right\} p_X(x)\,dx \\
&\le \delta \sup_u |G^{(2)}(u)| \int p_X(x)\,dx.
\end{aligned}$$

Hence, for instance, if G has uniformly bounded second-order derivatives, then the limit as $\delta \to 0$ exists. However, such a strong assumption is not needed in fact, and

a condition like (4.61) suffices (see Ho and Hsing 1996, Lemma 6.2, Wu 2003). We may thus write $G_0^{(r)}(y) = E[G^{(r)}(a_0\varepsilon_0 + y)]$ and

$$G_1^{(r)}(y) = E\big[G^{(r)}(a_0\varepsilon_0 + a_1\varepsilon_{-1} + y)\big] = E\big\{E\big[G^{(r)}(a_0\varepsilon_0 + a_1\varepsilon_{-1} + y)|\varepsilon_{-1}\big]\big\}$$
$$= E\big[G_0^{(r)}(a_1\varepsilon_{-1} + y)\big].$$

Therefore, it is intuitively clear that properties of $G_0^{(r)}$ are transferred to $G_1^{(r)}$ and by induction to any of $G_K^{(r)}$, $K \geq 1$.

Example 4.8 Consider $G(u) = 1\{u \leq x_0\}$ for a fixed x_0. Then $G_\infty(y) = E[1\{X + y \leq x_0\}] = P(X \leq x_0 - y)$, and

$$G_\infty^{(1)}(0) = \frac{d}{dy} P(X \leq x_0 - y)|_{y=0} = -p_X(x_0 - y)|_{y=0} = -p_X(x_0),$$

where p_X is the density of X.

What is the consequence of the theorem above? Take $p = 1$. We obtain $\|T_n(G;1)\|_2^2 = O(\max\{n, n^{4d}\})$. Recall now Theorem 4.6 that describes convergence of partial sums $\sum_{t=1}^n X_t$. We conclude that the limiting behaviour of

$$n^{-(\frac{1}{2}+d)} L_1^{-1/2}(n) \sum_{t=1}^n \big\{G(X_t) - E\big(G(X_1)\big)\big\}$$

is the same as that of

$$n^{-(\frac{1}{2}+d)} L_1^{-1/2}(n) G_\infty^{(1)}(0) \sum_{t=1}^n X_t,$$

where $L_1(n) = (d(2d+1))^{-1} L_\gamma(n)$, and $L_\gamma(n)$ given in (4.39). If the power rank is greater than one, then one has to apply a higher-order expansion ($p \geq 2$). The limiting behaviour of the partial sum follows from the corresponding limit theorem for $U_{n,p}$. The latter was considered in (4.51) and (4.52).

Corollary 4.3 *Let $X_t = \sum_{j=0}^\infty a_j \varepsilon_{t-j}$ ($t \in \mathbb{Z}$) be a linear process defined by (4.37) with coefficients satisfying assumption* (B1), *i.e. $a_j \sim L_a(j) j^{d-1}$, $d \in (0, 1/2)$. Assume that G has the power rank m. If $m(1-2d) < 1$, then, under the conditions of Theorem 4.9,*

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) \sum_{t=1}^n \big\{G(X_t) - E\big(G(X_1)\big)\big\} \xrightarrow{d} G_\infty^{(m)}(0) Z_{m,H}(1),$$

where

$$L_m(n) = m! C_m L_\gamma^m(n),$$

$$C_m = \frac{2}{((2d-1)m+1)((2d-1)m+2)},$$

and L_γ *is given by* (4.39):

$$\begin{aligned} L_\gamma(k) &= L_a^2(k)\sigma_\varepsilon^2 \int_0^\infty v^{d-1}(v+1)^{d-1}\,dv \\ &= L_a^2(k)\sigma_\varepsilon^2 B(1-2d,d). \end{aligned}$$

Let us apply Corollary 4.3 to X_t^2, where X_t is a linear process such that $E(X_1^2) = 1$. The example shows that in a sense, the power rank method is distribution free. In contrast, limiting results for Appell polynomials are not directly applicable to $X_t^2 - 1$, unless X_t are Gaussian.

Example 4.9 Consider a linear process $X_t = \sum_{j=0}^\infty a_j \varepsilon_{t-j}$ $(t \in \mathbb{Z})$ such that $\sum_{k=0}^\infty a_k^2 = 1$ and $E[\varepsilon_1^2] = 1$. Let $G(x) = x^2$. Then recall from Example 4.6 that

$$\sum_{t=1}^n \big(X_t^2 - 1\big) = \sum_{t=1}^n \sum_{j=0}^\infty a_j^2\big(\varepsilon_{t-j}^2 - 1\big) + \sum_{t=1}^n \sum_{k,l=0;\ k\neq l}^\infty a_k a_l \varepsilon_{t-k}\varepsilon_{t-l}.$$

The first term can be represented as $\sum_{t=1}^n Y_t$, where Y_t $(t \in \mathbb{Z})$ is the linear process $Y_t = \sum_{j=0}^\infty c_k \xi_{t-j}$, $\xi_{t-j} = \varepsilon_{t-j}^2 - 1$, with summable coefficients $c_j = a_j^2$. Using Theorem 4.5, we have

$$n^{-1/2} \sum_{t=1}^n \sum_{j=0}^\infty a_j^2\big(\varepsilon_{t-j}^2 - 1\big) \xrightarrow{d} N\big(0, v^2\big),$$

where $v^2 = \sigma_Y^2 + 2\sum_{k=1}^\infty \gamma_Y(k)$. The second term can be recognized as $U_{n,2}$, see (4.51), (4.52) and (4.53). Therefore,

$$n^{-2d} L_2^{-1/2}(n) U_{n,2} \xrightarrow{d} Z_{2,H}(1)$$

if $d \in (1/4, 1/2)$, where $Z_{2,H}(u)$ is the Hermite–Rosenblatt process with $H = d + 1/2$. On the other hand,

$$n^{-1/2} U_{n,2} \xrightarrow{d} \sigma_S N(0,1)$$

if $d < 1/4$. Furthermore, the terms in (4.63) are uncorrelated. Therefore, if $d > 1/4$, then

$$n^{-2d} L_2^{-1/2}(n) \sum_{t=1}^n \big(X_t^2 - 1\big) \xrightarrow{d} Z_{2,H}(1).$$

Otherwise, if $d < 1/4$,

$$n^{-1/2} \sum_{j=1}^n \big(X_t^2 - 1\big) \xrightarrow{d} N\big(0, v + \sigma_S^2\big). \tag{4.63}$$

Example 4.10 (ARFIMA) Assume that X_t ($t \in \mathbb{N}$) is a FARIMA$(0, d, 0)$ process as in Examples 4.4 and 4.5. Then

$$\gamma_X(k) \sim c_\gamma k^{2d-1}, \qquad c_\gamma = \frac{\sigma_\varepsilon^2}{\pi} \Gamma(1-2d)\sin(\pi d).$$

Hence, for $d \in (1/4, 1/2)$,

$$n^{-2d} L_2^{-1/2}(n) \sum_{t=1}^{n} \left(X_t^2 - 1\right) \xrightarrow{d} Z_{2,H}(1),$$

where

$$L_2(n) = 2C_2 c_\gamma^2, \quad C_2 = \frac{1}{(2(2d-1)+1)(2d+1)}.$$

Of course, this is comparable to the Gaussian case, see Example 4.1.

4.2.5.5 Technical Details for Theorem 4.9

We write the proof for $p = 1$ only, leaving out some technical details. They can be found in Ho and Hsing (1996, 1997) and Wu (2003). Using the notation $\mathcal{V}_t = (\varepsilon_t, \varepsilon_{t-1}, \dots,)$, we may write $T_n(G; 1) = \sum_{t=1}^{n} U(\mathcal{V}_t)$, where $U(\cdot)$ is a suitable function. Let P_K be the conditional expectation operator

$$P_K Y = E[Y|\mathcal{V}_K] - E[Y|\mathcal{V}_{K-1}].$$

Noting that $P_K T_n(G; 1) = 0$ if $K > n$, we can write down the orthogonal decomposition

$$T_n(G; 1) = \sum_{K=-\infty}^{n} P_K T_n(G; 1).$$

Furthermore,

$$\begin{aligned} P_K T_n(G; 1) &= \sum_{t=1}^{n} \left\{ E\left(U(\mathcal{V}_t)|\mathcal{F}_K\right) - E\left(U(\mathcal{V}_t)|\mathcal{F}_{K-1}\right) \right\} \\ &= \sum_{t=\max\{K,1\}}^{n} \left\{ E\left(U(\mathcal{V}_t)|\mathcal{F}_K\right) - E\left(U(\mathcal{V}_t)|\mathcal{F}_{K-1}\right) \right\} \\ &= \sum_{t=\max\{K,1\}}^{n} P_K U(\mathcal{V}_t), \end{aligned}$$

since the terms corresponding to $t \le K-1$ vanish. Therefore,

$$\big\|T_n(G;1)\big\|_2^2 = \sum_{K=-\infty}^{n} \big\|P_K T_n(G;1)\big\|_2^2 = \sum_{K=-\infty}^{n} \Big\| \sum_{t=\max\{K,1\}}^{n} P_K U(\mathcal{V}_t) \Big\|_2^2 .$$

Now, for any stationary sequence Y_t ($t \in \mathbb{N}$), we have $\| \sum_{t=1}^n Y_t \|_2 \le \sum_{t=1}^n \|Y_t\|_2$. Therefore, if we define

$$\psi_{t-K}^2 = \big\|P_K U(\mathcal{V}_t)\big\|_2^2 = \big\|P_{-(t-K)} U(\mathcal{V}_0)\big\|_2^2$$

and use Lemma 4.17 below, we obtain

$$\big\|T_n(G;1)\big\|_2^2 \le \sum_{K=-\infty}^{n} \Bigg(\sum_{t=\max\{K,1\}}^{n} \big\|P_{-(t-K)} U(\mathcal{V}_0)\big\|_2 \Bigg)^2 \tag{4.64}$$

$$\le \sum_{K=-\infty}^{n} \Bigg(\sum_{t=\max\{K,1\}}^{n} (t-K)^{2(d-1)+1/2} \Bigg)^2 . \tag{4.65}$$

A rough bound for this expression can be established as follows:

$$\begin{aligned}
&\sum_{K=-\infty}^{n} \Bigg(\sum_{t=\max\{K,1\}}^{n} (t-K)^{2(d-1)+1/2} \Bigg)^2 \\
&\quad \approx \int_{-\infty}^{n} \Bigg(\int_{\max\{s,0\}}^{n} (v-s)^{2(d-1)+1/2}\, dv \Bigg)^2 ds \\
&\quad = \int_{-\infty}^{0} \Bigg(\int_{0}^{n} (v-s)^{2(d-1)+1/2}\, dv \Bigg)^2 ds + \int_{0}^{n} \Bigg(\int_{s}^{n} (v-s)^{2(d-1)+1/2}\, dv \Bigg)^2 ds .
\end{aligned}$$

Let us evaluate the first term only:

$$\begin{aligned}
&\int_{-\infty}^{0} \Bigg(\int_{0}^{n} (v-s)^{2(d-1)+1/2}\, dv \Bigg)^2 ds \\
&\quad = C \int_{-\infty}^{0} \big((n-s)^{2(d-1)+3/2} - (-s)^{2(d-1)+3/2} \big)^2 ds \\
&\quad = \int_{0}^{\infty} \big((n+s)^{2(d-1)+3/2} - s^{2(d-1)+3/2} \big)^2 ds = O\big(n^{4(d-1)+3+1}\big) = O\big(n^{4d}\big).
\end{aligned}$$

This is statement (4.62) of Theorem 4.9 when $p=1$. We note that the integral above is well defined. For example, as $s \to \infty$, the integrand behaves like $\{s^{2(d-1)+1/2}\}^2$, which is integrable since $d < 1/2$. A detailed computation can be found in Lemma 5 in Wu (2003).

To finish the proof of Theorem 4.9, we have to prove the following lemma.

Lemma 4.17 *Assume that the conditions of Theorem 4.9 are satisfied. Then*

$$\left\| P_{-K} U(\mathcal{V}_0) \right\|_2^2 = O\big(K^{4(d-1)+1}\big), \quad K \geq 0.$$

Proof We have

$$\begin{aligned} P_{-K} U(\mathcal{V}_0) &= E\big[G(X_0)|\mathcal{F}_{-K}\big] - E\big[G(X_0)|\mathcal{F}_{-(K+1)}\big] \\ &\quad - G_\infty^{(1)}(0)\big\{E[X_0|\mathcal{F}_{-K}] - E[X_0|\mathcal{F}_{-(K+1)}]\big\}. \end{aligned}$$

Now we use the decomposition $X_0 = X_{0,K-1} + \tilde{X}_{0,K-1}$ and note that $X_{0,K-1}$ is independent of $\mathcal{F}_{-K}$, whereas $\tilde{X}_{0,K-1}$ is measurable w.r.t. this sigma field. Thus, recalling that $E(\varepsilon_1) = 0$, the second term in $P_{-K} U(\mathcal{V}_0)$ yields

$$E[X_0|\mathcal{F}_{-K}] - E[X_0|\mathcal{F}_{-(K+1)}] = \tilde{X}_{0,K-1} - \tilde{X}_{0,K} = a_K \varepsilon_{-K}.$$

The first term in $P_{-K} U(\mathcal{V}_0)$ is

$$G_{K-1}(\tilde{X}_{0,K-1}) - G_K(\tilde{X}_{0,K}).$$

Applying (4.57) and (4.58) with $(\xi_A, \xi_B, \mathcal{F}) = (\tilde{X}_{0,K-1}, X_{0,K-1}, \mathcal{F}_{0-K})$ and $(\xi_A, \xi_B, \mathcal{F}) = (\tilde{X}_{0,K}, X_{0,K}, \mathcal{F}_{0-(K+1)})$, our goal is to evaluate the bound

$$\left\| P_{-K} U(\mathcal{V}_0) \right\|_2^2 = \left\| G_{K-1}(\tilde{X}_{0,K-1}) - G_K(\tilde{X}_{0,K}) - G_\infty^{(1)}(0) a_K \varepsilon_{0-K} \right\|_2^2.$$

In the first step, we will replace G_{K-1} by G_K. Note first that for any $y \in \mathbb{R}$,

$$\begin{aligned} G_K(y) &= E\big[G(X_{0,K} + y)\big] = E\big[G(X_{0,K-1} + a_K \varepsilon_{-K} + y)\big] \\ &= E\big\{E\big[G(X_{0,K-1} + a_K \varepsilon_{-K} + y)|\varepsilon_{-K}\big]\big\} = E\big[G_{K-1}(y + a_K \varepsilon_{-K})\big]. \end{aligned} \tag{4.66}$$

Taking into account that $E(\varepsilon_{-K}) = 0$ and applying a Taylor expansion, we therefore obtain

$$\begin{aligned} G_{K-1}(y) - G_K(y) &= E\big[G_{K-1}(y) - G_{K-1}(y + a_K \varepsilon_{-K})\big] \\ &= E\big[G_{K-1}(y) - G_{K-1}(y + a_K \varepsilon_{j-K}) + G_{K-1}^{(1)}(y) a_K \varepsilon_{-K}\big] \\ &\leq a_K^2 E\big(\varepsilon_{-K}^2\big) \sup_y \big|G_{K-1}^{(2)}(y)\big|. \end{aligned}$$

Therefore,

$$\begin{aligned} \left\| P_{-K} U(\mathcal{V}_0) \right\|_2^2 &\leq C\big\{\left\| G_K(\tilde{X}_{0,K-1}) - G_K(\tilde{X}_{0,K}) + G_\infty^{(1)}(0) a_K \varepsilon_{-K} \right\|_2^2 + a_K^4\big\} \\ &\leq C\big\{\left\| G_K(\tilde{X}_{0,K-1}) - G_K(\tilde{X}_{0,K}) + G_K^{(1)}(\tilde{X}_{0,K}) a_K \varepsilon_{-K} \right\|_2^2 + a_K^4\big\} \\ &\quad + C\left\| G_\infty^{(1)}(0) a_K \varepsilon_{-K} - G_K^{(1)}(\tilde{X}_{0,K}) a_K \varepsilon_{-K} \right\|_2^2 =: I_1 + I_2. \end{aligned}$$

The first term I_1 is treated again using a Taylor approximation: it is bounded by $a_K^4 E^2(\varepsilon_1^2) \sup_y |G_K^{(2)}(y)|$. As for the second term, since $\tilde{X}_{0,K}$ and ε_{-K} are independent, we have

$$I_2 = a_K^2 E\big[\varepsilon^2\big] \big\| G_\infty^{(1)}(0) - G_K^{(1)}(\tilde{X}_{0,K}) \big\|_2^2.$$

Thus, in analogy to (4.66), by conditioning on $\tilde{X}_{0,K}$,

$$G_\infty^{(1)}(y) = E\big[G^{(1)}(X+y)\big] = E\big[G_K^{(1)}(\tilde{X}_{0,K}+y)\big]. \tag{4.67}$$

Furthermore, for any two random variables η_A and η_B, we have $E[(\eta_A - E[\eta_B])^2] \le E[(\eta_A - \eta_B)^2]$. Therefore, using (4.67) with $\tilde{Y}_{0,K}$, an independent copy of $\tilde{X}_{0,K}$, we obtain

$$\begin{aligned} I_2 &\le a_K^2 E\big(\varepsilon_{-K}^2\big) \big\| G_K^{(1)}(\tilde{Y}_{0,K}) - G_K^{(1)}(\tilde{X}_{0,K}) \big\|_2^2 \\ &\le 2a_K^2 E\big(\varepsilon_{-K}^2\big) \big\| G_K^{(1)}(\tilde{X}_{0,K}) - G_K^{(1)}(0) \big\|_2^2 \le C a_K^2 E\big(\tilde{X}_{0,K}^2\big) \sup_y \big|G_K^{(2)}(y)\big|. \end{aligned}$$

Hence,

$$I_2 \le C a_K^2 \sum_{j=K+1}^{\infty} a_j^2 \sim C a_K^2 \sum_{j=K+1}^{\infty} j^{2(d-1)} \sim C K^{4(d-1)+1}.$$

This finishes the proof of the lemma.

Note that we had to assume that, for $p = 1$,

$$\max_{r=1,2} \sup_y \big|G_K^{(r)}(y)\big| < \infty.$$

This explains the conditions of Theorem 4.9. □

4.2.6 *Stochastic Volatility Models and Their Modifications*

In this section we consider limit theorems for partial sums of stochastic volatility models. Let $X_t = \sigma_t \xi_t$ $(t \in \mathbb{N})$, where

$$\sigma_t = \sigma(\zeta_t), \quad \zeta_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j},$$

and $\sigma(\cdot)$ is a positive function. It is assumed that (ξ_t, ε_t) $(t \in \mathbb{Z})$ is a sequence of i.i.d. random vectors and $E(\varepsilon_1) = 0$. The linear process ζ_t is assumed to have long memory with autocovariance function $\gamma_\zeta(k) \sim L_\gamma(k) k^{2d-1}$, $d \in (0, 1/2)$. However, we do not assume at the moment that $E(\xi_1) = 0$. If the sequences ξ_t and ε_t are mutually independent, then the model is called LMSV (Long-Memory Stochastic Volatility), but for the purpose of this section, we do not need to make this assumption.

Let $\mathcal{G}_j$ be the sigma field generated by $\xi_l, \varepsilon_l, l \le j$. We consider partial sums

$$S_n(u) = \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \quad \big(u \in [0,1]\big),$$

where G is a measurable function such that $E[G^2(X_1)] < \infty$.

The asymptotic behaviour of partial sums is described in the following theorem. For simplicity, we formulate it in a Gaussian setting; however, it can be extended to linear processes, using the results of Sect. 4.2.5 instead of Theorem 4.4.

Theorem 4.10 *Consider the stochastic volatility model described above with* $v^2 = \mathrm{var}(G(X_1)) < \infty$ *(but possibly* $E(\xi_1) \neq 0$*). Assume in addition that* ε_t $(t \in \mathbb{Z})$ *are standard normal.*

- *If* $E[G(X_1)|\mathcal{G}_0] = 0$, *then*

$$n^{-1/2} S_n(u) \Rightarrow v B(u), \tag{4.68}$$

  where $B(u)$ $(u \in [0,1])$ *is a standard Brownian motion.*
- *If* $E[G(X_1)|\mathcal{G}_0] \neq 0$, *then*

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \Rightarrow \frac{J(m)}{m!} Z_{m,H}(u), \tag{4.69}$$

  where $\Rightarrow$ *denotes weak convergence in* $D[0,1]$, $Z_{m,H}(u)$ $(u \in [0,1])$ *is the Hermite–Rosenblatt process*, m *is the Hermite rank of*

$$\tilde{G}(y) = \int G\big(s\sigma(y)\big)\, dF_\xi(s)$$

  with F_ξ *denoting the distribution of* ξ, $L_m(n) = m! C_m L_\gamma^m(n)$ *(cf.* (4.39), (4.21), (4.22)*) and* $J(m) = E[\tilde{G}(\zeta_1) H_m(\zeta_1)]$.

Proof Note that σ_t is measurable w.r.t. $\mathcal{G}_{t-1}$, whereas ξ_t is independent of $\mathcal{G}_{t-1}$. Thus,

$$\begin{aligned}
&\sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \\
&\quad = \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_t)|\mathcal{G}_{t-1}\big] \big\} \\
&\qquad + \sum_{t=1}^{[nu]} \big\{ E\big[G(X_t)|\mathcal{G}_{t-1}\big] - E\big[G(X_t)\big] \big\} =: M_n(u) + R_n(u).
\end{aligned}$$

Note that the first part is a martingale. For this part, it suffices to verify the conditions of the martingale central limit theorem; see Lemma 4.2. Set $X_{t,n} = n^{-1/2}G(X_t)$. The Lindeberg condition is clearly satisfied since

$$E\big[\tilde{X}_{t,n}^2 1\{|\tilde{X}_{t,n}| > \delta\}\big] \le 4E\big[X_{t,n}^2 1\{|X_{t,n}| > \delta\}\big] \to 0$$

on account of $E[G^2(X_1)] < \infty$, where $\tilde{X}_{t,n} = X_{t,n} - E[X_{t,n}|\mathcal{G}_{t-1}]$. Furthermore, $E[G^2(X_t)|\mathcal{G}_{t-1}]$ is a measurable function of the random variable ζ_t and hence of the i.i.d. sequence $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$. Therefore, the sequence $E[G^2(X_t)|\mathcal{G}_{t-1}]$ $(t \ge 1)$ is ergodic, and $n^{-1}\sum_{t=1}^{n} E[G^2(X_t)|\mathcal{G}_{t-1}]$ converges in probability to $E[G^2(X_1)]$. Therefore, we conclude (4.68) for the martingale part $M_n(u)$.

On the other hand, the second part $R_n(u)$ can be written as

$$R_n(t) = \sum_{t=1}^{[nu]} \big\{\tilde{G}(\zeta_t) - E\big[\tilde{G}(\zeta_t)\big]\big\},$$

and (4.69) can be concluded using Theorem 4.4. □

Several comments have to be made here. We note that the proof of (4.68) does not involve a particular structure of the model. Consider for example the standard stochastic volatility model where $E(\xi_1) = 0$. If we take $G(x) = x$, then $n^{-1/2}\sum_{t=1}^{[nu]} X_t$ converges to a Brownian motion without the assumption of Gaussianity on ε_t. Furthermore, it is worth mentioning that this approach works (in the case (4.68) only) for partial sums of GARCH, ARCH(∞) or LARCH(∞) models; for the latter, see Beran (2006).

Example 4.11 Assume that $G(y) = y^2$. Then $\tilde{G}(y) = E[\xi_1^2]\sigma^2(y)$. Therefore, m is the Hermite rank of $\sigma^2(y)$. In particular, if $\sigma(y) = \exp(y)$, then $m = 1$. We conclude

$$n^{-(d+1/2)} L_1^{-1/2}(n) \sum_{t=1}^{[nu]} \big(X_t^2 - E\big(X_1^2\big)\big) \Rightarrow J(1) B_H(u),$$

where $J(1) = E(\zeta_1 \exp(2\zeta_1))E(\xi_1^2)$. This is analogous to Surgailis and Viano (2002); note however that the authors considered general linear processes.

If $E(\xi_1) \neq 0$ and $G(x) = x$, then (4.68) is no longer valid; rather (4.69) holds with $m = 1$.

Example 4.12 (Long-Memory Stochastic Duration, LMSD) For the purpose of this example, we assume that random variables ξ_t $(t \in \mathbb{N})$ are strictly positive and hence non-centred. Furthermore, it is assumed that the sequences ξ_t and σ_t are independent. Then $X_t = \xi_t \sigma_t$ inherits the dependence structure from σ_t, i.e.

$$cov(X_0, X_k) = E(X_0 X_k) - E(X_0)E(X_k) = E^2[\xi_1] cov(\sigma_0, \sigma_k).$$

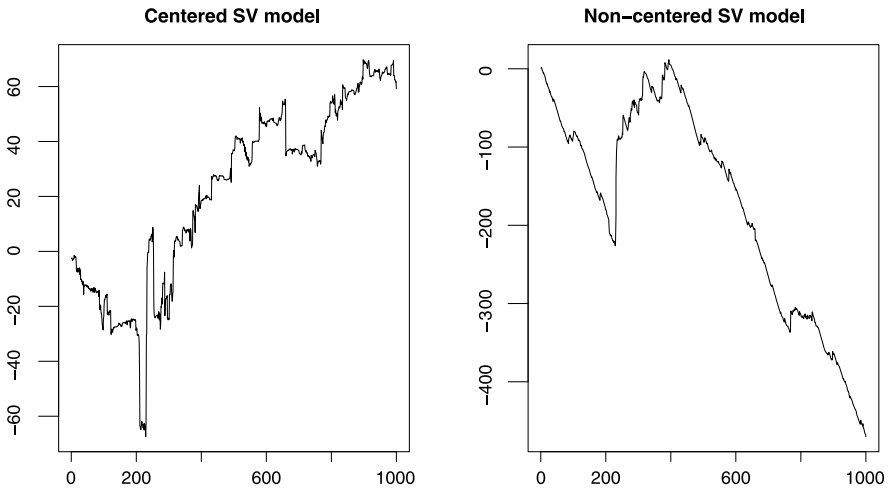

Fig. 4.3 Partial sums for a centred and a non-centred stochastic volatility model

Assume that $G(x) = x$ and $\sigma(x) = \exp(x)$. Then $\tilde{G}(y) = E(\xi_1)\exp(y)$ and $m = 1$. Application of Theorem 4.10 yields

$$n^{-(d+1/2)} L_1^{-1/2}(n) \sum_{t=1}^{[nu]} \big(X_t - E(X_1)\big) \Rightarrow J(1) B_H(u)$$

weakly in $D[0, 1]$, where $B_H(\cdot)$ is a fractional Brownian motion with $H = d + 1/2$, and $J(1) = E[\zeta_1 \exp(\zeta_1)]E[\xi_1]$.

Example 4.13 We illustrate the centering effect with a simulation example. First, we generate $n = 1000$ i.i.d. standard normal random variables ξ_t. Then we simulate independently $n = 1000$ observations ζ_t from a Gaussian FARIMA$(0, d, 0)$ process with $d = 0.4$ and compute $\sigma_t = \exp(\zeta_t)$. Then, we construct two stochastic volatility models: a centred one, $X_t = \xi_t \sigma_t$ and a non-centred one, $\tilde{X}_t = (\xi_t + 1)\sigma_t$. Finally, we plot the partial sum sequences $S_k = \sum_{t=1}^{k} X_t$ and $\tilde{S}_k = \sum_{t=1}^{k} (\tilde{X}_t - E(\tilde{X}_1))$, $k = 1, \ldots, n$. The corresponding partial sum processes are plotted in Fig. 4.3. The smoother path in the second, non-centred, case indicates an influence of long memory (cf. Fig. 4.1).

4.2.7 ARCH(∞) *Models*

Recall from Definition 2.1 that the ARCH(∞) model has the form $X_t = \sigma_t \xi_t$, where ξ_t $(t \in \mathbb{Z})$ are i.i.d. zero mean random variables with variance σ_ξ^2. Also,

$$\sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2.$$

Furthermore, if $\sigma_\xi^2 \sum_{j=1}^{\infty} b_j < 1$, then X_t $(t \in \mathbb{Z})$ is stationary, and $E(X_1^2) < \infty$. The sequence X_t $(t \in \mathbb{Z})$ is a martingale. Using the martingale central limit theorem (see Lemma 4.2), we conclude the following result. It can also be stated in a functional form (as convergence to a Brownian motion).

Corollary 4.4 *Consider an* ARCH(∞) *model as in Definition 2.1. Assume that* $\sigma_\xi^2 \sum_{j=1}^{\infty} b_j < 1$. *Then*

$$n^{-1/2} \sum_{t=1}^{n} X_t \xrightarrow{d} N\big(0, \sigma_X^2\big),$$

where

$$\sigma_X^2 = \frac{\sigma_\xi^2 b_0}{1 - \sigma_\xi^2 \sum_{j=1}^{\infty} b_j}.$$

Next, we are interested in the asymptotic behaviour of

$$S_n = \sum_{t=1}^{n} \big(X_t^2 - E\big(X_1^2\big)\big).$$

To deal with this, we will use the general Definition 2.2 of ARCH(∞) models and set $Y_t = X_t^2 = v_t \zeta_t = \sigma_t^2 \xi_t^2$. In contrast to X_t $(t \in \mathbb{Z})$, the squared sequence is not a martingale. However, we recall from Theorem 2.3 that, under the existence condition $\mu_\zeta^{1/2} \sum_{j=1}^{\infty} b_j < 1$ (which guarantees $E(Y^2) < \infty$), we have the summability of the covariances, $\sum_{k=\infty}^{\infty} |\gamma_Y(k)| < \infty$. Thus, we may expect a central limit for partial sum S_n with the rate $n^{-1/2}$. Indeed, we will argue that the ARCH(∞) model $Y_t = v_t \zeta_t$, $v_t = b_0 + \sum_{j=1}^{\infty} b_j Y_{t-j}$, can be written using the Wold decomposition with respect to a martingale difference.

To see this, assume that $E(\zeta_1) = E(\xi_1^2) = 1$ and let $\psi(z) = 1 - \sum_{j=1}^{\infty} b_j z^j$. Since $\sum_{j=1}^{\infty} b_j < 1$, we conclude that $\psi(\cdot)$ is analytic on $\{z : |z| < 1\}$ and has no zeros in $\{z : |z| \le 1\}$. Hence, it is invertible, and $\psi^{-1}(z) = \sum_{j=0}^{\infty} \tilde{b}_j z^j$ with $\sum_{j=0}^{\infty} |\tilde{b}_j| < \infty$. Now, $v_t = b_0 + (1 - \psi(B))Y_t$, which leads to

$$\psi(B)Y_t = Y_t - v_t + b_0 = v_t(\zeta_t - 1) + b_0.$$

On the other hand,

$$E(Y_1) = E(v_1)E(\zeta_1) = E(v_1)$$
$$= \frac{b_0}{1 - \sum_{j=1}^{\infty} b_j},$$

so that

$$E(Y_1)\psi(B) = E(Y_1)\psi(1) = b_0.$$

Hence, $\psi(B)(Y_j - E(Y_1)) = v_t(\zeta_t - 1)$ and

$$Y_t - E(Y_1) = \sum_{j=0}^{\infty} \tilde{b}_j v_t(\zeta_t - 1).$$

We note that $v_t(\zeta_t - 1)$ $(t \in \mathbb{Z})$ is a martingale difference sequence. Therefore, the centred Y_t has a Wold decomposition with summable coefficients $\sum_{j=0}^{\infty} \tilde{b}_j$, where the innovations $v_t(\zeta_t - 1)$ are uncorrelated and martingale differences. Consequently, we could in principle apply the same method as in the proof of Theorem 4.5, provided that it can be generalized to possibly dependent innovations that are martingale differences. Since this is possible, we can conclude the following result.

Theorem 4.11 *Consider an* ARCH(∞) *process as in Definition 2.2. Assume that* $\sqrt{E[\xi_1^2]}\sum_{j=1}^{\infty} b_j < 1$. *Then*

$$n^{-1/2}\sum_{t=1}^{n}\big(Y_t - E(Y_1)\big) \xrightarrow{d} N\big(0, \sigma_Y^2\big),$$

where $\sigma_Y^2 = \sum_{k=-\infty}^{\infty} \gamma_Y(k)$.

4.2.8 LARCH Models

Recall that a LARCH(∞) process is defined as

$$X_t = \sigma_t \xi_t,$$

$$\sigma_t = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j},$$

where $b_0 \neq 0$, and ξ_t $(t \in \mathbb{Z})$ are i.i.d. zero mean random variables with $\sigma_\xi^2 = E(\xi_1^2) = 1$. As in the case of ARCH(∞) processes, the sequence X_t is a martingale difference. Therefore, the statement of Corollary 4.4 still holds with $\sigma_X^2 = E[\sigma_1^2] = b_0^2/(1 - \|b\|_2^2)$ (cf. (2.51)).

The situation is different when we consider X_t^2. We can use the decomposition (cf. (2.56))

$$\sum_{t=1}^{n}\big(X_t^2 - E\big(X_1^2\big)\big) = \sum_{t=1}^{n}\big(\sigma_t^2 - E\big(\sigma_1^2\big)\big) + \sum_{t=1}^{n}\big(\xi_t^2 - 1\big)\sigma_t^2. \tag{4.70}$$

The second term is a martingale and therefore of the order $O_P(\sqrt{n})$. Therefore, in the case of a long-memory LARCH(∞) process, the asymptotic behaviour of $\sum_t (X_t^2 - E(X_t^2))$ is the same as that of $\sum_t (\sigma_t^2 - E(\sigma_t^2))$. On the other hand, (2.57) of Theorem 2.7 suggests that $\sum_t (\sigma_t^2 - E(\sigma_1^2))$ behaves (up to a constant) like $\sum_t (\sigma_t - E(\sigma_1))$. This will be justified below. We then obtain the following result.

Theorem 4.12 *Consider a* LARCH(∞) *process. Let* $\mu_p = E[|\xi_1|^p] < \infty$. *Assume that* $11\mu_4^{1/2} b^2 < 1$, *where* $b = \sum_{j=1}^{\infty} b_j^2$, *and that*

$$b_j \sim c_b j^{d-1} \quad (j \to \infty), \tag{4.71}$$

where $c_b > 0$, $d \in (0, 1/2)$. *Then*

$$n^{-(d+1/2)} \sum_{t=1}^{[nu]} \big(X_t^2 - E\big(X_1^2\big)\big) \Rightarrow 2b_0^{-1} E\big(\sigma_1^2\big) c_1 \left(\frac{1}{d(2d+1)}\right)^{1/2} B_H(u),$$

where $\Rightarrow$ *denotes weak convergence in* $D[0,1]$, $B_H(u)$ *is a fractional Brownian motion with the Hurst parameter* $H = d + 1/2$, *and*

$$c_1 = \left(\frac{b_0^2}{1 - \|b\|^2}\right)^{1/2} \sqrt{B(d, 1-2d)} c_b.$$

Remark 4.1 According to Theorem 2.7, the condition $11\mu_4^{1/2} b^2 < 1$ implies that the fourth moment of X_t is finite.

Proof

Step 1: First, we look at $\sum_{t=1}^{[nu]} (\sigma_t - E(\sigma_1))$. It can be written as

$$\sum_{t=1}^{n} \big(\sigma_t - E(\sigma_1)\big) = \sum_{t=1}^{n} \sum_{l=1}^{\infty} b_l \sigma_{t-l} \xi_{t-l} = \sum_{t=1}^{n} \sum_{l=-\infty}^{t-1} b_{t-l} \sigma_l \xi_l.$$

We note that $\sigma_t \xi_t$ ($t \in \mathbb{Z}$) are uncorrelated and martingale differences. Therefore, we have the partial sum of a process $\sum b_{t-l} \sigma_l \xi_l$ that is a weighted linear sum with innovations being martingale differences. This is similar, though not identical, to the sum studied in Sect. 4.2.5 (the difference is that the innovations are only uncorrelated, not independent, i.e. we do not have a linear process). To identify asymptotic constants, rewrite the sum as $\sum_{t=1}^{n} \sum_{l=1}^{\infty} b_l \xi_{t-l} \sigma_{t-l}$. Then for $t < t'$,

$$cov\left(\sum_{l=1}^{\infty} b_l \xi_{t-l} \sigma_{t-l}, \sum_{l=1}^{\infty} b_l \xi_{t'-l} \sigma_{t'-l}\right) = \operatorname{var}(\xi_1 \sigma_1) \sum_{l=1}^{\infty} b_l b_{l+t'-t}.$$

If (4.71) holds, then, as $|j' - j| \to \infty$, the covariance behaves like

$$\begin{aligned} &\operatorname{var}(\xi_1 \sigma_1) c_b^2 \int_0^{\infty} v^{d-1} (1+v)^{d-1} \, dv \big|j' - j\big|^{2d-1} \\ &\quad = \operatorname{var}(\xi_1 \sigma_1) c_b^2 B(d, 1-2d) \big|j' - j\big|^{2d-1}. \end{aligned}$$

Using known results for linear processes (see Lemma 4.9), we obtain, as $n \to \infty$,

$$\operatorname{var}\left(\sum_{t=1}^{n} \sigma_t\right) \sim \operatorname{var}(\xi_1 \sigma_1) \frac{1}{d(2d+1)} c_b^2 B(d, 1-2d) n^{2d+1}$$

(note that these results are applicable as long as the innovations are uncorrelated). Now,

$$\operatorname{var}(\xi_0 \sigma_0) = \frac{b_0^2}{1 - \|b\|_2^2}.$$

Theorem 4.6 can be generalized to the case where innovations are martingale differences. Setting

$$c_1 = \left(\frac{b_0^2}{1 - \|b\|^2}\right)^{1/2} \big(B(d, 1-2d)\big)^{1/2} c_b,$$

one then can apply the generalized version of Theorem 4.6 to obtain

$$\frac{1}{n^{d+1/2}} \sum_{t=1}^{[nu]} \big(\sigma_t - E(\sigma_1)\big) \Rightarrow c_1 \left(\frac{1}{d(2d+1)}\right)^{1/2} B_H(u). \tag{4.72}$$

Step 2: To deal with $\sum_{t=1}^{[nu]} (\sigma_t^2 - E(\sigma_1^2))$, we recall that (cf. (2.57))

$$cov\big(\sigma_t^2, \sigma_{t+k}^2\big) \sim \left(\frac{2E(\sigma_1^2)}{b_0}\right)^2 cov(\sigma_t, \sigma_{t+k}) \quad (k \to \infty).$$

The implication is that the asymptotic behaviour of the partial sum is the same as that of

$$2b_0^{-1} E\big[\sigma_1^2\big] \sum_{t=1}^{[nu]} \big(\sigma_t - E(\sigma_1)\big)$$

(though more detailed arguments are required to obtain a similar linear representation as for σ_t). Hence,

$$n^{-(d+1/2)} \sum_{t=1}^{[nu]} \big(\sigma_t^2 - E\big(\sigma_1^2\big)\big) \Rightarrow 2b_0^{-1} E\big(\sigma_1^2\big) c_1 \left(\frac{1}{d(2d+1)}\right)^{1/2} B_H(u).$$

Using this and decomposition (4.70), we obtain the result. □

4.2.9 *Summary of Limit Theorems for Partial Sums*

We summarize the main results for partial sums under long memory in Table 4.1. For simplicity, the slowly varying functions are assumed to be constant in this summary. Also, only X_t^2 is considered as a representative of nonlinear transformations.

Table 4.1 Limits for partial sums with finite moments

	Partial sums—finite moments	
	$S_n(u) = \sum_{t=1}^{[nu]} X_t$	$T_n(u) = \sum_{t=1}^{[nu]} (X_t^2 - E(X_1^2))$
Linear processes	$n^{-(1/2+d)} S_n(u) \Rightarrow cB_H(u)$ (Theorems 4.2, 4.6)	$n^{-1/2} T_n(u) \Rightarrow cB(u)$ $(d \in (0, 1/4))$ $n^{-2d} T_n(u) \Rightarrow cZ_{2,H}(u)$ $(d \in (1/4, 1/2))$ (Theorem 4.3, Corollary 4.3, Examples 4.1, 4.9)
Stochastic volatility $X_t = \xi_t \sigma_t$, $E[\xi_t] = 0$	$n^{-1/2} S_n(u) \Rightarrow cB(u)$ (Theorem 4.10)	$n^{-(1/2+d)} T_n(u) \Rightarrow cB_H(u)$ (Theorem 4.10)
LARCH	$n^{-1/2} S_n(u) \Rightarrow cB(u)$	$n^{-(1/2+d)} T_n(u) \Rightarrow cB_H(u)$ (Theorem 4.12)

4.3 Limit Theorems for Sums with Infinite Moments

4.3.1 Introduction

In this section we present limit theorems for partial sums of long-memory processes with infinite moments. Although the theory is quite well understood for weakly dependent random variables (Davis and Resnick 1985, Davis and Hsing 1995, Denker and Jakubowski 1989, Dabrowski and Jakubowski 1994, Bartkiewicz et al. 2011), the case of long memory is less well developed yet, except in the linear case. Results for linear processes with long memory were proven already several decades ago in Astrauskas (1983) and Kasahara and Maejima (1988). Subordinated linear processes were studied in Hsing (1999), Koul and Surgailis (2001), Surgailis (2002, 2004), Vaičiulis (2003). Surprisingly, the martingale decomposition method, used for finite-variance random variables in Theorem 4.9, works also here. Subordinated Gaussian processes were considered for instance in Davis (1983) and Sly and Heyde (2008). Limiting results for infinite-variance stochastic volatility models with long memory are almost non-existing; see McElroy and Politis (2007), Surgailis (2008), Kulik and Soulier (2012). In particular, both subordinated Gaussian processes and stochastic volatility models can be treated using a point process methodology. A complete list of the meanwhile quite extended literature would be too long to be included here. However, some important results and more references can be found for instance in Astrauskas et al. (1991), Benassi et al. (2002), Heath et al. (1998), Houdré and Kawai (2006), Kokoszka and Taqqu (1995a, 1995b, 1996, 1997, 1999), Koul and Surgailis (2001), Samorodnitsky (2004), Samorodnitsky and Taqqu (1994), Surgailis (2004), Zhou and Wu (2010).

First, we will summarize (with some details) results on regularly varying distributions, stable laws and point processes, referring the reader for details to standard textbooks such as Bingham et al. (1989), Feller (1971), Kallenberg (1997), Resnick (2007), Samorodnitsky and Taqqu (1994), Embrechts et al. (1997).

4.3.2 General Tools: Regular Variation, Stable Laws and Point Processes

4.3.2.1 Regular Variation

Let X_t ($t \in \mathbb{N}$) be an i.i.d. sequence whose marginal distribution has regularly varying tails:

$$P(X_1 > x) \sim \frac{1+\beta}{2} x^{-\alpha} L_X(x), \qquad P(X_1 < -x) \sim \frac{1-\beta}{2} x^{-\alpha} L_X(x) \quad (x \to \infty), \tag{4.73}$$

where $L_X(\cdot)$ is slowly varying at infinity, and $\beta \in [-1, 1]$. Condition (4.73) is the balanced tail condition. It is equivalent to $P(|X_1| > x) \sim x^{-\alpha} L_X(x)$ and

$$\lim_{x\to\infty} \frac{P(X_1 > x)}{P(|X_1| > x)} = \frac{1+\beta}{2}, \qquad \lim_{x\to\infty} \frac{P(X_1 < -x)}{P(|X_1| > x)} = \frac{1-\beta}{2}.$$

A typical example is a random variable with Cauchy density $p_X(x) = \pi(1+x^2)^{-1}$. This random variable is symmetric, and $P(X_1 > x) \sim (\pi x)^{-1}$, $x > 0$. Therefore, the Cauchy distribution is regularly varying with index $\alpha = 1$. Another example is a (two-sided) Pareto distribution where

$$P\big(|X_1| > x\big) = x^{-\alpha} \quad (x > 1).$$

We note that if $\alpha \in (0, 2)$, then random variable X has an infinite second moment. The case $\alpha = 2$ requires special attention.

Example 4.14 Assume that $L_X(x) \equiv 1$ and that for $x > x_0 > 0$, we have $\bar{F}_{|X|}(x) := P(|X| > x) = x^{-\alpha}$ with $\alpha = 2$. Then

$$\int_{x_0}^{\infty} x \bar{F}_{|X|}(x)\, dx = \int_{x_0}^{\infty} x x^{-\alpha}\, dx = \int_{x_0}^{\infty} x^{-1}\, dx = +\infty.$$

On the other hand, if $L_X(x) = (\log x)^{-2}$, then

$$\int_{x_0}^{\infty} x x^{-\alpha} \frac{1}{(\log x)^2}\, dx = \int_{x_0}^{\infty} \frac{1}{x(\log x)^2}\, dx = \int_{\log x_0}^{\infty} \frac{1}{u^2}\, du < +\infty.$$

Therefore, we have infinite and finite variance, respectively, in the first and the second case. This means that for $\alpha = 2$, the slowly varying function plays an important role.

The following result is the appropriately modified Karamata theorem. It provides extremely useful estimates for truncated moments (see e.g. Resnick 2007, pp. 25, 36).

Lemma 4.18 *Assume that X is a random variable such that (4.73) holds. Let $\bar{F}(x) = P(X > x)$.*

- *If $\alpha < \eta$, then*

$$E\big[X^\eta 1\{|X| \le x\}\big] \sim \frac{\alpha}{\eta - \alpha} x^\eta \bar{F}(x).$$

Finally note that

$$c_n = \inf\big\{x : P\big(|X| > x\big) \le n^{-1}\big\} \tag{4.74}$$

will be the appropriate normalization sequence used to establish convergence of partial sums and point process convergence. In particular, this sequence can be chosen as $c_n = n^{1/\alpha} L(n)$, where L is a slowly varying at infinity. If $L_X(x) \equiv A$ (i.e. L is constant), then $c_n = A^{1/\alpha} n^{1/\alpha}$.

4.3.2.2 Stable Random Variables

Stable random variables can be considered as a special case of (4.73). There are several equivalent definitions of stable random variables.

Definition 4.2 A random variable X is stable if for any $n \ge 2$, there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that

$$X_1 + \cdots + X_n \stackrel{d}{=} c_n X + d_n,$$

where $X_1, X_2, \ldots$ are independent copies of X. Necessarily, $c_n = n^{1/\alpha}$, where $\alpha \in (0, 2]$. If $d_n = 0$, then X is called strictly stable.

Equivalently, stable random variables are characterized in terms of *domains of attraction*:

Definition 4.3 A random variable X is stable if there exists an i.i.d. sequence Y_t ($t \in \mathbb{N}$) and constants $c_n > 0$, $d_n \in \mathbb{R}$ such that

$$\frac{Y_1 + \cdots + Y_n}{c_n} + d_n \stackrel{d}{\to} X.$$

The characteristic function of a stable random variable X is given by

$$E\big[e^{i\theta X}\big] = \begin{cases} \exp(-\eta^\alpha |\theta|^\alpha (1 - i\beta \operatorname{sign}(\theta) \tan \frac{\pi\alpha}{2}) + i\mu\theta) & \text{if } \alpha \ne 1, \\ \exp(-\eta|\theta|(1 + i\beta \frac{2}{\pi} \operatorname{sign}(\theta) \ln(\theta)) + i\mu\theta) & \text{if } \alpha = 1. \end{cases}$$

Here, $0 < \alpha \le 2$, $\eta > 0$ is the scale parameter, $-1 \le \beta \le 1$ is a skewness, and $\mu \in \mathbb{R}$ a shift parameter. We write $X \sim S_\alpha(\eta, \beta, \mu)$. In particular, X is symmetric α-stable (written as $X \sim S\alpha S$) if $X \sim S_\alpha(\eta, 0, 0)$. If $\beta = 1$, then the random variable X is

called *totally skewed to the right*. If $\alpha \in (1,2]$, then $-\infty < \mu = E(X) < \infty$. In what follows, we will omit the case $\alpha = 1$ from our discussion.

If $\alpha \in (0,2)$, then stable random variables are heavy tailed in the sense of (4.73). Indeed, if $X \sim S_\alpha(\eta, \beta, \mu)$, then

$$\lim_{x\to\infty} x^\alpha P(X > x) = C_\alpha \frac{1+\beta}{2}\eta^\alpha, \qquad \lim_{x\to\infty} x^\alpha P(X < -x) = C_\alpha \frac{1-\beta}{2}\eta^\alpha, \tag{4.75}$$

where

$$C_\alpha = \left(\int_0^\infty x^{-\alpha}\sin x\right)^{-1} = \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} \quad (\alpha \neq 1).$$

Therefore, (4.73) holds with $L_X(x) \equiv C_\alpha \eta^\alpha$. If $\eta = 1$, then the scaling constant c_n defined in (4.74) is $c_n = C_\alpha^{1/\alpha} n^{1/\alpha}$.

In what follows, we will use several properties of stable random variables. They can be obtained by considering the characteristic function. If $X_j \stackrel{\mathrm{d}}{=} S_\alpha(\eta_j, \beta_j, \mu_j)$ $(j = 1, 2)$ are independent, then

$$X_1 + X_2 \stackrel{\mathrm{d}}{=} S_\alpha\left(\left(\eta_1^\alpha + \eta_2^\alpha\right)^{1/\alpha}, \frac{\beta_1\eta_1^\alpha + \beta_2\eta_2^\alpha}{\eta_1^\alpha + \eta_2^\alpha}, \mu_1 + \mu_2\right) \tag{4.76}$$

and

$$cX_1 \stackrel{\mathrm{d}}{=} S_\alpha\big(|c|\eta_1, \mathrm{sign}(c)\beta_1, c\mu_1\big). \tag{4.77}$$

Due to the scaling property, it is sufficient to consider $S_\alpha(1, \beta, \mu)$ random variables.

4.3.2.3 Stable Convergence

Stable random variables play a crucial in the asymptotic theory for heavy-tailed random variables (with $\alpha \in (0,2)$; see Gnedenko and Kolmogorov 1968, Feller 1971). Assume that X_t $(t \in \mathbb{N})$ is an i.i.d. sequence of $S_\alpha(1, \beta, \mu)$ random variables. Using (4.76) and (4.77), we have

$$n^{-1/\alpha}\sum_{t=1}^n X_t \stackrel{\mathrm{d}}{=} S_\alpha\left(1, \beta, \frac{n\mu}{n^{1/\alpha}}\right).$$

Thus, if $\alpha \in (0,1)$, then $n/n^{1/\alpha} \to 0$ and

$$n^{-1/\alpha}\sum_{t=1}^n X_t \stackrel{\mathrm{d}}{\to} S_\alpha(1, \beta, 0). \tag{4.78}$$

If $\alpha \in (1,2)$, a centering is required:

$$n^{-1/\alpha}\sum_{t=1}^n (X_t - \mu_n) \stackrel{\mathrm{d}}{=} S_\alpha\left(1, \beta, \frac{n(\mu - \mu_n)}{n^{1/\alpha}}\right).$$

Thus, we may choose $\mu_n = \mu$ (recall from Definition 4.3 that for $\alpha \in (1,2)$, we have $\mu = E(X)$) to obtain

$$n^{-1/\alpha} \sum_{t=1}^{n} (X_t - \mu) \xrightarrow{\mathrm{d}} S_\alpha(1, \beta, 0). \tag{4.79}$$

However, we may also choose $\mu_n = E[X1\{|X| < n^{1/\alpha}\}]$. Then from the Karamata theorem, as $n \to \infty$,

$$\frac{n(\mu - \mu_n)}{n^{1/\alpha}} = \frac{nE[X \cdot 1\{|X| \geq n^{1/\alpha}\}]}{n^{1/\alpha}} \to C_\alpha \frac{\alpha}{\alpha - 1}.$$

Consequently,

$$n^{-1/\alpha} \sum_{t=1}^{n} \left(X_t - E\left[X \cdot 1\left\{|X| < n^{1/\alpha}\right\}\right]\right) \xrightarrow{\mathrm{d}} S_\alpha\left(1, \beta, C_\alpha \frac{\alpha}{\alpha - 1}\right).$$

Of course, we can restate these results using $c_n = C_\alpha^{1/\alpha} n^{1/\alpha}$ instead of $n^{1/\alpha}$. The convergence results can be proven formally using the characteristic functions.

More generally, a classical result by Skorokhod (1957) states that if the i.i.d. random variables X_t ($t \in \mathbb{N}$) fulfill (4.73) with $L_X(x) \equiv A$, then

$$n^{-1/\alpha} S_n(u) := n^{-1/\alpha} \sum_{t=1}^{[nu]} (X_t - \mu) \Rightarrow A^{1/\alpha} C_\alpha^{-1/\alpha} Z_\alpha(u), \tag{4.80}$$

where $Z_\alpha(\cdot)$ is an α-stable Lévy motion with $Z_\alpha(u) \stackrel{\mathrm{d}}{=} u^{1/\alpha} S_\alpha(1, \beta, 0)$, $\Rightarrow$ denotes weak convergence in $D[0,1]$ w.r.t. J_1 topology, and $\mu = E(X)$ if $\alpha \in (1,2)$ and $\mu = 0$ if $\alpha \in (0,1)$. We say then that random variables X_t ($t \in \mathbb{N}$) are in the domain of attraction of the α-stable law. Of course, if the random variables X_t are stable $S_\alpha(1, \beta, 0)$ and $u = 1$, then (4.80) reduces to (4.79) since then $A = C_\alpha$.

4.3.2.4 Point Processes

Point processes are a useful tool to study limit theorems for partial sums, sample covariances and some other functionals such as extremes. Here, we summarize (with some details) results on convergence of point processes. For a detailed exposition, the reader is referred to Resnick (2007) or Embrechts et al. (1997).

Let X_t ($t \in \mathbb{N}$) be a stationary sequence, and c_n a sequence of constants. Define the point process as

$$N_n = \sum_{t=1}^{n} \delta_{(t/n, c_n^{-1} X_t)}.$$

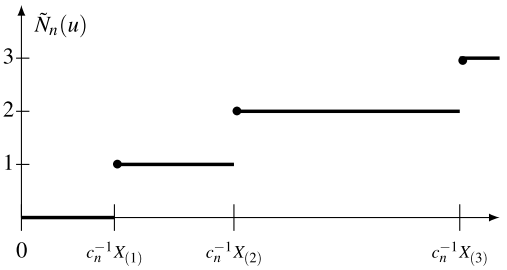

Fig. 4.4 Counting process: $X_{(1)} \le X_{(2)} \le X_{(3)}$ are the smallest observations in the sample $X_1, \dots, X_n$

Here, δ is a Dirac measure, which means that $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. A point process N_n can be viewed as a random element defined on $[0, 1] \times (-\infty, \infty)$, with values in $\mathbb{N}$. In other words, this is a random element with values in $M_p(E)$, the set of all Radon point measures on $E = \mathbb{R}^2$. In particular, if we choose a set $U = [0, 1] \times (0, u)$, then $N_n(U) = \tilde{N}_n(u) = \sum_{t=1}^{n} 1\{0 < c_n^{-1} X_t < u\}$ counts points $c_n^{-1} X_t$ that lie between 0 and u. The process $\tilde{N}_n(u)$ $(u \in \mathbb{R}_+)$ is called a *counting process* and is depicted on Fig. 4.4.

There are several ways to establish convergence of point processes. The first one is referred to as Kallenberg's theorem (see Theorem 14.17 in Kallenberg 1997, or Theorem 5.2.2 in Embrechts et al. 1997).

Proposition 4.2 *Let N_n, $n \in \mathbb{N}$, and N be point processes on $\mathbb{R}^d$ such that N has no multiple points. Assume that*

$$\lim_{n\to\infty} E\big[N_n(U)\big] = E\big[N(U)\big], \tag{4.81}$$

$$\lim_{n\to\infty} P\big(N_n(U) = 0\big) = P\big(N(U) = 0\big) \tag{4.82}$$

for $U = \bigcup_{i=1}^{K} (k_i, l_i) \times (s_i, t_i)$, $K \ge 1$, $0 \le k_i < l_i \le 1$, and arbitrary relatively compact open intervals (s_i, t_i) of $(-\infty, 0) \cup (0, \infty)$. Then N_n converges weakly to N in $M_p(\mathbb{R}^d)$.

We illustrate this theorem by proving convergence of point processes based on i.i.d. sequences. The proof will be easily adapted to models with (long-range) dependence, such as stochastic volatility or subordinated Gaussian sequences. Define the measure λ on $(-\infty, \infty) \setminus \{0\}$ by

$$d\lambda(x) = \alpha\left[\frac{1+\beta}{2} x^{-(\alpha+1)} 1\{0 < x < \infty\} + \frac{1-\beta}{2}(-x)^{-(\alpha+1)} 1\{-\infty < x < 0\}\right] dx, \tag{4.83}$$

where $\beta \in [-1, 1]$. We say that $ds \times d\lambda(x)$ is an intensity measure of a Poisson process N on $[0, 1] \times (-\infty, \infty)$ if for any $A \subset [0, 1]$, $B \subset (-\infty, \infty)$, we have

$$E\big[N(A \times B)\big] = \int_B \int_A d\lambda(x)\, ds.$$

In particular, we note that $E[N([0, 1] \times (-\infty, \infty))] < \infty$.

Theorem 4.13 *Let X_t $(t \in \mathbb{N})$ be a sequence of i.i.d. random variables such that (4.73) holds. Let*

$$P\big(|X_1| > c_n\big) \sim n^{-1}.$$

Then N_n converges weakly in $M_p([0,1] \times \mathbb{R})$ to a Poisson process N on $[0,1] \times ((-\infty, \infty) \setminus \{0\})$ with intensity measure $ds \times d\lambda(x)$.

Before we prove this result, let us state some of its consequences. First, the result can be restated as

$$\sum_{t=1}^{n} \delta_{c_n^{-1} X_t} \Rightarrow \sum_{l=0}^{\infty} \delta_{j_l},$$

where $\Rightarrow$ denotes weak convergence in $M_p(\mathbb{R})$, and j_l are points of a Poisson process with intensity measure $d\lambda(x)$. If $\alpha \in (0,1)$, then the continuous mapping theorem yields that

$$c_n^{-1} \sum_{j=1}^{n} X_t \xrightarrow{\mathrm{d}} \sum_{l=0}^{\infty} j_l.$$

If we assume for a moment that X_t $(t \in \mathbb{N})$ fulfill (4.73) with $L_X \equiv A$, then the scaling constants defined in (4.74) become $c_n = n^{1/\alpha} A^{1/\alpha}$, and so

$$n^{-1/\alpha} \sum_{t=1}^{n} X_t \xrightarrow{\mathrm{d}} A^{1/\alpha} \sum_{l=0}^{\infty} j_l.$$

For the α-stable random variables X_t, we have $A = C_\alpha$. Comparing this expression with (4.78) and using the scaling property (4.77), we conclude that $\sum_{l=0}^{\infty} j_l$ is a series representation of $S_\alpha(C_\alpha^{-1/\alpha}, \beta, 0)$. However, this consideration is not valid for the case where $\alpha \in (1,2)$.

Analogously,

$$\sum_{t=1}^{n} \delta_{c_n^{-2} X_t^2} \Rightarrow \sum_{l=0}^{\infty} \delta_{j_l^2},$$

and for $\alpha \in (0,2)$,

$$c_n^{-2} \sum_{t=1}^{n} X_t^2 \xrightarrow{\mathrm{d}} \sum_{l=0}^{\infty} j_l^2 = S_{\alpha/2}\big(C_{\alpha/2}^{-2/\alpha}, 1, 0\big),$$

or

$$n^{-2/\alpha} \sum_{t=1}^{n} X_t^2 \xrightarrow{\mathrm{d}} A^{2/\alpha} S_{\alpha/2}\big(C_{\alpha/2}^{-2/\alpha}, 1, 0\big).$$

We note that for X_t^2, the skewness parameter is $\beta = 1$. Then the stable random variable is called *totally skewed to the right*. This means that the heavy-tailed property

(4.75) of the limiting stable distribution is related to the heavy-tailed behaviour of

$$P\big(X^2 > x\big) = P(X > \sqrt{x}) + P(X < -\sqrt{x}) \sim Ax^{-\alpha},$$

which is valid for positive values of x only. In contrast, when considering X_t, the heavy-tailed behaviour of the limiting random variable $S_\alpha(C_\alpha^{-1/\alpha}, \beta, 0)$ is attributed to the heavy-tailed behaviour of $P(X > x)$ $(x > 0)$ and $P(X < x)$ $(x < 0)$.

Proof of Theorem 4.13 We verify (4.81). It is enough to consider $U = \bigcup_{i=1}^K (k_i, l_i) \times (s_i, t_i)$ for $K = 1$. We have

$$\begin{aligned} E\big[N_n(U)\big] &= \sum_{t=1}^n E[\delta_{(t/n, c_n^{-1} X_t)}] \\ &= (l_1 - k_1) P\big(c_n^{-1} X_t \in (s_1, t_1)\big) \\ &\to (k_1 - l_1)\lambda\big((s_1, t_1)\big), \end{aligned}$$

where we recall that $\lambda((s_i, t_i)) = \int_{s_i}^{t_i} d\lambda(x)$, and the measure $\lambda(\cdot)$ is given by (4.83). To prove (4.82), write

$$\begin{aligned} P\big(N_n(U) = 0\big) &= P\left(\sum_{i=1}^K \sum_{nk_i < t < nl_i}^n 1\big\{c_n^{-1} X_t \in (s_i, t_i)\big\} = 0\right) \\ &= \prod_{i=1}^K \prod_{nk_i < t < nl_i}^n P\big(c_n^{-1} X_t \notin (s_i, t_i)\big). \end{aligned}$$

Let

$$Q_n = \prod_{i=1}^K \prod_{nk_i < t < nl_i} e^{-n^{-1}\lambda((s_i, t_i))}$$

and note that

$$\begin{aligned} Q_n &= \exp\left(-\sum_{i=1}^K n^{-1} \sum_{nk_i < t < nl_i} \lambda\big((s_i, t_i)\big)\right) \to \exp\left(-\sum_{i=1}^K (l_i - k_i)\lambda\big((s_i, t_i)\big)\right) \\ &= P\big(N(U) = 0\big) \end{aligned}$$

as $n \to \infty$. Recall the two elementary inequalities

$$\left|\prod_{i=1}^K (s_i - t_i)\right| \le \sum_{i=1}^K |s_i - t_i| \quad \text{and} \quad \big|1 - e^{-x} - x\big| \le x^{1+\varepsilon}$$

for any $\varepsilon > 0$. Then we obtain

$$
\begin{aligned}
&\left|P\big(N_n(U)=0\big)-Q_n\right| \\
&\quad=\left|\prod_{i=1}^{K}\prod_{nk_i<t<nl_i}\big(1-P\big(c_n^{-1}X\in(s_i,t_i)\big)\big)-\prod_{i=1}^{K}\prod_{nk_i<t<nl_i}e^{-n^{-1}\lambda((s_i,t_i))}\right| \\
&\quad\le\sum_{i=1}^{K}(l_i-k_i)n\left|\big(1-P\big(c_n^{-1}X\in(s_i,t_i)\big)\big)-e^{-n^{-1}\lambda((s_u,t_u))}\right| \\
&\quad\le\sum_{i=1}^{K}(l_i-k_i)\left|nP\big(c_n^{-1}X\in(s_i,t_i)\big)-\lambda\big((s_i,t_i)\big)\right| \\
&\qquad+\sum_{i=1}^{K}n(l_i-k_i)\left|1-e^{-n^{-1}\lambda((s_i,t_i))}-\frac{\lambda((s_i,t_i))}{n}\right| \\
&\quad=o(1)+Cn^{-\varepsilon}=o(1)
\end{aligned}
$$

for some $\varepsilon>0$. □

Another result, due to Davis and Resnick (1988, Proposition 2.1), is useful when studying processes that can be approximated by sequences with finite memory. Their result is stated in fact in a much more general setting, which is omitted here.

We say that a sequence ν_n of measures converges vaguely to ν ($\nu_n \xrightarrow{v} \nu$) if for all continuous functions $g: E\to\mathbb{R}^d$ with compact support (written as $g\in C^+(E)$), we have

$$\int g(x)\nu_n(dx)\to\int g(x)\nu(dx).$$

We refer to Appendix A for additional precise notions related to vague convergence.

Proposition 4.3 *Assume that X_t ($t\in\mathbb{N}$) is a stationary K-dependent sequence with values in $\mathbb{R}^d$ and $c_n\to\infty$ is a sequence of constants such that for the marginal distribution, we have*

$$nP\big(c_n^{-1}X\in\cdot\big)\xrightarrow{v}\lambda(\cdot).$$

Furthermore, assume that for any $g\in C^+(\mathbb{R}^d)$,

$$\lim_{k\to\infty}\limsup_{n\to\infty}n\sum_{t=2}^{[n/k]}E\big[g\big(c_n^{-1}X_1\big)g\big(c_n^{-1}X_t\big)\big]=0.$$

Then

$$N_n=\sum_{t=1}^{n}\delta_{(t/n,c_n^{-1}X_t)}$$

converges weakly in $M_p([0,1]\times\mathbb{R})$ *to a Poisson process* N *on* $[0,1]\times(-\infty,\infty)$ *with intensity measure* $ds\times d\lambda(x)$.

This result is applicable to sequences X_t with regularly varying tails as in (4.73). In fact (see Theorem 3.6 in Resnick 2007), the vague convergence of $nP(c_n^{-1}X\in\cdot)$ is equivalent to regular variation of the distribution of X.

4.3.3 Sums of Linear and Subordinated Linear Processes

In this section we discuss limit theorems for partial sums of linear processes

$$X_t=\sum_{j=0}^{\infty}a_j\varepsilon_{t-j},$$

where $a_j\sim c_a j^{d-1}$, $d\in(0,1/2)$, and ε_t $(t\in\mathbb{Z})$ are i.i.d. random variables such that

$$P(\varepsilon_1>x)\sim A\frac{1+\beta}{2}x^{-\alpha},\qquad P(\varepsilon_1<-x)\sim A\frac{1-\beta}{2}x^{-\alpha}. \tag{4.84}$$

In both, the coefficients a_j and the tail $P(\varepsilon_1>x)$, we assume for simplicity that possible slowly varying functions are constant. If $\alpha\in(1,2)$, we assume also that $E(\varepsilon_1)=0$.

The infinite series above converges if $\sum_{j=0}^{\infty}|a_j|^{\delta}<\infty$ for some $\delta<\alpha$ (see e.g. Avram and Taqqu 1992). In our case this is possible if an only if $\alpha(d-1)<-1$ and hence $d<1-1/\alpha$. Thus, if $\alpha\in(0,1)$, then the existence condition implies that $\sum_{j=0}^{\infty}|a_j|<\infty$. Consequently, for $\alpha\in(0,1)$, long memory (in the sense of non-summability of the coefficients) is excluded.

Linear processes are the easiest models to describe the interplay between dependence and heavy tails. The asymptotic theory for partial sums is well developed and includes approaches such as convergence of stochastic integrals (Astrauskas 1983, Kasahara and Maejima 1986, 1988) or K-dependent approximations, together with the point process methodology (Davis and Resnick 1985, Davis and Hsing 1995). Interesting results on functional convergence are given in Avram and Taqqu (1992), among others.

4.3.3.1 Tail Behaviour

First, we analyse the tail behaviour of linear processes. We note that if ε_t $(t\in\mathbb{Z})$ are $S_\alpha(1,0,0)$, so that (4.84) holds with $\beta=0$ and $A=C_\alpha$, then

$$X_1\overset{\mathrm{d}}{=}\left(\sum_{j=0}^{\infty}|a_j|^{\alpha}\right)^{1/\alpha}S_\alpha(1,0,0)=:D_\alpha^{1/\alpha}S_\alpha(1,0,0)\overset{\mathrm{d}}{=}D_\alpha^{1/\alpha}\varepsilon_1,$$

which follows directly from properties (4.76) and (4.77). Therefore, we may conclude that, as $x \to \infty$,

$$P\big(|X_1| > x\big) \sim P\big(D_\alpha^{1/\alpha}|\varepsilon_1| > x\big) \sim D_\alpha C_\alpha x^{-\alpha} \sim D_\alpha P\big(|\varepsilon_1| > x\big).$$

This property is valid in fact under the general assumption (4.84).

Lemma 4.19 *Assume that X_t $(t \in \mathbb{N})$ is a linear process, ε_t $(t \in \mathbb{Z})$ are i.i.d. random variables such that (4.84) holds, and $E(\varepsilon_1) = 0$ if $\alpha \in (1, 2)$.*

- *If for some $\delta < \alpha$,*

$$\sum_{j=0}^{\infty} |a_j| + \sum_{j=0}^{\infty} |a_j|^\delta < \infty, \tag{4.85}$$

  then

$$\lim_{x\to\infty} \frac{P(|X_1| > x)}{P(|\varepsilon_1| > x)} = \sum_{j=0}^{\infty} |a_j|^\alpha. \tag{4.86}$$

- *If $a_j \sim c_a j^{d-1}$, $d \in (0, 1 - 1/\alpha)$, and ε_t $(t \in \mathbb{Z})$ are symmetric with $\alpha \in (1, 2)$, then (4.86) holds.*

Note that in the second part of the theorem, the coefficients a_j are not absolutely summable, however $\sum |a_j|^\alpha$ is finite. This turns out to be sufficient. The first part was proven in Cline (1983); see also Davis and Resnick (1985). The second part was proven (under special assumptions with symmetry of the innovations) in Kokoszka and Taqqu (1996).

4.3.3.2 Point Process Convergence

In what follows we show that, under the conditions of Lemma 4.19, a point process based on X_t $(t \in \mathbb{N})$ converges. Its behaviour is the same under short memory (4.85) and under long memory.

Theorem 4.14 *Under the assumptions of Lemma 4.19, we have*

$$\sum_{t=1}^{n} \delta_{c_n^{-1}(X_t,\ldots,X_{t-K})} \Rightarrow \sum_{l=1}^{\infty}\sum_{r=0}^{\infty} \delta_{j_l(a_r, a_{r-1},\ldots,a_{r-K})}$$

in $M_p(\mathbb{R}^{K+1})$, where c_n is such that $P(|\varepsilon_1| > c_n) \sim n^{-1}$, i.e. $c_n \sim A^{1/\alpha} n^{1/\alpha}$.

Proof We give the proof for $K = 0$ only. For details, we refer to Davis and Resnick (1985, Theorem 2.4). We note that the authors prove the results under condition (4.85). However, a crucial part of the proof relies on (4.86) only, which due to

Lemma 4.19 is valid under more general conditions on a_j. We restate Theorem 4.13 in terms of i.i.d. random variables ε_t $(t \in \mathbb{Z})$,

$$\sum_{t=1}^{n} \delta_{c_n^{-1}\varepsilon_t} \Rightarrow \sum_{l=1}^{\infty} \delta_{j_l}$$

where $c_n \sim A^{1/\alpha} n^{1/\alpha}$. Moreover (see Theorem 2.2. in Davis and Resnick 1985), this convergence can be extended to

$$\sum_{t=1}^{n} \delta_{c_n^{-1}(\varepsilon_t,\ldots,\varepsilon_{t-K})} \Rightarrow \sum_{l=1}^{\infty} \sum_{r=0}^{K} \delta_{j_l \mathbf{e}_r}, \tag{4.87}$$

where $\mathbf{e}_r$ is a unit vector in $\mathbb{R}^{K+1}$ with the rth coordinate equal to one. In other words, the limiting process has the following structure. It is a Poisson process with values in $\{0,\ldots,K\} \times \mathbb{R}$ such that it is a univariate Poisson process on the horizontal line $\{0\} \times \mathbb{R}$ and its points are repeated on the other horizontal lines. Since the mapping $(z_t,\ldots,z_{t-K}) \to \sum_{r=0}^{K} b_k z_{t-k}$ from $M_p(\mathbb{R}^{K+1})$ to $M_p(\mathbb{R} \setminus \{0\})$ is continuous, (4.87) implies

$$\sum_{t=1}^{n} \delta_{c_n^{-1} X_{t,K}} \Rightarrow \sum_{l=1}^{\infty} \sum_{r=0}^{K} \delta_{j_l a_r},$$

where $X_{t,K} = \sum_{r=0}^{K} a_r \varepsilon_{t-k}$. Letting $K \to \infty$, we obtain

$$\sum_{l=1}^{\infty} \sum_{r=0}^{K} \delta_{j_l a_r} \xrightarrow{\mathrm{p}} \sum_{l=1}^{\infty} \sum_{r=0}^{\infty} \delta_{j_l a_r}.$$

Therefore, to apply Proposition 4.1, we need to verify that the sequence X_t can be approximated by the K-dependent sequence $X_{t,K}$, in the sense that for each $\gamma > 0$,

$$\lim_{K\to\infty} \limsup_{n\to\infty} P\Big(c_n^{-1} \sup_{1\le t\le n} |X_t - X_{t,K}| > \gamma\Big) = 0.$$

The latter probability is bounded by $nP(c_n^{-1}|X_0 - X_{0,K}| > \gamma)$. Since $P(|\varepsilon_1| > c_n) \sim n^{-1}$, applying (4.86), we have, as $n \to \infty$,

$$nP\big(c_n^{-1}|X_0 - X_{0,K}| > \gamma\big) \sim \frac{P(|X_0 - X_{0,K}| > c_n\gamma)}{P(|\varepsilon_1| > c_n)} = \gamma^{-\alpha} \sum_{r=K+1}^{\infty} |a_r|^{\alpha}.$$

The last expression converges to zero as $K \to \infty$. □

4.3.3.3 Convergence of Partial Sums

Recall our comments following Theorem 4.13. If the innovations ε_t have tail index $\alpha \in (0,1)$, then we may conclude directly from Theorem 4.14 that

$$c_n^{-1} \sum_{t=1}^{n} X_t \xrightarrow{\mathrm{d}} \left(\sum_{j=0}^{\infty} a_j \right) \sum_{l=1}^{\infty} j_l \stackrel{\mathrm{d}}{=} \left(\sum_{j=0}^{\infty} a_j \right) S_\alpha \left(C_\alpha^{-1/\alpha}, \beta, 0 \right),$$

where j_l are points of a Poisson process, and $\sum_{l=1}^{\infty} j_l$ is a series representation of $S_\alpha(C_\alpha^{-1/\alpha}, \beta, 0)$. Equivalently,

$$n^{-1/\alpha} \sum_{t=1}^{n} X_t \xrightarrow{\mathrm{d}} A^{1/\alpha} \left(\sum_{j=0}^{\infty} a_j \right) S_\alpha \left(C_\alpha^{-1/\alpha}, \beta, 0 \right) \stackrel{\mathrm{d}}{=} A^{1/\alpha} C_\alpha^{-1/\alpha} \left(\sum_{j=0}^{\infty} a_j \right) S_\alpha (1, \beta, 0).$$

The situation is more complicated for $\alpha \in (1,2)$. Convergence of partial sums does not follow directly from point process convergence (however, as in Davis and Resnick 1985, an implication of point process convergence may serve as an intermediate tool—this will be illustrated for stochastic volatility models in the following section). In particular, for a long-memory sequence, the scaling for partial sums $\sum_{t=1}^{n} X_t$ of linear processes may differ from c_n.

Theorem 4.15 *Assume that X_t $(t \in \mathbb{Z})$ is a linear process such that $a_j \sim c_a j^{d-1}$, $d \in (0, 1/2)$ and ε_t $(t \in \mathbb{Z})$ are i.i.d random variables such that (4.84) holds with $\alpha \in (1,2)$ and $E(\varepsilon_1) = 0$.*

- *If for some $\delta < \alpha$,*

$$\sum_{j=0}^{\infty} |a_j| + \sum_{j=0}^{\infty} |a_j|^{\delta} < \infty, \tag{4.88}$$

  then

$$n^{-1/\alpha} S_n(u) = n^{-1/\alpha} \sum_{t=1}^{[nu]} X_t \xrightarrow{\text{f.d.}} A^{1/\alpha} C_\alpha^{-1/\alpha} \left(\sum_{j=0}^{\infty} a_j \right) Z_\alpha(u),$$

  where $Z_\alpha(\cdot)$ is an α-stable Lévy motion (with independent increments) such that $Z_\alpha(1) \stackrel{\mathrm{d}}{=} S_\alpha(1, \beta, 0)$, and $\xrightarrow{\text{f.d.}}$ denotes finite-dimensional convergence.
- *If $0 < d < 1 - 1/\alpha$, then*

$$n^{-H} S_n(u) = n^{-H} \sum_{t=1}^{[nu]} X_t \Rightarrow A^{1/\alpha} C_\alpha^{-1/\alpha} \frac{c_a}{d} \tilde{Z}_{H,\alpha}(u),$$

  where $H = d + \alpha^{-1}$, $\tilde{Z}_{H,\alpha}(\cdot)$ is a Linear Fractional stable motion, and $\Rightarrow$ denotes weak convergence in $D[0,1]$ w.r.t. the Skorokhod J_1-topology.

Before we present a proof, we make several comments.

Remark 4.2 If condition (4.88) holds, then the scaling factor and the limiting process are (up to a constant) the same as for i.i.d. random variables; see (4.80). The limiting Lévy process has independent increments and discontinuous sample paths. Thus, in this case the particular structure of the coefficients a_j is not really important. On the other hand, if $d \in (0, 1 - 1/\alpha)$, then the scaling factor involves the memory parameter d. This is one reason why such a process is said to have long-range dependence. Also, the limiting process has dependent increments but continuous sample paths. We illustrate this in Example 4.15. Note also that the theorem can be stated more generally by allowing slowly varying functions in both a_j and the tail of ε_1.

Remark 4.3 It should be pointed out that in the long-memory case ($d \in (0, 1-1/\alpha)$) we have weak convergence w.r.t. the standard J_1-topology and the limiting process has continuous paths. In contrast, in the case of summable coefficients we have finite-dimensional convergence only, and this cannot be extended to J_1-convergence. This can be seen as follows. Assume for a moment that $X_t = b_0\varepsilon_t + b_1\varepsilon_{t-1}$ ($t \in \mathbb{N}$). The limiting behaviour of $S_n = \sum_{t=1}^n X_t$ is determined by large values of X_t ($t \in \mathbb{N}$). Now, there is a small chance that both ε_t and ε_{t+1} are large since $P(\varepsilon_t > x, \varepsilon_{t+1} > x) = o(P(\varepsilon_1 > x))$ as $x \to \infty$. Therefore, we have one large value of a particular ε_{t^*}, say which implies $X_{t^*} \approx b_0\varepsilon_{t^*}$ and $X_{t^*+1} \approx b_1\varepsilon_{t^*}$. This produces two "clustered" large jumps in the limiting process, which contradicts a heuristic explanation of J_1-topology in the Appendix A. However, it is possible to have weak convergence w.r.t. different topologies. We refer to Avram and Taqqu (1992).

Proof In the case of weak dependence (i.e. where (4.88) holds), the proof mimics the one for normal convergence (see Theorem 4.5). Let $X_{t,K} = \sum_{j=0}^K a_j\varepsilon_{t-j}$. Note that (4.80) can be restated for $u = 1$ as

$$n^{-1/\alpha}\left(\sum_{t=1}^n \varepsilon_t, \ldots, \sum_{t=1}^n \varepsilon_{t-m}\right) \xrightarrow{\mathrm{d}} A^{1/\alpha}C_\alpha^{-1/\alpha}\big(Z_\alpha(1), \ldots, Z_\alpha(1)\big).$$

The continuous mapping theorem implies

$$n^{-1/\alpha}\sum_{t=1}^n X_{t,K} \xrightarrow{\mathrm{d}} A^{1/\alpha}C_\alpha^{-1/\alpha}\left(\sum_{j=0}^K a_j\right)Z_\alpha(1).$$

Furthermore, $(\sum_{j=0}^K a_j)Z_\alpha(1) \xrightarrow{\mathrm{p}} (\sum_{j=0}^\infty a_j)Z_\alpha(1)$. We finish the proof by verifying

$$\limsup_{K\to\infty} \lim_{n\to\infty} P\big(n^{-1/\alpha}\big|S_n(1) - S_{n,K}(1)\big| > \gamma\big) = 0$$

for each $\gamma > 0$. This requires precise calculations on the tail behaviour of X_t. In particular, (4.86) plays a crucial role. We refer to Davis and Resnick (1985) for details. The result then follows from Proposition 4.1.

As for the long-memory case, we assume for simplicity that ε_t $(t \in \mathbb{Z})$ are $S_\alpha(1, \beta, 0)$. We may write

$$S_n = \sum_{t=1}^{n} X_t = \sum_{l=-\infty}^{n} \varepsilon_l \sum_{j=1-l}^{n-l} a_j =: \sum_{l=-\infty}^{n} \tilde{a}_{l,n} \varepsilon_l$$

with $\tilde{a}_{l,n} = \sum_{j=1-l}^{n-l} a_j$. If $a_j \sim c_a j^{d-1}$, then

$$\tilde{a}_{l,n} \sim \frac{c_a}{d} \left\{ (n-l)^d - (1-l)^d \right\}.$$

Therefore, since S_n is a sum of independent stable random variables, on account of (4.76), we expect that

$$\sum_{l=-\infty}^{n} \tilde{a}_{l,n} \varepsilon_l \stackrel{\mathrm{d}}{=} S_\alpha(\eta_n, \beta, 0)$$

with the scale parameter such that

$$\begin{aligned} \eta_n^\alpha = \sum_{l=-\infty}^{n} \tilde{a}_{l,n}^\alpha &= \left(\frac{c_a}{d} \right)^\alpha \sum_{l=-\infty}^{n} \left\{ (n-l)^d - (1-l)^d \right\}^\alpha \\ &\sim \left(\frac{c_a}{d} \right)^\alpha \frac{1}{n^{d\alpha+1}} \int_{-\infty}^{1} \left\{ (1-v)_+^d - (-v)_+^d \right\}^\alpha dv. \end{aligned}$$

Here, note that the integral above is defined only if $0 < d < 1 - 1/\alpha$. Therefore, with $b_n = (c_\alpha / d) n^H$ (recall that now $C_\alpha = A$ since we consider stable innovations), the distribution of $b_n^{-1} S_n(1)$ agrees asymptotically with the distribution of a stable random variable with the scale

$$\eta = \left(\int_{-\infty}^{1} \left\{ (1-v)_+^d - (-v)_+^d \right\}^\alpha dv \right)^{1/\alpha}$$

and skewness β. Now, if we have a stable integral $\int g(x)\, dM(x)$, then it is a stable random variable with the scale $(\int |g(x)|^\alpha \, dx)^{1/\alpha}$. Thus, for each u, the Linear Fractional Stable Motion $\tilde{Z}_{H,\alpha}(\cdot)$ (see Sect. 3.7.2 for additional details)

$$\int_{-\infty}^{u} \left\{ (u-v)_+^{H-1/\alpha} - (-v)_+^{H-1/\alpha} \right\} dZ_\alpha(v)$$

is a stable random variable with the scale

$$u^{1/\alpha} \left(\int_{-\infty}^{1} \left\{ (u-v)_+^{H-1/\alpha} - (-v)_+^{H-1/\alpha} \right\}^\alpha dv \right)^{1/\alpha}.$$

Consequently, the result follows for $u = 1$. In this argument we replaced the coefficients $\tilde{a}_{l,n}$ by the asymptotically equivalent expressions. This approximation can be made more precise by computing the characteristic function. □

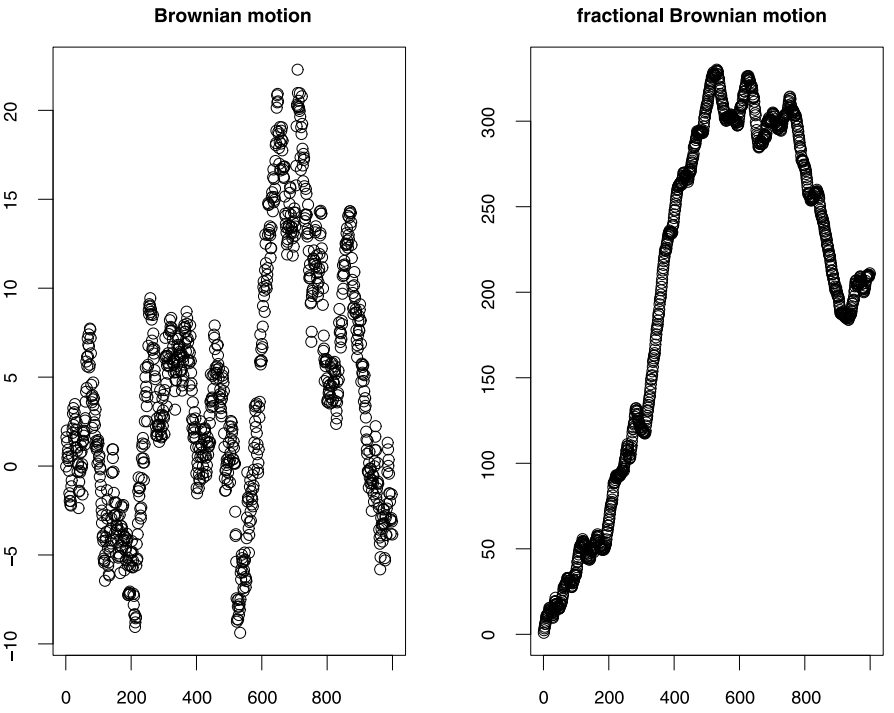

Fig. 4.5 Paths of a partial sum sequence $S_k = \sum_{t=1}^{k} X_t$ with X_t i.i.d. $N(0,1)$ (*left*) and X_t generated by a FARIMA$(0, 0.4, 0)$ process

Example 4.15 We illustrate Theorem 4.15 by a simulation study. First, as in Example 4.2, we generate $n = 1000$ i.i.d. standard normal random variables X_t and plot the partial sum sequence $S_k = \sum_{t=1}^{k} X_t$, $k = 1, \ldots, n$. This procedure is repeated for Gaussian FARIMA$(0, d, 0)$ process with $d = 0.4$. The path of the fractional Brownian motion is much smoother than of the Brownian motion. This is due to the influence of long memory. The corresponding partial sum processes are plotted in Fig. 4.5. For comparison, we simulate i.i.d. random variables from a t-distribution with $3/2$ degrees of freedom (hence, with a finite mean and infinite variance) and a FARIMA$(0, 0.4, 0)$ process where the innovations have a t-distribution with $3/2$ degrees of freedom. The partial sum processes are depicted on Fig. 4.6. In the i.i.d. case, the process has clearly discontinuous sample paths, whereas this effect does not seem to be present in the long-memory case.

4.3.3.4 Subordinated Case

Consider the partial sum

$$S_{n,G}(u) = \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \quad \big(u \in [0,1]\big),$$

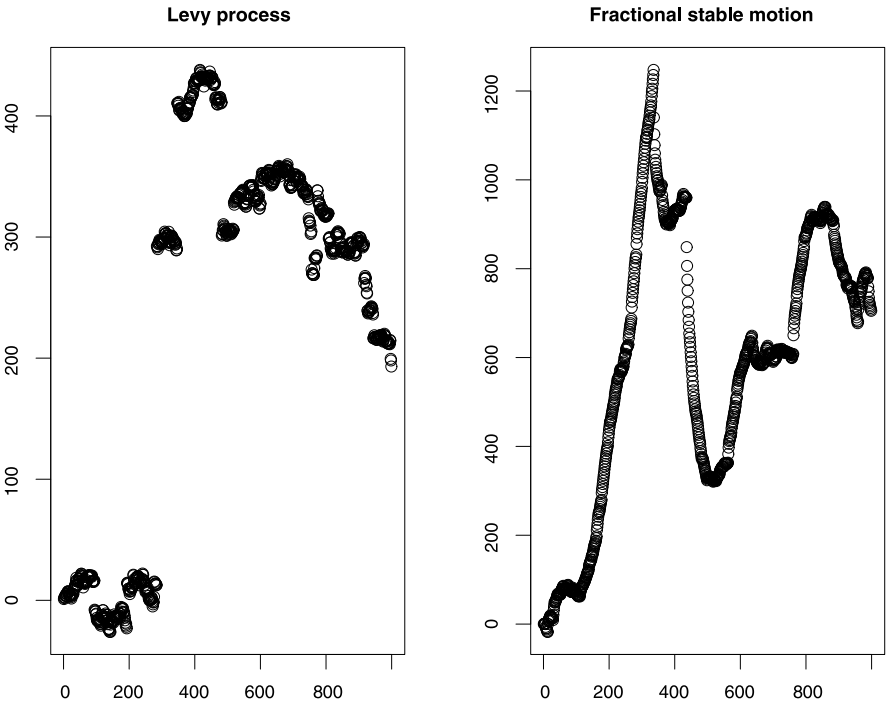

Fig. 4.6 Paths of a Lévy stable motion and a fractional stable motion with Hurst parameter $H = d + 1/\alpha$, $d = 0.4$, $\alpha = 3/2$

where G is a measurable function. Subordinated linear processes with infinite second moments were studied in Hsing (1999), Koul and Surgailis (2001), Surgailis (2002, 2004), Vaičiulis (2003). Surprisingly, the martingale decomposition method, used in Theorem 4.9 for variables with finite variance, works also here.

We start with the simple case of polynomials. Let us focus on a quadratic function $G(x) = x^2$. If $\alpha \in (0, 2)$, then we can repeat the argument following point process convergence in Theorem 4.14. First (see the discussion following Theorem 4.13), we can also write

$$\sum_{t=1}^{n} \delta_{c_n^{-2} X_t^2} \Rightarrow \sum_{j=0}^{\infty} \sum_{l=0}^{\infty} \delta_{j_l^2 a_j^2}.$$

This is valid as long as the conditions of Lemma 4.19 hold. Now, if $\alpha \in (0, 2)$, the random variables X_t^2 ($t \in \mathbb{N}$) have infinite means. Therefore, for $\alpha \in (0, 2)$,

$$c_n^{-2} \sum_{t=1}^{n} X_t^2 \xrightarrow{\mathrm{d}} \left(\sum_{j=0}^{\infty} a_j^2 \right) \sum_{l=0}^{\infty} j_l^2 = \left(\sum_{j=0}^{\infty} a_j^2 \right) S_{\alpha/2}\big(C_{\alpha/2}^{-2/\alpha}, 1, 0\big),$$

or equivalently,

$$n^{-2/\alpha}\sum_{t=1}^{n} X_t^2 \xrightarrow{\mathrm{d}} \left(\sum_{j=0}^{\infty} a_j^2\right) A^{2/\alpha} C_{\alpha/2}^{-2/\alpha} S_{\alpha/2}(1,1,0).$$

The case $\alpha \in (0,1)$ was proven in Davis and Resnick (1985, Theorem 4.2), whereas the case $\alpha \in (1,2)$ is addressed in Kokoszka and Taqqu (1996, Theorem 2.1). In other words, long memory does not influence the limiting behaviour.

Now, the situation changes when $2 < \alpha < 4$. The partial sum

$$S_{n,G}(u) = \sum_{t=1}^{[nu]} \left(X_t^2 - E\left(X_1^2\right)\right)$$

can be decomposed as (cf. Example 4.9)

$$S_{n,G,1}(u) + S_{n,G,2}(u) := \sum_{t=1}^{[nu]} \sum_{j=0}^{\infty} a_j^2 \left(\varepsilon_{t-j}^2 - E\left(\varepsilon_1^2\right)\right) + \sum_{t=1}^{[nu]} \sum_{j,k=0;\ j\neq k}^{\infty} a_j a_l \varepsilon_{t-j}\varepsilon_{t-k}.$$

The first part $S_{n,G,1}(u)$ is a partial sum process based on the linear process with summable coefficients a_j^2. Therefore, on account of the first part of Theorem 4.15,

$$n^{-2/\alpha} S_{n,G,1}(u) \xrightarrow{\text{f.d.}} A^{2/\alpha} C_{\alpha/2}^{-2/\alpha} \left(\sum_{j=0}^{\infty} a_j^2\right) Z_{\alpha/2}(u),$$

where $Z_{\alpha/2}(\cdot)$ is a Lévy process such that $Z_{\alpha/2}(1) \stackrel{\mathrm{d}}{=} S_{\alpha/2}(1,1,0)$, i.e. $Z_{\alpha/2}(1)$ is an $\alpha/2$-stable random variable that is completely skewed to the right.

Convergence of the second term follows exactly as in Example 4.9. First, since $2 < \alpha < 4$, the random variables ε_t have a finite variance where under the assumption $a_j \sim c_a j^{d-1}$ we have $\gamma_X(k) = cov(X_t, X_{t+k}) \sim L_\gamma(k) k^{2d-1}$ with

$$L_\gamma(k) = c_a^2 \sigma_\varepsilon^2 \int_0^\infty v^{d-1}(v+1)^{d-1}\, dv,$$

see Lemma 4.13. If $1/4 < d < 1/2$, then

$$n^{-2d} L_2^{-1/2}(n) S_{n,G,2}(u) \Rightarrow Z_{2,H}(u),$$

where $H = d + 1/2$, $Z_{2,H}(u)$ is the Hermite–Rosenblatt process, and

$$L_2(n) = m! C_m L_\gamma^m(n).$$

Otherwise, if $0 < d < 1/4$, then $n^{-1/2} S_{n,G,2}(u) = O_P(1)$. Therefore, we have a dichotomous behaviour depending on a relation between the "memory parameter" d and tails. Such consideration can be carried out for instance for Appell polynomials

(see Vaičiulis 2003). Before we state our theorem, we recall for convenience the heavy-tail condition (4.84):

$$P(\varepsilon_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha}. \tag{4.89}$$

Theorem 4.16 *Assume that X_t $(t \in \mathbb{Z})$ is a linear process such that $a_j \sim c_a j^{d-1}$, $d \in (0, 1/2)$ and ε_t $(t \in \mathbb{Z})$ are i.i.d. random variables such that (4.89) holds with $\alpha \in (2, 4)$. Also, assume that $E(\varepsilon_1) = 0$.*

- *If $0 < d < 1/\alpha$, then*

$$n^{-2/\alpha}\sum_{t=1}^{[nu]}\left(X_t^2 - E\left(X_1^2\right)\right) \overset{\text{f.d.}}{\to} A^{2/\alpha}C_{\alpha/2}^{-2/\alpha}\left(\sum_{j=0}^{\infty}a_j^2\right)Z_{\alpha/2}(u),$$

  where $Z_{\alpha/2}(\cdot)$ is an $\alpha/2$-stable Lévy motion such that $Z_{\alpha/2}(1) \overset{\text{d}}{=} S_{\alpha/2}(1, 1, 0)$.
- *If $1/\alpha < d < 1/2$, then*

$$n^{-2d}L_2^{-1/2}(n)\sum_{t=1}^{[nu]}\left(X_t^2 - E\left(X_1^2\right)\right) \Rightarrow Z_{2,H}(u),$$

  where $\Rightarrow$ denotes weak convergence in $D[0, 1]$, $Z_{2,H}(\cdot)$ is the Hermite–Rosenblatt process, and $H = d + 1/2$.

The next theorem follows from Theorem 4.15 and a reduction principle along the lines of Theorem 4.9. We assume that the innovations in the linear process are symmetric.

Theorem 4.17 *Assume that X_t $(t \in \mathbb{Z})$ is a linear process such that $a_j \sim c_a j^{d-1}$, $d \in (-\infty, 1/2)$, ε_t $(t \in \mathbb{Z})$ are i.i.d. symmetric random variables such that (4.89) holds with $\alpha \in (1, 2)$ and $\beta = 0$, i.e.*

$$P(\varepsilon_1 > x) \sim \frac{A}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim \frac{A}{2}x^{-\alpha}.$$

Furthermore, assume that the distribution F_ε of ε_1 fulfills

$$\left|F_\varepsilon^{(2)}(x)\right| \le C\left(1 + |x|\right)^{-\alpha}, \qquad \left|F_\varepsilon^{(2)}(x) - F_\varepsilon^{(2)}(y)\right| \le C|x - y|\left(1 + |x|\right)^{-\alpha},$$

where $|x - y| < 1$, $x \in \mathbb{R}$.

- *If $0 < d < 1 - 1/\alpha$ and G is bounded, then*

$$n^{-H}\sum_{t=1}^{[nu]}\left\{G(X_t) - E\left[G(X_1)\right]\right\} \Rightarrow A^{1/\alpha}C_\alpha^{-1/\alpha}\frac{c_a}{d}G_\infty^{(1)}(0)\tilde{Z}_{H,\alpha}(u), \tag{4.90}$$

where $\Rightarrow$ *denotes weak convergence in* $D[0,1]$, *and* $\tilde{Z}_{H,\alpha}(\cdot)$ *is a linear fractional stable motion with* $H = d + \alpha^{-1}$ *such that* $\tilde{Z}_{H,\alpha}(1)$ *is a symmetric* α*-stable random variable with scale*

$$\eta = \left(\int_{-\infty}^{1} \left\{ (1-v)_+^d - (-v)_+^d \right\}^\alpha dv \right)^{1/\alpha}$$

and $G_\infty(x) = E[G(X+x)]$.

- *If* $1 - 2/\alpha < d < 0$ *and* $A = 1$ *in* (4.89) *and* G *is bounded, then*

$$n^{-1/\alpha(1-d)} \sum_{t=1}^{[nu]} \left\{ G(X_t) - E\left[G(X_1)\right] \right\} \Rightarrow c_G^+ \tilde{Z}^+_{\alpha(1-d)}(u) + c_G^- \tilde{Z}^-_{\alpha(1-d)}(u), \quad (4.91)$$

where $\tilde{Z}^+_{\alpha(1-d)}(\cdot)$, $\tilde{Z}^-_{\alpha(1-d)}(\cdot)$ *are independent copies of an* $\alpha(1-d)$*-stable Lévy motion such that* $Z_{\alpha(1-d)}(1) \stackrel{\mathrm{d}}{=} S_{\alpha(1-d)}(1,1,0)$ *and*

$$c_G^\pm = C_{\alpha(1-d)}^{-1/\alpha(1-d)} \frac{c_a^{1/(1-d)}}{1-d} \int_0^\infty \left[G_\infty(\pm v) - G_\infty(0) \right] v^{-1-1/(1-d)} \, dv,$$

where $G_\infty(x) = E[G(X_1 + x)]$.

- *If* $-\infty < d < 1 - 2/\alpha$ *and* G *is bounded, then*

$$n^{-1/2} \sum_{t=1}^{[nu]} \left\{ G(X_t) - E\left[G(X_1)\right] \right\} \Rightarrow \sigma_S B(u), \quad (4.92)$$

where $B(\cdot)$ *is a standard Brownian motion, and* σ_S *is a finite positive constant.*

This theorem was proven in Koul and Surgailis (2001), Surgailis (2002) and Hsing (1999). Remarkably, in (4.90) and (4.91), we may obtain a stable limit arising from a summation of bounded random variables. The convergence in (4.90) can be thought of as a *long-memory-type behaviour* since the scaling involves the memory parameter d and the limiting process has dependent increments. The convergence in (4.91) is a sort of an *intermediate case*: the scaling involves d, but the limiting process has independent increments. Finally, (4.92) represents a *standard behaviour*: as in the i.i.d. case, the limiting process is a Brownian motion since $\mathrm{var}(G(X_1))$ is finite.

Below, we give an outline of the proof of (4.90). As for (4.91), the limiting process has independent increments, but the scaling factor involves the memory parameter d. The reason for this is that the process $S_{n,G}(u)$ can be approximated by a sum $\sum_{t=1}^n \eta_G(\varepsilon_t)$ of i.i.d. random variables, where

$$\eta_G(\varepsilon_t) = \sum_{j=0}^{\infty} \left\{ G_\infty(a_j \varepsilon_t) - E\left[G_\infty(a_j \varepsilon_1) \right] \right\},$$

and the variables η_G have a tail decaying like $|x|^{-\alpha(1-d)}$.

In (4.90) it may happen that the quantity $G_\infty^{(1)}(0)$ vanishes. It is an open question, whether it is possible to obtain a nondegenerate limit in this case with $1 < \alpha < 2$. Let us recall that in the case of linear processes with finite moments the solution to this problem is given for example in Theorem 4.4. In the case of infinite moments, this question was studied in Surgailis (2004) under the assumption $2 < \alpha < 4$. It may happen that the limit is an $\alpha(1-d)$-Lévy stable motion, Hermite–Rosenblatt process or Brownian motion.

Proof of Theorem 4.17 Recall the notation from the proof of Theorem 4.9. We denote by $\mathcal{V}_t$ the sigma field generated by $(\varepsilon_t, \varepsilon_{t-1}, \ldots)$ and set

$$T_n(G;1) = \sum_{t=1}^{n} \big(G(X_t) - E\big[G(X_1)\big] - G_\infty^{(1)}(0) X_t\big)$$

and $P_K Y = E(Y|\mathcal{V}_K) - E(Y|\mathcal{V}_{K-1})$. We can repeat the computation there, using the rth norm with $r < \alpha$ instead of $r = 2$:

$$\begin{aligned} \big\|T_n(G;1)\big\|_r^r &\le 2 \sum_{K=-\infty}^{n} \Big\| \sum_{t=\max\{K,1\}}^{n} P_K U(\mathcal{V}_t) \Big\|_r^r \\ &\le \sum_{K=-\infty}^{n} \Bigg(\sum_{t=\max\{K,1\}}^{n} \big\| P_{-(t-K)} U(\mathcal{V}_0) \big\|_r \Bigg)^r . \end{aligned}$$

The first inequality follows from a result for martingale differences Y_t $(t \in \mathbb{N})$, namely

$$\Big\| \sum_{t=1}^{n} Y_t \Big\|_r^r \le 2 \sum_{t=1}^{n} \|Y_j\|_r^r$$

for any $1 \le r \le 2$. The second one is the norm inequality used in the proof of Theorem 4.9. Now, instead of Lemma 4.17, we use

$$\big\| P_{-(t-K)} U(\mathcal{V}_0) \big\|_r \le (t-K)^{-(1-d)(1+\gamma)},$$

where $(1+\gamma)r < \alpha$. Computations leading to this expression are quite involved; we refer the reader to Koul and Surgailis (2001). Then one obtains

$$\big\|T_n(G;1)\big\|_r^r \le C \sum_{K=-\infty}^{n} \Bigg(\sum_{t=K\vee 1}^{n} (t-K)^{-(1-d)(1+\gamma)} \Bigg)^r \le C n^{r+1} n^{-(1-d)(1+\gamma)r}$$

by similar calculations as those leading to (4.64), (4.65). Choosing γ sufficiently close to 0, we conclude that

$$\big\|T_n(G;1)\big\|_r^r = o\big(n^{r(d+1/\alpha)}\big).$$

In particular, $\|T_n(G;1)\|_r^r = o(v_n^r)$, where

$$v_n = C_\alpha^{-1/\alpha} A^{1/\alpha} \frac{c_a}{d} n^H$$

with $H = d + \frac{1}{\alpha}$. Therefore, on account of Theorem 4.15, the limiting behaviour of

$$v_n^{-1} \sum_{t=1}^{n} \{G(X_t) - E[G(X_1)]\}$$

is the same as that of $v_n^{-1} G_\infty^{(1)}(0) \sum_{t=1}^n X_t$. □

4.3.4 Stochastic Volatility Models

In this section we consider Long-Memory Stochastic Volatility (LMSV) sequences with infinite moments. Let $X_t = \sigma_t \xi_t$ $(t \in \mathbb{N})$, where

$$\sigma_t = \sigma(\zeta_t), \quad \zeta_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j},$$

$\sigma(\cdot)$ is a positive function, $\sum_{j=1}^{\infty} a_j^2 < \infty$, and ε_t $(t \in \mathbb{Z})$ are i.i.d. random variables. It is further assumed that ξ_t $(t \in \mathbb{Z})$ is a sequence of i.i.d. random variables such that

$$P(\xi_1 > x) \sim A \frac{1+\beta}{2} x^{-\alpha}, \qquad P(\xi_1 < -x) \sim A \frac{1-\beta}{2} x^{-\alpha}. \tag{4.93}$$

Also, we assume that the sequences ε_t $(t \in \mathbb{Z})$ and ξ_t $(t \in \mathbb{Z})$ are mutually independent. At the moment we do not assume anything about the mean of ξ_t.

Limiting results for infinite-variance volatility models with long memory are almost non-existing; see Kulik and Soulier (2012) or Surgailis (2008); the latter in a quadratic LARCH case. In particular, we will show below that stochastic volatility models can be treated using a point process methodology.

4.3.4.1 Tail Behaviour

The first question we have to answer is the following. If ξ is like in (4.93), what is the consequence on the tail of X? The next lemma shows that if the random variables ε and σ are independent, then $\sigma\varepsilon$ is still regularly varying. The result is often referred to as Breiman's lemma (Breiman 1965), and a proof can be found for example in Resnick (2007, Proposition 7.5).

Lemma 4.20 *Assume that* (4.93) *holds. If* σ_1 *is a positive random variable independent of* ξ_1 *and such that for some* $\delta > 0$,

$$E\left(\sigma_1^{\alpha+\delta}\right) < \infty, \tag{4.94}$$

then the distribution of $\sigma\xi$ *is regularly varying, and*

$$\lim_{x\to\infty} \frac{P(\sigma_1\xi_1 > x)}{P(|\xi_1| > x)} = \frac{1+\beta}{2} E\left(\sigma_1^{\alpha}\right), \qquad \lim_{x\to\infty} \frac{P(\sigma_1\xi_1 < -x)}{P(|\xi_1| > x)} = \frac{1-\beta}{2} E\left(\sigma_1^{\alpha}\right). \tag{4.95}$$

Lemma 4.20 implies for the LMSV model and arbitrary $p > 0$ that

$$P\left(|X_1|^p > x\right) = P\left(X_1 > x^{1/p}\right) + P\left(X_1 < -x^{1/p}\right) \sim A\, E\left(\sigma_1^{\alpha}\right) x^{-\alpha/p}. \tag{4.96}$$

Thus, if we consider the LMSV model, we may take ξ_t as in (4.93), $\sigma(x) = e^x$ and ζ_t $(t \in \mathbb{N})$ to be e.g. long-memory Gaussian. Then the random variables X_t $(t \in \mathbb{N})$ have heavy tails and long memory.

4.3.4.2 Point Process Convergence

Point process convergence results play a crucial role when proving asymptotic results for partial sums based on infinite-variance sequences. Here, we assume that the reader is familiar with material presented in Sect. 4.3.2.4.

We start with a simple generalization of Theorem 4.13 to the LMSV model. Recall the intensity measure

$$d\lambda(x) = \alpha\left[\frac{1+\beta}{2} x^{-(\alpha+1)} 1\{0 < x < \infty\} + \frac{1-\beta}{2}(-x)^{-(\alpha+1)} 1\{-\infty < x < 0\}\right] dx,$$

where $\beta \in [-1, 1]$, and consider the point processes

$$N_n = \sum_{t=1}^{n} \delta_{(t/n, c_n^{-1} X_t)},$$

where c_n is chosen to fulfill $P(|\xi_1| > c_n) \sim n^{-1}$, i.e.

$$c_n = A^{1/\alpha} n^{1/\alpha}.$$

The next result shows that the point process based on the LMSV sequence X_t behaves as if the random variables were independent. It will be clear from the proof that the same applies to $|X_t|^r$ where r is any power. Furthermore, we do not really need the particular structure $\sigma_t = \sigma(\zeta_t)$, where ζ_t $(t \in \mathbb{Z})$ is a linear process. Only the ergodicity of σ_t $(t \in \mathbb{N})$ is needed.

Theorem 4.18 *Consider the LMSV model $X_t = \sigma_t \xi_t$ $(t \in \mathbb{N})$ such that (4.93) and Breiman's condition (4.94) hold. Then N_n converges weakly in $M_p([0,1] \times \mathbb{R})$ to a Poisson process N with intensity measure $E(\sigma_1^\alpha) ds \times d\lambda(x)$.*

Proof (Personal communication with P. Soulier) The proof is basically the same as in the i.i.d. case, see Theorem 4.13. We also use the same notation as in Theorem 4.13. Let $U = \bigcup_{i=1}^K (k_i, l_i) \times (s_i, t_i)$. Then

$$P\big(N_n(U) = 0\big) = P\left(\sum_{i=1}^K \sum_{nk_i < t < nl_i}^n 1\big\{c_n^{-1} X_t \in (s_i, t_i)\big\} = 0\right)$$
$$= E\left[\prod_{i=1}^K \prod_{nk_i < t < nl_i}^n P\big(c_n^{-1} X_t \notin (s_i, t_i) | \mathcal{F}_\sigma\big)\right] =: m E[P_n],$$

where $\mathcal{F}_\sigma$ is the sigma field generated by the entire sequence σ_t. Let $\theta_t((s_i, t_i))$ be the limit of $nP(c_n^{-1} X_t \in (s_i, t_i) | \mathcal{F}_\sigma)$ and write

$$Q_n = \prod_{i=1}^K \prod_{nk_i < t < nl_i} \exp\big\{-n^{-1} \theta_t\big((s_i, t_i)\big)\big\}.$$

Note that θ_t is a random variable since it depends on the sequence σ_t $(t \in \mathbb{N})$. Therefore, the only difference between the LMSV setting and the i.i.d. one is that Q_n here is a random variable and $\lambda((s_i, t_i))$ is replaced by $\theta_t((s_i, t_i))$. Nevertheless, Q_n converges in probability to

$$\exp\left\{-E\big(\sigma_1^\alpha\big) \sum_{i=1}^K (l_i - k_i) \lambda\big((s_i, t_i)\big)\right\} = P\big(N(U) = 0\big).$$

It remains to prove that $|P_n - Q_n|$ converges in probability to 0 and apply the bounded convergence theorem. To prove that $|P_n - Q_n| \to_P 0$, we proceed as in Theorem 4.13:

$$E|P_n - Q_n|$$
$$\le \sum_{i=1}^K (l_i - k_i) E\big[\big|nP\big(c_n^{-1} X_1 \in (s_i, t_i) | \mathcal{F}_\sigma\big) - \theta_1\big((s_i, t_i)\big)\big|\big]$$
$$+ \sum_{i=1}^K n(l_i - k_i) E\left[\left|1 - e^{-n^{-1}\theta_1((s_i,t_i))} - \frac{\theta_1((s_i, t_i))}{n}\right|\right].$$

For the second term, we have

$$nE\left[\left|1 - e^{-n^{-1}\theta_1((s_i,t_i))} - \frac{\theta_1((s_i, t_i))}{n}\right|\right] \le C n^{-\delta} E\big[\sigma_1^{\alpha+\delta}\big].$$

Furthermore, let us recall the so-called Potter's bound (see Theorem 1.5.6. in Bingham et al. 1989), namely: for $v > 0$,

$$nP\big(c_n^{-1} v\xi_1 \in (s_i, t_i)\big) \leq C\big(\max\{v, 1\}\big)^{\alpha+\delta},$$

where $\delta > 0$. For the first term, we apply Potter's bound to get

$$nP\big(c_n^{-1} X_1 \in (s_i, t_i)|\mathcal{F}_\sigma\big) = nP\big(c_n^{-1}\xi_1\sigma_1 \in (s_i, t_i)|\mathcal{F}_\sigma\big) \leq \big(\max\{\sigma_1, 1\}\big)^{\alpha+\delta},$$

and the same bound holds for $\theta_1(s_i, t_i)$. We then can apply bounded convergence to get

$$\lim_{n\to\infty} E\Big[\big|nP\big(c_n^{-1} X_1 \in (s_i, t_i)\big) - \theta_1\big((s_i, t_i)\big)\big|\Big] = 0. \qquad \square$$

4.3.4.3 Convergence of Partial Sums

Having established point process convergence, we proceed with its consequences for partial sums. Assume that ξ_1 fulfills (4.93) and $E(\xi_1) = 0$ or ξ_1 is symmetric if $\alpha \in (0, 1)$. Define

$$S_n(u) = \sum_{t=1}^{[nu]} X_t$$

and

$$S_{n,p}(u) = \sum_{t=1}^{[nu]} \big(|X_t|^p - E\big[|X_1|^p\big]\big),$$

assuming that $E[|X_1|^p] < \infty$ but $E[|X_1|^{2p}] = \infty$. Due to Lemma 4.20, this is achieved when $p < \alpha < 2p$. In the next theorem we show that depending on an interplay between long memory and tails, partial sums based on the LMSV sequence may converge either to a Lévy process (weakly dependent behaviour) or to a Hermite process (long-memory behaviour).

Theorem 4.19 *Consider the LMSV model $X_t = \sigma_t\xi_t$ $(t \in \mathbb{N})$ and assume that the conditions of Theorem 4.18 hold. In addition, we assume that $\alpha > 1$, $E(\xi_1) = 0$ and ζ_t $(t \in \mathbb{N})$ is a Gaussian linear process with coefficients a_j satisfying* (B1), *i.e. $a_j = L_a(j)j^{d-1}$, $d \in (0, 1/2)$, and covariance function $\gamma_\zeta(k) \sim L_\gamma(k)k^{2d-1}$. Let $m \geq 1$ be the Hermite rank of the function $\sigma^p(\cdot)$ and assume further that $E(\sigma_1^{2\alpha+2\varepsilon}) < \infty$.*

- *If $1 < \alpha < 2$, then*

$$n^{-1/\alpha} S_n(u) \Rightarrow A^{1/\alpha} C_\alpha^{-1/\alpha} \big(E\big[\sigma_1^\alpha\big]\big)^{1/\alpha} Z_\alpha(u), \tag{4.97}$$

where $Z_\alpha(\cdot)$ is an α-stable Lévy process such that $Z_\alpha(1) \stackrel{\mathrm{d}}{=} S_\alpha(1, \beta, 0)$, and $\Rightarrow$ denotes weak convergence in $D[0, 1]$.

- *If $p < \alpha < 2p$ and $1 - m(1/2 - d) < p/\alpha$, then*

$$n^{-p/\alpha} S_{n,p}(u) \Rightarrow A^{p/\alpha} C_{\alpha/p}^{-p/\alpha} \left(E\left[\sigma_1^{\alpha}\right]\right)^{p/\alpha} Z_{\alpha/p}(u), \tag{4.98}$$

  where $Z_\alpha(\cdot)$ is an α/p-stable Lévy process such that $Z_\alpha(1) \stackrel{\mathrm{d}}{=} S_{\alpha/p}(1, 1, 0)$, and $\Rightarrow$ denotes weak convergence in $D[0, 1]$.
- *If $p < \alpha < 2p$ and $1 - m(1/2 - d) > p/\alpha$, then*

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) S_{n,p}(u) \Rightarrow \frac{J(m) E[|\xi_1|^p]}{m!} Z_{m,H}(u), \tag{4.99}$$

  where $Z_{m,H}(\cdot)$ is a Hermite process of order m, $H = d + \frac{1}{2}$,

$$L_m(n) = m! C_m L_\gamma(n),$$

  $J(m)$ is the Hermite coefficient of $\sigma^p(\cdot)$, and $\Rightarrow$ denotes weak convergence in $D[0, 1]$.

When $\alpha \in (1, 2)$, the partial sum $S_n(u)$ is a martingale because $E(X_t) = E(\xi_t)E(\sigma_t) = 0$. Hence, only the stable Lévy limit arises, and (4.97) holds. This can be concluded from a general theory by Surgailis (2008). If $S_{n,p}(\cdot)$ is considered, then we observe a dichotomous behaviour. Assume for simplicity that $m = 1$. If long memory is strong enough, then it influences the limiting behaviour. Interestingly, the infinite variance sequence $|X_t|^p$ yields a limiting process with finite variance. Furthermore, results are readily extendable to the case where ζ_t is a general linear process. Instead of Theorem 4.4, one has to use corresponding results for subordinated linear processes; see Theorem 4.6. Furthermore, in contrast to Theorem 4.15 for linear processes with infinite variance, we note that we have weak convergence w.r.t. J_1-topology in all three cases.

Example 4.16 (Cf. Example 4.11) Assume that $X_t = \xi_t \exp(\zeta_t)$, where ζ_t is a standard normal sequence with covariance $\gamma_\zeta(k) \sim L_\gamma k^{2d-1}$, $d \in (0, 1/2)$. If $\alpha \in (2, 4)$ and $d + 1/2 < 2/\alpha$, then $n^{-2/\alpha} S_{n,2}(u)$ converges to a Lévy process. Otherwise, if $\alpha \in (2, 4)$ and $d + 1/2 > 2/\alpha$, then

$$n^{-(1/2+d)} L_1(n)^{-1/2}(n) S_{n,2}(u) \Rightarrow J(1) E\left(\xi_1^2\right) B_H(u),$$

where $L_1(n) = (d(2d + 1))^{-1} L_\gamma(n)$ and $J(1) = E[\zeta_1 \exp(2\zeta_1)]$.

In the spirit of Example 4.12, if $\alpha \in (1, 2)$ and $E(\xi_t) \neq 0$, then long memory appears already in $\sum_{t=1}^{[nu]} X_t$.

Example 4.17 (LMSD with Infinite Variance) As in Example 4.12, we assume that the random variables ξ_t ($t \in \mathbb{N}$) are strictly positive. Suppose that we have heavy tails

$$P(\xi_1 > x) \sim A x^{-\alpha} \quad (x \to \infty)$$

with $\alpha \in (1,2)$. Furthermore, it is assumed that the sequences ξ_t and ζ_t are independent and the covariance of ζ_t is of the asymptotic form $\gamma_\zeta(k) \sim L_\gamma(k)k^{2d-1}$, $d \in (0,1/2)$. Let $G(x) = x$ and $\sigma(x) = \exp(x)$, so that the Hermite rank $m = 1$. Then we have a dichotomous behaviour for $S_n(u) := \sum_{t=1}^{[nu]}(X_t - E(X_1))$. Specifically, (4.98) and (4.99) hold with $p = 1$:

- If $1/2 + d < 1/\alpha$, then

$$n^{-1/\alpha} S_n(u) \Rightarrow A^{1/\alpha} C_\alpha^{-1/\alpha} \big(E\big[\sigma_1^\alpha\big]\big)^{1/\alpha} Z_\alpha(u), \tag{4.100}$$

  where $Z_\alpha(\cdot)$ is an α-stable Lévy process such that $Z_\alpha(1) \stackrel{\mathrm{d}}{=} S_\alpha(1,1,0)$.
- If $1/2 + d > 1/\alpha$, then

$$n^{-(1/2+d)} L_1^{-1/2}(n) S_n(u) \Rightarrow J(1) E[\xi_1] B_H(u), \tag{4.101}$$

  where $B_H(\cdot)$ is a fractional Brownian motion, $H = d + \frac{1}{2}$, $L_1(n) = C_1 L_\gamma(n)$ and $J(1) = E[\zeta_1 \exp(\zeta_1)]$.

Proof of Theorem 4.19 Let $\mathcal{F}_t$ be a sigma field generated by ξ_j, ε_j $(j \le t)$. We start by studying $S_{n,p}(\cdot)$. Write

$$\sum_{t=1}^{[nu]} \big(|X_t|^p - E\big[|X_t|^p\big]\big) = \sum_{t=1}^{[nu]} \big(|X_t|^p - E\big[|X_t|^p | \mathcal{F}_{t-1}\big]\big) + \sum_{t=1}^{[nu]} \big(E\big[|X_t|^p | \mathcal{F}_{t-1}\big] - E\big[|X_1|^p\big]\big) =: M_n(u) + R_n(u).$$

Note that $E[|X_t|^p | \mathcal{F}_{t-1}] = E(|\xi_1|^p)\sigma^p(\zeta_t)$ is a function of ζ_t and does not depend on ξ_t. Therefore, for the long-memory part $R_n(u)$, we have

$$n^{-(1-m(\frac{1}{2}-d))} L_1^{-1/2}(n) R_n(u) \Rightarrow \frac{J(m) E[|\xi_1|^p]}{m!} Z_{m,H}(u) \tag{4.102}$$

if $m(1/2 - d) < 1$, where $Z_{m,H}(\cdot)$ is a Hermite–Rosenblatt process, and L_1 is a slowly varying function defined in Theorem 4.4. If $m(1/2 - d) > 1$, then

$$n^{-1/2} R_n(u) \Rightarrow v E\big[|\xi_1|^p\big] B(u), \tag{4.103}$$

where $B(\cdot)$ is a standard Brownian motion, and v is a constant.

We will show that under the assumptions we have,

$$c_n^{-p} M_n(u) \Rightarrow C_{\alpha/p}^{-p/\alpha} \big(E\big[\sigma_1^\alpha\big]\big)^{p/\alpha} Z_{\alpha/p}(u), \tag{4.104}$$

or equivalently,

$$n^{-1/\alpha} M_n(u) \Rightarrow A^{p/\alpha} C_{\alpha/p}^{-p/\alpha} \big(E\big[\sigma_1^\alpha\big]\big)^{p/\alpha} Z_{\alpha/p}(u).$$

From (4.102), (4.103) and (4.104) we conclude the proof of the theorem. First we prove (4.104). The proof is very similar to the proof of convergence of the partial sum of an i.i.d. sequence in the domain of attraction of a stable law to a Lévy stable process. The difference consists of some additional technicalities (see e.g. the proof of Theorem 71 in Resnick 2007 for additional details).

Step 1: For $0 < \varepsilon < 1$, decompose $M_n(u)$ further as

$$\begin{aligned} M_n(u) &= \sum_{t=1}^{[nu]} \big(|X_t|^p 1\{|X_t| < \varepsilon c_n\} - E\big[|X_t|^p 1\{|X_t| < \varepsilon c_n\}|\mathcal{F}_{t-1}\big]\big) \\ &\quad + \sum_{t=1}^{[nu]} \big(|X_t|^p 1\{|X_t| > \varepsilon c_n\} - E\big[|X_t|^p 1\{|X_t| > \varepsilon c_n\}|\mathcal{F}_{t-1}\big]\big) \\ &=: M_n^{(\varepsilon)}(u) + \tilde{M}_n^{(\varepsilon)}(u). \end{aligned}$$

The term $\tilde{M}_n^{(\varepsilon)}(\cdot)$ is treated using point process convergence. It excludes *small jumps* X_t defined by $c_n^{-1}|X_t| < \varepsilon$. The reason for this is that the summation functional is not continuous on the entire real line; one has to exclude small jumps. For any $\varepsilon > 0$, the summation functional is an almost surely (with respect to the distribution of the Poisson point process, see e.g. p. 215 in Resnick 2007) continuous mapping from the set of Radon measures on $[0, 1] \times [\varepsilon, \infty)$ to $D([0, 1], \mathbb{R})$. From Theorem 4.18 we then conclude

$$c_n^{-p} \sum_{t=1}^{[nu]} |X_t|^p 1\{|X_t| > \varepsilon c_n\} \Rightarrow \sum_{k: t_k \le u} |j_k|^p 1\{|j_k| > \varepsilon\} \tag{4.105}$$

in $([0, 1], \mathbb{R})$, where we recall that (t_k, j_k) are points of the limiting Poisson process. Taking expectations in (4.105), we obtain

$$\lim_{n\to\infty} [nu] c_n^{-p} E\big[|X_1|^p 1\{|X_1| > \varepsilon c_n\}\big] = u \int_{|x|>\varepsilon} |x|^p \, d\lambda(x)$$

uniformly with respect to $u \in [0, 1]$, since this is a sequence of increasing functions with a continuous limit. Furthermore, we claim that

$$c_n^{-p} \left| \sum_{t=1}^{[nu]} \big(E\big[|X_1|^p 1\{|X_1| > \varepsilon c_n\}\big] - E\big[|X_t|^p 1\{|X_t| > \varepsilon c_n\}|\mathcal{F}_{t-1}\big]\big) \right| \xrightarrow{\text{p}} 0$$

uniformly in $u \in [0, 1]$. The variance of the last expression is in fact bounded by

$$\begin{aligned} &c_n^{-2p} [nu]^2 \gamma_\zeta^m([nu]) \operatorname{var}\big(E\big[|X_1|^p 1\{|X_1| > \varepsilon c_n\}|\mathcal{F}_0\big]\big) \\ &\quad \le c_n^{-2p} [nu]^2 \gamma_\zeta^m([nu]) E\big[E^2\big[|X_1|^p 1\{|X_1| > \varepsilon c_n\}|\mathcal{F}_0\big]\big], \end{aligned}$$

where $\gamma_\zeta(k)$ is the covariance function of the Gaussian sequence ζ_t ($t \in \mathbb{Z}$), and m is the Hermite rank of $\sigma^p(\cdot)$. Recall Potter's bound (see Theorem 1.5.6. in Bingham et al. 1989): for $v > 0$,

$$nP\big(c_n^{-1} v\xi_1 \in (s_i, t_i)\big) \le C\big(\max\{v, 1\}\big)^{\alpha+\delta},$$

where $\delta > 0$. Now, if $p < \alpha < 2p$, then we combine Karamata's theorem with Potter's bound to obtain

$$\begin{aligned} E\big[\sigma^p(x)|\xi_1|^p 1\big\{\big|\sigma(x)\xi_1\big| > \varepsilon c_n\big\}\big] &\le Cn^{-1} c_n^p \frac{\bar{F}_\xi(\varepsilon c_n/\sigma(x))}{\bar{F}_\xi(c_n)} \\ &\le Cn^{-1} c_n^p \sigma^{\alpha+\varepsilon}(x). \end{aligned}$$

Since by assumption $E[\sigma_1^{2\alpha+2\varepsilon}] < \infty$ for some $\varepsilon > 0$, we have for each t,

$$\begin{aligned} &\operatorname{var}\left(c_n^{-p} \sum_{j=1}^{[nu]} \big\{E\big[|X_0|^p 1\big\{|X_0| > \varepsilon c_n\big\}\big] - E\big[|X_t|^p 1\big\{|X_t| > \varepsilon c_n\big\}\big|\mathcal{F}_{j-1}\big]\big\}\right) \\ &\quad \le Cn^{-2}[nu]^2 \gamma_\zeta\big([nu]\big) \le Cn^{2-2H+\varepsilon} u^{2H-\varepsilon}, \end{aligned} \tag{4.106}$$

where the last bound is obtained for some $\varepsilon > 0$ by Potter's bound. This proves the convergence of finite-dimensional distributions to 0 and tightness in $D([0, 1])$. We now argue that the bounds obtained above imply

$$c_n^{-p} \tilde{M}_n^{(\varepsilon)}(u) \Rightarrow C_{\alpha/p}^{-p/\alpha} \big(E\big[\sigma_1^\alpha\big]\big)^{p/\alpha} Z_{\alpha/p}^{(\varepsilon)}(u)$$

and also $Z_{\alpha/p}^{(\varepsilon)}(u) \Rightarrow Z_{\alpha/p}(u)$ as $\varepsilon \to 0$. Therefore, it is suffices to show the negligibility of $c_n^{-p} M_n^{(\varepsilon)}$, i.e. that small jumps are negligible. By Doob's martingale inequality we obtain

$$\begin{aligned} &E\left[\left(\sup_{u\in[0,1]} c_n^{-p} \sum_{t=1}^{[nu]} \big\{|X_t|^p 1\big\{|X_t| < \varepsilon c_n\big\} - E\big[|X_t|^p 1\big\{|X_t| < \varepsilon c_n\big\}\big|\mathcal{F}_{t-1}\big]\big\}\right)^2\right] \\ &\quad \le Cnc_n^{-2p} E\big[\big(|X_1|^p 1\big\{|X_1| < \varepsilon c_n\big\} - E\big[|X_1|^p 1\big\{|X_1| < \varepsilon c_n\big\}\big|\mathcal{F}_0\big]\big)^2\big] \\ &\quad \le 4Cnc_n^{-2p} E\big[\big(|X_1|^{2p} 1\big\{|X_1| < \varepsilon c_n\big\}\big)\big]. \end{aligned}$$

Recall that $\alpha < 2p$. By Karamata's theorem (Lemma 4.18),

$$E\big[|X_1|^{2p} 1\big\{|X_1| < \varepsilon c_n\big\}\big] \sim \frac{2\alpha}{2p-\alpha} (\varepsilon c_n)^{2p} \bar{F}_X(\varepsilon c_n) \sim \frac{2\alpha}{2p-\alpha} \varepsilon^{2p-\alpha} c_n^{2p} n^{-1}.$$

Applying this and letting $\varepsilon \to 0$, we conclude that $c_n^{-p} M_n^{(\varepsilon)}$ is uniformly negligible in L^2 and therefore also in probability. Thus,

$$c_n^{-p} M_n(u) \Rightarrow C_{\alpha/p}^{-p/\alpha} \big(E\big[\sigma_1^\alpha\big]\big)^{p/\alpha} Z_{\alpha/p}(u).$$

This finishes the proof of (4.98) and (4.99).

As for the sum S_n, the long-memory part R_n vanishes since $E(X_1) = E(\xi_1)E(\sigma_1) = 0$. Thus, in this case also only the stable limit arises. □

The reader is referred to Kulik and Soulier (2012) for more discussion, a detailed proof and extensions to stochastic volatility with leverage.

4.3.5 Subordinated Gaussian Processes with Infinite Variance

Previously (see Theorem 4.16 or Theorem 4.19, Eq. (4.99)) we have seen that it is possible to obtain limiting distributions with finite variance although we start with innovations with infinite second moments. In this section we illustrate that this type of behaviour can also be achieved in the context of Gaussian subordination with infinite variance. This rather peculiar result depends on specific circumstances to be explained below.

Let X_t ($t \in \mathbb{Z}$) be a stationary centred Gaussian process with covariance $\gamma_X(K) \sim L_\gamma(k)k^{2d-1}$, $d \in (0, 1/2)$. Assume that G is a function such that, as $x \to \infty$,

$$P\big(G(X_1) > x\big) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P\big(G(X_1) < -x\big) \sim A\frac{1-\beta}{2}x^{-\alpha}, \quad (4.107)$$

where $\beta \in [-1, 1]$. If $\alpha \in (0, 2)$, then $G(X_t)$ have infinite (or non-existing) variance. Furthermore, if $\alpha \in (0, 1)$, then $E(|G(X_1)|) = +\infty$. A typical example is $G(x) = |x|^{-1/\alpha}$. After the transformation $|x|^{-1/\alpha}$ the mass from zero is "sent" to infinity (since for a standard normal density, $\phi(0) \neq 0$). Another example is $G(x) = b\exp(cx^2)$ for some constants $b \in \mathbb{R}$ and $c > 0$.

In this section we shall assume that $\alpha \in (1, 2)$. Again we consider

$$S_{n,G}(u) = \sum_{t=1}^{[nu]} \big\{G(X_t) - E\big[G(X_1)\big]\big\}.$$

With a similar trick as in the proof of Theorem 4.19, i.e. the decomposition into a martingale and a long-memory part, $S_{n,G}$ will be studied using techniques available for weakly dependent processes with infinite variance (see $M_n(\cdot)$ in the proof of Theorem 4.19) and finite-variance subordinated Gaussian processes (see Sect. 4.2.3). This method was used in Sly and Heyde (2008) for $\alpha \in (1, 2)$. The result for $\alpha \in (0, 1)$ was proven in Davis (1983).

4.3.5.1 Point Process Convergence

Assume that $\alpha \in (1, 2)$, so that $\mathrm{var}(G(X_t)) < \infty$. As in case of the LMSV model, we start with the convergence of point processes

$$N_n = \sum_{t=1}^{n} \delta_{(t/n, c_n^{-1}G(X_t))},$$

where in the present context

$$c_n = \inf\{x : P(|G(X_1)| > x) \le n^{-1}\}.$$

Recall that

$$d\lambda(x) = \alpha\left[\frac{1+\beta}{2}x^{-(\alpha+1)}1\{0 < x < \infty\} + \frac{1-\beta}{2}(-x)^{-(\alpha+1)}1\{-\infty < x < 0\}\right].$$

We state the following result without proof. In principle, as in the LMSV case, it says that the random variables $G(X_t)$ behave as if they were independent.

Theorem 4.20 *Consider a Gaussian sequence X_t ($t \in \mathbb{N}$) and a real-valued function G such that (4.107) holds. Then N_n converges weakly in $M_p([0,1] \times \mathbb{R})$ to a Poisson process N with intensity measure $ds \times d\lambda(x)$.*

4.3.5.2 Hypercontraction Principle for Gaussian Random Variables

We shall explain how it is possible to obtain a finite-variance random variable from infinite-variance variables $G(X_t)$. Recall that for a function G such that $E[G^2(X_1)] < \infty$, we have the following expansion:

$$G(x) = E[G(X_1)] + \sum_{l=m}^{\infty} \frac{J(l)}{l!} H_l(x),$$

where m is the Hermite rank of G, and $J(l) = E[G(X_1)H_l(X_1)]$. This expansion is also valid for a function G with $E[|G(X_1)|^{1+\theta}] < \infty$, where $\theta \in (0,1)$. Indeed, the Hermite coefficients $J(l)$ are still well defined. Applying the Hölder inequality, we obtain with $r = (1+\theta)/\theta$,

$$|J(l)| \le E^{\frac{1}{1+\theta}}\left[|G(X_1)|^{1+\theta}\right] E^{\frac{1}{r}}\left[|H_l(X_1)|^r\right] = \|G\|_{1+\theta}\|H_l\|_r < \infty, \qquad (4.108)$$

where $\|G\|_r^r = \int G^r(u)\phi(u)\,du$. Now, let $X = a_1X_1 + \theta X_2$, where $a_1^2 + \theta^2 = 1$, and X_1, X_2 are independent standard normal random variables. Let $\mathcal{F}$ be the sigma field generated by X_2. We will argue below that although $E[G^2(X)] = +\infty$, we have

$$\operatorname{var}\left(E[G(X_1)|\mathcal{F}]\right) < \infty.$$

We start with the following result.

Lemma 4.21 *Assume that $E[|G(X_1)|^{1+\theta}] < \infty$, where $\theta \in (0,1)$. Then*

$$\sum_{l=m}^{\infty} \frac{J^2(l)}{l!}\theta^{2l} < \infty.$$

Proof From Lemma 3.1 in Taqqu (1977) we have the following bound:

$$\|H_l\|_r \leq (r-1)^{l/2}\sqrt{l!}.$$

Applying (4.108) (recall that $r = (1+\theta)/\theta$), we obtain

$$\frac{J^2(l)\theta^{2l}}{l!} \leq \frac{\theta^{2l}}{l!}\|G\|^2_{1+\theta}(r-1)^l l! = \theta^{2l}\|G\|^2_{1+\theta}\theta^{-l} = \|G\|^2_{1+\theta}\theta^l. \qquad \square$$

The consequence of this simple lemma is quite remarkable. Applying formula (3.16) and recalling that X_2 is $\mathcal{F}$-measurable and Hermite polynomials H_l ($l \geq 1$) are centred, we obtain

$$E\big[H_l(X)|\mathcal{F}\big] = E\big[H_l(a_1X_1 + \theta X_2)|\mathcal{F}\big] = \sum_{j=0}^{l}\binom{l}{j}a_1^j\theta^{l-j}E\big[H_j(X_1)H_{l-j}(X_2)|\mathcal{F}\big]$$

$$= \sum_{j=0}^{l}\binom{l}{j}a_1^j\theta^{l-j}H_{l-j}(X_2)E\big[H_j(X_1)|\mathcal{F}\big] = \theta^l H_l(X_2).$$

We recall that $E[H_l^2(X_2)] = l!$. From Lemma 4.21 we have

$$\sum_{l=m}^{\infty}\left(\frac{J(l)}{l!}\right)^2\theta^{2l}l! < \infty.$$

This expression is however equal to

$$\mathrm{var}\left(\sum_{l=m}^{\infty}\frac{J(l)}{l!}\theta^l H_l(X_2)\right) = \mathrm{var}\left(\sum_{l=m}^{\infty}\frac{J(l)}{l!}E\big[H_l(X)|\mathcal{F}\big]\right).$$

Thus, $\sum_{l=m}^{\infty}E[H_l(X)|\mathcal{F}]J(l)/l!$ is a well-defined Hermite expansion of a function

$$\tilde{g}(X_2) := E\big[G(X)|\mathcal{F}\big] = E\big[\tilde{g}(X_2)\big] + \sum_{l=m}^{\infty}\frac{J(l)}{l!}\theta^l H_l(X_2)$$

with finite variance. Note also that, since X_2 is $\mathcal{F}$-measurable,

$$E\big[\tilde{g}(X_2)H_l(X_2)\big] = E\big\{E\big[G(X)|\mathcal{F}\big]H_l(X_2)\big\} = E\big[G(X)H_l(X_2)\big].$$

4.3.5.3 Partial Sums Convergence

Theorem 4.21 *Assume that X_t ($t \in \mathbb{Z}$) is a stationary standard normal sequence with covariance $\gamma_X(K) \sim L_\gamma(k)k^{2d-1}$, $d \in (0, 1/2)$. Let G be a function with Hermite rank m such that (4.107) holds with $1 < \alpha < 2$.*

- *If* $1 < \alpha < 2$ *and* $1 - m(1/2 - d) < 1/\alpha$, *then*

$$n^{-1/\alpha} \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \stackrel{\text{f.d.}}{\to} A^{1/\alpha} C_\alpha^{-1/\alpha} Z_\alpha(u), \tag{4.109}$$

  where $Z_\alpha(\cdot)$ *is an* α*-stable Lévy process such that* $Z_\alpha(1) \stackrel{\text{d}}{=} S_\alpha(1, \beta, 0)$.
- *If* m *is the Hermite rank of* G *and* $1 - m(\frac{1}{2} - d) > 1/\alpha$, *then*

$$n^{-(1-m(\frac{1}{2}-d))} L_m^{-1/2}(n) \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \Rightarrow Z_{m,H}(u) \quad \big(u \in [0,1]\big),$$

  where $H = d + \frac{1}{2}$, $L_m(n) = m! C_m L_\gamma^m(n)$, $Z_{m,H}(u)$ *is the Hermite–Rosenblatt process, and* $\Rightarrow$ *denotes weak convergence in* $D[0,1]$.

Proof We present just a short heuristic derivation. The Gaussian sequence can be written as a linear process $X_t = \sum_{j=0}^\infty a_j \varepsilon_{t-j}$, where ε_t $(t \in \mathbb{Z})$ are i.i.d. standard normal, and $\sum_{j=0}^\infty a_j^2 = 1$. Let $\mathcal{F}_t = \sigma(\varepsilon_t, \varepsilon_{t-1}, \ldots)$. Then

$$\begin{aligned}
&\sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_1)\big] \big\} \\
&\quad = \sum_{t=1}^{[nu]} \big\{ G(X_t) - E\big[G(X_t)|\mathcal{F}_{t-l}\big] \big\} + \sum_{t=1}^{[nu]} \big\{ E\big[G(X_t)|\mathcal{F}_{t-l}\big] - E\big[G(X_1)\big] \big\} \\
&\quad =: M_n(u) + R_n(u),
\end{aligned}$$

where l is such that $\theta := \sqrt{\sum_{j=l}^\infty a_j^2} < \alpha - 1$. The first part $M_n(\cdot)$ is a martingale. Therefore, its limiting properties are studied in the very same way as $M_n(\cdot)$ in the proof of Theorem 4.19. As for the second part, write

$$X_t := \sum_{j=0}^{l-1} a_j \varepsilon_{t-j} + \theta \tilde{X}_{t,l},$$

where $\tilde{X}_{t,l} := \theta^{-1} \sum_{j=l}^\infty a_j \varepsilon_{t-j}$. The random variables $\tilde{X}_{t,l}$ $(t \in \mathbb{N})$ are standard normal. Applying Lemma 4.21, the function

$$g(\tilde{X}_{t,l}) := E\big[G(X_t)|\mathcal{F}_{t-l}\big] - E\big[G(X_1)\big]$$

has finite variance. Therefore, the convergence of the second part $R_n(u)$ follows from Theorem 4.4. □

4.3.6 Quadratic LARCH Models

We recall (cf. (2.58)) that the quadratic LARCH(∞) (or LARCH$_+$) process is the unique solution of

$$X_t = b_0\eta_t + \xi_t \sum_{j=1}^{\infty} b_j X_{t-j}, \tag{4.110}$$

where (η_t, ξ_t) $(t \in \mathbb{Z})$ is a sequence of i.i.d. random vectors. We assume that $b_j \sim c_b j^{d-1}$ $(d \in (0, 1/2))$ and that the random variables η_t are heavy tailed in the sense that

$$P\big(|\eta_1| > x\big) \sim Ax^{-\alpha}$$

for some $\alpha \in (2, 4)$. In other words, $E(\eta_1^2) < \infty$, but $E(\eta_1^4) = \infty$. Furthermore, we assume that $E(\xi_1^4 + \xi_1^2\eta_1^2) < \infty$. Surgailis (2008) considers convergence of the sum of the squares and proves that under appropriate technical assumptions we have a dichotomous behaviour as in case of the stochastic volatility model (cf. Theorem 4.19) or the subordinated Gaussian sequence with heavy tails (cf. Theorem 4.21): if $d + \frac{1}{2} < 2/\alpha$, then

$$n^{-2/\alpha} \sum_{t=1}^{[nu]} \big(X_t^2 - E\big(X_1^2\big)\big)$$

converges in a finite-dimensional sense to a Lévy process. Otherwise, if $d + \frac{1}{2} > 2/\alpha$, then

$$n^{-(d+\frac{1}{2})} \sum_{t=1}^{[nu]} \big(X_t^2 - E\big(X_1^2\big)\big)$$

converges to a fractional Brownian motion.

Also, if $\alpha \in (1, 2)$, then $n^{-1/\alpha} \sum_{t=1}^{n} X_t$ converges to a stable limit. As in the case of LMSV processes (see Sect. 4.3.4), this can be concluded from a general theory by Surgailis (2008).

4.3.7 Summary of Limit Theorems for Partial Sums

We summarize the main limit theorems. We consider centred linear process $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ such that, as $x \to \infty$,

$$P(\varepsilon_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha}$$

Table 4.2 Limits for partial sums with infinite moments

	Partial sums—infinite moments	
	$S_n(u) = \sum_{t=1}^{[nu]} X_t$	$T_n(u) = \sum_{t=1}^{[nu]} (X_t^2 - E(X_1^2))$
Linear processes	$n^{-1/\alpha} S_n(1) \xrightarrow{\mathrm{d}} c\tilde{Z}_\alpha(1)$ if $\sum \lvert a_j \rvert < \infty$ $n^{-(d+1/\alpha)} S_n(u) \Rightarrow c\tilde{Z}_{H,\alpha}(u)$ if $0 < d < 1 - 1/\alpha$ (Theorem 4.15)	$n^{-2/\alpha} T_n(1) \xrightarrow{\mathrm{d}} c\tilde{Z}_{\alpha/2}(1)$ if $d \in (0, 1/\alpha)$ $n^{-2d} T_n(u) \Rightarrow cZ_{2,H}(u)$ if $d \in (1/\alpha, 1/2)$ (Theorem 4.16)
Stochastic volatility	$n^{-1/\alpha} S_n(u) \Rightarrow c\tilde{Z}_\alpha(u)$ (Theorem 4.19)	$n^{-2/\alpha} T_n(u) \Rightarrow c\tilde{Z}_{\alpha/2}(u)$ if $d \in (0, 2/\alpha - 1/2)$ $n^{-(1/2+d)} T_n(u) \Rightarrow cB_H(u)$ if $d \in (2/\alpha - 1/2, 1/2)$ (Theorem 4.19)

with $\alpha \in (1, 2)$ and appropriate regularity conditions (that assure the existence of the process) hold. When the sum of the squares X_t^2 is considered, then we assume instead that α is in the range $\alpha \in (2, 4)$.

Another class of processes considered above are stochastic volatility models with infinite second moments. As a representative, we look at $X_t = \xi_t \exp(\sum_{j=1}^{\infty} a_j \varepsilon_{t-j})$, where the sequences ξ_t and ε_t are mutually independent. We assume that

$$P(\xi_1 > x) \sim A \frac{1+\beta}{2} x^{-\alpha}, \qquad P(\xi_1 < -x) \sim A \frac{1-\beta}{2} x^{-\alpha}$$

with $\alpha \in (1, 2)$ and $E[\xi_1] = 0$. Again, if the sum of X_t^2 is considered, then this tail behaviour is assumed to hold for $\alpha \in (2, 4)$. Furthermore, the random variables ε_t are assumed to be standard normal. We use the notation $B(\cdot)$ for a Brownian motion on $[0, 1]$, $B_H(\cdot)$ denotes a fractional Brownian motion on $[0, 1]$, $Z_{2,H}(\cdot)$ is the Hermite–Rosenblatt process on $[0, 1]$, and $\tilde{Z}_{H,\alpha}$ is a linear fractional stable motion with Hurst parameter $H = d + 1/\alpha$. Furthermore, c is a generic constant. We summarize the results for partial sums in Table 4.2. For simplicity, the slowly varying functions are assumed to be constant.

4.4 Limit Theorems for Sample Covariances

In a preliminary analysis of a time series, sample autocovariances play a crucial role. Moreover, limit theorems for quadratic forms can often be deduced from those for sample covariances. In this section we therefore study the limiting behaviour of sample covariances and, more generally, of multivariate functions applied to long-memory sequences. Surprisingly, this theory is not well developed beyond Gaussian (Rosenblatt 1979; Ho and Sun 1987, 1990; Arcones 1994) and linear processes with finite (Hosking 1996; Horváth and Kokoszka 2008) and infinite moments (Kokoszka

and Taqqu 1996; Horváth and Kokoszka 2008). Some recent results were developed for stochastic volatility models (Davis and Mikosch 2001; McElroy and Politis 2007; Kulik and Soulier 2012).

4.4.1 Gaussian Sequences

In what follows, all vectors are considered as column vectors. Consider a stationary centred sequence of Gaussian vectors

$$\mathbf{X}_t = \left(X_t^{(1)}, \ldots, X_t^{(q)}\right)^T \quad (t \in \mathbb{Z})$$

with the marginal covariance matrix Σ and autocovariance function $\gamma_{i,j}(k) = E[X_0^{(i)} X_k^{(j)}]$ $(i, j = 1, \ldots, q)$, and assume either

$$\sum_{k=-\infty}^{\infty} \left|\gamma_{i,j}(k)\right| < \infty \tag{4.111}$$

or the existence of a parameter $d \in (0, 1/2)$ and a slowly varying function L_γ such that

$$\gamma_{i,j}(k) \sim a_{i,j} k^{2d-1} L_\gamma(k) \quad (i, j = 1, 2, \ldots, q), \tag{4.112}$$

where the constants $a_{i,j}$ are not all equal to zero. We will then use the same notation $\gamma(k) = k^{2d-1} L_\gamma(k)$ as in the univariate case.

Example 4.18 Let $q = 2$ and assume that $\tilde{X}_t^{(1)}$ $(t \in \mathbb{N})$ and $\tilde{X}_t^{(2)}$ $(t \in \mathbb{N})$ are mutually independent long-memory standard Gaussian sequences with the same covariances $\gamma_X(k) = \gamma_{\tilde{X}}(k) = \gamma(k)$. Then (4.112) holds with $a_{1,1} = a_{2,2} = 1$ and $a_{1,2} = a_{2,1} = 0$.

Example 4.19 Let X_t $(t \in \mathbb{N})$ be a stationary standard Gaussian sequence with covariance $\gamma_X(k) = c_\gamma k^{2d-1}$. Fix $s > 0$, and let

$$\left(X_t^{(1)}, X_t^{(2)}\right)^T = (X_t, X_{t+s})^T \quad (t \in \mathbb{N}).$$

Then

$$\gamma_{1,1}(k) = \gamma_{2,2}(k) = E[X_0 X_k] = \gamma_X(k),$$

so that $a_{1,1} = a_{2,2} = 1$. Furthermore,

$$\gamma_{1,2}(k) = E[X_0 X_{s+k}] = \gamma_X(k+s) \sim \gamma_X(k)$$

as $k \to \infty$, so that $a_{1,2} = 1$. Similarly, $a_{2,1} = 1$.

Example 4.20 Assume that $\tilde{X}_t^{(1)}$ and $\tilde{X}_t^{(2)}$ ($t \in \mathbb{N}$) are as in Example 4.18. Fix $s > 0$, and let

$$\left(X_t^{(1)}, X_t^{(2)}\right)^T = \left(\tilde{X}_t^{(1)}, \rho \tilde{X}_t^{(2)} + \sqrt{1-\rho^2}\tilde{X}_t^{(2)}\right)^T,$$

where $\rho = \gamma_X(s)$. Note that for a fixed t, the vectors $(X_t^{(1)}, X_t^{(2)})^T$ in Example 4.19 and here have the same covariance matrix. Now, $a_{1,1} = a_{2,2} = 1$, whereas

$$\gamma_{1,2}(k) = \rho \gamma_X(k),$$

so that $a_{1,2} = \rho$. Similarly, $a_{2,1} = \rho$.

After explaining basic structures of dependent Gaussian vectors, we turn our attention to limit theorems. It turns out that limit theorems for multivariate Gaussian vectors can be reduced to the case where the vectors have the identity covariance matrix I_q. Therefore, we start with the case of independent components.

4.4.1.1 Independent Components

Consider the collection $\{\tilde{X}_t^{(l)}, l \in \mathbb{N}, t \in \mathbb{N}\}$ of long-memory Gaussian sequences. For any $l \neq k$, the sequences $X_t^{(l)}$ and $X_t^{(k)}$ ($t \in$) are assumed to be independent. Recall the following notation from Sect. 4.2.3 (see also Sect. 4.1.3) the following notation. Assume for a moment that $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ is the Gaussian process, where ε_t ($t \in \mathbb{Z}$) are i.i.d. standard normal random variables. Consider the following random measures: $M_\varepsilon(\cdot)$ is a Gaussian random measure with independent increments, associated with the sequence ε_t, that is $E[|dM_\varepsilon(\lambda)|^2] = \sigma_\varepsilon^2/(2\pi)\, d\lambda$, $dM_0(\lambda) = \sqrt{2\pi}\, dM_\varepsilon(\lambda)$,

$$dM_X(\lambda) = \left(\sum_{j=0}^{\infty} a_j e^{-ij\lambda}\right) dM_\varepsilon(\lambda) = A\left(e^{-i\lambda}\right) dM_\varepsilon(\lambda) = a(\lambda)\, dM_0(\lambda)$$

is the spectral random measure associated with a sequence X_t ($t \in \mathbb{N}$). Recall further that $n^{1/2} M_0(n^{-1}A)$ is another Gaussian random measure with the same distribution as $M_0(A)$. Then

$$\frac{L_f^{1/2}((n\lambda)^{-1})}{L_f^{1/2}(n^{-1})} |\lambda|^{-d} n^{1/2}\, dM_0\left(n^{-1}\lambda\right)$$

converges vaguely to $W_X(d\lambda) := |\lambda|^{-d}\, dM_0(\lambda)$.

As in Sect. 4.2.3, we can represent the Gaussian sequences $\tilde{X}_t^{(l)}$ ($t \in \mathbb{N}$) as (cf. (4.28))

$$\tilde{X}_t^{(l)} = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_{\tilde{X}^{(l)}}(\lambda) \quad (t \geq 1),$$

where

$$dM_{\tilde{X}^{(l)}}(\lambda) = a^{(l)}(\lambda)\, dM_0^{(l)}(\lambda),$$

and $M_0^{(l)}(\cdot)$ $(l \geq 1)$ are independent Gaussian random measures. Furthermore, $|a^{(l)}(\lambda)|^2 = f_{\tilde{X}^{(l)}}(\lambda)$, where $f_{(l)} = f_{\tilde{X}^{(l)}}$ is the spectral density associated with the sequence $\tilde{X}_t^{(l)}$ $(t \in \mathbb{N})$. Also, $n^{1/2} M_0^{(l)}(n^{-1}A) \overset{\mathrm{d}}{=} M_0(A)$, and

$$\frac{L_{f_{(l)}}^{1/2}((n\lambda)^{-1})}{L_{f_{(l)}}(n^{-1})} |\lambda|^{-d} n^{1/2}\, dM_0^{(l)}\big(n^{-1}\lambda\big) \tag{4.113}$$

converges vaguely to a measure $dW_{\tilde{X}^{(l)}}(\lambda) = |\lambda|^{-d}\, dM_0^{(l)}(\lambda)$.

As in the alternate proof of Theorem 4.2 (see also the proof of Theorem 4.3), we may write

$$\begin{aligned}
\sum_{t=0}^{n-1} \tilde{X}_t^{(1)} \tilde{X}_t^{(2)} &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{e^{in(\lambda_1+\lambda_2)} - 1}{e^{i(\lambda_1+\lambda_2)} - 1} a^{(1)}(\lambda_1) a^{(2)}(\lambda_2)\, dM_0^{(1)}(\lambda_1)\, dM_0^{(2)}(\lambda_2) \\
&= \int_{-n\pi}^{n\pi}\int_{-n\pi}^{n\pi} D_n\big((\lambda_1+\lambda_2)/n\big) \\
&\quad \times \prod_{l=1}^{2} a^{(l)}\left(\frac{\lambda_l}{n}\right) n^{1/2}\, dM_0^{(1)}\big(n^{-1}\lambda_1\big) n^{1/2}\, dM_0^{(2)}\big(n^{-1}\lambda_2\big)
\end{aligned}$$

with

$$D_n(\lambda) = \frac{e^{i\lambda n} - 1}{n(e^{i\lambda} - 1)} 1\big\{|\lambda| < \pi n\big\}.$$

The functions above converge to

$$D(\lambda) = \frac{e^{i\lambda} - 1}{i\lambda}.$$

Thus, if

$$a^{(l)}(\lambda) = a_{l,l} L_f^{1/2}\big(\lambda^{-1}\big) |\lambda|^{-d} \quad (l = 1, 2),$$

then we may conclude that for $d \in (1/4, 1/2)$,

$$\begin{aligned}
&n^{-2d} L_f^{-1}\big(n^{-1}\big) \sum_{t=0}^{n-1} \tilde{X}_t^{(1)} \tilde{X}_t^{(2)} \\
&\quad \xrightarrow{\mathrm{d}} a_{1,1} a_{2,2} \int_{\mathbb{R}^2} D(\lambda_1+\lambda_2) \prod_{l=1}^{2} \frac{1}{|\lambda_l|^d}\, dM_0^{(1)}(\lambda_1)\, dM_0^{(2)}(\lambda_2).
\end{aligned}$$

This convergence can be extended to nonlinear functionals. The following theorem is adapted from Arcones (1994). For simplicity, we assume that all $a_{l,l}$ in (4.112) are one. (Recall from Example 4.18 that the terms $a_{i,l}$, $i \neq l$, vanish.)

Theorem 4.22 *Let $\tilde{X}_t = (\tilde{X}_t^{(1)}, \dots, \tilde{X}_t^{(q)})^T$ $(t \in \mathbb{N})$, be a stationary sequence of centred Gaussian vectors with the marginal covariance matrix I_q, such that (4.112) holds. Let $G : \mathbb{R}^q \to \mathbb{R}$ be a function with the Hermite rank $m = \tilde{m}(G)$. If $m(1 - 2d) > 1$, then*

$$n^{-(1-m(1/2-d))} L_f^{m/2}(n^{-1}) \sum_{t=1}^{n} \{G(\tilde{X}_t) - E[G(\tilde{X}_1)]\}$$

$$\xrightarrow{\mathrm{d}} \sum_{r_1,\dots,r_m=1}^{q} \tilde{c}_{r_1,\dots,r_m} \tilde{Z}_{(r_1,\dots,r_m),H}(1),$$

where

$$\tilde{Z}_{(r_1,\dots,r_m),H}(1) = \int_{\mathbb{R}^m} D(\lambda_1 + \cdots + \lambda_m) \prod_{l=1}^{m} \frac{1}{|\lambda_l|^{r_l}} \, dM_0^{(r_1)}(\lambda_1) \cdots dM_0^{(r_m)}(\lambda_m),$$

$\int_{\mathbb{R}^m}$ is the m-fold multiple Wiener–Ito integral, and

$$\tilde{c}_{r_1,\dots,r_m} = \frac{1}{m!} E\left[G(\tilde{X}_1) \prod_{l=1}^{q} H_{k(r_1,\dots,r_m)}\big(\tilde{X}_1^{(l)}\big) \right],$$

where $k(r_1, \dots, r_m)$ is the number of components among $r_1, \dots, r_m$ that are equal to l.

Again, as in (4.33), the limiting random variable $\tilde{Z}_{(r_1,\dots,r_m),H}(1)$ can be expressed as

$$\int_{\mathbb{R}^m} \frac{e^{iu(\lambda_1+\cdots+\lambda_m)} - 1}{i(\lambda_1 + \cdots + \lambda_m)} \, dW_{\tilde{X}^{(r_1)}}(\lambda_1) \cdots dW_{\tilde{X}^{(r_m)}}(\lambda_m), \tag{4.114}$$

where $dW_{\tilde{X}^{(r)}}(\lambda) = |\lambda|^{-d} \, dM_0^{(r)}(\lambda)$.

Example 4.21 Consider $G(y_1, y_2) = H_2(y_2)H_2(y_2)$. Then (see Example 3.8) its Hermite rank with respect to a vector $\tilde{X}_1 = (\tilde{X}_1^{(1)}, \tilde{X}_1^{(2)})^T$ of independent standard normal random variables is $m(G) = 4$. Then

$$c_{1,1,2,2} = \frac{1}{4!} E\big[G(\tilde{X}_1) H_2\big(\tilde{X}_1^{(1)}\big) H_2\big(\tilde{X}_1^{(2)}\big)\big] = \frac{1}{4!} \tilde{J}\big(G, (2,2)\big) = \frac{4}{4!}.$$

Also, this computation is invariant under permutation of indices $(1,1,2,2)$. All other coefficients c_{r_1,r_2,r_3,r_4} vanish. Note that $k(1,1,2,2)=2$ for $l=1,2$. Thus,

$$n^{-(1-4(1/2-d))}L_f^{4/2}\big(n^{-1}\big)\sum_{t=1}^{n}H_2\big(\tilde{X}_t^{(1)}\big)H_2\big(\tilde{X}_t^{(2)}\big)$$

converges in distribution to

$$\frac{6\times 4}{4!}\int_{\mathbb{R}^4}\frac{e^{iu(\lambda_1+\cdots+\lambda_4)}-1}{i(\lambda_1+\cdots+\lambda_4)}\,dW_{\tilde{X}^{(1)}}(\lambda_1)\,dW_{\tilde{X}^{(1)}}(\lambda_2)\,dW_{\tilde{X}^{(2)}}(\lambda_3)\,dW_{\tilde{X}^{(2)}}(\lambda_4).$$

This can be also seen by expanding

$$\sum_{t=1}^{n}H_2\big(\tilde{X}_t^{(1)}\big)H_2\big(\tilde{X}_t^{(2)}\big)$$

and using a representation for $H_m(X_t)$, see the proof of Theorem 4.3. The convergence is valid for $d\in(1/4,1/2)$.

Example 4.22 Let $G(y)=H_m(y)$. Then one can see that $Z_{m,H}(1)$ in Theorem 4.22 is exactly the Hermite–Rosenblatt random variable.

4.4.1.2 From Independent to Dependent Components

In general, let $X_t=(X_t^{(1)},\ldots,X_t^{(q)})^T$ $(t\in\mathbb{N})$ be a long-memory Gaussian sequence with cross-autocovariance function $\gamma_{i,j}(k)=E(X_0^{(i)}X_k^{(j)})$ as in (4.112) and marginal covariance matrix Σ. Then the statement of Theorem 4.22 remains valid if we replace $m=\tilde{m}(G)$ by $m=m(G,X_1)$, where $m(G,X_1)$ is the Hermite rank of G with respect to the Gaussian vector X_1; the spectral measures $W_{\tilde{X}^{(r_l)}}$ are replaced by the so-called joint spectral measure

$$\big(dW_{X^{(1)}}(\lambda_1),\ldots,dW_{X^{(q)}}(\lambda_q)\big),$$

and

$$c_{r_1,\ldots,r_m}=\frac{1}{m!}E\left[G(X_1)\prod_{l=1}^{q}H_{k(r_1,\ldots,r_m)}\big(X_1^{(l)}\big)\right].$$

We do not provide details here; the reader is referred to Arcones (1994). However, we will consider the special case of the covariance matrix Σ since this leads to study of sample covariances.

Example 4.23 Recall Example 3.13. We consider the function

$$G(X_t,X_{t+s})=e^{pX_t}e^{pX_{t+s}}.$$

Then the Hermite rank is one. Thus, we have to evaluate c_{r_1}, $r_1 = 1, 2$. We compute

$$c_1 = E\big[G(X_t, X_{t+s})X_t\big] = p\big(1 + \gamma_X(s)\big)e^{p^2(1+\gamma_X(s))}.$$

Also, $c_2 = E[G(X_t, X_{t+s})X_{t+s}] = c_1$. Thus,

$$n^{-(d+1/2)}L_f^{-1/2}\big(n^{-1}\big)\sum_{t=1}^{n} e^{pX_t}e^{pX_{t+s}} \xrightarrow{d} 2c_1\int D(\lambda)\,dW_X(\lambda),$$

where W_X is the spectral random measure associated with X_t ($t \in \mathbb{N}$), see (4.34).

4.4.1.3 From Independent to Dependent Components: Sample Covariances

We go back to the original problem of sample covariances. Our vectors $X_t = (X_1^{(1)}, X_t^{(2)})^T$ are as in Example 4.19:

$$\big(X_t^{(1)}, X_t^{(2)}\big)^T = (X_t, X_{t+s})^T \quad (t \in \mathbb{N}).$$

We write

$$X_t = \int_{-\pi}^{\pi} e^{ij\lambda}a(\lambda)\,dM_0(\lambda) = \int_{-\pi}^{\pi} e^{ij\lambda}\,dM_X(\lambda),$$
$$X_{t+s} = \int_{-\pi}^{\pi} e^{it\lambda}e^{is\lambda}a(\lambda)\,dM_0(\lambda) = \int_{-\pi}^{\pi} e^{it\lambda}e^{is\lambda}\,dM_X(\lambda).$$

Recall now the proof of Theorems 4.2 and 4.3. Like in the proof of Theorem 4.3

$$\begin{aligned}
&\sum_{t=0}^{n-1}\big(X_tX_{t+s} - E(X_tX_{t+s})\big)\\
&= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{e^{in(\lambda_1+\lambda_2)}-1}{e^{i(\lambda_1+\lambda_2)}-1}\prod_{r=1}^{2}a(\lambda_r)e^{is\lambda_2}\,dM_0(\lambda_1)\,dM_0(\lambda_2)\\
&= \int_{-n\pi}^{n\pi}\int_{-n\pi}^{n\pi} D_n\big((\lambda_1+\lambda_2)/n\big)e^{is\lambda_2/n}\\
&\quad\times\prod_{r=1}^{2}a\left(\frac{\lambda_r}{n}\right)n^{1/2}\,dM_0\big(n^{-1}\lambda_1\big)n^{1/2}\,dM_0\big(n^{-1}\lambda_2\big). \qquad (4.115)
\end{aligned}$$

Note that, as $n \to \infty$, $e^{is\lambda_2/n} \to 1$. Therefore, omitting technical details, the limiting behaviour of

$$n^{-2d}L_f^{-1}\big(n^{-1}\big)\sum_{t=0}^{n-1}\big(X_tX_{t+s} - E(X_tX_{t+s})\big)$$

or, equivalently, of

$$n^{-2d}L_2^{-1/2}(n^{-1})\sum_{t=0}^{n-1}\big(X_tX_{t+s}-E(X_tX_{t+s})\big)$$

is the same as that of $n^{-2d}L_2^{-1/2}(n^{-1})\sum_{t=0}^{n-1}(X_t^2-E(X_1^2))$, i.e. it does not involve s. Hence, using Theorem 4.3 with $m=2$, one can argue that for $d\in(1/4,1/2)$,

$$\begin{aligned}&n^{1-2d}L_2^{-1/2}(n^{-1})\big(\hat\gamma_n(1)-\gamma_X(1),\dots,\hat\gamma_n(K)-\gamma_X(K)\big)\\&\quad\xrightarrow{\mathrm{d}}\big(Z_{2,H}(1),\dots,Z_{2,H}(1)\big),\end{aligned}\tag{4.116}$$

where

$$\hat\gamma_n(s)=\frac{1}{n}\sum_{t=0}^{n-s}X_tX_{t+s}\quad(s=1,\dots,K)$$

is the sample covariance at lag s and $H=d+1/2$. Thus, the limiting random vector has totally dependent components.

We extend this to arbitrary Hermite polynomials. Recall Example 3.15. One can derive the equation (see Lemma 3.4 in Fox and Taqqu 1985)

$$H_m(X_t)H_m(X_{t+s})=m!\gamma_X^m(s)+\sum_{r=1}^{m}(m-r)!\binom{m}{r}^2\gamma_X^{m-r}(s)K_r(t,t+s),\tag{4.117}$$

where

$$K_r(j,l)=\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}e^{ij(\lambda_1+\cdots+\lambda_r)+il(\lambda_{r+1}+\cdots+\lambda_{2r})}\prod_{l=1}^{2r}a(\lambda_l)\,dM_0(\lambda_1)\cdots dM_0(\lambda_{2r}).$$

For $m=1$, the formula reduces to the formula for X_tX_{t+s}, used in deriving (4.115). For $m=2$, the formula yields

$$\begin{aligned}&2\gamma_X^2(s)+4\gamma_X(s)\int\int e^{ij\lambda_1+is\lambda_2}\prod_{r=1}^{2}a(\lambda_r)\,dM_0(\lambda_1)\,dM_0(\lambda_2)\\&\quad+\int\cdots\int e^{ij(\lambda_1+\lambda_2)+i(j+s)(\lambda_3+\lambda_4)}\prod_{r=1}^{4}a(\lambda_r)\,dM_0(\lambda_1)\cdots dM_0(\lambda_4).\end{aligned}$$

The important feature of decomposition (4.117) is that under the condition $d\in(1/4,1/2)$ only the term with $r=1$ will contribute. In other words, the limiting behaviour of

$$\hat\gamma_n(s;H_m):=\frac{1}{n}\sum_{t=1}^{n-s}H_m(X_t)H_m(X_{t+s})$$

is up to a constant the same for each $m \geq 1$. Noting that $(m-1)!\binom{m}{1}^2 = m!m$ and using (4.117), we have for $d \in (1/4, 1/2)$,

$$n^{1-2d} L_2^{-1}(n^{-1})\big(\hat{\gamma}_n(1; H_m) - m!\gamma_X^m(1), \ldots, \hat{\gamma}_n(K; H_m) - m!\gamma_X^m(K)\big)$$
$$\xrightarrow{d} m!m\big(\gamma_X^{m-1}(1), \ldots, \gamma_X^{m-1}(K)\big) Z_{2,H}(1), \tag{4.118}$$

where $H = d + 1/2$.

4.4.2 Linear Processes with Finite Moments

In this section we consider second-order stationary linear processes $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ $(t \in \mathbb{N})$, where ε_t $(t \in \mathbb{Z})$ are i.i.d. random variables such that $E(\varepsilon_1) = 0$, $E(\varepsilon_1^2) = \sigma_\varepsilon^2 = 1$ and $E(\varepsilon_1^4) = \eta < \infty$.

Let

$$\hat{\gamma}_n(s) = \frac{1}{n} \sum_{t=0}^{n-s} X_t X_{t+s}.$$

It converges in probability to the population covariance

$$\gamma_X(s) = E(X_0 X_s) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} a_j a_{j+s}.$$

Classical results for weakly dependent sequences under $E(\varepsilon_1^4) < \infty$ were obtained in Anderson (1971, p. 478); see also Brockwell and Davis (1991, Proposition 7.3.3). For long-memory linear processes, they were obtained in Hosking (1996) and Horváth and Kokoszka (2008).

Theorem 4.23 *Let $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ $(t \in \mathbb{N})$ be a linear process such that $E(\varepsilon_1) = 0$, $E(\varepsilon_1^2) = \sigma_\varepsilon^2 = 1$ and $E(\varepsilon_1^4) = \eta < \infty$. Furthermore, assume that $\sum_{j=0}^{\infty} a_j^2 = 1$.*

(a) *If $a_j \sim L_a(j) j^{d-1}$, $d \in (0, 1/4)$ or $\sum_{j=0}^{\infty} |a_j| < \infty$, then*

$$n^{1/2}\big(\hat{\gamma}_n(s) - \gamma_X(s)\big) \xrightarrow{d} N\big(0, v^2\big),$$

where the variance is

$$v^2 = (\eta - 3)\gamma_X^2(s) + \sum_{k=-\infty}^{\infty} \big(\gamma_X^2(k) + \gamma_X^2(k+s)\big).$$

(b) *If* $a_j \sim L_a(j)j^{d-1}$ *and* $d \in (1/4, 1/2)$, *then*

$$n^{1-2d}L_2^{-1/2}(n)\big(\hat{\gamma}_n(s) - \gamma_X(s)\big) \xrightarrow{d} Z_{2,H}(1),$$

where $Z_{2,H}(u)$ *is a Hermite–Rosenblatt process,* $L_2(n) = 2C_2L_\gamma^2(n)$,

$$C_2 = \big[\big(2(2d-1)+1\big)(2d+1)\big]^{-1},$$

and $L_\gamma(n)$ *is given in* (4.39).

This theorem can be formulated in a multivariate setup. In the first case the limiting distribution is multivariate normal (with dependent components):

$$n^{1/2}\big(\hat{\gamma}_n(0) - \gamma_X(0), \ldots, \hat{\gamma}_n(q) - \gamma_X(q)\big) \xrightarrow{d} (G_0, \ldots, G_q), \tag{4.119}$$

where $(G_0, \ldots, G_q)$ is a zero-mean Gaussian vector with covariance

$$E[G_sG_t] = (\eta - 3)\gamma_X(s)\gamma_X(t) + \sum_{k=-\infty}^{\infty}\big(\gamma_X(k)\gamma_X(k+s-t) + \gamma_X(k+s)\gamma_X(k+t)\big). \tag{4.120}$$

In the second case, $d \in (1/4, 1/2)$, the limit has the form $(Z_{2,H}(1), \ldots, Z_{2,H}(1))$.

Proof For part (a), we use the standard truncation argument as illustrated in the proof of Theorem 4.5. Let

$$X_{t,K} = \sum_{j=0}^{K} a_j\varepsilon_{t-j},$$

$$\hat{\gamma}_n^{(K)}(s) = \frac{1}{n}\sum_{t=0}^{n-s} X_{t,K}X_{t+s,K}, \qquad \gamma_X^{(K)}(s) = E[X_{0,K}X_{s,K}] = \sigma_\varepsilon^2\sum_{j=0}^{K} a_ja_{j+s}.$$

First, since the sequence $X_{t,K}X_{t+s,K}$ is $(K+s)$-dependent, its convergence is described by

$$n^{1/2}\big(\hat{\gamma}_n^{(K)}(s) - \gamma_X^{(K)}(s)\big) \xrightarrow{d} N\big(0, \nu_K^2\big),$$

where

$$\nu_K^2 = (\eta - 3)\big(\gamma_X^{(K)}(s)\big)^2 + \sum_{k=-\infty}^{\infty}\big[\big(\gamma_X^{(K)}(k)\big)^2 + \big(\gamma_X^{(K)}(k+s)\big)^2\big].$$

Since $\nu_K^2 \to \nu^2$ as $K \to \infty$, we also have $N(0, \nu_K^2) \xrightarrow{d} N(0, \nu^2)$. It suffices to verify that for all $\delta > 0$,

$$\lim_{K\to\infty}\limsup_{n\to\infty} P\big(\big|n^{1/2}\big(\hat{\gamma}_n^{(K)}(s) - \gamma_X^{(K)}(s)\big) - n^{1/2}\big(\hat{\gamma}_n(s) - \gamma_X(s)\big)\big| > \delta\big) = 0.$$

By Markov's inequality, to do this, it suffices to verify that

$$\lim_{K\to\infty}\lim_{n\to\infty} n\cdot \operatorname{var}\bigl(\hat\gamma_n^{(K)}(s)-\hat\gamma_n(s)\bigr)=0.$$

In the case of Theorem 4.5 this was handled by introducing the random variable $\bar X_{t,K}=X_t-X_{t,K}$. In our situation here this is not straightforward since

$$\sum_{j,j'=0}^{\infty} a_j a_{j+s}-\sum_{j,j'=0}^{K} a_j a_{j+s}\neq \sum_{j,j'=K+1}^{\infty} a_j a_{j+s}.$$

We have to verify that

$$\lim_{n\to\infty} n\cdot\operatorname{var}\bigl(\hat\gamma_n(s)\bigr)=\nu^2,$$

$$\lim_{K\to\infty}\lim_{n\to\infty} n\cdot\operatorname{var}\bigl[\hat\gamma_n^{(K)}(s)\bigr]=\nu^2,\qquad \lim_{K\to\infty}\lim_{n\to\infty} n\cdot cov\bigl(\hat\gamma_n^{(K)}(s),\hat\gamma_n(s)\bigr)=\nu^2.$$

We prove the first part only. The expression is

$$\sum_{k=-(n-1)}^{n-1}\left(1-\frac{|k|}{n}\right)\left[(\eta-3)\sigma_\varepsilon^2\sum_{j=0} a_j a_{j+s}a_{j+k}a_{j+k+s}+\gamma_X^2(k)+\gamma_X^2(k+s)\right].$$

Then the relation follows by the dominated convergence theorem. For this, one needs, in particular, $\sum_k \gamma_X^2(k)<\infty$, which is achieved if $d\in(0,1/4)$ or $\sum_{j=0}^{\infty}|a_j|<\infty$.

As for part (b), we use the following decomposition:

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}\bigl(X_tX_{t+s}-E(X_tX_{t+s})\bigr)\\&\quad=\frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{\infty}a_ja_{j+s}\bigl(\varepsilon_{t-j}^2-\sigma_\varepsilon^2\bigr)+\frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{\infty}\sum_{l=0;\ l\neq j+s}^{\infty}a_ja_l\varepsilon_{t-j}\varepsilon_{t-l}\\&\quad=:M_n+R_n.\end{aligned}$$

We may write the first part as $M_n=n^{-1}\sum_{t=1}^{n}Y_t$, where Y_t $(t\in\mathbb{N})$ is the linear process $Y_t=\sum_{j=0}^{\infty}c_j(\varepsilon_{t-j}-\sigma_\varepsilon^2)$ with summable coefficients $c_j=a_ja_{j+s}$. Indeed, by the Cauchy–Schwarz inequality,

$$\sum|c_j|\le\Bigl(\sum a_j^2\Bigr)^{1/2}\Bigl(\sum a_{j+s}^2\Bigr)^{1/2}<\infty.$$

Thus, $n^{1/2}M_n$ converges to a normal distribution on account of Theorem 4.5.

As for the second part, we may recognize that it has almost the same form as the therm $U_{n,2}$ in (4.51), so that its limiting distribution is of Hermite–Rosenblatt type.

If $d \in (1/4, 1/2)$, then

$$n^{1-2d} L_2^{-1/2}(n) R_n \xrightarrow{d} Z_{2,H}(1).$$

Thus, the second part dominates if $d \in (1/4, 1/2)$.

Note that formally the limit in part (b) may depend on s. However, this is not the case; a precise computation is given in Horváth and Kokoszka (2008). □

4.4.3 Linear Processes with Infinite Moments

Here we consider the same linear processes as in Sect. 4.4.2, however, instead of assuming $E[\varepsilon_1^4] < \infty$, we impose the regularly varying condition:

$$P(\varepsilon_1 > x) \sim A \frac{1+\beta}{2} x^{-\alpha}, \qquad P(\varepsilon_1 < -x) = A \frac{1+\beta}{2} x^{-\alpha}, \tag{4.121}$$

where $A > 0$, $\beta \in [-1, 1]$ and $\alpha \in (1, 4)$. In particular, $E[|\varepsilon_1|] < \infty$, $E[\varepsilon_1^4] = +\infty$.

There is a vast literature on sample covariances for weakly dependent linear processes with regularly varying innovations. Kanter and Steiger (1974) considered AR(p) models, Davis and Resnick (1985, 1986) considered processes with infinite variance and with finite variance, but infinite fourth moment, respectively. In the latter papers, the authors used point process techniques, as described in the section on partial sums with infinite moments; see Sect. 4.3. This technique was successfully applied to bilinear processes with infinite moments (Davis and Resnick 1996; Basrak et al. 1999) and to GARCH models (Davis and Mikosch 1998; Basrak et al. 2002)

As for long-memory linear processes, Kokoszka and Taqqu (1996) generalized the results by Davis and Resnick (1985) for $\alpha \in (1, 2)$, whereas Horváth and Kokoszka (2008) generalized Davis and Resnick (1986) for $\alpha \in (2, 4)$. (Recall that there is no long memory if $\alpha \in (0, 1)$).

Recall that the sample covariance is defined as

$$\hat{\gamma}_n(s) = \frac{1}{n} \sum_{t=1}^{n-s} X_t X_{t+s} \quad (s = 1, \ldots, q).$$

The first result deals with $\alpha \in (1, 2)$. There is no influence of long memory.

Theorem 4.24 *Assume that X_t ($t \in \mathbb{N}$) is a linear process and ε_t ($t \in \mathbb{Z}$) are i.i.d. random variables such that (4.121) holds with $\alpha \in (1, 2)$ and $E(\varepsilon_1) = 0$. If $\alpha \in$*

$(1, 2)$, *then*

$$n^{1-2/\alpha}\big(\hat{\gamma}_n(0), \dots, \hat{\gamma}_n(q)\big) \xrightarrow{d} A^{2/\alpha} C_{\alpha/2}^{-2/\alpha} \left(\sum_{j=0}^{\infty} a_j a_{j+0}, \dots, \sum_{j=0}^{\infty} a_j a_{j+q} \right) S_{\alpha/2}(1, 1, 0), \quad (4.122)$$

where $S_\alpha(1, 1, 0)$ *is a stable random variable.*

Proof The proof is given in Davis and Resnick in the weakly dependent case (4.88); however it applies to the long-memory situation as long as the conditions of Theorem 4.24 are fulfilled. The reason for this is that under the condition $\sum_j a_j^2 < \infty$, the quantity $\sum_j a_j a_{j+s}$ is also finite. We give a sketch of the proof for $\hat{\gamma}_n(q)$ only. Recall from Theorem 4.14 that

$$\sum_{t=1}^{n} \delta_{c_n^{-1}(X_t, \dots, X_{t-K})} \Rightarrow \sum_{l=1}^{\infty} \sum_{r=0}^{\infty} \delta_{j_l(a_r, a_{r-1}, \dots, a_{r-K})},$$

where j_l are points of the limiting Poisson process, c_n is such that $P(|\varepsilon_1| > c_n) \sim n^{-1}$, i.e. $c_n \sim A^{1/\alpha} n^{1/\alpha}$. The continuous mapping theorem yields

$$c_n^{-2} \sum_{t=1}^{n} X_t X_{t+q} 1\big\{|X_t| > c_n\gamma \text{ or } |X_{t+h}| > c_n\gamma\big\} \xrightarrow{d} \sum_{l=0}^{\infty} \sum_{t=0}^{\infty} a_j a_{j+q} j_l^2 1\big\{|j_l| > \min\big\{a_j^{-1}, a_{j+q}^{-1}\big\}\gamma\big\}.$$

As $\gamma \to 0$, the latter random variable converges to

$$\left(\sum_{j=0}^{\infty} a_j a_{j+q} \right) \sum_{l=0}^{\infty} j_l^2 \stackrel{d}{=} \left(\sum_{j=0}^{\infty} a_j a_{j+q} \right) S_{\alpha/2}\big(C_{\alpha/2}^{-2/\alpha}, 1, 1\big).$$

It remains to show that

$$\lim_{\gamma \to 0} \limsup_{n \to \infty} P\left(c_n^{-2} \left| \sum_{t=1}^{n} X_t X_{t+q} 1\big\{|X_t| < c_n\gamma, |X_{t+q}| < c_n\gamma\big\} \right| > \gamma \right) = 0.$$

This probability is bounded by

$$\frac{n}{c_n^2 \gamma} E\big[|X_1^2| 1\big\{|X_1| < \gamma c_n\big\}\big].$$

We conclude the proof by applying Karamata's theorem (Lemma 4.18) together with the tail estimates in Lemma 4.19. □

The situation is different for $\alpha \in (2, 4)$. We have a dichotomous behaviour, depending on the interplay between tails and memory.

Theorem 4.25 *Assume that X_t $(t \in \mathbb{N})$ is a linear process such that $a_j \sim c_a j^{d-1}$, $d \in (0, 1/2)$ (so that $\gamma_X(k) \sim L_\gamma(k)k^{2d-1}$, see (4.39)) and ε_t $(t \in \mathbb{Z})$ are i.i.d. random variables such that (4.121) holds with $\alpha \in (2, 4)$ and $E(\varepsilon_1) = 0$.*

- *If $\alpha \in (2, 4)$ and $0 < d < 1/\alpha$, then (4.122) holds.*
- *If $\alpha \in (2, 4)$ and $1/\alpha < d < 1/2$, then*

$$n^{1-2d}L_2^{-1/2}(n)\big(\hat{\gamma}_n(s) - \gamma_X(s)\big) \xrightarrow{d} Z_{2,H}(1),$$

where $Z_{2,H}(u)$ is a Hermite–Rosenblatt process, and $L_2(n) = 2!C_2L_\gamma^2(n)$.

Proof Consider the decomposition $M_n + R_n$ from the proof of Theorem 4.23:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\big(X_tX_{t+s} - E(X_tX_{t+s})\big)\\
&\quad= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{\infty}a_ja_{j+s}\big(\varepsilon_{t-j}^2 - \sigma_\varepsilon^2\big) + \frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{\infty}\sum_{l=0;\ l\neq j+s}^{\infty}a_ja_l\varepsilon_{t-j}\varepsilon_{t-l}\\
&\quad=: M_n + R_n.
\end{aligned}$$

Since the random variables ε_t have a finite variance, we again have

$$n^{1-2d}L_2^{-1/2}(n)R_n \xrightarrow{d} Z_{2,H}(1)$$

if $d \in (1/4, 1/2)$ and $n^{-1/2}R_n = O_P(1)$ if $d \in (0, 1/4)$. The first part, M_n, is the partial sum of a linear process with summable coefficients and infinite variance, and hence we can conclude the stable limit for M_n. □

4.4.4 Stochastic Volatility Models

Some recent results were developed for stochastic volatility models (McElroy and Politis 2007, Kulik and Soulier 2012). In the latter paper, the authors show differences between LMSV and models with a leverage.

Consider a stochastic volatility model $X_t = \sigma_t\xi_t$ $(t \in \mathbb{N})$ such that the sequences σ_t $(t \in \mathbb{N})$ and ξ_t $(t \in \mathbb{N})$ are independent. Assume that $E(\xi_1) = 0$. We are interested in sample covariances of X_t and X_t^2. For the first one, we note that

$$\hat{\gamma}_n(s) = \frac{1}{n}\sum_{t=1}^{n-s}\xi_t\xi_{t+s}\sigma_t\sigma_{t+s}$$

is a martingale w.r.t. sigma field generated by (σ_j, ξ_j), $j \le t$. Therefore, if we assume additionally $E[\xi_1^2] < \infty$, then

$$\sqrt{n}\hat{\gamma}_n(s) \xrightarrow{d} N\big(0, v^2\big),$$

where $v^2 = E[\sigma_0^2\sigma_s^2]E^2[\xi_1^2]$. The more interesting situation happens in the second case of squares. Assume that $E[\xi_1^4] < \infty$. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\big(\xi_t^2\xi_{t+s}^2\sigma_t^2\sigma_{t+s}^2 - E\big[\xi_t^2\xi_{t+s}^2\big]E\big[\sigma_j^2\sigma_{t+s}^2\big]\big)\\
&\quad= \frac{1}{n}\sum_{t=1}^{n}\sigma_t^2\sigma_{t+s}^2\big(\xi_t^2\xi_{t+s}^2 - E\big[\xi_t^2\xi_{t+s}^2\big]\big) + E^2\big[\xi_1^2\big]\frac{1}{n}\sum_{t=1}^{n}\big(\sigma_t^2\sigma_{t+s}^2 - E\big[\sigma_t^2\sigma_{t+s}^2\big]\big)\\
&\quad=: M_n + R_n.
\end{aligned}$$

Again, the first part is a martingale, and therefore it is $O_P(n^{-1/2})$. The second part is a possible long-memory contribution of the bivariate sequence $\sigma_t\sigma_{t+s}$ $(t \in \mathbb{N})$. For example, if we consider $\sigma_t = \exp(p\zeta_t)$, where ζ_t $(t \in \mathbb{N})$ is the long-memory Gaussian process as in Example 4.23, then for $d \in (1/4, 1/2)$ (refer to Example 4.23 for the precise notation),

$$n^{-(d+1/2)}L_f^{-1/2}(n)R_n \xrightarrow{d} 2E^2\big[\xi_1^2\big]c_1\int D(\lambda)\,dW_\zeta(\lambda),$$

where W_ζ is the spectral random measure associated with ζ_t $(t \in \mathbb{N})$. Therefore, since the second part R_n dominates, the limiting distribution for

$$n^{1-(d+1/2)}L_2^{-1/2}\big(n^{-1}\big)\hat{\gamma}_n(s)$$

is the same as for R_n. If on the other hand $d \in (0, 1/4)$, then both terms M_n and R_n are of the same order.

This consideration can be extended to random variables ξ_t such that (4.121) holds with $\alpha \in (2, 4)$. Then, we have again a dichotomous behaviour: the limit can be either a stable random variable or a Hermite–Rosenblatt random variable. The situation becomes complicated though when one considers models with leverage. We refer to Davis and Mikosch (2001) and Kulik and Soulier (2012).

4.4.5 *Summary of Limit Theorems for Sample Covariances*

We consider a centred linear process $X_t = \sum_{j=0}^{\infty} a_j\varepsilon_{t-j}$ such that either $E(\varepsilon_1^4) < \infty$ or

$$P(\varepsilon_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha}$$

with $\alpha \in (1, 4)$ and appropriate regularity conditions (that assure existence of the process). In the table, $Z_{2,H}(\cdot)$ is a Hermite–Rosenblatt process on $[0, 1]$, and $\tilde{S}_{\alpha/2}$ is an $\alpha/2$-stable random variable. Furthermore, c is a generic constant. The main results are summarized in Table 4.3.

Table 4.3 Limits for sample covariances

	Sample covariances	
	Finite moments	Infinite moments
Linear processes	$n^{1/2}(\hat{\gamma}_n(s) - \gamma_X(s)) \xrightarrow{\mathrm{d}} cN(0,1)$ if $d \in (0, 1/4)$ $n^{1-2d}(\hat{\gamma}_n(s) - \gamma_X(s)) \xrightarrow{\mathrm{d}} cZ_{2,H}(1)$ if $d \in (1/4, 1/2)$ (Theorem 4.23)	$\alpha \in (1,2)$ $n^{1-2/\alpha}(\hat{\gamma}_n(s) - \gamma_X(s)) \xrightarrow{\mathrm{d}} c\tilde{S}_{\alpha/2}$ $\alpha \in (2,4)$ $n^{1-2/\alpha}(\hat{\gamma}_n(s) - \gamma_X(s)) \xrightarrow{\mathrm{d}} c\tilde{S}_{\alpha/2}$ if $d \in (0, 1/\alpha)$ $n^{1-2d}(\hat{\gamma}_n(s) - \gamma_X(s)) \xrightarrow{\mathrm{d}} cZ_{2,H}(1)$ if $d \in (1/\alpha, 1/2)$ (Theorems 4.25, 4.24)

4.5 Limit Theorems for Quadratic Forms

In this section we consider quadratic forms,

$$Q_n(u) := \sum_{t,s=1}^{[nu]} b_{t-s}\big\{G(X_t, X_s) - E\big[G(X_t, X_s)\big]\big\}, \qquad Q_n := Q_n(1), \quad (4.123)$$

where b_k ($k \in \mathbb{Z}$) is a sequence of constants, and $G : \mathbb{R}^2 \to \mathbb{R}$. We are interested in asymptotic properties of $Q_n(u)$.

In the Gaussian case, such studies were conducted in Rosenblatt (1979), Fox and Taqqu (1985, 1987), Avram (1988), Terrin and Taqqu (1990), Beran and Terrin (1994), among others. For linear processes, classical limit theorems for weakly dependent sequences are given in Brillinger (1969) and Hannan (1970) (and references therein); also see Klüppelberg and Mikosch (1996). They follow directly from limit theorems for sample covariances, proven in Theorem 4.23. For long memory such studies were initiated by Giraitis and Surgailis (1990). The authors concluded a weakly dependent behaviour, using approximation of a quadratic form by another quadratic form with weakly dependent variables. Other results along these lines were proven for instance in Horváth and Shao (1999) and Bhansali et al. (1997). The case of the multivariate Appell polynomials is studied in Terrin and Taqqu (1991), Giraitis and Taqqu (1997, 1998, 1999a, 2001), Giraitis et al. (1998). Kokoszka and Taqqu (1997) discuss quadratic forms for infinite-variance processes. We also refer to Giraitis and Taqqu (1999b) for an overview.

There are two principal applications of quadratic forms. First, we can derive the limiting behaviour of the periodogram and the Whittle estimator (see Sect. 5.5 for results and references), or we can use quadratic forms to test for possible changes in the long-memory parameter (see e.g. Beran and Terrin 1996, Horváth and Shao 1999).

4.5.1 Gaussian Sequences

In this section we shall assume that X_t $(t \in \mathbb{Z})$ is a centred Gaussian sequence with autocovariance function $\gamma_X(k) = L_\gamma(k)k^{2d-1}$. First, we exploit the relation between sample covariances and quadratic forms. Using results obtained in Sect. 4.4, we obtain a *long-memory behaviour I* (i.e. of "type I") of $Q_n(u)$ for $d \in (1/4, 1/2)$ directly from limit theorems for sample covariances. The result was proven in Fox and Taqqu (1985) and is presented in Theorem 4.26. For $d \in (0, 1/4)$, we obtain convergence with rate $n^{-1/2}$, as proven in Fox and Taqqu (1985) as well. The result is presented in Theorem 4.27 and is referred to as *weakly dependent behaviour I*.

These results are very similar to those for partial sums $\sum_{t=1}^{[nu]}(X_t^2 - 1)$. These sums were studied in Sect. 4.2.3, and we recall the dichotomous behaviour: convergence to the Hermite–Rosenblatt process or Brownian motion for $d \in (1/4, 1/2)$ and $d \in (0, 1/4)$ respectively.

In Theorem 4.26 the limiting process will be degenerated if $\sum_l b_l = 0$, as it happens for Fourier coefficients. Another type of weakly dependent behaviour is obtained if in addition to $\sum_l b_l = 0$ the coefficients also decay to zero fast enough. Then, the coefficients b_l *compensate* for long memory, and $Q_n(\cdot)$ converges at rate $n^{1/2}$ for all $d \in (0, 1/2)$ (*weakly dependent behaviour II*). Such results were proven in Fox and Taqqu (1985, Theorem 3; 1987), Avram (1988), Beran and Terrin (1994) (also Beran 1986). The authors use the method of cumulants; see the proof of Theorem 4.28. On the other hand, if the coefficients b_l do not compensate for long memory, then Terrin and Taqqu (1990) prove that the limiting process is neither Gaussian nor Hermite–Rosenblatt (*long-memory behaviour II*). The authors use multiple Wiener–Itô integrals; see the proof of Theorem 4.29.

4.5.1.1 Long Memory Behaviour I

Recall that the sample covariances for the sequence X_t $(t \in \mathbb{Z})$ are defined by

$$\hat{\gamma}_n(s) = \frac{1}{n}\sum_{t=1}^{n-|s|} X_t X_{t+|s|}.$$

Reorganizing indices, we may write

$$Q_n(1) = \sum_{t,s=1}^{n} b_{t-s}\big(X_t X_s - E(X_t X_s)\big) = n \sum_{|l|\le n-1} b_l\big(\hat{\gamma}_n(l) - \gamma_X(l)\big).$$

Recall that for $d \in (1/4, 1/2)$ (see (4.116)),

$$n^{1-2d}L_2^{-1/2}(n)\big(\hat{\gamma}_n(1) - \gamma_X(1), \ldots, \hat{\gamma}_n(K) - \gamma_X(K)\big) \xrightarrow{\mathrm{d}} \big(Z_{2,H}(1), \ldots, Z_{2,H}(1)\big). \tag{4.124}$$

This, together with the continuous mapping theorem, implies that for any fixed integer $K > 0$,

$$n^{-2d}L_2^{-1/2}(n)\,Q_{n,K}(1) := n^{-2d}L_2^{-1/2}(n)n\sum_{|l|\le K} b_l\big(\hat{\gamma}_n(l) - \gamma_X(l)\big)$$
$$\xrightarrow{\mathrm{d}} \left(\sum_{l=-K}^{K} b_l\right) Z_{2,H}(1).$$

Clearly, $(\sum_{l=-K}^{K} b_l)Z_{2,H}(1) \xrightarrow{\mathrm{p}} (\sum_{l=-\infty}^{\infty} b_l)Z_{2,H}(1)$. Furthermore,

$$\lim_{K\to\infty}\limsup_{n\to\infty} P\big(n^{-2d}L_2^{-1/2}(n)\big|Q_{n,K}(1) - Q_n(1)\big| > \delta\big) = 0$$

for each $\delta > 0$. The reader is referred to Fox and Taqqu (1985, Theorem 1) for details on the latter approximation and tightness. This leads to the following result, which is formulated more generally in a functional form.

Theorem 4.26 *Assume that X_t $(t \in \mathbb{Z})$ is a stationary sequence of standard normal random variables such that $\gamma_X(k) \sim L_\gamma(k)k^{2d-1}$, $d \in (1/4, 1/2)$. If $\sum_{l=-\infty}^{\infty} |b_l| < \infty$, then*

$$n^{-2d}L_2^{-1/2}(n)\,Q_n(u) = n^{-2d}L_2^{-1/2}(n)\sum_{t,s=1}^{[nu]} b_{t-s}\big(X_tX_s - E(X_tX_s)\big)$$
$$\Rightarrow \left(\sum_{l=-\infty}^{\infty} b_l\right) Z_{2,H}(u),$$

where $L_2(n) = 2!C_2L_\gamma^2(n)$ (cf. (4.22)), $H = d + \frac{1}{2}$, $\Rightarrow$ denotes weak convergence, and $Z_{2,H}(\cdot)$ is the Hermite–Rosenblatt process.

This result has been proven in fact in a more general setting Fox and Taqqu (1985). Consider

$$Q_n(u; H_m) := \sum_{t,s=1}^{[nu]} b_{t-s}\big\{H_m(X_t)H_m(X_s) - E\big[H_m(X_t)H_m(X_s)\big]\big\}.$$

The same methodology as above works, given that we use (4.118) instead of (4.124):

$$n^{1-2d}L_2^{-1/2}(n)\big(\hat{\gamma}_n(1; H_m) - m!\gamma_X^m(1), \ldots, \hat{\gamma}_n(K; H_m) - m!\gamma_X^m(K)\big)$$
$$\xrightarrow{\mathrm{d}} m!m\big(\gamma_X^{m-1}(1), \ldots, \gamma_X^{m-1}(K)\big)Z_{2,H}(1).$$

We conclude for $d \in (1/4, 1/2)$ and under the condition $\sum_{l=-\infty}^{\infty} |b_l| < \infty$,

$$n^{-2d} L_{A_2}^{-1/2}(n) Q_n(1; H_m) \xrightarrow{\mathrm{d}} m! m \left(\sum_{l=-\infty}^{\infty} b_l \gamma_X^{m-1}(l) \right) Z_{2,H}(1).$$

4.5.1.2 Weakly Dependent Behaviour I

Theorem 4.26 above requires $d \in (1/4, 1/2)$. What about $d \in (0, 1/4)$? As in the case of partial sums $\sum_{t=1}^{[nu]} (X_t^2 - 1)$, one obtains a weakly dependent behaviour, i.e. a central limit theorem with scaling $n^{-1/2}$ Fox and Taqqu (1985).

Theorem 4.27 *Assume that X_t $(t \in \mathbb{Z})$ is a stationary sequence of standard normal random variables such that $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$, $d \in (0, 1/4)$. Then*

$$n^{-1/2} Q_n(u) = n^{-1/2} \sum_{t,s=1}^{[nu]} b_{t-s} \big(X_t X_s - E(X_t X_s) \big) \Rightarrow \sigma_0 B(u),$$

where $B(\cdot)$ is a standard Brownian motion, and $\sigma_0 > 0$.

The constant σ_0 is given in a complicated form, and we refer to Fox and Taqqu (1985) for a precise formula.

4.5.1.3 Weakly Dependent Behaviour II

In Theorem 4.26 it may happen that $\sum_{l=-\infty}^{\infty} b_l = 0$ and hence the limit will be degenerated. This can happen when b_l are Fourier coefficients of a real-valued function g. Specifically, let

$$b_l = \int_{-\pi}^{\pi} e^{il\lambda} g(\lambda)\, d\lambda =: 2\pi \hat{g}_l, \qquad g(\lambda) \sim c_g |\lambda|^{-\gamma} \quad \text{as } |\lambda| \to 0. \tag{4.125}$$

To assure the existence of Fourier coefficients, we assume that $\gamma < 1$. Then, $b_l \sim c_b l^{\gamma-1}$, $c_b = 2c_g \Gamma(1-\gamma) \sin(\pi \frac{\gamma}{2})$. The following result was proven in Fox and Taqqu (1987); see also Theorem 3 in Fox and Taqqu (1985) and Avram (1988).

Theorem 4.28 *Assume that X_t $(t \in \mathbb{Z})$ is a stationary sequence of standard normal random variables such that $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$, $d \in (0, 1/2)$. If*

$$2d + \gamma < 1/2, \tag{4.126}$$

then

$$n^{-1/2} Q_n(1) \xrightarrow{d} \sigma_Q Z, \tag{4.127}$$

where

$$\sigma_Q^2 := 16\pi^3 \int_{-\pi}^{\pi} \big(f(\lambda)g(\lambda)\big)^2 \, d\lambda,$$

$f = f_X$ is the spectral density of X_t ($t \in \mathbb{Z}$), and Z is a standard normal random variable.

Let us comment on condition (4.126). First, it assures that σ_Q^2 is finite. Second, it means that the coefficients b_l decay appropriately fast, to compensate for long memory in X_t ($t \in \mathbb{Z}$).

Proof We present a modified version of the proof in Avram (1988). Let $\Sigma = [\gamma_X(j-l)]_{j,l=1}^n$ and $B = [b_{j-l}]_{j,l=0}^{n-1}$. Then,

$$Q_n(1) = (X_1, \ldots, X_n) B (X_1, \ldots, X_n)^T$$

has the pth cumulant equal to (see Grenander and Szegö 1958, p. 218)

$$\mathrm{cum}_p\big(Q_n(1)\big) = 2^{p-1}(p-1)!\mathrm{Trace}(\Sigma B)^p.$$

Note that

$$\gamma_X(j-l) = \int_{-\pi}^{\pi} e^{i(j-l)\lambda} f_X(\lambda) \, d\lambda =: 2\pi \hat{f}_{j-l},$$

where $\hat{f}_{j-l}$ is the Fourier coefficient of the spectral density $f = f_X$. Furthermore, $B = 2\pi[\hat{g}_{j-l}]_{j,l=0}^{n-1}$. Recall that the trace of a matrix is the sum of its diagonal elements. We have

$$\frac{1}{n}\mathrm{Trace}(\Sigma) = \frac{2\pi}{n}(\hat{f}_0 + \cdots + \hat{f}_0) = 2\pi \hat{f}_0 = \int_{-\pi}^{\pi} f_X(\lambda) \, d\lambda.$$

Of course, f_X is integrable given $d < 1/2$. Analogously, recall that the trace can be written as a Hadamard product: $\mathrm{Trace}(\Sigma B) = \sum_{j,l} \gamma_X(j-l) B_{j,l}$. Since $\hat{f}_l \hat{g}_l$ is summable, we then obtain

$$\frac{1}{n}\mathrm{Trace}(\Sigma B) = 4\pi^2 \frac{1}{n} \sum_{j,l=1}^{n} \hat{f}_{j-l}\hat{g}_{j-l} = 4\pi^2 \frac{1}{n} \sum_{l=-(n-1)}^{n-1} \big(n - |l|\big) \hat{f}_l \hat{g}_l$$

$$\approx 4\pi^2 \sum_{l=-(n-1)}^{n-1} \hat{f}_l \hat{g}_l \to 4\pi^2 \sum_{l=-\infty}^{\infty} \hat{f}_l \hat{g}_l$$

as $n \to \infty$. By the Parseval identity and since g is real,

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Trace}(\Sigma B) = 4\pi^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} f_X(\lambda)\bar{g}(\lambda) \, d\lambda = 2\pi \int_{-\pi}^{\pi} f_X(\lambda) g(\lambda) \, d\lambda.$$

On the other hand, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of ΣB, then we can write alternatively

$$\lim_{n\to\infty} \frac{1}{n}\text{Trace}(\Sigma B) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \lambda_j \to \frac{4\pi^2}{2\pi}\int_{-\pi}^{\pi} f_X(\lambda) g(\lambda)\, d\lambda.$$

The matrix $(\Sigma B)^p$ has eigenvalues λ_j^p, $j = 1, \ldots, n$. One can then argue analogously that

$$\lim_{n\to\infty} \frac{1}{n}\text{Trace}(\Sigma B)^p = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \lambda_j^p = \frac{(4\pi^2)^p}{2\pi}\int_{-\pi}^{\pi} (f_X(\lambda) g(\lambda))^p\, d\lambda.$$

Thus,

$$\text{cum}_p\big(n^{-1/2} Q_n(1)\big) = n^{-p/2}\text{cum}_p\big(Q_n(1)\big) = \frac{2^{p-1}}{n^{p/2-1}}\frac{(p-1)!}{n}\text{Trace}(\Sigma B)^p.$$

Consequently, $\lim_{n\to\infty} \text{cum}_p(n^{-1/2} Q_n(1)) = 0$ if $p > 2$, and

$$\lim_{n\to\infty} \text{cum}_2\big(n^{-1/2} Q_n(1)\big) = 16\pi^3 \int_{-\pi}^{\pi} \big(f_X(\lambda) g(\lambda)\big)^2\, d\lambda,$$

which provides the limiting variance. Application of the method of cumulants (see Theorem 4.1) then yields the result. □

4.5.1.4 Long-Memory Behaviour II

In contrast to Theorem 4.28, if the coefficients b_l do not compensate for long memory (i.e., when (4.126) fails to hold), then we have the following result, due to Terrin and Taqqu (1990). Recall that $g(\lambda) \sim c_g|\lambda|^{-\gamma}$ as $\lambda \to 0$ (see (4.125)) and that $M_0(\cdot)$ is a random measure that appears in the spectral representation of the linear Gaussian sequence; see Sect. 4.1.3.

Theorem 4.29 *Assume that X_t ($t \in \mathbb{Z}$) is a stationary sequence of standard normal random variables such that $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$, $d \in (0, 1/2)$. If*

$$1/2 < 2d + \gamma < 1, \tag{4.128}$$

then

$$n^{-(2d+\gamma)} L_f^{-1}\big(n^{-1}\big) Q_n(u) \Rightarrow c_g Z(u), \tag{4.129}$$

where

$$Z(u) = \iint \psi_u(\lambda_1, \lambda_2) \frac{1}{\lambda_1}\frac{1}{\lambda_2}\, dM_0(\lambda_1)\, dM_0(\lambda_2),$$

and

$$\psi_u(\lambda_1,\lambda_2)=\int_{\mathbb{R}}\frac{e^{iu(\lambda_1-\lambda)}-1}{i(\lambda_1+\lambda)}\frac{e^{iu(\lambda_2+\lambda)}-1}{i(\lambda_2-\lambda)}|\lambda|^{-\gamma}\,d\lambda.$$

The limiting process is self-similar with $H=2d+\gamma\in(\frac{1}{2},2)$, but neither Gaussian nor Hermite–Rosenblatt.

We note that for $\gamma=0$, we have $b_l=1$ for $l=0$ and 0 otherwise. In this case the result of Theorem 4.29 reduces to the asymptotic behaviour of $\sum_{t=1}^{[nu]}(X_t^2-1)$, see Theorem 4.3.

Proof The proof is sketched here. It follows the same idea as in the case of partial sums $\sum_{t=1}^{n}H_m(X_t)$. Recall that the multiple Wiener–Itô integral "removes" the diagonal (see Appendix A). We write

$$X_tX_s-E(X_tX_s)=\int_{[-\pi,\pi]^2\setminus\{\lambda_1=\lambda_2\}}e^{it\lambda_1}e^{is\lambda_2}a(\lambda_1)a(\lambda_2)\,dM_0(\lambda_1)\,dM_0(\lambda_2),$$

where $|a(\lambda)|^2=f_X(\lambda)$.

Thus,

$$\begin{aligned}Q_n(1)&=\sum_{t,s=0}^{n-1}\int_{-\pi}^{\pi}e^{i(t-s)\lambda}g(\lambda)\,d\lambda\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}e^{it\lambda_1}e^{is\lambda_2}a(\lambda_1)a(\lambda_2)\,dM_0(\lambda_1)\,dM_0(\lambda_2)\\&=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}a(\lambda_1)a(\lambda_2)\\&\quad\times\left(\int_{-\pi}^{\pi}\frac{e^{in(\lambda_1+\lambda)}-1}{e^{i(\lambda_1+\lambda)}}\frac{e^{in(\lambda_2-\lambda)}-1}{e^{i(\lambda_2-\lambda)}}g(\lambda)\,d\lambda\right)dM_0(\lambda_1)\,dM_0(\lambda_2)\\&=c_gn^{\gamma}\int_{-n\pi}^{n\pi}\int_{-n\pi}^{n\pi}a\left(\frac{\lambda_1}{n}\right)a\left(\frac{\lambda_2}{n}\right)\\&\quad\times\psi_1(\lambda_1,\lambda_2;n)n^{1/2}\,dM_0\big(n^{-1}\lambda_1\big)n^{1/2}\,dM_0\big(n^{-1}\lambda_2\big),\end{aligned}$$

where

$$\begin{aligned}\psi_1(\lambda_1,\lambda_2;n)&=\left(\int_{-n\pi}^{n\pi}D_n\left(\frac{\lambda_1+\lambda}{n}\right)D_n\left(\frac{\lambda_2-\lambda}{n}\right)g(\lambda)\,d\lambda\right),\\D_n(\lambda)&=\frac{e^{i\lambda n}-1}{n(e^{i\lambda}-1)}1\{|\lambda|\le\pi n\}.\end{aligned}$$

Thus, $Q_n(1)$ equals in distribution to

$$\int_{-n\pi}^{n\pi}\int_{-n\pi}^{n\pi}a\left(\frac{\lambda_1}{n}\right)a\left(\frac{\lambda_2}{n}\right)\psi_1(\lambda_1,\lambda_2;n)\,dM_0(\lambda_1)\,dM_0(\lambda_2).$$

Clearly, $\lim_{n\to\infty} \psi_1(\lambda_1, \lambda_2; n) = \psi_1(\lambda_1, \lambda_2)$, and as in the alternative proof of Theorem 4.2, one can argue that the convergence is uniform. Therefore, the same method as in Theorem 4.2 applies, and the result (4.129) follows for $u = 1$. A proof of functional convergence is omitted here. □

4.5.2 Linear Processes

As in the case of partial sums, the results on quadratic forms for Gaussian LRD sequences have a counterpart for general linear sequences

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \quad (t \in \mathbb{Z}), \tag{4.130}$$

where $\sum_{j=0}^{\infty} a_j^2 = 1$, ε_t $(t \in \mathbb{Z})$ are i.i.d. zero mean random variables with $\mathrm{var}(\varepsilon_1) = \sigma_\varepsilon^2 = 1$. We will assume that either $\sum_{j=0}^{\infty} |a_j| < \infty$ or $a_j \sim L_a(j) j^{d-1}$ with $d \in (0, 1/2)$.

Results for quadratic forms

$$Q_n(u) = \sum_{t,s=1}^{[nu]} b_{t-s}\big(X_t X_s - E(X_t X_s)\big)$$

based on weakly dependent linear processes are classical (see Brillinger 1969; Hannan 1970; also see Klüppelberg and Mikosch 1996) and follow directly from limit theorems for sample covariances, as proven before in Theorem 4.23.

For long memory, such studies had been initiated by Giraitis and Surgailis (1990). The authors concluded a weakly dependent behaviour, similar to that of Theorem 4.28, using an approximation of the quadratic form by another quadratic form with weakly dependent variables. Other results along this line can be found in Horváth and Shao (1999) and Bhansali et al. (1997).

When one replaces $Q_n(u)$ by

$$Q_n(u; P_{m_1,m_2}) = \sum_{t,s=1}^{[nu]} b_{t-s}\big\{P_{m_1,m_2}(X_t X_s) - E\big[P_{m_1,m_2}(X_t, X_s)\big]\big\},$$

where P_{m_1,m_2} is a multivariate Appell polynomial, then limit theorems are very complicated; see Terrin and Taqqu (1991), Giraitis and Taqqu (1997, 1998, 1999a, 2001). We refer to Giraitis and Taqqu (1999b) for an overview.

4.5.2.1 Weakly Dependent Processes

Assume that $\sum_{j=0}^{\infty} |a_j| < \infty$. Recall Theorem 4.23 and the multivariate convergence (4.120):

$$n^{1/2}\big(\hat{\gamma}_n(0) - \gamma_X(0), \ldots, \hat{\gamma}_n(K) - \gamma_X(K)\big) \xrightarrow{\mathrm{d}} (G_0, \ldots, G_K),$$

where $(G_0, \ldots, G_K)$ is a Gaussian vector. We apply a similar method as in the proof of Theorem 4.26. There we concluded long-memory behaviour of quadratic forms from long-memory behaviour of sample covariances. Here, we will conclude short-memory behaviour of quadratic forms from short memory-behaviour of sample covariances.

We have

$$Q_n(1) = \sum_{t,s=1}^{n} b_{t-s}\big(X_t X_s - E(X_t X_s)\big) = n \sum_{|l|\le n-1} b_l\big(\hat{\gamma}_n(l) - \gamma_X(l)\big).$$

The continuous mapping theorem implies

$$n^{-1/2} Q_{n,K}(1) := n^{-1/2} n \sum_{|l|\le K} b_l\big(\hat{\gamma}_n(l) - \gamma_X(l)\big) \xrightarrow{\mathrm{d}} b_0 G_0 + 2\sum_{l=1}^{K} b_l G_l.$$

To apply Proposition 4.1, we need to show that

$$\lim_{K\to\infty} \limsup_{n\to\infty} P\left(\sqrt{n}\left|\sum_{l=K+1}^{n-1} b_l\big(\hat{\gamma}_n(l) - \gamma_X(l)\big)\right| > \delta\right) = 0.$$

This is straightforward since the correlations between $\hat{\gamma}_n(l)$ $(l \ge 1)$ are absolutely summable. Therefore, we may apply Chebyshev inequality in a suitable way to finish the proof. □

4.5.2.2 Long-Memory Sequences

The following result is a counterpart to Theorem 4.28.

Theorem 4.30 *Assume that X_t $(t \in \mathbb{N})$ is a linear process with long-range dependence defined in (4.130), with spectral density $f_X(\lambda) \sim c_f |\lambda|^{-2d}$. Assume that the coefficients b_l are given by (4.125), i.e. $b_l \sim c_b l^{\gamma-1}$. Let κ_4 be the fourth cumulant of ε_1. If*

$$2d + \gamma < 1/2, \tag{4.131}$$

then

$$n^{-1/2} Q_n(1) = n^{-1/2} \sum_{t,s=1}^{n} b_{t-s}\big(X_t X_s - E(X_t X_s)\big) \xrightarrow{\mathrm{d}} \sigma_Q Z, \tag{4.132}$$

where Z is standard normal, and

$$\sigma_Q^2 := 16\pi^3 \int_{-\pi}^{\pi} \big(f_X(\lambda)g(\lambda)\big)^2 \, d\lambda + \kappa_4 \bigg(2\pi \int_{-\pi}^{\pi} f_X(\lambda)g(\lambda)\, d\lambda\bigg)^2.$$

Of course, if the innovations ε_t are normal, then $\kappa_4 = 0$, and the result reduces to Theorem 4.28.

Proof To prove this theorem, Giraitis and Surgailis (1990) do not use the method of cumulants. Instead, they approximate $Q_n = Q_n(1)$ by a weakly dependent sequence. A similar approach is also used in Bhansali et al. (1997), and we present a sketch of the method there.

Write $Q_{n,X} = \sum_{t,s=1}^n b_{t-s} X_t X_s$ and $Q_{n,\varepsilon} = \sum_{t,s=1}^n v_{t-s}\varepsilon_t\varepsilon_s$, where

$$v_l = 2\pi \int_{-\pi}^{\pi} g(\lambda) f_X(\lambda) e^{il\lambda}\, d\lambda.$$

Since $Q_{n,\varepsilon}$ is a quadratic form of independent random variables, it is much easier to derive its asymptotic distribution, namely (see Bhansali et al. 1997, Theorem 4.1):

$$\frac{1}{\sqrt{\operatorname{var}(Q_{n,\varepsilon})}}\big(Q_{n,\varepsilon} - E(Q_{n,\varepsilon})\big) \xrightarrow{\mathrm{d}} N(0,1),$$

where

$$\operatorname{var}(Q_{n,\varepsilon}) = v_0^2 n \cdot \sigma_\varepsilon^4 + 2 \sum_{j,l=1;\ j\neq l}^{n} v_{j-l}^2$$

and $\sigma_\varepsilon^2 = \operatorname{var}(\varepsilon_t)$. Under our assumptions,

$$g(\lambda) f_X(\lambda) \sim c_g |\lambda|^{-\gamma} c_f |\lambda|^{-2d}$$

as $\lambda \to 0$. Therefore, the coefficients v_l satisfy

$$v_l \sim c_v l^{2d+\gamma-1}, \quad c_v = 2c_f c_g \Gamma\big(1-(2d+\gamma)\big) \sin\bigg(\pi \frac{2d+\gamma}{2}\bigg).$$

Furthermore, $Q_{n,X} - Q_{n,\varepsilon} = o_P(1)$. Evaluation of this is quite challenging, and the reader is referred to Giraitis and Surgailis (1990). Once this is verified, the convergence of $Q_{n,X}$ follows from the convergence of $Q_{n,\varepsilon} - E(Q_{n,\varepsilon})$. □

The limiting behaviour of quadratic forms becomes more involved if one considers nonlinear functionals. Recall the definition of bivariate Appell polynomials. Redefine Q_n as

$$Q_n(u) = Q_n(u; P_{m_1,m_2}) = \sum_{t,s=1}^{[nu]} b_{t-s}\big\{P_{m_1,m_2}(X_t, X_s) - E\big[P_{m_1,m_2}(X_t, X_s)\big]\big\}.$$

Table 4.4 Panorama of limits for quadratic forms of Gaussian sequences

Quadratic forms—Gaussian sequences (notation: $g(\lambda) = \frac{1}{2\pi}\sum b_l e^{-il\lambda}$)		
$g(0) = \sum_l b_l \neq 0$ $\sum \lvert b_l \rvert < \infty$	$d \in (0, 1/2)$ $n^{-1/2} Q_n(u) \Rightarrow c B(u)$ Theorem 4.27	$d \in (1/4, 1/2)$ $n^{-2d} Q_n(u) \Rightarrow c\, Z_{2,H}(u)$ Theorem 4.26
$g(\lambda) \sim c_g \lvert\lambda\rvert^{-\gamma}$ $(\lambda \to 0)$	$d \in (0, 1/2)$ and $2d + \gamma < 1/2$ $n^{-1/2} Q_n(1) \xrightarrow{\mathrm{d}} c\, B(1)$ Theorem 4.28	
$g(\lambda) \sim c_g \lvert\lambda\rvert^{-\gamma}$ $(\lambda \to 0)$	$d \in (0, 1/2)$ and $1/2 < 2d + \gamma < 1$ $n^{-(2d+\gamma)} Q_n(u) \Rightarrow c Z(u)$ Theorem 4.29	

Let $B = [b_{j-l}]_{j,l=1}^n$ and $\Sigma^{(m)} = [\gamma_X^m(j-l)]_{j,l=1}^n$. Also, let h^{*m} be the m-fold convolution of a function h. Giraitis and Taqqu (1997) showed that if

$$\lim_{n\to\infty} \frac{\mathrm{Trace}(\Sigma^{(m_1)} B \Sigma^{(m_2)} B)}{n} = \int_{-\pi}^{\pi} f_X^{*m_1}(\lambda) f_X^{*m_2}(\lambda) g^2(\lambda)\, d\lambda < \infty, \qquad (4.133)$$

then $n^{-1/2} Q_n$ converges in distribution to a normal random variable; however the formula for the limiting variance is quite complicated. Condition (4.133) holds if

$$\max\bigl(1 - m_1(1-2d), 0\bigr)/2 + \max\bigl(1 - m_2(1-2d), 0\bigr)/2 + \gamma < 1/2. \qquad (4.134)$$

In particular, if $m_1 = m_2 = 1$, then this is equivalent to $2d + \gamma < 1/2$, so that we recover (4.131). On the other hand, if $m_1 = 1$, $m_2 = 2$, then the condition reads: $3d - 1 + \gamma < 1/2$ if $d \in (1/4, 1/2)$; $d + \gamma < 3/2$ if $d \in (0, 1/4)$.

If (4.134) does not hold, then there is a variety of different possible limits, as presented in Giraitis and Taqqu (1999b). The proofs involve the familiar method based on the multiple Wiener–Itô integrals.

4.5.3 Summary of Limit Theorems for Quadratic Forms

We summarize the main results for quadratic forms of Gaussian sequences in Table 4.4. We assume that X_t $(t \in \mathbb{Z})$ is a centred Gaussian sequence with covariance $\gamma_X(k) \sim c_\gamma k^{2d-1}$, $d \in (0, 1/2)$, so that a slowly varying function can be omitted. In what follows, $B(\cdot)$ is a Brownian motion on $[0, 1]$, $Z_{2,H}(\cdot)$ is a Hermite–Rosenblatt process on $[0, 1]$, and $Z(\cdot)$ is the self-similar process with Hurst parameter $H = 2d + \gamma$, as in Theorem 4.29. Furthermore, c is a generic constant.

4.6 Limit Theorems for Fourier Transforms and the Periodogram

In this section we present some basic properties of the Discrete Fourier Transform (DFT) and the periodogram. We analyse their second-order properties showing a remarkable difference between weakly dependent and long-memory linear processes. In particular, the DFT and the periodogram computed at Fourier frequencies are asymptotically independent under short memory but asymptotically dependent under long memory. To achieve asymptotic independence in the latter case, one has to consider the DFT at appropriately high frequencies. The asymptotic dependence of the DFT and the periodogram ordinates implies a different limiting behaviour of the DFT under short and long memory respectively.

4.6.1 Periodogram and Discrete Fourier Transform (DFT)

For an observed second-order stationary time series $X_1, \ldots, X_n$, let $\bar{x} = \bar{x}_n = n^{-1} \sum_{t=1}^{n} X_t$ and define by

$$\hat{\gamma}_X(k) = \frac{1}{n} \sum_{t=1}^{n-|k|} (X_t - \bar{x})(X_{t+|k|} - \bar{x}) \quad \big(|k| \le n-1\big),$$

$$\hat{\gamma}_X(k) = 0 \quad \big(|k| \ge n\big),$$

the sample autocovariances. Also, define the (centred) periodogram by

$$I_{n,X}^{\text{centred}}(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \hat{\gamma}_X(k) e^{-ik\lambda} = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \hat{\gamma}_X(k) e^{-ik\lambda}$$

$$= \frac{1}{2\pi n} \left| \sum_{t=1}^{n} (X_t - \bar{x}) e^{-ik\lambda} \right|^2.$$

If $E[X_1] = \mu = 0$, then $I_{n,X}^{\text{centred}}(\lambda)$ can be approximated by

$$I_{n,X}(\lambda) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} X_t e^{-it\lambda} \right|^2.$$

For Fourier frequencies $\lambda_j = 2\pi j/n$ ($j = 1, \ldots, N_n$; $N_n = [(n-1)/2]$), we have the exact identity $I_{n,X}^{\text{centred}}(\lambda_j) = I_{n,X}(\lambda_j)$ since $\sum_{t=1}^{n} e^{-it\lambda_j} = 0$. Therefore, in most applications the non-centred periodogram $I_{n,X}$ is used. The non-centred periodogram can be written in terms of the discrete Fourier transform (DFT). Let

$$d_{n,X}(\lambda) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t e^{it\lambda}.$$

Then clearly $I_{n,X}(\lambda) = |d_{n,X}(\lambda)|^2$.

4.6.2 Second-Order Properties of the Fourier Transform and the Periodogram

4.6.2.1 Mean and Covariance of the DFT and the Periodogram

We are interested in a general expression for the expected value and covariance of the DFT and the periodogram ordinates $I_{n,X}(\lambda_j)$, where λ_j are Fourier frequencies.

Lemma 4.22 *Assume that X_t ($t \in \mathbb{Z}$) is a second-order stationary sequence with mean 0, covariance function γ_X and spectral density f_X. Then $E[d_{n,X}(\lambda_j)] = 0$,*

$$E\left(\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)}\right) = \frac{1}{f_X(\lambda_j)} \int_{-\pi}^{\pi} K_n(\lambda_j - \lambda) f_X(\lambda)\, d\lambda$$

and

$$E\left[d_{n,X}(\lambda_j)\,\overline{d_{n,X}(\lambda_j)}\right] = \int_{-\pi}^{\pi} K_n(\lambda - \lambda_j) f_X(\lambda)\, d\lambda, \tag{4.135}$$

where

$$K_n(\lambda) = \frac{1}{2\pi n}\left(\frac{\sin(n\lambda/2)}{\sin(\lambda/2)}\right)^2$$

is the Féjer kernel.

Proof The formula is classical (see Priestley 1981 p. 419), but we give a proof for completeness. We have

$$\begin{aligned} E\left[I_{n,X}(\lambda_j)\right] &= \frac{1}{2\pi n} \sum_{t=1}^{n} \sum_{s=1}^{n} e^{-i(t-s)\lambda_j} E(X_t X_s) \\ &= \frac{1}{2\pi n} \sum_{k=-(n-1)}^{n-1} \big(n - |k|\big) e^{-ik\lambda_j} \gamma_X(k) \\ &= \frac{1}{2\pi n} \int_{-\pi}^{\pi} \left(\sum_{k=-(n-1)}^{n-1} \big(n - |k|\big) e^{-ik(\lambda - \lambda_j)} \right) f_X(\lambda)\, d\lambda. \end{aligned}$$

Furthermore,

$$\begin{aligned} \sum_{t=1}^{n} \sum_{s=1}^{n} e^{-i(t-s)u} &= \sum_{k=-(n-1)}^{n-1} \big(n - |k|\big) e^{iku} \\ &= \frac{1}{2\pi n}\left(\frac{\sin(nu/2)}{\sin(u/2)}\right)^2 = K_n(u). \end{aligned}$$

Similarly, (4.135) follows from

$$\begin{aligned} E\big[d_{n,X}(\lambda_j)\overline{d_{n,X}(\lambda_j)}\big] &= \frac{1}{2\pi n}\sum_{t,s=1}^{n} e^{-i(t-s)\lambda_j}\gamma_X(t-s) \\ &= \frac{1}{2\pi n}\sum_{t,s=1}^{n} e^{-i(t-s)\lambda_j}\int_{-\pi}^{\pi} e^{i(t-s)\lambda} f_X(\lambda)\,d\lambda \\ &= \int_{-\pi}^{\pi} K_n(\lambda-\lambda_j) f_X(\lambda)\,d\lambda. \end{aligned}$$

□

Note that the Féjer kernel is also defined by

$$K_n(\lambda) = \frac{1}{2\pi n}\sum_{t,s=1}^{n} e^{-i(t-s)\lambda} = \frac{1}{2\pi n}\big|D_n(\lambda)\big|^2,$$

where

$$D_n(\lambda) = \sum_{t=1}^{n} e^{it\lambda} = \frac{e^{i(n+1)\lambda} - e^{i\lambda}}{e^{i\lambda} - 1}$$

is (a version of) the Dirichlet kernel.

4.6.2.2 Weakly Dependent Sequences

Assume that X_t ($t \in \mathbb{Z}$) is a second-order stationary weakly dependent time series with mean 0. Then (see e.g. Brockwell and Davis 1991) the following holds:

- The periodogram is an asymptotically unbiased estimator of the spectral density:

$$E\big[I_{n,X}(\lambda_j) - f_X(\lambda_j)\big] = O\big(n^{-1}\big) \tag{4.136}$$

  uniformly in $j = 1, \ldots, [n/2]$.
- The periodogram ordinates at Fourier frequencies are asymptotically uncorrelated with correlations converging to zero uniformly:

$$\big|cov\big(I_{n,X}(\lambda_j), I_{n,X}(\lambda_l)\big)\big| \le C_1 n^{-1} \tag{4.137}$$

  with some finite constant C_1.
-

$$\left(\frac{I_{n,X}(\lambda_{j_1})}{f_X(\lambda_{j_1})}, \ldots, \frac{I_{n,X}(\lambda_{j_k})}{f_X(\lambda_{j_k})}\right) \xrightarrow[d]{} (Z_1, \ldots, Z_k), \tag{4.138}$$

  where $Z_1, \ldots, Z_k$ are i.i.d. standard exponential random variables, and λ_{j_1}, $\ldots, \lambda_{j_k}$ are distinct Fourier frequencies.

On the other hand, it will be shown in a subsequent section that these properties are no longer valid for linear time series with long memory.

Of course, the main tool to establish (4.137) and (4.138) is Lemma 4.22. Note that (cf. Gradshteyn and Rhyzhik 1965, p. 414) $\int_{-\pi}^{\pi} K_n(\lambda_j - \lambda)\, d\lambda = 1$. Thus, if $X_t = \varepsilon_t$ is a centred i.i.d. sequence, then $f_\varepsilon(\lambda) = \sigma_\varepsilon^2/(2\pi)$, and hence,

$$E\left(\frac{I_{n,\varepsilon}(\lambda_j)}{f_\varepsilon(\lambda_j)}\right) = 1 \quad \big(j = 1, \ldots, [n/2]\big), \tag{4.139}$$

independently of the chosen Fourier frequency λ_j. This justifies (4.137) for an i.i.d. sequence. It should be mentioned, though, that this equality is valid at Fourier frequencies only. Furthermore, if ε_t ($t \in \mathbb{Z}$) are i.i.d. with mean zero and variance σ_ε^2, then we have, for distinct Fourier frequencies λ_k, λ_l ($k \neq l$),

$$E\big[d_{n,\varepsilon}(\lambda_k)\,\overline{d_{n,\varepsilon}(\lambda_l)}\,\big] = \frac{\sigma_\varepsilon^2}{2\pi} \sum_{t=1}^{n} e^{it(\lambda_k - \lambda_l)} = 0. \tag{4.140}$$

If in addition the random variables ε_t are standard Gaussian, then the discrete Fourier transform at different Fourier frequencies is also jointly Gaussian and hence independent. Consequently, the periodogram ordinates $I_{n,\varepsilon}(\lambda_j) = |d_{n,\varepsilon}(\lambda_j)|^2$ computed at distinct Fourier frequencies are independent. Moreover, $2\pi\, I_{n,\varepsilon}(\lambda_j)$ ($j = 1, \ldots, N_n$; $N_n = [(n-1)/2]$) have a standard exponential distribution. In particular,

$$E\big[2\pi\, I_{n,\varepsilon}(\lambda_j)\big] = 1, \qquad \mathrm{var}\big(2\pi\, I_{n,\varepsilon}(\lambda_j)\big) = 1. \tag{4.141}$$

If the random variables ε_t are not Gaussian, then $d_{n,\varepsilon}(\lambda_k)$, $d_{n,\varepsilon}(\lambda_l)$ are uncorrelated (i.e. (4.140) still holds), but they are no longer independent. For the periodogram, we have

$$cov\big(I_{n,\varepsilon}(\lambda_k),\, I_{n,\varepsilon}(\lambda_l)\big) = \frac{\kappa_4}{4\pi^2 n}, \tag{4.142}$$

where κ_4 is the fourth cumulant. Note that in the Gaussian case $\kappa_4 = 0$. Nevertheless, the periodogram ordinates are *asymptotically* independent and have the standard exponential distribution. This way one obtains (4.138).

4.6.2.3 Linear Long-Memory Sequences

Properties (4.136), (4.137) and (4.138) are not valid in the case of linear process with long memory. The behaviour of the periodogram at frequencies converging to zero can be formulated as follows (Künsch 1986; Hurvich and Beltrao 1993, 1994a, 1994b; Robinson 1995a):

Theorem 4.31 *Let $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ be a second-order stationary linear process and assume that $f_X(\lambda) \sim c_f |\lambda|^{-2d}$ as $|\lambda| \to 0$ with $d \in (0, 1/2)$. Define*

$$\mu(j; d) = |2\pi j|^{2d} \frac{2}{\pi} \int_{-\infty}^{\infty} \frac{\sin^2(\lambda/2)}{(2\pi j - \lambda)^2} |\lambda|^{-2d} \, d\lambda.$$

Then for any fixed positive integer j,

$$\lim_{n\to\infty} E\left[\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)}\right] = \mu(j; d).$$

Proof We use Lemma 4.22. Using the assumption $f_X(\lambda) \sim c_f |\lambda|^{-2d}$, we have

$$\begin{aligned} E\left(\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)}\right) &= \frac{1}{n} \int_{-n\pi}^{n\pi} K_n\left(\frac{2\pi j}{n} - \frac{\lambda}{n}\right) \frac{f_X(\lambda/n)}{f_X(2\pi j/n)} \, d\lambda \\ &\approx \left(\frac{2\pi j}{n}\right)^{2d} \frac{1}{n} \int_{-n\pi}^{n\pi} K_n\left(\frac{2\pi j - \lambda}{n}\right) \left|\frac{\lambda}{n}\right|^{-2d} d\lambda \\ &= \frac{1}{n} \int_{-n\pi}^{n\pi} K_n\left(\frac{2\pi j - \lambda}{n}\right) \left|\frac{2\pi j}{\lambda}\right|^{2d} d\lambda. \end{aligned} \tag{4.143}$$

It is easy to see that, as $n \to \infty$, the functions

$$\begin{aligned} g_n(\lambda) &:= \frac{1}{n} K_n\left(\frac{2\pi j - \lambda}{n}\right) \left|\frac{2\pi j}{\lambda}\right|^{2d} \\ &= \frac{1}{2\pi n^2} \frac{\sin^2(\frac{2\pi j - \lambda}{2})}{\sin^2(\frac{2\pi j - \lambda}{2n})} \left|\frac{2\pi j}{\lambda}\right|^{2d} \end{aligned}$$

converge pointwise to

$$\left|\frac{2\pi j}{\lambda}\right|^{2d} \frac{2}{\pi} \frac{\sin^2(\lambda/2)}{(2\pi j - \lambda)^2}.$$

Thus,

$$\lim_{n\to\infty} E\left(\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)}\right) = |2\pi j|^{2d} \frac{2}{\pi} \int_{-\infty}^{\infty} \frac{\sin^2(\lambda/2)}{(2\pi j - \lambda)^2} |\lambda|^{-2d} \, d\lambda,$$

given that we can exchange limit with integration (which follows from Lebesgue dominated convergence) and that integration over $(-\infty, -n\pi) \cup (n\pi, \infty)$ is negligible. □

Detailed calculations can be found in Hurvich and Beltrao (1993). The authors considered a more general spectral density $f_X(\lambda) = |\lambda|^{-2d} f_*(\lambda)$ with a smooth function f_*. In fact, this computation is valid for $d \in (-0.5, 1.5)$; however, if $d > 0.5$, f_X is not a spectral density since the model is not stationary (Hurvich

and Ray 1995). What is important here is that the normalized periodogram at Fourier frequencies depends on both j and d, as opposed to the i.i.d. case described in (4.139).

Furthermore, using the same argument as for the mean, Hurvich and Beltrao (1993) argue that for any two integers $l \neq k$,

$$\lim_{n\to\infty} E\left[\frac{d_{n,X}(\lambda_k)\,\overline{d_{n,X}(\lambda_l)}}{\sqrt{f_X(\lambda_k)\,f_X(\lambda_l)}}\right] =: \gamma_w(l,k;d),$$

where

$$\gamma_w(l,k;d) = (-1)^{l+k+1}|2\pi k|^d|2\pi l|^d\frac{2}{\pi}\int_{-\infty}^{\infty}\frac{\sin^2(\lambda/2)}{(2\pi k-\lambda)(2\pi l+\lambda)}|\lambda|^{-2d}\,d\lambda.$$

Furthermore, if the random variables X_t are Gaussian, then

$$\lim_{n\to\infty} cov\left(\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)},\frac{I_{n,X}(\lambda_k)}{f(\lambda_k)}\right) = \gamma_w^2(j,k;d)+\gamma_w^2(j,-k;d)\quad (j\neq k),$$

$$\lim_{n\to\infty} \operatorname{var}\left(\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)}\right) = 2\gamma_w^2(j,j;d).$$

Thus, unlike the i.i.d. case, the DFTs and the normalized periodogram ordinates are not asymptotically independent.

4.6.2.4 Refined Covariance Bounds for Long-Memory Sequences

One can obtain the following asymptotic independence of the DFT and periodogram ordinates if the Fourier frequencies λ_j are not too close to zero.

Recall that $f_X(\lambda) \sim c_f|\lambda|^{-2d}$ and let

$$d^0_{n,X}(\lambda) = \frac{d_{n,X}(\lambda)}{\sqrt{c_f\lambda^{-2d}}}$$

and $\gamma_X(k) = cov(X_t, X_{t+k})$. Then the following holds.

Theorem 4.32 *Let $X_t = \sum_{j=0}^{\infty} a_j\varepsilon_{t-j}$ be a second-order stationary linear process with*

$$f_X(\lambda) = \left|1-\exp(-i\lambda)\right|^{-2d} f_*(\lambda) \approx |\lambda|^{-2d} f_*(\lambda) \approx c_f|\lambda|^{-2d} \tag{4.144}$$

and such that

$$f_X(\lambda) = c_f|\lambda|^{-2d} + O\left(\lambda^{\rho-2d}\right) \tag{4.145}$$

for some $0 < \rho \le 2$ *and* $-\frac{1}{2} < d < \frac{1}{2}$. *Let* j_n, k_n *be positive integer-valued sequences such that* $j_n/n \to 0$ *and* $j_n > k_n$. *Then,*

$$\begin{aligned} \operatorname{var}\big(d^0_{n,X}(\lambda_{j_n})\big) &= E\big[d^0_{n,X}(\lambda_{j_n})\,\overline{d^0_{n,X}(\lambda_{j_n})}\,\big] \\ &= 1 + O\left(\frac{\log j_n}{j_n}\right) + O\left(\left(\frac{j_n}{n}\right)^{\rho}\right) \end{aligned} \tag{4.146}$$

and

$$cov\big(d^0_{n,X}(\lambda_{j_n}), d^0_{n,X}(\lambda_{k_n})\big) = O\left(\frac{\log j_n}{k_n}\right). \tag{4.147}$$

Before we proceed with the proof, we comment on assumption (4.145). This is a smoothness condition for f_*. For example, if $\rho = 2$, then f_* is twice differentiable in the neighbourhood of the origin. This type of condition is crucial in studying for example semiparametric estimators of d.

Proof The essential arguments can be seen by considering (4.146). Condition (4.145) implies

$$\begin{aligned} f_X(\lambda_j) - c_f\lambda_j^{-2d} &= f_X(\lambda_j)\left[1 - \left(\frac{f_X(\lambda_j)}{c_f\lambda_j^{-2d}}\right)^{-1}\right] \\ &= f_X(\lambda_j)\left[1 - \frac{1}{1 + O(\lambda_j^{\rho})}\right] \\ &= c_f\lambda_j^{-2d}\big[1 + O\big(\lambda_j^{\rho}\big)\big] = O\big(\lambda_j^{\rho-2d}\big), \end{aligned}$$

so that

$$\frac{f_X(\lambda_j)}{c_f\lambda_j^{-2d}} = 1 + O\left(\left(\frac{j}{n}\right)^{\rho}\right).$$

In a second step, one shows

$$E\big[d_{n,X}(\lambda_j)\,\overline{d_{n,X}(\lambda_j)}\,\big] = f_X(\lambda_j) + O\left(\lambda_j^{-2d}\frac{\log j}{j}\right), \tag{4.148}$$

so that

$$E\left[\frac{d_{n,X}(\lambda_j)\,\overline{d_{n,X}(\lambda_j)}}{f_X(\lambda_j)}\right] = 1 + O\left(\frac{\log j}{j}\right).$$

To show (4.148), we use the general formula for the covariance of DFT; see (4.135). Since K_n is 2π-periodic with $\int_{-\pi}^{\pi} K_n(u)\,du = 1$, we obtain

$$E\big[d_{n,X}(\lambda_j)\,\overline{d_{n,X}(\lambda_j)}\,\big] - f_X(\lambda_j) = \int_{-\pi}^{\pi}\big[f_X(\lambda) - f_X(\lambda_j)\big]K_n(\lambda - \lambda_j)\,d\lambda. \tag{4.149}$$

Now, for n large enough, λ_j is smaller than $\delta/2$, so that

$$f_X(\lambda_j) \le c_\delta \lambda_j^{-2d}, \qquad \big|f_X'(\lambda_j)\big| \le c_\delta \lambda_j^{-2d-1}$$

for a suitable finite constant c_δ. Noting that $K_n(u) = O(n^{-1})$ for $\delta/2 < u \le \pi$, we obtain

$$\begin{aligned}\int_{|\lambda|\ge\delta} \big|f_X(\lambda) - f_X(\lambda_j)\big| K_n(\lambda-\lambda_j)\,d\lambda &\le O\big(n^{-1}\big)\cdot\left[\int_{-\pi}^{\pi} f_X(\lambda)\,d\lambda + 2\pi c_\delta \lambda_j^{-2d}\right] \\ &= O\big(n^{-1}\big) + O\big(n^{-1}\lambda_j^{-2d}\big).\end{aligned}$$

For $0 < d < \frac{1}{2}$, this is of order $O((j/n)^{1-2d}\cdot j^{-1}) = o(j^{-1}\log j)$. Similarly, for $-\frac{1}{2} < d < \frac{1}{2}$, the overall order is $O(n^{-1}) = O((j/n)j^{-1}) = o(j^{-1}\log j)$. Therefore, the only relevant range of integration in (4.143) is $-\delta \le \lambda \le \delta$. There are two asymptotic poles that are approached asymptotically on the right-hand side of (4.149): a pole in f_X for $\lambda_j \to 0$ and an asymptotic singularity in $K_n(\lambda-\lambda_j)$ for $\lambda = \lambda_j$. The largest order is obtained for the integral over $\Delta_n = [\frac{1}{2}\lambda_j, 2\lambda_j]$. There, we have

$$\begin{aligned}&\int_{\lambda\in\Delta_n} \big|f_X(\lambda) - f_X(\lambda_j)\big| K_n(\lambda-\lambda_j)\,d\lambda \\ &\quad\le \max_{\lambda_j/2\le\lambda\le 2\lambda_j} \big|f_X'(\lambda)\big| \underbrace{\int_{\lambda_j/2}^{2\lambda_j} |\lambda-\lambda_j| K(\lambda-\lambda_j)\,d\lambda}_{J(\lambda_j)} = O\big(\lambda_j^{-1-2d}\big)\cdot J(\lambda_j).\end{aligned}$$

Since $|D_n(u)| \le 2|u|^{-1}$ $(0 < |u| < \pi)$, we have

$$\int_{-c\lambda_j}^{c\lambda_j} \big|D_n(\lambda)\big|\,d\lambda = O(\log j)$$

for any fixed $c > 0$. Moreover, $\lim_{\lambda\to\lambda_j} |\lambda-\lambda_j| K(\lambda-\lambda_j) = 0$, and we obtain

$$\begin{aligned}|\lambda-\lambda_j| K(\lambda-\lambda_j) &\le (2\pi n)^{-1} |\lambda-\lambda_j| \cdot 2|\lambda-\lambda_j|^{-1} \cdot \big|D_n(\lambda-\lambda_j)\big| \\ &= \pi^{-1} n^{-1} \big|D_n(\lambda-\lambda_j)\big|,\end{aligned}$$

and thus,

$$J(\lambda_j) = O\big(n^{-1}\log j\big).$$

Putting the orders together, we have

$$\begin{aligned}\int_{\lambda\in\Delta_n} \big|f_X(\lambda) - f_X(\lambda_j)\big| K_n(\lambda-\lambda_j)\,d\lambda &= O\big(\lambda_j^{-1-2d}\cdot n^{-1}\log j\big) \\ &= O\left(\lambda_j^{-2d}\cdot\frac{\log j}{j}\right),\end{aligned}$$

as required in (4.148). □

4.6.3 Limiting Distribution

4.6.3.1 Fourier Transform and Periodogram for Long-Memory Sequences

Now, we will describe the limiting distribution for the DFT and the periodogram ordinates. Let us write $d_{n,X}(\lambda_j) = A(\lambda_j) + iB(\lambda_j)$, where

$$A(\lambda) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t \cos(t\lambda), \qquad B(\lambda) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t \sin(t\lambda).$$

Then $I_{n,X}(\lambda_j) = A^2(\lambda_j) + B^2(\lambda_j)$. Assume for simplicity that X_t is a Gaussian process. It follows from (4.147) that for each fixed K,

$$\left(\frac{d_{n,X}(\lambda_j)}{\sqrt{f_X(\lambda_j)}}, j = 1, \ldots, K \right)$$

converges to a multivariate Gaussian distribution with *dependent* components and covariance matrix $[\gamma_w(l,k;d)]_{k,l=1,\ldots,K}$. Furthermore, for each fixed j, the cosine and the sine parts $A(\lambda_j)$ and $B(\lambda_j)$ are uncorrelated with *different variances*. Therefore,

$$\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} = \frac{A^2(\lambda_j)}{f_X(\lambda_j)} + \frac{B^2(\lambda_j)}{f_X(\lambda_j)} \xrightarrow[d]{} a\chi_1^2(1) + b\chi_1^2(2), \tag{4.150}$$

where a, b are constants, and $\chi_1^2(j)$, $j = 1, 2$, are independent χ^2 random variables with one degree of freedom. Thus, in contrast to the i.i.d. case, the normalized periodogram ordinates have a different asymptotic distribution at each frequency. Moreover, the limiting distribution has dependent components.

4.6.3.2 Sum of Periodogram Ordinates

Let ϕ be a deterministic, real-valued function and consider the partial sum

$$S_{n,X}(\phi) = \sum_{j=1}^{N_n} \phi\big(I_{n,X}(\lambda_j)\big),$$

where $N_n = [(n-1)/2]$. If $X_t = \varepsilon_t$ are i.i.d., then (cf. (4.141))

$$\operatorname{var}\left(\sum_{j=1}^{N_n} 2\pi I_{n,\varepsilon}(\lambda_j) \right) \approx n(1 + \kappa_4/2).$$

Also,

$$n^{-1/2} \sum_{j=1}^{N_n} 2\pi I_{n,\varepsilon}(\lambda_j) \xrightarrow[d]{} N(0, 1 + \kappa_4/2).$$

These asymptotic results are obvious when ε_t are Gaussian since the periodogram ordinates are independent. If $\phi = \log$ and ε_t are Gaussian, then

$$\operatorname{var}\big(\log\big(2\pi I_{n,\varepsilon}(\lambda_j)\big)\big) = \operatorname{var}\big(\log\big(I_{n,\varepsilon}(\lambda_j)/f_\varepsilon(\lambda_j)\big)\big) = \operatorname{var}\big(\log(Z)\big),$$

where Z is standard exponential. We compute

$$\begin{aligned}\operatorname{var}(\log Z) &= \int_0^\infty e^{-x}(\log x)^2\,dx - \left[\int_0^\infty e^{-x}(\log x)\,dx\right]^2 \\ &= \left(\frac{\pi^2}{6} + \eta^2\right) - (-\eta)^2 = \frac{\pi^2}{6}. \end{aligned} \tag{4.151}$$

Therefore, in the Gaussian i.i.d. case,

$$n^{-1/2}\sum_{j=1}^{N_n} \log\big(2\pi I_{n,\varepsilon}(\lambda_j)\big) \underset{d}{\to} N\big(0, \pi^2/6\big).$$

In the long-memory case, the periodogram ordinates are asymptotically dependent, so that these convergence results are not valid. However, for a proper choice of asymptotically negligible constants $c_{n,k}$, it is possible to obtain asymptotic normality of $\sum c_{n,k}\phi(I_{n,X}(\lambda_k))$ regardless whether X_t is weakly or strongly dependent. We will illustrate this in the context of semiparametric estimation of the long-memory parameter d.

4.7 Limit Theorems for Wavelets

4.7.1 Introduction

In this section we discuss limit theorems for the discrete wavelet transform of long-memory stochastic processes. We refer to Sect. 3.5 for basic definitions of wavelets. At this point we recall that for a scaling function ϕ and a wavelet function ψ, dilated and translated functions are defined as

$$\phi_{j,k}(x) = 2^{j/2}\phi\big(2^j x - k\big), \qquad \psi_{j,k}(x) = 2^{j/2}\psi\big(2^j x - k\big).$$

However, it is not necessary that the wavelet functions are constructed using the multiresolution analysis, nor that they are orthogonal.

4.7.2 Discrete Wavelet Transform of Stochastic Processes

Assume first that $Y(u)$ $(u \in \mathbb{R})$ is a continuous-time stochastic process. Define

$$d^Y_{j,k} = \int_{\mathbb{R}} Y(u)\psi_{j,k}(u)\,du, \qquad a^Y_{j,k} = \int_{\mathbb{R}} Y(u)\phi_{j,k}(u)\,du \quad (j,k \in \mathbb{Z}).$$

In other words, $d^Y_{j,k}$ and $a^Y_{j,k}$ are (random) wavelet coefficients of the continuous-time process $Y(u)$ $(u \in \mathbb{R})$. If the continuous-time process has mean zero, then clearly $E(d_{j,k}) = 0$ for each j, k. For simplicity, we write in the following $a_{j,k}$, $d_{j,k}$ instead of $a^Y_{j,k}$, $d^Y_{j,k}$.

Assume further that $Y(u)$ $(u \in \mathbb{R})$ has stationary increments. For each fixed resolution level j, the process $d_{j,k}$ $(k \in \mathbb{Z})$ is stationary. Indeed, we may verify, for instance, that the marginal distributions are invariant under translation: the random coefficient

$$\begin{aligned} d_{j,k+l} &= \int Y(u)\psi_{j,k+l}(u)\,du = \int Y(u+l)\psi_{j,k}(u)\,du \\ &= \int \big(Y(u+l) - Y(l)\big)\psi_{j,k}(u)\,du \end{aligned}$$

is equal in distribution to

$$\int \big(Y(u) - Y(0)\big)\psi_{j,k}(u)\,du = \int Y(u)\psi_{j,k}(u)\,du = d_{j,k}.$$

The same applies to the scaling coefficients $a_{j,k} = \int Y(u)\phi_{j,k}(u)\,du$. A more rigorous proof of stationarity can be found in e.g. Houdré (1994). See also Masry (1993) and Cambanis and Houdré (1995) for the DWT of stochastic processes.

If moreover, the process $Y(u)$ is H-self-similar, then for each j, k,

$$d_{j,k} \overset{\mathrm{d}}{=} 2^{-j(H+1/2)} d_{0,k}.$$

Indeed, heuristically,

$$\begin{aligned} d_{j,k} &= \int Y(u)\psi_{j,k}(u) = 2^{j/2} \int Y(u)\psi\big(2^j u - k\big)\,du \\ &= 2^{-j/2} \int Y\big(2^{-j}u\big)\psi(u-k)\,du \overset{\mathrm{d}}{=} 2^{-j/2}2^{-jH} \int Y(u)\psi_{j,k}(u) \\ &= 2^{-j(H+1/2)} d_{0,k}. \end{aligned}$$

Hence, if the continuous-time process $Y(u)$ $(u \in \mathbb{R})$ is self-similar with stationary increments (H-SSSI), then

$$E\big[d^2_{j,k+l}\big] = 2^{-j(2H+1)} E\big[d^2_{0,k}\big] = 2^{-j(2H+1)} E\big[d^2_{0,0}\big].$$

This applies, in particular, to fractional Brownian motion. As we will see later, these formulas can be used to define a wavelet-based estimator of the self-similarity parameter H.

4.7.3 Second-Order Properties of Wavelet Coefficients

Now, we turn our attention to stationary processes $X(u)$ $(u \in \mathbb{R})$. For example, $X(u) = Y(u) - Y(u-1)$ $(u \in \mathbb{R})$ can be defined as increments of the H-SSSI process considered above. Define analogously wavelet and scaling coefficients:

$$d_{j,k} = d_{j,k}^X = \int_{\mathbb{R}} X(u)\psi_{j,k}(u)\,du,$$
$$a_{j,k} = a_{j,k}^X = \int_{\mathbb{R}} X(u)\phi_{j,k}(u)\,du \quad (j,k \in \mathbb{Z}).$$

Then $d_{j,k}$ and $a_{j,k}$ $(k \in \mathbb{Z})$ form stationary sequences. We verify for instance that the marginal distributions are shift-invariant: for $l \in \mathbb{Z}$, we have

$$\begin{aligned} d_{j,k+l} &= \int_{-\infty}^{\infty} X(u)\psi_{j,k+l}(u)\,du = 2^{j/2}\int_{-\infty}^{\infty} X(u)\psi\big(2^j u - (k+l)\big)\,du \\ &= 2^{j/2}\int_{-\infty}^{\infty} X\big(v + 2^{-j}l\big)\psi\big(2^j v - k\big)\,dv \stackrel{\mathrm{d}}{=} 2^{j/2}\int_{-\infty}^{\infty} X(v)\psi\big(2^j v - k\big)\,dv \\ &= d_{j,k}. \end{aligned}$$

Hence, we can analyse the covariance structure of the stationary sequence $d_{j,k}$ $(k \in \mathbb{Z})$. Assume that the process $X(u)$ $(u \in \mathbb{R})$ is centred, has the covariance function $\gamma_X(s)$ $(s \in \mathbb{R})$ and the spectral density

$$f_X(\lambda) = \int_{-\infty}^{\infty} \gamma_X(s)e^{-i\lambda s}\,ds.$$

Assume further that

$$f_X(\lambda) = \lambda^{-2d} f_*(\lambda), \quad \lambda \to 0,$$

where $\lim_{\lambda\to 0} f_*(\lambda) = c_f \in (0,\infty)$ and $d \in [0, 1/2)$. For example, $X(u)$ could be fractional Gaussian noise, i.e. increments of fractional Brownian motion with Hurst parameter $H = d + \frac{1}{2}$.

One of the most intriguing properties of DWT is the *decorrelation* (*whitening*) *property*. Specifically, if the wavelet ψ has M vanishing moments, then we will argue below that

$$cov(d_{j,0}, d_{j,k}) = O\big(k^{-2M+2d-1}\big) \quad (k \to \infty).$$

That is, the stationary sequence $d_{j,k}$ $(k \in \mathbb{Z})$ is *weakly* dependent (i.e. has summable covariances) if $M \geq 1$. For example, the *whitening property* applies to fractional Gaussian noise $X(u) = B_H(u) - B_H(u-1)$, where $B_H(u)$ is a fractional Brownian motion with Hurst parameter $H \in (1/2, 1)$. This phenomenon is discussed for instance in Flandrin (1992), Tewfik and Kim (1992), Abry et al. (1998) or Mielniczuk and Wojdyłło (2007a).

To justify the whitening property, recall that

$$\hat{\psi}(\lambda) = \int_{-\infty}^{\infty} \psi(x) e^{-i\lambda x}\, dx$$

is the Fourier transform of ψ. Hence,

$$\begin{aligned}\hat{\psi}_{j,k}(\lambda) &= \int_{-\infty}^{\infty} e^{-i\lambda x} \psi_{j,k}(x)\, dx = 2^{j/2} \int_{-\infty}^{\infty} e^{-i\lambda x} \psi\big(2^j x - k\big)\, dx \\ &= 2^{-j/2} e^{-i2^{-j}\lambda k} \int_{-\infty}^{\infty} e^{-i\lambda 2^{-j} x} \psi(x)\, dx = 2^{-j/2} e^{-i2^{-j}\lambda k} \hat{\psi}\big(2^{-j}\lambda\big).\end{aligned}$$

We can then evaluate covariance structure of the wavelet coefficients of the process $X(\cdot)$ as

$$\begin{aligned}&cov(d_{j,k}, d_{j',k'}) \\ &\quad = \int\int \gamma_X(v-u) \psi_{j,k}(v) \psi_{j',k'}(u)\, du\, dv \\ &\quad = \int_{-\infty}^{\infty} f_X(\lambda) \hat{\psi}_{j,k}(\lambda) \hat{\psi}_{j',k'}(\lambda)\, d\lambda \\ &\quad = 2^{-j/2} 2^{-j'/2} \int_{-\infty}^{\infty} f_X(\lambda) \hat{\psi}\big(2^{-j}\lambda\big) \overline{\hat{\psi}\big(2^{-j'}\lambda\big)} e^{-i2^{-j}\lambda k} e^{i2^{j'}\lambda k'}\, d\lambda. \qquad (4.152)\end{aligned}$$

This formula is crucial to evaluate the variance and covariance structure of the wavelet coefficients for stochastic processes with long memory. A change of variables $\omega = 2^{-(j+j')/2}\lambda$,

$$\lambda = 2^{(j+j')/2}\omega,$$

and the form $f(\lambda) = \lambda^{-2d} f_*(\lambda)$ of the spectral density yield

$$\begin{aligned}&cov(d_{j,k}, d_{j\prime,k\prime}) \\ &\quad = \int_{-\infty}^{\infty} f_X\big(2^{(j+j')/2}\omega\big) \hat{\psi}\big(2^{(j'-j)/2}\omega\big) \overline{\hat{\psi}\big(2^{(j-j')/2}\omega\big)} e^{-i2^{(j-j')/2}\omega k} e^{i2^{(j'-j)/2}\omega k'}\, d\omega \\ &\quad = 2^{-(j+j')d} \int_{-\infty}^{\infty} \omega^{-2d} f_*\big(2^{(j+j')/2}\omega\big) \hat{\psi}\big(2^{(j'-j)/2}\omega\big) \overline{\hat{\psi}\big(2^{(j-j')/2}\omega\big)} e^{-ir\omega}\, d\omega,\end{aligned}$$

where

$$r = \big|2^{(j-j')/2}k - 2^{(j'-j)/2}\omega k'\big|.$$

When $j, j' \to -\infty$ (i.e. we are considering coarse resolution levels or "low frequencies"), then $2^{(j+j')/2}\omega \to 0$, so that

$$f_*\big(2^{(j+j')/2}\omega\big) \sim f_*(0) = c_f.$$

This motivates the following definition:

$$\Psi_{j,j'}(k,k') := \int_{-\infty}^{\infty} \omega^{-2d} \hat{\psi}\big(2^{(j'-j)/2}\omega\big)\overline{\hat{\psi}\big(2^{(j-j')/2}\omega\big)} e^{-ir\omega}\, d\omega. \tag{4.153}$$

We note that if $j \neq j'$, $k \neq k'$ and $d = 0$, then, due to orthogonality, the covariances vanish *if the wavelet family $\psi_{j,k}$ is constructed using the MRA*. As we will see below, in the case of long memory, orthogonality of wavelets is not crucial at all. The most important property is the number M of vanishing moments of the wavelet function ψ.

To see this, let $d > 0$ and consider $j = j'$ and $k' = 0$. Then

$$cov(d_{j,0}, d_{j,k}) = 2^{-2jd} \int \omega^{-2d} f_*\big(2^j\omega\big)\big|\hat{\psi}(\omega)\big|^2 e^{-ik\omega}\, d\omega.$$

Again, as $j \to -\infty$, we approximate this integral as

$$cov(d_{j,0}, d_{j,k}) = 2^{-2jd} f_*(0) \int \omega^{-2d}\big|\hat{\psi}(\omega)\big|^2 e^{-ik\omega}\, d\omega.$$

Next, recall now from Sect. 3.5 that if the wavelet function ψ has M vanishing moments, then

$$\big|\hat{\psi}(\lambda)\big| = \big|\hat{\psi}^{(M)}(0)\big||\lambda|^M + o\big(|\lambda|^M\big) \quad (\lambda \to 0).$$

Thus, if k is large enough, then we have to analyse the following integral in a neighbourhood $(-\varepsilon/k, \varepsilon/k)$ of the origin:

$$2^{-2jd} c_f \big\{\hat{\psi}^{(M)}(0)\big\}^2 \int_{\varepsilon/k}^{\varepsilon/k} \omega^{-2d}\omega^{2M} e^{-ik\omega}\, d\omega.$$

The change of variables $\lambda = k\omega$ yields the approximation

$$2^{-2jd} c_f \big\{\hat{\psi}^{(M)}(0)\big\}^2 k^{-2M+2d-1} \int_{-\varepsilon}^{\varepsilon} \lambda^{2M-2d} e^{-i\lambda}\, d\lambda.$$

The integral is finite as long as $2M - 2d > -1$. Of course, in these computations several simplifications and informal approximations are used. Nevertheless, we have obtained heuristically the following *decorrelation property*.

Lemma 4.23 *Assume that $X(u)$ $(u \in \mathbb{R})$ is a stationary centred process such that its spectral density is given by $f_X(\lambda) = |\lambda|^{-2d} f_*(\lambda)$, $\lambda \in \mathbb{R}$, $d \in (0, 1/2)$ and $\lim_{\lambda\to 0} f_*(\lambda) = c_f \in (0, \infty)$. Then for each $j \in \mathbb{Z}$,*

$$cov(d_{j,0}, d_{j,k}) = O\big(k^{-2M+2d-1}\big) \quad (k \to \infty).$$

The same result carried over to series X_t $(t \in \mathbb{Z})$ in discrete time, when transformed into their continuous-time versions as discussed in the introduction to

wavelets. In particular, the restrictions $d < \frac{1}{2}$ and $M \geq 1$ imply that we always have $cov(d_{j,0}, d_{j,k}) = o(k^{-2})$. This means that

$$\sum_{k=-\infty}^{\infty} \left|cov(d_{j,0}, d_{j,k})\right| < \infty$$

and the wavelet coefficients $d_{j,k}$ $(k \in \mathbb{Z})$ are weakly dependent. Moreover, if the process $X(u)$ $(u \in \mathbb{R})$ is Gaussian, then the wavelet coefficients are Gaussian as well. Also, in the Gaussian case we have

$$cov\big(d_{j,0}^2, d_{j,k}^2\big) = 2cov^2(d_{j,0}, d_{j,k}),$$

so that these autocovariances converge as well.

As indicated above, a very useful property is also (4.153) because for large enough scales, i.e. for $j,\ j' \to -\infty$,

$$cov(d_{j,k}, d_{j',k'}) \approx 2^{-(j+j')d} f_*(0) \Psi_{j,j'}\big(k, k'\big).$$

Thus, the weak dependence extends to the wavelet coefficients at different resolution levels $j \neq j'$.

To evaluate the variance of $d_{j,k}$, set $j = j', k = k'$ in (4.152). Then

$$\begin{aligned}\sigma_j^2 := \operatorname{var}(d_{j,k}) &= 2^{-j} \int f_X(\lambda) \big|\hat{\psi}\big(2^{-j}\lambda\big)\big|^2 d\lambda \\ &= 2^{-2jd} \int |\lambda|^{-2d} f_*\big(2^j\lambda\big) \big|\hat{\psi}(\lambda)\big|^2 d\lambda.\end{aligned}$$

Again, we approximate $f_*(2^j\lambda) \approx f_*(0) = c_f$ (for $j \to -\infty$) and hence

$$\operatorname{var}(d_{j,k}) \approx 2^{-2jd} c_f \int |\lambda|^{-2d} \big|\hat{\psi}(\lambda)\big|^2 d\lambda =: 2^{-2jd} c_f \Psi(2d), \tag{4.154}$$

where

$$\Psi(\gamma) = \int \lambda^{-\gamma} \big|\hat{\psi}(\lambda)\big|^2 d\lambda.$$

This heuristic approximation has been derived in Abry et al. (1998). More precise bounds have been obtained in Lemma 1 in Bardet et al. (2000) or Theorem 1 in Moulines et al. (2007a). A bound that requires a semiparametric assumption on the spectral density similar to the one used for the DFT is for instance:

Lemma 4.24 *Assume that for some* $d \in (0, 1/2)$,

$$f_X(\lambda) = \lambda^{-2d}\big(f_*(0) + O\big(|\lambda|^{\rho}\big)\big).$$

Under appropriate regularity conditions, we have, as $j \to -\infty$,

$$\left|\operatorname{var}(d_{j,k}) - 2^{-2jd} c_f \Psi(2d)\right| \leq 2^{-2jd} 2^{j\rho} \Psi(2d - \rho).$$

Proof In the proof, we omit several details, referring to the papers mentioned above. We note that

$$\left|\operatorname{var}(d_{j,k}) - 2^{-2jd} c_f \Psi(2d)\right| \le 2^{-2jd} \int |\lambda|^{-2d} \left|\left\{f_*\left(2^j\lambda\right) - f_*(0)\right\}\right| \left|\hat{\psi}(\lambda)\right|^2 d\lambda.$$

Under the assumption

$$f_*(\lambda) = |\lambda|^{-2d}\left(f_*(0) + O\left(|\lambda|^{\rho}\right)\right),$$

the bound is

$$2^{-2jd} \int |\lambda|^{-2d} \left\{2^j\lambda\right\}^{\rho} \left|\hat{\psi}(\lambda)\right|^2 d\lambda = 2^{-2jd} 2^{j\rho} \Psi(2d-\rho). \qquad \square$$

4.8 Limit Theorems for Empirical and Quantile Processes

4.8.1 Linear Processes with Finite Moments

The empirical distribution function plays an essential role in statistical inference. Many statistics that are concerned with inference for the marginal distribution of a process can be written as functionals of the (marginal) empirical distribution function $F_n(x)$. Therefore, in principle, their distribution follows "automatically", once the empirical distribution function is characterized asymptotically. Sometimes, the functionals are quite involved however so that the derivation requires some additional work. Relatively simple functionals occur for instance in goodness-of-fit tests, and even more directly in quantile estimation. For obvious reasons, limiting results for quantile processes follow directly from those for the empirical distribution function.

Recall that for a stationary process X_t $(t \in \mathbb{Z})$ with marginal distribution function $F_X(x) = P(X \le x)$, a simple nonparametric estimator of F_X is the (marginal) empirical distribution function

$$F_{n,X}(x) = \frac{1}{n} \sum_{t=1}^{n} 1\{X_t \le x\} \quad (x \in \mathbb{R}). \tag{4.155}$$

Under very general assumptions (for example ergodicity of the sequence), $F_{n,X}$ is a uniformly consistent estimator of F_X, which means that, as $n \to \infty$,

$$\sup_{x\in\mathbb{R}} \left|F_{n,X}(x) - F_X(x)\right| \underset{p}{\to} 0. \tag{4.156}$$

Furthermore, if X_t $(t \in \mathbb{Z})$ are i.i.d., then the classical Donsker invariance principle states

$$\sqrt{n} E_{n,X}(x) := \sqrt{n}\left[F_{n,X}(x) - F_X(x)\right] \Rightarrow \tilde{B}\left(F_X(x)\right), \tag{4.157}$$

where $\Rightarrow$ denotes weak convergence in $D[0,\infty)$, and $\tilde{B}(u)$ $(u \in [0,1])$ is a Brownian bridge, i.e. $\tilde{B}(u) = B(u) - uB(u)$ where $B(u)$ is standard Brownian motion. In other words, the appropriately normalized empirical processes $E_{n,X}(x)$ converge weakly to the time-changed Brownian bridge. An analogous result, with the same normalizing rate but a different limiting process, holds for weakly dependent processes under very general conditions. The situation is quite different, however, under long memory. This can be seen as follows. The indicator function is a very specific transformation of X, i.e. we consider

$$G(X;x) = 1\{X \le x\} - F_X(x).$$

Let $p_X = F_X'$ be the density of X. With the function $y \to G(y;x)$ we can associate the Appell coefficients $a_{\mathrm{app},j}$ $(j \ge 1)$:

$$\begin{aligned} a_{\mathrm{app},j} &= (-1)^j \int G(y;x)\, p_X^{(j)}(y)\, dy \\ &= (-1)^j \left[\int_{-\infty}^{x} p_X^{(j)}(y)\, dy - F_X(x) \int_{-\infty}^{\infty} p_X^{(j)}(y)\, dy \right] \\ &= (-1)^j \int_{-\infty}^{x} p_X^{(j)}(y)\, dy = (-1)^j\, p_X^{(j-1)}(x). \end{aligned}$$

Furthermore, recall also (see Definition 4.1) that $G_\infty(y) = E[G(X+y)]$. Applying this to $G(y;x) = 1\{y \le x\}$, we obtain $G_\infty(y) = P(X \le x - y)$, and hence,

$$G_\infty^{(1)}(0) = -p_X(x-y)|_{y=0} = -p_X(x).$$

Therefore, the theory for partial sums of subordinated long-memory processes (considered e.g. in Sects. 4.2, 4.3) will imply the limiting behaviour for the empirical distribution $F_{n,X}(x)$ function when x is fixed.

The asymptotic behaviour of the empirical process based on long-memory linear processes with finite variance was studied in Dehling and Taqqu (1989b), Giraitis and Surgailis (1999), Ho and Hsing (1996), Giraitis et al. (1997), Wu (2003) and Csörgő et al. (2006), Csörgő and Kulik (2008a, 2008b). Here, we state the result under the assumptions that are needed to apply the martingale expansion technique of Ho and Hsing (1996) and Wu (2003), as considered in Theorem 4.9. When dealing with linear processes, this technique seems to be superior to the Appell expansion.

Theorem 4.33 *Let X_t $(t \in \mathbb{Z})$ be a linear process $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ with coefficients satisfying assumption* (B1), *i.e. $a_j \sim L_a(j) j^{d-1}$, $d \in (0, 1/2)$ (so that $\gamma_X(k) \sim L_\gamma(k) k^{2d-1}$). Also, assume that $E(|\varepsilon_1|^{4+\gamma}) < \infty$ for some $\gamma > 0$ and that p_ε, the density of the innovations, is such that*

$$\sup_{x \in \mathbb{R}} \left| p_\varepsilon^{(r)}(x) \right| + \int \left| p_\varepsilon^{(r)}(x) \right|^2 dx < \infty \quad (r = 0, 1, 2). \tag{4.158}$$

Then we have the uniform reduction principle

$$n^{\frac{1}{2}-d} L_1^{-\frac{1}{2}}(n) \sup_{x\in\mathbb{R}} \left| F_{n,X}(x) - F_X(x) + p_X(x)\bar{x} \right| \to_p 0. \tag{4.159}$$

Consequently,

$$n^{\frac{1}{2}-d} L_1^{-\frac{1}{2}}(n) \left[F_{n,X}(x) - F_X(x) \right] \Rightarrow p_X(x) Z, \tag{4.160}$$

where $L_1(n) = (d(2d+1))^{-1} L_\gamma(n)$, $\Rightarrow$ *denotes weak convergence in* $D(-\infty, \infty)$, *and* Z *is a standard normal random variable.*

Remark 4.4 Condition (4.158) implies that the same holds for the density p_X. In particular, the conditions on $p_X^{(1)}(x)$ and $p_X^{(2)}(x)$ are required to control a remainder term in the second-order expansion leading to (4.159). Note also that the assumptions of the theorem can be modified to $E(|\varepsilon_1|^{2+\gamma}) < \infty$ and

$$\left| E\left[\exp(is\varepsilon_1) \right] \right| \leq C \left(1 + |s| \right)^{\delta} \tag{4.161}$$

for some $\delta > 0$, $0 < C < \infty$. Condition (4.161) means in principle that p_X is infinitely often differentiable. These assumptions were used in Giraitis and Surgailis (1999). The authors were also able to deal with double-sided linear processes, however, at the cost of additional moment assumptions.

Remark 4.5 Under the conditions of Theorem 4.33, the finite-dimensional convergence in (4.160) follows directly from Theorem 4.9 and Corollary 4.3. Tightness is usually not proven directly, but rather follows from the reduction principle (4.159). For the latter, we refer to Dehling and Taqqu (1989b) or Csörgö, Szyszkowicz and Wang in the Gaussian case and to Ho and Hsing (1996) and Wu (2003) in the linear case.

Proof We repeat the martingale approximation argument presented before Theorem 4.9, adapting it to the indicator function $G(y; x) = 1\{y \leq x\}$. Recall that $\mathcal{F}_K = \sigma(\varepsilon_j, j \leq K)$ is the σ-algebra generated by ε_j $(j \leq K)$. We start with an orthogonal expansion of the indicator function,

$$1\{X_t \leq x\} - F_X(x) \underset{L_X^2(\Omega)}{=} \sum_{j=0}^{\infty} \zeta_t(j),$$

where

$$\zeta_t(j) = P(X_t \leq x | \mathcal{F}_{t-j}) - P(X_t \leq x | \mathcal{F}_{t-j-1}).$$

Note that $\zeta_t(0) = 1\{X_t \leq x\} - P(X_t \leq x | \mathcal{F}_{t-1})$. As before, the nice feature of this expansion is that, for fixed t, $\zeta_t(j)$ $(j = 0, 1, 2, \ldots)$ is a martingale difference, so that we indeed obtain orthogonality in the sense that for $j \neq j^*$,

$$\left\langle \zeta_t(j), \zeta_t\left(j^*\right) \right\rangle = cov\left(\zeta_t(j), \zeta_t\left(j^*\right) \right) = 0.$$

In more concrete terms, we have

$$P(X_i \le x|\mathcal{F}_{t-j}) = P\left(\sum_{s=0}^{j-1} a_s\varepsilon_{t-s} \le x - \sum_{s=j}^{\infty} a_s\varepsilon_{t-s}\right) = F_j(u_j),$$

where, given $\mathcal{F}_{t-j}$, the argument

$$u_j = x - \sum_{s=j}^{\infty} a_s\varepsilon_{t-s}$$

is fixed (of course, u_j depends on t as well, but this dependence is omitted). Similarly,

$$F_{j+1}(u_{j+1}) = P(X_t \le x|\mathcal{F}_{t-j-1}) = P\left(\sum_{s=0}^{j} a_s\varepsilon_{t-s} \le x - \sum_{s=j+1}^{\infty} a_s\varepsilon_{t-s}\right).$$

Note that $u_{j+1} = u_j - a_j\varepsilon_{t-j}$ and

$$\zeta_t(j) = F_j(u_j) - F_{j+1}(u_{j+1}).$$

A heuristic argument leads to the idea how one may obtain a linearization. We will use the notation $p_j(u) = F_j'(u)$ for the probability density function of $\sum_{s=0}^{j-1} a_s\varepsilon_{t-s}$ and $F_\varepsilon(y) = P(\varepsilon \le y)$. For $F_{j+1}(u_{j+1})$, we can write

$$F_{j+1}(u_{j+1}) = \int p_j(y)F_\varepsilon\big(q_j(x,y)\big)\,dy$$

with

$$q_j(x,y) = \frac{u_{j+1}(x) - y}{a_j}.$$

For the sake of argument, assume that $a_j > 0$ for j large enough. Since $a_j \to 0$ (as $j \to \infty$), we have $q_j \to \infty$ and $F_\varepsilon(q_j(x,y)) \to 1$ if $y < u_{j+1}(x)$. On the other hand, $q_j \to -\infty$ and $F_\varepsilon(q_j(x,y)) \to 0$, if $y > u_{j+1}$. Therefore, as $j \to \infty$,

$$F_{j+1}(u_{j+1}) \approx \int_{-\infty}^{u_{j+1}} p_j(y)\,dy = F_j(u_{j+1}).$$

Furthermore, using $u_j = u_{j+1} - a_j\varepsilon_{t-j}$ with $a_j\varepsilon_{t-j} \to 0$ in probability as $j \to \infty$, we obtain in first approximation

$$F_j(u_j) \approx F_j(u_{j+1}) - p_j(u_{j+1})a_j\varepsilon_{t-j},$$

so that

$$\begin{aligned}\zeta_t(j) &= F_j(u_j) - F_{j+1}(u_{j+1}) \\ &\approx \big[F_j(u_{j+1}) - p_j(u_{j+1})a_j\varepsilon_{t-j}\big] - F_j(u_{j+1}) \\ &= -p_j(u_{j+1})a_j\varepsilon_{t-j}.\end{aligned}$$

Finally, as $j \to \infty$, F_j converges to F_X (and p_j to f_X) and u_{j+1} to x, so that we may hope to obtain the following approximation:

$$\begin{aligned}F_{n,X}(x) - F_X(x) &= \frac{1}{n}\sum_{t=1}^{n}\big[1\{X_t \le x\} - F_X(x)\big] \\ &\approx \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{j=0}^{\infty} -p_j(u_{j+1})a_j\varepsilon_{t-j}\right) \\ &\approx -p_X(x)\frac{1}{n}\sum_{t=1}^{n}\left(\sum_{j=0}^{\infty} a_j\varepsilon_{t-j}\right) = -p_X(x)\bar{x}.\end{aligned}$$

A precise computation establishes the rate in (4.159). □

Taking into account higher-order terms in the Taylor expansions above, a complete orthogonal decomposition can be obtained:

$$F_{n,X}(x) - F_X(x) = \frac{1}{n}\sum_{t=1}^{n}\sum_{r=1}^{\infty}(-1)^k F_X^{(r)}(x)V_{t,r} \tag{4.162}$$

with

$$V_{t,r} = \sum_{0\le j_1<j_2<\cdots<j_r}^{\infty}\prod_{s=1}^{r} a_{j_s}\varepsilon_{t-j_s},$$

already defined in (4.51).

Theorem 4.33 is remarkable not only because of the slower rate of convergence under long memory, but also because the asymptotic process $p_X(x)Z$ (in x) is degenerate. The entire sample path is determined by one normal variable Z and a deterministic function $p_X(x)$. In other words, all sample paths have the shape of $p_X(x)$! This is in sharp contrast to the case of weak memory where the asymptotic process is proportional to a Brownian bridge (see (4.157) above).

The convergence (4.160) can be extended further. In addition to (4.158), assume that the condition holds with $r = 3$. Then the following holds:

- If $d \in (1/4, 1/2)$, then

$$n^{1-2d}L_2^{-1/2}(n)\big[F_{n,X}(x) - F_X(x) + p_X(x)\bar{x}\big] \Rightarrow p_X^{(1)}(x)Z_{2,H}(1), \tag{4.163}$$

where $Z_{2,H}(1)$ is the Hermite–Rosenblatt random variable, and $H = d + 1/2$.

- If $d \in (0, 1/4)$, then

$$\sqrt{n}\big[F_{n,X}(x) - F_X(x) + p_X(x)\bar{x}\big] \Rightarrow Z(x), \tag{4.164}$$

where $Z(\cdot)$ is a Gaussian process.

Essentially, these convergence results are very similar to the case of nonlinear functionals. The asymptotic behaviour of

$$F_{n,X}(x) - F_X(x) + p_X(x)\bar{x}$$

is determined by $\frac{1}{2}p_X^{(1)}(x)n^{-1}U_{n,2}$, where

$$U_{n,2} = 2!\sum_{t=1}^{n}\sum_{0=j_1<j_2}^{\infty} a_{j_1}a_{j_2}\varepsilon_{t-j_1}\varepsilon_{t-j_2}$$

is defined in (4.51).

Furthermore, Theorem 4.33 can be extended to subordinated processes $Y_j = \tilde{G}(X_t)$. As expected from Theorem 4.4 (Gaussian case) or Theorem 4.8 (the linear case), the rate of convergence and the asymptotic distribution depends on the Appell (or, equivalently, the power) rank of

$$G(X; x) = 1\big\{\tilde{G}(X) \le x\big\} - F_Y(x).$$

The limiting process is a Hermite–Rosenblatt random variable multiplied by a deterministic function.

4.8.2 Applications and Extensions

4.8.2.1 Quantile Processes and Trimmed Sums

Weak convergence (4.160) for empirical processes based on LRD linear sequences has immediate implications for sample quantiles. For $y \in (0, 1)$, define the quantile function

$$Q_X(y) = F_X^{-1}(y) = \inf\big\{x : F_X(x) \ge y\big\}.$$

We will assume that F_X and Q_X are differentiable, so that

$$Q_X(y) = \inf\big\{x : F_X(x) = y\big\}.$$

In an analogous manner, the empirical quantile function is defined as $Q_{n,X}(y) = F_{n,X}^{-1}(y)$ with $F_{n,X}$ defined in (4.155). By definition, $Q_{n,X}$ is left-continuous. Noting that for $x = Q_X(y)$,

$$Q_X'(y) = \frac{1}{p_X(x)},$$

(4.160) implies

$$L_1^{-\frac{1}{2}}(n)n^{\frac{1}{2}-d}\big[Q_{n,X}(y) - Q_X(y)\big] \Rightarrow Z, \tag{4.165}$$

where Z is a standard normal random variable, and the convergence is in $D[a, b]$ equipped with the sup-norm for $0 < a < b < 1$. It is remarkable that the limiting variable does not depend on y (this is of course due to the degenerate structure of the limiting process in (4.160)). A detailed evaluation and further extensions can be found in Ho and Hsing (1996), Wu (2005), Csörgő et al. (2006), Youndjé and Vieu (2006), Csörgő and Kulik (2008a, 2008b) or Coeurjolly (2008a, 2008b).

The result for the quantile function can be extended to trimmed sums

$$T_{n,h} := \frac{1}{n - 2[nh]} \sum_{t=[nh]+1}^{n-[nh]} X_{t:n}, \tag{4.166}$$

where $h \in (0, 1/2)$, and $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ are the order statistics. Then

$$L_1^{-\frac{1}{2}}(n)n^{\frac{1}{2}-d}T_{n,h} \to_d Z.$$

See Ho and Hsing (1996), Wu (2003) or Kulik and Ould Haye (2008).

Note, however, that the weak convergence (4.165) cannot be extended to $(0, 1)$. Similarly, the result (4.166) does not hold for sums of extremes $\sum_{t=1}^{[nh]} X_{t:n}$ or $\sum_{t=n-[nh]}^{n} X_{t:n}$. There, the limiting behaviour depends on an interplay between the dependence parameter d and the heaviness of tails of the random variables X_t. We refer to Kulik (2008a) for details. Similar issues will be discussed in Sect. 4.8.5 in connection with tail empirical processes.

4.8.2.2 Goodness-of-Fit Test

An immediate consequence for statistical inference is for instance an unusual behaviour of the Kolmogorov–Smirnov statistic, namely

$$L_1^{-\frac{1}{2}}(n)n^{\frac{1}{2}-d}T_{\mathrm{KS},n} := L_1^{-\frac{1}{2}}(n)n^{\frac{1}{2}-d}\sup_{x\in\mathbb{R}}\big|F_{n,X}(x) - F_X(x)\big| \underset{d}{\to} |Z| \sup_{x\in\mathbb{R}} p_X(x), \tag{4.167}$$

given that $\sup_{x\in\mathbb{R}} p_X(x) < \infty$. Therefore, we may approximate p-values by

$$P(T_{KS,n} > u) \approx 2\bar{\Phi}\left(\frac{u}{\sup_{x\in\mathbb{R}} p_X(x)} L_1^{\frac{1}{2}}(n)n^{d-\frac{1}{2}}\right), \tag{4.168}$$

where $u \ge 0$, Φ is the cumulative standard normal distribution, and $\bar{\Phi} = 1 - \Phi$. Note in particular that for a given density, the value $\sup_{x\in\mathbb{R}} p_X(x)$ is known. Of course, in general one has to estimate the dependence parameter d.

In contrast, for weakly dependent processes, the supremum of the transformed Brownian bridge $\tilde{B} \circ F$ over the interval $[0, 1]$ is required.

4.8.3 Empirical Processes with Estimated Parameters

Consider the assumptions of Theorem 4.33. As mentioned previously, a direct statistical application of the limiting behaviour of the empirical process is the Kolmogorov–Smirnov statistic, as established in (4.167). As explained in (4.168), this result can be used, in principle, to test whether the marginal distribution F_X of an observed series $X_1, \ldots, X_n$ is equal to a specific distribution F^0. Usually, however, one needs to test whether F_X belongs to a certain type of distributions, instead of one fixed F^0. For instance, we would like to test whether F_X is in a parametric family $\{F_X(\cdot, \theta), \theta \in \mathbb{R}\}$, without specifying the parameter θ a priori. The nuisance parameter θ has to be estimated from the observed series. Thus, instead of $T_{\mathrm{KS}}(\theta) = T_{\mathrm{KS},n}(\theta)$, one considers

$$T_{\mathrm{KS}}(\hat{\theta}) = \sup_{x \in \mathbb{R}} \left| F_{n,X}(x) - F_X(x; \hat{\theta}) \right|,$$

where $\hat{\theta}$ is a suitable estimate of θ. If the observations are i.i.d., then the rate of convergence for both, the original Kolmogorov–Smirnov statistics $T_{\mathrm{KS}} = T_{\mathrm{KS}}(\theta)$ and $T_{\mathrm{KS}}(\hat{\theta})$, is the same, though the variances of the limiting distributions are different.

To show what may happen in the long-memory case, let us consider a sequence $Y_t = X_t + \mu$ $(t \in \mathbb{N})$. Clearly, $F_Y(x) = F_X(x; \mu) = F_X(x - \mu)$. The empirical processes

$$E_{n,X}(x) = F_{n,X}(x) - F_X(x) = \frac{1}{n} \sum_{t=1}^{n} 1\{X_t \le x\} - F_X(x)$$

and

$$E_{n,Y}(x; \mu) := F_{n,Y}(x) - F_Y(x) = \frac{1}{n} \sum_{t=1}^{n} 1\{Y_t \le x\} - F_Y(x)$$

are related by

$$E_{n,Y}(x; \mu) = E_{n,X}(x - \mu). \tag{4.169}$$

On account of (4.160), $L_1^{-\frac{1}{2}}(n) n^{\frac{1}{2}-d} E_{n,Y}(x)$ converges weakly to $p_X(x - \mu)Z$. Now, consider instead

$$E_{n,Y}(x; \hat{\mu}) = F_{n,Y}(x) - F_X(x; \hat{\mu}).$$

We will use the estimate $\hat{\mu} = \bar{y}$, so that $\hat{\mu} - \mu = \bar{x}$. We then write

$$\begin{aligned} E_{n,Y}(x; \hat{\mu}) &= F_{n,Y}(x) - F_Y(x) + F_Y(x) - F_X(x; \hat{\mu}) \\ &= E_{n,X}(x - \mu) + F_X(x; \mu) - F_X(x; \hat{\mu}). \end{aligned}$$

Now, we apply Taylor's expansion to obtain

$$\begin{aligned} F_X(x; \mu) - F_X(x; \hat{\mu}) &= p_X(x - \mu)(\hat{\mu} - \mu) - \frac{1}{2} p_X^{(1)}(x - \mu)(\hat{\mu} - \mu)^2 + R_n \\ &= p_X(x - \mu)\bar{x} - \frac{1}{2} p_X^{(1)}(x - \mu)\bar{x}^2 + R_n, \end{aligned}$$

where R_n is of a smaller order than $\bar{x}^2$. Furthermore, the reduction principle (4.159) implies

$$n^{\frac{1}{2}-d} L_1^{-\frac{1}{2}}(n) \sup_{x\in\mathbb{R}} \left| E_{n,X}(x-\mu) + p_X(x-\mu)\bar{x} \right| \to_p 0.$$

Thus,

$$\begin{aligned} & n^{\frac{1}{2}-d} L_1^{-\frac{1}{2}}(n) E_{n,Y}(x;\hat{\mu}) \\ & \quad = o_P(1) - \frac{1}{2} n^{\frac{1}{2}-d} L_1^{-\frac{1}{2}}(n) \left(p_X^{(1)}(x-\mu)\bar{x}^2 + R_n \right) = o_P(1), \end{aligned}$$

where the bound $o_P(1)$ is uniform in x given that $\sup_{x\in\mathbb{R}} |p_X^{(2)}(x)| < \infty$. In other words, the empirical processes $E_{n,Y}(\cdot;\mu)$ and $E_{n,Y}(\cdot;\hat{\mu})$ have different rates of convergence. Surprisingly, plugging in the parameter estimate improves the rate of convergence of the empirical process and therefore of goodness-of-fit tests such as the Kolmogorov–Smirnov or Anderson–Darling tests (Beran and Ghosh 1991; Ho 2002; Kulik 2009). The precise convergence rates are described in the following theorem.

Theorem 4.34 *Assume that the conditions of Theorem 4.33 are fulfilled. Additionally, assume that (4.158) holds with $r = 3$.*

- *If $d \in (1/4, 1/2)$ then*

$$n^{1-2d} L_1^{-1/2}(n) E_{n,Y}(x;\hat{\mu}) \Rightarrow p_X^{(1)}(x-\mu) \left(Z_2 - \frac{1}{2} Z_1^2 \right), \tag{4.170}$$

  where Z_1 and Z_2 are uncorrelated random variables, $Z_1 \sim N(0,1)$, and $Z_2 = Z_{2,H}(1)$ is the Hermite–Rosenblatt variable.
- *If $d \in (0, 1/4)$ then*

$$\sqrt{n} E_{n,Y}(x;\hat{\mu}) \Rightarrow Z(x-\mu), \tag{4.171}$$

  where $Z(\cdot)$ is a Gaussian process.

Remark 4.6 The limiting Gaussian process has a rather complicated covariance structure. Nevertheless, the result (4.171) suggests that for $d \in (0, 1/4)$, we can apply standard resampling techniques available for weakly dependent data, see Chap. 10.

To shed some light on the results of Theorem 4.34, consider the case $d \in (1/4, 1/2)$. The expression for the limiting process follows essentially from the approximation

$$E_{n,Y}(x;\hat{\mu}) \approx \left\{ E_{n,X}(x-\mu) + p_X(x-\mu)\bar{x} \right\} + \frac{1}{2} p_X^{(1)}(x-\mu)\bar{x}^2.$$

Now, the result follows from (4.163) and the limiting behaviour of the sample mean.

Furthermore, the limiting behaviour may change if different estimators of the mean μ are considered or if one considers a location-scale family $Y = \mu + \sigma X$ (see Beran and Ghosh 1991; Ho 2002; Kulik 2009).

4.8.4 Linear Processes with Infinite Moments

As noticed above, finite-dimensional convergence of the appropriately scaled empirical process $E_{n,X} = F_{n,X} - F_X(x)$ follows from the result for partial sums of subordinated linear processes, by considering the function $y \to G(y; x) = 1\{y \le x\}$. We will apply the same idea to linear processes $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ with i.i.d. symmetric infinite variance innovations, i.e.

$$P(\varepsilon_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha} \tag{4.172}$$

with $\beta = 0$. The general result mimics Theorem 4.17. We established there that for $0 < d < 1 - 1/\alpha$, we have

$$n^{-H}\sum_{t=1}^{[nu]}\big\{G(X_t) - E\big[G(X_1)\big]\big\} \Rightarrow A^{1/\alpha}C_\alpha^{-1/\alpha}\frac{c_a}{d}G_\infty^{(1)}(0)\tilde{Z}_{H,\alpha}(u),$$

where $\tilde{Z}_{H,\alpha}(\cdot)$ is a linear fractional stable motion with $H = d + \alpha^{-1}$ and $G_\infty(y) = E[G(X + y)]$. Setting $u = 1$ and evaluating $G_\infty(y) = P(X \le x - y)$, $G_\infty^{(1)}(0) = -p_X(x)$, we may conclude that for a fixed $x \in \mathbb{R}$,

$$n^{-H}\sum_{t=1}^{n}\big(1\{X_t \le x\} - P(X_1 \le x)\big) \xrightarrow{\mathrm{d}} A^{1/\alpha}C_\alpha^{-1/\alpha}\frac{c_a}{d}p_X(x)\tilde{Z}_{H,\alpha}(1).$$

This can be extended to convergence of the process $E_{n,X}(x)$ $(x \in \mathbb{R})$, see Koul and Surgailis (2001).

Theorem 4.35 *Assume that X_t $(t \in \mathbb{Z})$ is a linear process with $a_j \sim c_a j^{d-1}$,*

$$0 < d < 1 - 1/\alpha,$$

and ε_t $(t \in \mathbb{Z})$ are i.i.d. symmetric random variables such that (4.89) holds with $\alpha \in (1, 2)$ and $\beta = 0$:

$$P(\varepsilon_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha}.$$

Furthermore, assume that the distribution F_ε of ε_1 is such that

$$\big|F_\varepsilon^{(2)}(x)\big| \le C\big(1 + |x|\big)^{-\alpha}, \qquad \big|F_\varepsilon^{(2)}(x) - F_\varepsilon^{(2)}(y)\big| \le C|x - y|\big(1 + |x|\big)^{-\alpha},$$

where $|x-y|<1$, $x\in\mathbb{R}$. *Then*

$$n^{1-H}E_{n,X}(x)\Rightarrow A^{1/\alpha}C_\alpha^{-1/\alpha}\frac{c_a}{d}p_X(x)\tilde{Z}_{H,\alpha}(1), \tag{4.173}$$

where $\tilde{Z}_{H,\alpha}(1)$ *is a symmetric* α*-stable random variable with scale* η *given by*

$$\eta=\left(\int_{-\infty}^{1}\left\{(1-v)_+^d-(-v)_+^d\right\}^\alpha dv\right)^{1/\alpha}.$$

4.8.5 Tail Empirical Processes

Let X_t $(t\in\mathbb{Z})$ be a stationary sequence with marginal distribution F_X. More specifically, we shall assume that X_t is a stochastic volatility model considered in Sect. 4.3.4. Recall that the model is $X_t=\xi_t\sigma_t$ $(t\in\mathbb{Z})$, where

$$\sigma_t=\sigma(\zeta_t),\quad \zeta_t=\sum_{j=1}^{\infty}a_j\varepsilon_{t-j},$$

and $\sigma(\cdot)$ is a positive function. It is assumed that ξ_t $(t\in\mathbb{Z})$ is a sequence of i.i.d. random variables such that

$$P(\xi_1>x)\sim A\frac{1+\beta}{2}x^{-\alpha},\qquad P(\xi_1<-x)\sim A\frac{1-\beta}{2}x^{-\alpha}. \tag{4.174}$$

Also, we assume that the sequences ξ_t $(t\in\mathbb{Z})$ and ε_t $(t\in\mathbb{Z})$ are mutually independent. In particular (cf. Lemma 4.20), we have

$$P\big(|X_1|>x\big)\sim E\big(\sigma^\alpha(\zeta_1)\big)P\big(|\xi_1|>x\big),$$

provided that

$$E\big[\sigma^{\alpha+\delta}(\zeta_1)\big]<\infty \tag{4.175}$$

for some $\delta>0$. In Theorem 4.19 we saw that the limiting behaviour of partial sums depends on an interplay between the long-memory parameter d and the tail index α. Therefore, it is important to have reliable estimates of both parameters, d and α. With the help of the tail empirical process it is possible to prove asymptotic normality of the so-called Hill estimator of α.

We note first that the tail behaviour of X implies that, as $n\to\infty$,

$$T_n(x):=P\big(X_1>(1+x)u_n|X_1>u_n\big)=\frac{\bar{F}_X((1+x)u_n)}{\bar{F}_X(u_n)}\to T(x):=(1+x)^{-\alpha}$$

for any sequence of constants $u_n \to \infty$. The tail empirical distribution functions $\tilde{T}_n(s)$ and the tail empirical processes $e_n(s)$ are defined by

$$\tilde{T}_n(s) = \frac{1}{n\bar{F}_X(u_n)} \sum_{t=1}^{n} 1\{X_t > u_n(1+s)\}$$

and

$$e_n(s) = \tilde{T}_n(s) - T_n(s) \ \big(s \in [0,\infty)\big). \tag{4.176}$$

We note that for large values of u_n, only extreme observations are included in the sum. Hence the name "tail empirical".

Drees (1998, 2000) and Rootzén (2009) show that for weakly dependent observations X_t, scaled processes $w_n e_n$ converge weakly in $D[0,\infty)$ to a Gaussian process $w = B \circ T$, where B is a standard Brownian motion, and $w_n^2 = n\bar{F}_X(u_n)$. The situation changes in the long-memory case. The limiting behaviour depends on an interplay between the memory parameter d and the behaviour of u_n. If u_n grows sufficiently fast (that means that very few extremes are included in the tail empirical distribution), then long memory does not influence the limit: $w_n e_n \Rightarrow w$ with, as before, $w_n = \sqrt{n\bar{F}_X(u_n)}$ and $w = B \circ T$. However, if u_n grows at an appropriately slow rate, then long memory starts to play a role: $w_n e_n$ converge weakly to a degenerate limiting process $w(s) = CT(s)Z_{m,H}(1)$ (where C is a constant), and the scaling factor is different, namely $w_n = n^{m(\frac{1}{2}-d)}L(n)$, where L is a slowly varying function. The corresponding result is stated in Theorem 4.36.

In order to state the result, let us define the function G_n on $(-\infty,\infty) \times [0,\infty)$ by

$$G_n(x,s) = \frac{P(\sigma(x)\xi_1 > (1+s)u_n)}{P(\xi_1 > u_n)}. \tag{4.177}$$

This function converges pointwise to $T(s)G(x) = T(s)\sigma^\alpha(x)$. Furthermore, the Hermite coefficients $J_n(m,s)$ of the function $x \to G_n(x,s)$ converge (as $n \to \infty$) to $J(m)T(s)$, uniformly with respect to $s \geq 0$, where $J(m)$ is the m-th Hermite coefficient of G. This implies that for large n, the Hermite rank $m_n(s)$ of $G_n(\cdot,s)$ is not greater than the Hermite rank m of G. To avoid further complications, we impose the assumption $\inf_{s\geq 0} m_n(s) = m$ for sufficiently large n.

Theorem 4.36 *Consider the stochastic volatility model $X_t = \xi_t\sigma_t$ $(t \in \mathbb{Z})$ and assume that (4.174) and (4.175) hold. Additionally, we assume that ζ_j $(t \in \mathbb{Z})$ is a Gaussian linear process with coefficients a_j satisfying* (B1), *i.e. $a_j = L_a(j)\,j^{d-1}$, $d \in (0,1/2)$ (so that $\gamma_X(k) \sim L_\gamma(k)k^{2d-1}$). Let $m \geq 1$ be the Hermite rank of the function $\sigma^\alpha(\cdot)$, and set $H = d + 1/2$. Assume that $E[\sigma^{2\alpha+\delta}(X_1)] < \infty$.*

(i) *If $n\bar{F}_X(u_n) \to \infty$ and $n^{1-m(1-2d)}L_m(n)\bar{F}_X(u_n) \to 0$ as $n \to \infty$, then $\sqrt{n\bar{F}_X(u_n)}\,e_n$ converges weakly in $D[0,\infty)$ to the Gaussian process $B \circ T$, where B is a standard Brownian motion.*

(ii) *If* $n\bar{F}_X(u_n) \to \infty$ *and* $n^{1-m(1-2d)}L_m(n)\bar{F}_X(u_n) \to \infty$ *as* $n \to \infty$, *then*

$$n^{m(\frac{1}{2}-d)}L_m^{-1/2}(n)e_n(s) \Rightarrow \frac{J(m)T(s)}{E[\sigma^\alpha(\zeta_1)]}Z_{m,H}(1),$$

where $\Rightarrow$ *denotes weak convergence in* $D[0,\infty)$, $Z_{m,H}(\cdot)$ *is a Hermite–Rosenblatt process, and* $L_m(n) = m!C_m L_\gamma^m(n)$.

The practical application of these limit theorems for $e_n(\cdot)$ is not quite straightforward. First of all, $\bar{F}_X(u_n)$ is unknown. The second problem is that we would like to center the tail empirical distribution function by $T(s)$, not $T_n(s)$. The second question can be addressed by introducing the assumption

$$\lim_{n\to\infty} w_n \|T_n - T\|_\infty = 0, \tag{4.178}$$

where

$$\|T_n - T\|_\infty = \sup_{t\geq 1}\left|\frac{P(X_1 > u_n t)}{P(X_1 > u_n)} - t^{-\alpha}\right|,$$

and the scaling w_n is either $\sqrt{n\bar{F}_X(u_n)}$ or $n^{m(\frac{1}{2}-d)}L_m^{-1/2}(n)$ in cases (i) and (ii) respectively. In other words, we impose a condition that makes the bias $T_n - T$ negligible. This is related to the so-called second-order regular variation (see Drees 1998; Kulik and Soulier 2011), but we omit details here. As an example, assume for instance that

$$P(\xi_1 > x) = cx^{-\alpha}\left(1 + O\left(x^{-\beta}\right)\right) \quad (x \to \infty)$$

for some constant $c > 0$. Then the second-order regular variation refers to the second-order term $x^{-\beta}$ in the expansion for the tail of ξ_1.

Now, suppose that the second-order assumption holds. Let $X_{1:n} \leq \cdots \leq X_{n:n}$ be the order statistics of $X_1, \ldots, X_n$, define $k_n = n\bar{F}_X(u_n)$ and replace u_n by $X_{n-k:n}$ in the definition of the tail empirical distribution function. Implicitly, $k = k_n$ will become a user chosen number of extreme statistics such that $k_n \to \infty$ and $k_n = o(n)$. Thus, we define

$$\hat{T}_n(s) = \frac{1}{k}\sum_{t=1}^{n} 1\left\{X_t > X_{n-k:n}\cdot(1+s)\right\}$$

and the practically computable processes

$$\hat{e}_n^*(s) = \hat{T}_n(s) - T(s) \qquad \left(s \in [0,\infty)\right).$$

It follows from Rootzén (2009) and Kulik and Soulier (2011) that

$$w_n\,\hat{e}_n^*(s) \Rightarrow w^*(s) = w(s) - T(s)w(1).$$

In particular, if $w_n = \sqrt{n\bar{F}_X(u_n)} = \sqrt{k_n}$ and $w(s) = B(T(s))$, then $w^*(s) = \tilde{B}(T(s))$, where $\tilde{B}$ is a Brownian bridge. However, if $w_n = n^{m(\frac{1}{2}-d)}L(n)$ and

$w(s) = CT(s)Z_{m,H}(1)$, then $w^*(s) = 0$. This is a similar effect as for the standard empirical process with estimated parameters considered in Sect. 4.34. More surprisingly, we have the following result for the process $\hat{e}_n^*(s)$.

Theorem 4.37 *Assume that the conditions of Theorem 4.36 are fulfilled. Assume additionally that (4.178) holds. Then $\sqrt{k}\,\hat{e}_n^*(s)$ converges weakly in $D[0,\infty)$ to the Gaussian process $\tilde{B}(T(s))$, where $\tilde{B}$ is a standard Brownian bridge, regardless of the behaviour of $n^{1-m(1-2d)}L_m(n)\bar{F}_X(u_n)$.*

4.8.5.1 Application to Tail Index Estimation

One of the most important problems when dealing with heavy tails is to estimate the tail index α. The best known (though in many ways not always reliable) method is Hill's estimator. Using the notation $\gamma = \alpha^{-1}$, the Hill estimator of γ is defined by

$$\hat{\gamma}_n = \frac{1}{k}\sum_{j=1}^{k} \log\left(\frac{X_{n-j+1:n}}{X_{n-k:n}}\right).$$

Noting that

$$\begin{aligned}\int_0^\infty \frac{\hat{T}_n(s)}{1+s}\,ds &= \frac{1}{k}\sum_{t=1}^{n}\int_0^\infty \frac{1\{s < X_t/X_{n-k:n} - 1\}}{1+s}\,ds \\ &= \frac{1}{k}\sum_{t=1}^{n}\log\left(1+\max\left\{\frac{X_t}{X_{n-k:n}} - 1, 0\right\}\right),\end{aligned}$$

the estimator can also be written as

$$\hat{\gamma}_n = \int_0^\infty \frac{\hat{T}_n(s)}{1+s}\,ds.$$

Since $\gamma = \int_0^\infty (1+s)^{-1}T(s)\,ds$, we have

$$\hat{\gamma}_n - \gamma = \int_0^\infty \frac{\hat{e}_n^*(s)}{1+s}\,ds.$$

Thus we can apply Theorem 4.37 to obtain the asymptotic distribution of the Hill estimator. Heuristically,

$$\sqrt{k_n}(\hat{\gamma}_n - \gamma) \to_d \int_0^\infty \frac{\tilde{B}(T(s))}{1+s}\,ds.$$

This integral is a normal random variable with variance γ^2 (for details, see Kulik and Soulier 2011). In summary, we have the following result.

Corollary 4.5 *Under the assumptions of Theorem 4.37, $\sqrt{k}(\hat{\gamma}_n - \gamma)$ converges in distribution to a centred Gaussian distribution with variance γ^2.*

This result can be used to construct confidence intervals for γ. It is known that this result gives the best possible rate of convergence for the Hill estimator for i.i.d. data (see Drees 1998). The surprising result is that it is possible to achieve the same i.i.d. rates regardless of the dependence parameter d.

4.8.5.2 Proof of Theorem 4.36

Proof We follow a similar idea as in the proof of Theorem 4.19. Let $\mathcal{E}$ be the σ-field generated by the Gaussian process ζ_t ($t \in \mathbb{Z}$). Write

$$\begin{aligned} e_n(s) &= \frac{1}{n\bar{F}_X(u_n)} \sum_{t=1}^{n} \big\{ 1\big\{X_t > (1+s)u_n\big\} - P\big(X_t > (1+s)u_n|\mathcal{E}\big)\big\} \\ &\quad + \frac{1}{n\bar{F}_X(u_n)} \sum_{t=1}^{n} \big\{ P\big(X_t > (1+s)u_n|\mathcal{E}\big) - \bar{F}_X(u_n)\big\} \\ &=: M_n(s) + R_n(s). \end{aligned} \tag{4.179}$$

The difference between (4.179) and the decomposition used in the proof of Theorem 4.19 is that here the first part is the sum of conditionally independent random variables, instead of being a martingale. The second part is a function of the Gaussian sequence ζ_t ($t \in \mathbb{N}$) and does not depend on the sequence ξ_t ($t \in \mathbb{N}$).

For the first part, it can be shown that, using the conditional independence,

$$\log E\big[\exp\big(it\sqrt{n\bar{F}_X(u_n)}M_n(0)\big)\big|\mathcal{E}\big] \to_P -t^2/2.$$

The bounded convergence theorem implies

$$\sqrt{n\bar{F}_X(u_n)}M_n(0) \to_d T(0)Z,$$

where Z is standard normal. Using the Cramer–Wald device, it is extended to

$$\begin{aligned} &\sqrt{n\bar{F}_X(u_n)}\big(M_n(s_1), M_n(s_l) - R_n(s_{l-1}), l = 2, \ldots, K\big) \\ &\quad \to_d \big(N\big(0, T(s_1)\big), N\big(0, T(s_l) - T(s_{l-1})\big), l = 2, \ldots, K\big), \end{aligned} \tag{4.180}$$

where the normal random variables are independent. Computations are somewhat involved, but the idea is relatively easy. Since the random variables are conditionally independent, the characteristic function can be evaluated.

Recall that

$$G_n(x,s) = \frac{P(\sigma(x)\xi_1 > (1+s)u_n)}{P(\xi_1 > u_n)}$$

converges pointwise to $T(s)G(x) = T(s)\sigma^{\alpha}(x)$. Let us now write

$$\begin{aligned}
&\sum_{t=1}^{n}\big(G_n(\zeta_t,s) - E\big[G_n(\zeta_t,s)\big]\big) \\
&\quad = \sum_{t=1}^{n}\sum_{q=m}^{\infty}\frac{T(s)J(q)}{q!}H_q(\zeta_t) + \sum_{t=1}^{n}\sum_{q=m}^{\infty}\frac{J_n(q,s)-T(s)J(q)}{q!}H_q(\zeta_t) \\
&\quad =: T(s)R_n^* + \tilde{R}_n(s)
\end{aligned}$$

with $R_n^* = \sum_{t=1}^{n} G(\zeta_t)$. Convergence of $T(s)R_n^*$ is concluded in the very same way as in (4.102) and (4.103). For $m(1/2-d) < 1$ and $m(1/2-d) > 1$, we have, respectively,

$$n^{-(1-m(\frac{1}{2}-d))}L_m^{-1/2}(n)R_n^* \Rightarrow \frac{J(m)}{m!}Z_{m,H}(1)$$

and

$$n^{-1/2}R_n^* \Rightarrow vZ,$$

where v is a constant. The second part, $\tilde{R}_n(s)$ is of a smaller order than R_n^*, uniformly in $s \geq 0$. Since

$$R_n(s) = \frac{P(\xi_1 > u_n)}{n\bar{F}_X(u_n)}\sum_{t=1}^{n}\big(G_n(\zeta_t,s) - E\big[G_n(\zeta_t,s)\big]\big), \tag{4.181}$$

and $P(\xi_t > u_n)/\bar{F}_X(u_n) \to 1/E[\sigma^{\alpha}(\zeta_1)]$, we conclude that for $m(1/2-d) < 1$,

$$n^{m(\frac{1}{2}-d)}L_m^{-1/2}(n)R_n(s) \to_d \frac{J(m)T(s)}{E[\sigma^{\alpha}(\zeta_1)]}Z_{m,H}(1). \tag{4.182}$$

This convergence is easily extended to multivariate convergence. If $m(1/2-d) > 1$, then $R_n(s)$ is uniformly negligible w.r.t. the conditionally independent part $M_n(s)$. Therefore, (4.182) and (4.180) yield the finite-dimensional convergence. For details and proof of tightness, we refer to Kulik and Soulier (2011). □

4.8.5.3 Further Extensions

The results given above are extendable to stochastic volatility models with leverage. Instead of decomposing $e_n(s)$ into a conditionally i.i.d. part $M_n(s)$ and a long-memory part $R_n(s)$, we may apply the martingale decomposition as in the proof of Theorem 4.19. For details, see Luo (2011).

4.9 Limit Theorems for Counting Processes and Traffic Models

In this section we review limit theorems for counting processes and traffic models, such as renewal reward, ON–OFF, shot-noise and infinite source Poisson processes, considered in Sect. 2.2.4.

4.9.1 Counting Processes

Let X_j $(j \geq 1)$ be a stationary sequence of strictly positive random variables with distribution F and finite mean. Let τ_0 have the distribution $F^{(0)}$ and define

$$\tau_j = \tau_0 + \sum_{k=1}^{j} X_k \quad (j \geq 1)$$

and

$$S_n(t) = \sum_{j=1}^{[nt]} X_j.$$

Note that the notation X_j and $S_n(t)$ is different from what was used previously (which was $S(u) = \sum_{t=1}^{[nu]} X_t$). The reason is that here the natural time parameter is in the upper limit $[nt]$ of the sum.

Now, let $N(t)$ be the associated counting process. Since

$$N(t) = \max\{k \geq 0 : \tau_{k-1} \leq t\} = \min\{k \geq 0 : \tau_k > t\},$$

one can view $N(t)$ as the generalized inverse of the partial sums process $S_n(t)$. Consequently, if the limiting process for partial sums is Gaussian, Lemma 4.7 will imply the weak convergence of $N(t)$ from that of $S_n(t)$. In other words, we apply Lemma 4.7 to

- $y_n(t) = S_n(t)/(n\mu)$,
- $y_n^{-1}(t) = N_n(t)/n$, where $N_n(t) = N(n\mu t)$.

If $c_n^{-1}(S_n(t)/(n\mu) - t)$ converges to a process $S(t)$ with some constants c_n, then $c_n^{-1}(N(n\mu t)/n - t)$ converges to $-S(t)$. The same procedure applies to any stationary counting process associated with a stationary sequence X_j $(j \in \mathbb{N})$ with finite mean.

Example 4.24 Recall Theorem 4.5. There, X_j $(j \in \mathbb{N})$ is a linear process $X_j = \sum_{k=0}^{\infty} a_k \varepsilon_{j-k}$ with summable coefficients a_k and i.i.d. centred innovations ε_j $(j \in \mathbb{Z})$. We can reformulate Theorem 4.5 to accommodate $\mu = E(X_1) \neq 0$. We have

$$n^{-1/2} \sum_{j=1}^{[nt]} (X_j - \mu) \Rightarrow vB(t)$$

in $D[0,1]$, where $v^2 = \sigma_X^2 + 2\sum_{k=1}^{\infty} \gamma_X(k)$, and $B(t)$ ($t \in [0,1]$) is a standard Brownian motion. Equivalently,

$$\frac{S_n(t)/(n\mu) - t}{n^{-1/2}} \Rightarrow v\mu^{-1} B(t),$$

so that $S(t) = v\mu^{-1}B(t)$ and $c_n = n^{-1/2}$. Application of Lemma 4.7 yields

$$n^{-1/2}\big(N(n\mu t) - nt\big) \Rightarrow v\mu^{-1}B(t).$$

However, we cannot extend this to the situation of Theorem 4.6. The long-range dependent linear process must have zero mean and hence cannot be strictly positive.

Example 4.25 Recall Example 4.12. The model considered there is $X_j = \xi_j \sigma(\zeta_j)$, where ξ_j ($j \geq 1$) are strictly positive random variables with mean $E(\xi_1)$, and ζ_j is a centred Gaussian sequence with covariance $\gamma_\zeta(k) \sim L_\gamma(k)k^{2d-1}$, $d \in (0, 1/2)$. We established in Example 4.12 that for $G(x) = x$ and $\sigma(x) = \exp(x)$, we have

$$n^{-(d+1/2)} L_1^{-1/2}(n) \sum_{j=1}^{[nt]} \big(X_j - E(X_1)\big) \Rightarrow J(1) B_H(t)$$

weakly in $D[0,1]$, where $B_H(\cdot)$ is fractional Brownian motion with $H = d + 1/2$ and $J(1) = E(\zeta_1 \exp(\zeta_1))E(\xi_1)$. Hence, for the inverse processes, we obtain

$$n^{-H} L_1^{-1/2}(n)\big(N(n\mu t) - nt\big) \Rightarrow J(1)\mu^{-1} B_H(t).$$

Thus, long memory in the interpoint distances generates long-memory-type behaviour in the functional central limit theorem for the counting process.

Let now X_j ($t \in \mathbb{N}$) be an i.i.d. sequence of strictly positive random variables such that

$$P(X_1 > x) \sim Ax^{-\alpha} \quad (A > 0, \alpha > 1).$$

In Sect. 4.3 we saw that the appropriately centred and normalized $S_n(t)$ converges to an α-stable Lévy process with independent increments (cf. (4.80)):

$$c_n^{-1} \sum_{j=1}^{[nt]} (X_j - \mu) \Rightarrow C_\alpha^{-1/\alpha} Z_\alpha(t),$$

where $c_n = \inf\{s : P(X > x) \leq n^{-1}\}$, $c_n \sim A^{1/\alpha} n^{1/\alpha}$, and $Z_\alpha(t)$ is an α-stable Lévy motion such that $Z_\alpha(1) \stackrel{\mathrm{d}}{=} S_\alpha(1,1,0)$. The limiting process has discontinuous sample paths, and hence Lemma 4.7 is not applicable. However (see Theorem 7.3.2 in Whitt 2002), one can generalize Vervaat's result to cover the case of limiting processes with discontinuous sample paths. One has to mention though that although $S_n(t)$ may converge in the standard Skorokhod topology, the same does not apply

Table 4.5 Limits for counting processes—tails vs. dependence

Counting processes		
	Weak dependence	Strong dependence
Interarrival times with finite variance	Brownian motion (Example 4.24)	fBm (Example 4.25)
Interarrival times with infinite variance	Lévy process (Example 4.26)	fBm or Lévy process (Example 4.27)

to the counting process. One has to consider a weaker M_1 topology (see comments on p. 235 as well as Sects. 13.6 and 13.7 in Whitt 2002). Here, we just illustrate finite-dimensional convergence.

Example 4.26 In the situation described above,

$$c_n^{-1}\big(N(n\mu t) - nt\big) \stackrel{\text{fidi}}{\to} -C_\alpha^{-1/\alpha}\mu^{-1}Z_\alpha(t). \tag{4.183}$$

Thus, a heavy-tailed distribution of interarrival times X_j generates Long-Range count Dependence (LRcD) in the counting process (see Example 2.5). On the other hand, the limiting process has independent increments. Furthermore, in Example 2.5 we found out that $\mathrm{var}(N(t))$ is proportional to t^{2H} (as $t \to \infty$) with $H = (3-\alpha)/2$. On the other hand, $n^{-H}(N(n\mu t) - nt)$ converges to 0 in probability. Hence, $N(\cdot)$ is an example of a second-order stationary process where its standard deviation does not yield an appropriate scaling.

Example 4.27 Recall Example 4.17. If $d + 1/2 < 1/\alpha$, then by Whitt's approach

$$n^{-1/\alpha}\big(N(n\mu t) - nt\big) \stackrel{\text{fidi}}{\to} -A^{1/\alpha}C_\alpha^{-1/\alpha}\big\{E\big(\sigma_1^\alpha\big)\big\}^{1/\alpha}\mu^{-1}Z_\alpha(t). \tag{4.184}$$

If however $d + 1/2 > 1/\alpha$, we can use Vervaat's Lemma 4.7 to conclude

$$n^{-(d+1/2)}L_1^{-1/2}(n)\big(N(n\mu t) - nt\big) \Rightarrow J(1)E(\xi_1)\mu^{-1}B_H(t). \tag{4.185}$$

We summarize our findings in Table 4.5. It should be noted that in the case of strong dependence the results are just for the case in Examples 4.25, 4.27, not for all long-memory models.

4.9.2 Superposition of Counting Processes

Let $N^{(m)}(t)$ $(t \geq 0, m = 1, \ldots, M)$ be independent copies of a stationary renewal process $N(t)$ associated with a renewal sequence X_j $(j \in \mathbb{N})$. We assume that, as $x \to \infty$,

$$\bar{F}(x) = P(X_1 > x) \sim x^{-\alpha}L(x) \quad (1 < \alpha < 2),$$

and that $P(\tilde{X}_0 > x) = \mu^{-1} \int_x^\infty \bar{F}(u)\,du$, where $\mu = E[X_1] = \lambda^{-1}$. Application of Lemma 4.6 yields

$$\lim_{M\to\infty} \frac{1}{M^{1/2}} \sum_{m=1}^{M} \bigl(N^{(m)}(t) - \lambda t\bigr) \Rightarrow G(t), \tag{4.186}$$

where $G(\cdot)$ is a Gaussian process with stationary increments and the same covariance structure as $N(t)$. In particular (see Example 2.5),

$$\mathrm{var}\bigl(G(t)\bigr) = \mathrm{var}\bigl(N(t)\bigr) \sim \frac{2\lambda}{(\alpha-1)(2-\alpha)(3-\alpha)} t^{3-\alpha} L(t) =: \sigma_0^2 t^{3-\alpha} L(t).$$

Indeed, to apply Lemma 4.6, we verify that for $t > s$,

$$\mathrm{var}\bigl(N(t) - N(s)\bigr) = \mathrm{var}\bigl(N(t-s)\bigr) \sim C(t-s)^{2H}$$

and $2H > 1$. Also, the second condition of Lemma 4.6 is easily verified.

We recognize that the limiting process has up to a constant the same variance as a fractional Brownian motion with the Hurst index $H = (3-\alpha)/2$. Now, let us consider the time scaled process $N^{(m)}(Tt)$. For a fixed $T > 0$, application of (4.186) yields

$$\lim_{M\to\infty} \frac{1}{M^{1/2}} \sum_{m=1}^{M} \bigl(N^{(m)}(Tt) - \lambda Tt\bigr) \Rightarrow G(Tt) = \sigma_0 B_H(Tt)$$

and $\mathrm{var}(G(Tt)) \sim \sigma_0^2 T^{2H} t^{2H} L(Tt) \sim \sigma_0^2 T^{2H} t^{2H} L(T)$ as $T \to \infty$. Thus, applying H-self-similarity of fractional Brownian motion, we have

$$\lim_{T\to\infty} \frac{1}{T^H} \lim_{M\to\infty} \frac{1}{M^{1/2}} \sum_{m=1}^{M} \bigl(N^{(m)}(Tt) - \lambda Tt\bigr) \Rightarrow \sigma_0 B_H(t).$$

On the other hand, (4.183) yields

$$\lim_{T\to\infty} a_T^{-1} \bigl(N^{(m)}(Tt) - \lambda Tt\bigr) \overset{\text{fidi}}{\to} -\mu^{-1} C_\alpha^{-1/\alpha} Z_\alpha^{(m)}(\lambda t) \quad (m = 1, \dots, M),$$

where $Z^{(m)}(\cdot)$ $(m = 1, \dots, M)$ are independent Lévy processes, and $a_T \sim T^{1/\alpha} \ell(T)$. Consequently, since the sum of independent Lévy processes yields a Lévy process, we obtain

$$\lim_{M\to\infty} \frac{1}{M^{1/\alpha}} \lim_{T\to\infty} a_T^{-1} \sum_{m=1}^{M} \bigl(N^{(m)}(Tt) - \lambda Tt\bigr) \overset{\text{fidi}}{\to} -\lambda^{1+1/\alpha} C_\alpha^{-1/\alpha} Z_\alpha(t),$$

where $Z_\alpha(\cdot)$ is an α-stable Lévy process. The limiting constants were obtained by replacing t with λt and using $Z_\alpha(\lambda t) \overset{\mathrm{d}}{=} \lambda^{1/\alpha} Z_\alpha(t)$.

Table 4.6 Limits for superposition of counting processes—tails vs. dependence

Superposition of counting processes		
	Weak dependence	Strong dependence
Interarrival times with finite variance	$\lim_{M\to\infty}\lim_{T\to\infty} = \text{Bm}$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{Bm}$	$\lim_{M\to\infty}\lim_{T\to\infty} = \text{fBm}$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{fBm}$
Interarrival times with infinite variance	$\lim_{M\to\infty}\lim_{T\to\infty} = \text{Lévy}$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{fBm}$	

We observe that different limiting schemes yield different limiting processes. This feature will be also present in different traffic models.

In contrast, if the renewal sequence has a finite variance and short memory, then application of Example 4.24 yields that both procedures $\lim_{M\to\infty}\lim_{T\to\infty}$ and $\lim_{T\to\infty}\lim_{M\to\infty}$ produce the same limit, namely a Brownian motion. Likewise, in the case of strong dependence and a finite variance (as in Example 4.25), both procedures yield a fractional Brownian motion.

We summarize these observations in Table 4.6. We do not fill in the case of strong dependence and heavy tails (situation of Example 4.27). It is clear that there are four possible limits. If the counting process converges to fBm, then the limit for superpositions must be fBm as well. If the counting process converges to a Lévy process, then the superposition converges to either fBm or a Lévy process, depending on the order of taking these limits.

4.9.3 Traffic Models

Let $W(u)$ be a traffic model. It can be either a renewal reward, or ON–OFF, or infinite source Poisson or error duration process. In Sect. 2.2.4 we noted that the models have long memory in terms of non-integrable covariances or nonlinear growth of the variance of the integrated process. A very interesting feature is that long memory in a traffic process implies that the integrated process

$$W^*(t) = \int_0^t \big\{W(v) - E\big[W(v)\big]\big\}\,dv$$

converges in the sense of finite-dimensional distributions to an α-stable Lévy motion. The scaling factor has to be chosen as $T^{-1/\alpha}L(T)$, where L is a slowly varying function. In particular, this is another example of a second-order long-memory process where the variance grows at rate T^{2H}, but $T^{-H}W^*(Tt)$ converges to zero in probability as $T\to\infty$ (see e.g. Example 4.26). Furthermore, as in the case of counting processes, the convergence cannot hold in the $D[0,1]$ space equipped with the J_1-topology.With respect to J_1 the continuous process $W^*(Tt)$ must converge to a continuous limit, which is not the case here.

In the context of computer networks, these phenomena describe long memory of an individual source. However, they do not explain long memory at the level of teletraffic, which usually consists of a large number of sources. Assume now that we have M independent copies $W^{(m)}(\cdot)$ $(m = 1, \ldots, M)$ of the traffic process $W(t)$. Define

$$W^*_{T,M}(t) = \int_0^{Tt} \sum_{m=1}^{M} \bigl\{ W^{(m)}(v) - E\bigl[W(v)\bigr] \bigr\}\, dv = \sum_{m=1}^{M} W^{(m)*}(Tt),$$

where $W^{(m)*}(u)$, $m = 1, \ldots, M$, are i.i.d. copies of the cumulated process $W^{(m)}(t)$. The process $W^*_{T,M}(t)$ can be interpreted as (centred) total workload of M workstations at time t or as cumulative packet counts in the network by time t. We are interested in the limiting behaviour of the properly normalized cumulative process $W^*_{T,M}(t)$.

We will consider two limiting scenarios. First, we will analyse what happens if we let first $M \to \infty$ and then $T \to \infty$. In this setup, we will proceed as follows.

Step 1: Use Lemma 4.6 to establish that with some sequence a_M,

$$\lim_{M\to\infty} a_M^{-1} \sum_{m=1}^{M} \bigl\{ W^{(m)}(t) - E\bigl[W^{(m)}(t)\bigr] \bigr\}$$

converges to a process, say, $G(t)$. If the process is Gaussian, then its covariance structure is the same as that of $W(u)$.

Step 2: If the process $G(t)$ is Gaussian, then the integral $G^*(Tt) = \int_0^{Tt} G(u)\, du$ is Gaussian as well. We have

$$\operatorname{var}\bigl(G^*(Tt)\bigr) = \int_0^{Tt} \left(\int_0^{v} cov\bigl(W(0), W(s)\bigr)\, ds \right) dv.$$

From the form of the covariance function we will conclude either a Brownian motion or a fractional Brownian motion as limit.

Step 3: The sum of independent (fractional) Brownian motions yields (fractional) Brownian motion. We will conclude that

$$\lim_{t\to\infty} a_T^{-1} \lim_{M\to\infty} a_M^{-1} \int_0^{Tt} \sum_{m=1}^{M} \bigl(W^{(m)}(v) - E\bigl[W^{(m)}(v)\bigr] \bigr)\, dv$$

converges to a (fractional) Brownian motion, where a_T is proportional to $T^{1/2}$ or T^H $(H > 1/2)$, respectively.

As for the case $T \to \infty$ and then $M \to \infty$, we will proceed as follows.

Step 1: For each $m = 1, \ldots, M$, approximate

$$\lim_{T\to\infty} c_T^{-1} \int_0^{Tt} \bigl\{ W^{(m)}(v) - E\bigl[W^{(m)}(v)\bigr] \bigr\}\, dv \approx c_T^{-1} \sum_{j=1}^{N(Tt)} U_j \quad (T \to \infty),$$

where $N(\cdot)$ is an appropriate counting process, and U_j ($j \in \mathbb{N}$) is an appropriate i.i.d. sequence. Note that both N and U_j depend on m. If the random variables U_j have a finite variance, then for each m, the limiting process is a Brownian motion, and $c_T = T^{1/2}$. If the random variables U_j are regularly varying with index α, then we obtain a Lévy process as a limit and $c_T = T^{1/\alpha}$.

Step 2: The sum of independent Brownian motions (Lévy processes) is a Brownian motion (Lévy process). We conclude the convergence for

$$\lim_{M\to\infty} d_M^{-1} \lim_{T\to\infty} c_T^{-1} \int_0^{Tt} \sum_{m=1}^{M} \bigl(W^{(m)}(v) - E\bigl[W^{(m)}(v)\bigr]\bigr)\, dv$$

with some sequence d_M.

One has to mention though that the proofs are sketched, without verifying some technical details.

4.9.4 Renewal Reward Processes

Recall from Example 2.12 the renewal reward process

$$W(t) = Y_0 1\{0 < t < \tau_0\} + \sum_{j=1}^{\infty} Y_j 1\{\tau_{j-1} \le t < \tau_j\},$$

$X_j = \tau_j - \tau_{j-1}$. We assume for simplicity that Y_j ($j \in \mathbb{N}$) is a centred i.i.d. sequence, independent of the renewal sequence τ_0, X_j ($j \ge 1$), and also that $E[X_1] = \mu = \lambda^{-1}$ is finite. We are interested in the limiting behaviour of the cumulative process $W^*_{T,M}(t)$ defined above. For the purpose of the limiting regime $\lim_{M\to\infty} \lim_{T\to\infty}$, we represent the cumulative process as follows:

$$\int_0^{Tt} W(u)\, du = \min\{Tt, \tau_0\} Y_0 + \sum_{j=0}^{\infty} Y_{j+1} \bigl(\min\{Tt, \tau_{j+1}\} - \tau_j\bigr)_+. \tag{4.187}$$

Indeed, if $Tt < \tau_0$, then $\int_0^{Tt} W(u)\, du = Y_0 Tt$; if $\tau_0 < Tt < \tau_1$, then $\int_0^{Tt} W(u)\, du = Y_0 Tt + Y_1(Tt - \tau_0)$ etc.

An alternative representation will yield an approximation of the cumulative reward by a sum of i.i.d. random variables. For $Tt > \tau_0$, we may write

$$\int_0^{Tt} W(u)\, du = Y_0 \tau_0 + \sum_{j=1}^{N(Tt)} Y_j X_j - U, \tag{4.188}$$

where $N(t)$ is the renewal process associated with τ_j. The first two terms represent the renewal intervals that are at least partially included in $[0, Tt]$. For example, if

$\tau_0 < Tt < \tau_1$, then $N(Tt) = 1$, and the sum includes $Y_0\tau_0 + Y_1X_1$. However, not the entire renewal interval X_1 is included in $[0, Tt]$. We have to subtract a portion $(\tau_1 - Tt)Y_1$, and this is "hidden" in the variable U.

In most cases considered below, only $\sum_{j=1}^{N(Tt)} Y_jX_j$ contributes to the limiting behaviour of $\int_0^{Tt} W(u)\,du$.

We start with a standard limiting behaviour. Specifically, we assume first that $\mathrm{var}(X) = \sigma_X^2 < \infty$ and $\mathrm{var}(Y) = \sigma_Y^2 < \infty$. In particular, there is no LRcD in the counting process $N(t)$ and hence in the cumulative renewal reward process $\int_0^t W(u)\,du$.

Theorem 4.38 *Assume that*

- *Interarrival times have a finite variance*: $\mathrm{var}(X_1) = \sigma_X^2 < \infty$;
- *Rewards have a finite variance*: $\mathrm{var}(Y_1) = \sigma_Y^2 < \infty$.

Then,

$$\lim_{T\to\infty}\lim_{M\to\infty}\frac{W^*_{T,M}(t)}{T^{1/2}M^{1/2}} = \lim_{M\to\infty}\lim_{T\to\infty}\frac{W^*_{T,M}(t)}{T^{1/2}M^{1/2}} \stackrel{\mathrm{d}}{=} \sigma_{\mathrm{reward},1}B(t),$$

where $(B(t), t \in \mathbb{R})$ *is a standard Brownian motion*,

$$\sigma^2_{\mathrm{reward},1} = \frac{E[X_1^2]E[Y_1^2]}{E[X_1]},$$

and the convergence is to be understood as a finite-dimensional one.

Proof First, we consider the limit taken in the order $\lim_{M\to\infty}$ first, and then $\lim_{T\to\infty}$.

Step 1: Since $W^{(m)}$ $(m = 1, \ldots, M)$ are independent identically distributed processes with finite variance, application of Lemma 4.6 implies that for each T,

$$\lim_{M\to\infty}\frac{1}{M^{1/2}}\sum_{m=1}^{M} W^{(m)}(Tt) \Rightarrow G(Tt)$$

in $D[0,\infty)$, where $G(t)$ $(t \geq 0)$ is a centred stationary Gaussian process with covariance function $cov(W(0), W(u))$.

Step 2: The cumulative process $G^*(\cdot t) = \int_0^{\cdot t} G(t)\,du$ is still a Gaussian process with variance $\mathrm{var}(G^*(Tt)) = \mathrm{var}(\int_0^{Tt} W(u)\,du) = Tt\,E[X_1^2]E[Y_1^2]/\mu$ (see Examples 2.5 and 2.12).

Step 3: The form of the covariance function yields that the process $T^{-1/2}G^*(Tt)$ $(t \geq 0)$ is a Brownian motion.

Now, we consider the reverse order of taking the limits.

Step 1: We use an approximation induced by representation (4.188).

$$\frac{1}{T^{1/2}}\sum_{j=1}^{N(Tt)} Y_jX_j = \left(\frac{N(Tt)}{T}\right)^{1/2}\frac{1}{\sqrt{N(Tt)}}\sum_{j=1}^{N(Tt)} Y_jX_j.$$

Recall that for a stationary renewal process, $N(Tt)/T \to EE[N(t)] = \lambda t = \mu^{-1}t$. Thus, as $T \to \infty$,

$$\frac{1}{T^{1/2}} \sum_{j=1}^{N(Tt)} Y_j X_j \approx \frac{t^{1/2}}{\mu^{1/2}} \frac{1}{(Tt)^{1/2}} \sum_{j=1}^{Tt} Y_j X_j \Rightarrow \frac{1}{\mu^{1/2}} \sqrt{\mathrm{var}(Y_1 X_1)} B(t).$$

Since X_1 and Y_1 are independent and $E[Y_1] = 0$, we obtain $\mathrm{var}(Y_1 X_1) = E[Y_1^2]E[X_1^2]$.

Step 2: Hence, for each fixed $m = 1, \ldots, M$,

$$T^{-1/2} \int_0^{Tt} W^{(m)}(u)\, du \Rightarrow \sigma_{\mathrm{reward},1} B^{(m)}(t),$$

where $B^{(m)}(t)$ are independent standard Brownian motions. Hence, the superposition converges to a Brownian motion.

□

Next, we analyse what happens if the finite variance assumption on the rewards still holds, but the renewal process has intervals with an infinite variance. Recall that then the corresponding counting process $N(t)$ has the LRcD property (see Examples 2.5 and 2.12) since its variance grows faster than linear. Also (see Examples 2.5 and 2.12), the variance of the cumulative process $\int_0^{Tt} W(u)\, du$ grows faster than linear.

Theorem 4.39 *Assume that*

- *Interarrival times are regularly varying*: $P(X_1 > x) \sim C_X x^{-\alpha}$ $(\alpha \in (1,2))$ *as* $x \to \infty$;
- *Rewards have a finite variance* $\mathrm{var}(Y_1) = \sigma_Y^2 < \infty$, *and they are symmetric*.

Then,

$$\lim_{T\to\infty} \lim_{M\to\infty} \frac{W^*_{T,M}(t)}{T^H M^{1/2}} \stackrel{\mathrm{d}}{=} \sigma_{\mathrm{reward},2} B_H(t), \tag{4.189}$$

where $(B_H(t), t \in \mathbb{R})$ *is a standard fractional Brownian motion with Hurst index* $H = (3-\alpha)/2$, *and*

$$\sigma^2_{\mathrm{reward},2} = C_X \frac{2E[Y_1^2]}{E[X_1](\alpha-1)(2-\alpha)(3-\alpha)}.$$

On the other hand,

$$\lim_{M\to\infty} \lim_{T\to\infty} \frac{W^*_{T,M}(t)}{T^{1/\alpha} M^{1/\alpha}} \stackrel{\mathrm{d}}{=} C_{\mathrm{reward},1} Z_\alpha(t), \tag{4.190}$$

where $Z_\alpha(t) \stackrel{\mathrm{d}}{=} t^{1/\alpha} S_\alpha(1,0,0)$ *is a symmetric Lévy process*, *and*

$$C_{\mathrm{reward},1} = \mu^{-1/\alpha} E^{1/\alpha}\big[|Y_1|^\alpha\big] C_X^{1/\alpha} C_\alpha^{-1}.$$

Sketch of Proof First, we proceed with $\lim_{T\to\infty}\lim_{M\to\infty}$.

Step 1: As in the case of Theorem 4.38, Lemma 4.6 implies that for each T,

$$\lim_{M\to\infty}\frac{1}{M^{1/2}}\sum_{m=1}^{M}W^{(m)}(Tt)\Rightarrow G(Tt)$$

in $D[0,\infty)$, where $G(t)$ $(t\in\mathbb{R})$ is a centred stationary Gaussian process with covariance function $cov(W(0),W(t))$.

Step 2: The cumulative process $G^*(Tt)$ is Gaussian with variance $\sigma_{\text{reward},2}(Tt)^{2H}$, $H=(3-\alpha)/2$ (see Example 2.12).

Step 3: The form of the variance yields that the scaled process $T^{-H}G^*(Tt)$ is a fractional Brownian motion.

Next, we deal with the reversed order of limits.

Step 1: We have

$$\frac{1}{T^{1/\alpha}}\sum_{j=1}^{N(Tt)}Y_jX_j=\left(\frac{N(Tt)}{T}\right)^{1/\alpha}\frac{1}{(N(Tt))^{1/\alpha}}\sum_{j=1}^{N(Tt)}Y_jX_j\approx\frac{1}{\mu^{1/\alpha}}\frac{1}{T^{1/\alpha}}\sum_{j=1}^{Tt}Y_jX_j.$$

By applying Breiman lemma we note that

$$P(Y_1X_1>x)\sim E\big[Y_+^\alpha\big]P(X_1>x)\sim E\big[Y_+^\alpha\big]C_Xx^{-\alpha}$$

and

$$P(Y_1X_1<-x)\sim E\big[Y_-^\alpha\big]P(X_1>x)\sim E\big[Y_-^\alpha\big]C_Xx^{-\alpha}.$$

Thus, application of (4.80) yields

$$\frac{1}{T^{1/\alpha}}\sum_{j=1}^{N(Tt)}Y_jX_j\Rightarrow\mu^{-1/\alpha}E^{1/\alpha}\big[|Y_1|^\alpha\big]C_X^{1/\alpha}C_\alpha^{-1}Z_\alpha(t).$$

Step 2: The result follows by taking $d_M=M^{1/\alpha}$. □

Finally, we analyse the case where both interarrival times and rewards are heavy tailed. We separate both limiting regimes in two theorems below.

Theorem 4.40 *Assume that*

- *Interarrival times are regularly varying*: $P(X_1>x)\sim C_Xx^{-\alpha}$ $(\alpha\in(1,2))$ *as* $x\to\infty$;
- *Rewards are regularly varying*: $P(Y_1>x)\sim C_Yx^{-\beta}$ $(\beta\in(1,2))$ *as* $x\to\infty$; *and they are symmetric*.

We have the following limits as $\lim_{M\to\infty}\lim_{T\to\infty}$:

- *If* $\alpha<\beta<2$, *then* (4.190) *still holds*.

- *If* $\beta < \alpha < 2$, *then*

$$\lim_{M\to\infty}\lim_{T\to\infty}\frac{W^*_{T,M}(t)}{T^{1/\beta}M^{1/\beta}} \stackrel{\mathrm{d}}{=} C_{\mathrm{reward},2}Z_\beta(t), \tag{4.191}$$

where $Z_\beta(t) \stackrel{\mathrm{d}}{=} t^{1/\beta}S_\beta(1,0,0)$ *is a symmetric Lévy process, and*

$$C_{\mathrm{reward},2} = \mu^{-1/\beta}E^{1/\beta}\big[X_1^\beta\big]C_Y^{1/\beta}C_\beta^{-1}.$$

Proof The proof is very similar to that of Theorem 4.39. Recall that the limiting behaviour of $\int_0^{Tt} W(u)\,du$ is determined by $\sum_{j=1}^{N(Tt)} Y_jX_j$. If $\alpha < \beta$, we may proceed exactly in the same way as in Theorem 4.39. Otherwise, if $\beta < \alpha$, then

$$\frac{1}{T^{1/\beta}}\sum_{j=1}^{N(Tt)} Y_jX_j = \left(\frac{N(Tu)}{T}\right)^{1/\beta}\frac{1}{(N(Tt))^{1/\beta}}\sum_{j=1}^{N(Tt)} Y_jX_j \approx \frac{1}{\mu^{1/\beta}}\frac{1}{T^{1/\beta}}\sum_{j=1}^{Tt} Y_jX_j.$$

By applying Breiman lemma we have

$$P(Y_1X_1 > x) \sim E\big[X_1^\beta\big]P(Y_1 > x) \sim E\big[X_1^\beta\big]C_Yx^{-\beta}$$

and

$$P(Y_1X_1 < -x) \sim E\big[X_1^\beta\big]P(Y_1 < -x) \sim E\big[X_1^\beta\big]C_Yx^{-\beta}.$$

Thus, application of (4.80) yields

$$\frac{1}{T^{1/\beta}}\sum_{j=1}^{N(Tt)} Y_jX_j \Rightarrow \mu^{-1/\beta}E^{1/\beta}\big[X_1^\beta\big]C_Y^{1/\beta}C_\beta^{-1}Z_\beta(t).$$

□

We note also in passing that the case $\beta < \alpha$ above does not require that X_1 is regularly varying. Therefore, (4.191) holds also when $\beta < 2$ and $\mathrm{var}(X_1) < \infty$.

We consider now the case of the other limit.

Theorem 4.41 *Assume that*

- *Interarrival times consist of positive integers and are regularly varying*: $P(X_1 > x) \sim C_Xx^{-(\alpha+1)}$ $(\alpha \in (1,2))$ *as* $x \to \infty$;
- *Rewards are regularly varying and symmetric*: $P(Y_1 > x) \sim C_Y\beta x^{-\beta}$ $(\beta \in (1,2))$ *as* $x \to \infty$;

We have the following limits as $\lim_{T\to\infty}\lim_{M\to\infty}$:

- *If* $\beta < \alpha < 2$, *then* (4.191) *holds.*
- *If* $\alpha < \beta < 2$, *then*

$$\lim_{T\to\infty}\lim_{M\to\infty} T^{-(\beta-\alpha+1)/\beta}M^{-1/\beta}W^*_{T,M}(t) \stackrel{\mathrm{d}}{=} C_X^{1/\beta}C_Y^{1/\beta}Z^*_\beta(t), \tag{4.192}$$

where $Z^*_\beta(t)$ is symmetric β-stable process with characteristic function

$$E\left[\exp\left(i\sum_{l=1}^{h}\theta_l Z^*_\beta(t_l)\right)\right] = \exp\bigl(-\sigma^\beta(\theta,\mathbf{t})\bigr),$$

where $\mathbf{t} = (t_1,\dots,t_h)^T$, $\theta = (\theta_1,\dots,\theta_h^T)$,

$$\sigma^\beta(\theta,\mathbf{t}) = C_\beta^{-1}\bigl(I(\theta,\mathbf{t}) + J(\theta,\mathbf{t})\bigr),$$

$$I(\theta,\mathbf{t}) = \mu^{-1}\int_0^\infty \left|\sum_{l=1}^{h}\theta_l(t_j\wedge x)\right|^\beta x^{-\alpha}\,dx,$$

$$J(\theta,\mathbf{t}) = \mu^{-1}\alpha\int_0^\infty\int_0^\infty \left|\sum_{l=1}^{h}\theta_l(t_j\wedge u - x)\right|^\beta (u-x)_+^{-\alpha-1}\,dx.$$

We observe that if $\beta<\alpha$, the order of taking limits does not matter. However, if $\alpha<\beta$, we obtain the new process $Z^*_\beta(t)$. This process has stationary increments and is self-similar with self-similarity parameter $H=(\beta-\alpha+1)/\beta$. For details on this process, we refer to Levy and Taqqu (2000). Furthermore, note that the convergence to $Z^*_\beta(t)$ requires the additional technical assumption that the interarrival times assume positive integers only.

Sketch of Proof We note that the technique of the proofs of Theorems 4.38 or 4.39 does not work. We cannot apply Lemma 4.6 because the process does not have a finite variance. Instead, we present a simplified version of the proofs of Theorems 2.2 and 2.3 in Levy and Taqqu (2000).

We use representation (4.187). Assume for a moment that Y_k $(k\ge 0)$ are symmetric β-stable, $Y_1 \overset{\mathrm{d}}{=} S_\beta(\eta,0,0)$, $\eta>0$. Thus, its characteristic function is given by

$$\varphi_Y(\theta) = E\exp(i\theta Y_1) = \exp\bigl(-\eta^\beta|\theta|^\beta\bigr).$$

We compute the characteristic function of $R(Tu)=\int_0^{Tu}W(u)\,du$. Set $\tau_{-1}=0$. Then, by conditioning on the entire sequence τ_j and using the fact that the random variables Y_j $(j\ge 0)$ are i.i.d.,

$$E\left[\exp\left(i\sum_{l=1}^{h}\theta_l R(t_l)\right)\right]$$

$$= E\left[\exp\left(i\sum_{l=1}^{h}\theta_l\left(Y_0\bigl(\min\{t_l,\tau_0\}\bigr) + \sum_{j=0}^{\infty}Y_{j+1}\bigl(\min\{t_l,\tau_{j+1}\}-\tau_j\bigr)\right)\right)\right]$$

$$= \exp\left(-\eta^\beta E\left(\sum_{l=1}^{h} |\theta_l| \left(\min\{t_l, \tau_0\} + \sum_{j=0}^{\infty} \big(\min\{t_l, \tau_{j+1}\} - \tau_j\big)\right)\right)^\beta\right)$$

$$=: \exp\big(-\sigma^\beta(\theta, \mathbf{t}; \eta)\big).$$

Since $W^*_{T,M}(t)$ is the sum of independent copies of the process $R(Tu)$, we have

$$E\left[\exp\left(i \sum_{l=1}^{h} \theta_l M^{-1/\beta} W^*_{1,M}(t_l)\right)\right] = \exp\big(-\sigma^\beta(\theta, \mathbf{t})\big).$$

An additional limiting argument applied to random variables Y_j that are regularly varying as in the theorem yields

$$\lim_{M\to\infty} M^{-1/\beta} W^*_{T,M}(t) \stackrel{\mathrm{d}}{=} Z^*_{\beta,T}(t),$$

where $Z^*_{\beta,T}(t)$ $(t \in [0,1])$ is a symmetric β-stable process with characteristic exponent $\sigma^\beta(\theta, T\mathbf{t}; C_Y/C_\beta)$. This process is neither self-similar, nor has it stationary increments.

More technical details are required to establish

$$T^{-(\beta-\alpha+1)/\beta} \sigma^\beta(\theta, T\mathbf{t}; C_Y/C_\beta) \to \sigma^\beta(\theta, T\mathbf{t}).$$

This implies the finite-dimensional convergence of $T^{-(\beta-\alpha+1)/\beta} Z^*_{\beta,T}(t)$ to $Z^*_\beta(t)$. □

Several bibliographical notes are in place here. Theorem 4.38 was proven in Taqqu and Levy (1986, Theorem 5). Theorem 4.39 was proven in Taqqu and Levy (1986). Theorem 4.40 was proven in Levy and Taqqu (1987), whereas Theorem 4.41 can be found in Levy and Taqqu (2000) and Pipiras and Taqqu (2000b). In particular, in the latter paper, the authors showed that the limiting process $Z^*_\beta(t)$ is not a linear fractional stable motion. Also see Taqqu (2002) and Willinger et al. (2003) for an overview.

A summary of the results discussed here is given in Table 4.7.

4.9.5 Superposition of ON–OFF Processes

Assume now that we have M independent copies $W^{(m)}(\cdot)$ $(m = 1, \ldots, M)$ of the ON–OFF process $W(t)$ defined in (2.77).

We shall assume that the ON and OFF periods in each model have the same distributions: $P(X_{j,\mathrm{on}}(m) > x) = \bar{F}_{\mathrm{on}}(x)$, $P(X_{j,\mathrm{off}}(m) > x) = \bar{F}_{\mathrm{off}}(x)$, where $X_{j,\mathrm{on}}(m)$, $X_{j,\mathrm{off}}(m)$ $(t \in \mathbb{Z})$ are the consecutive ON and OFF periods, respectively, in the mth ON–OFF process $(m = 1, \ldots, M)$. Since $W^{(m)}(u)$ are stationary and have the same distribution for each m, we obtain

$$E\left[\int_0^{Tt} \sum_{m=1}^{M} W^{(m)}(u)\, du\right] = TME\big[W(0)\big]t = TM\frac{\mu_{\mathrm{on}}}{\mu_{\mathrm{on}} + \mu_{\mathrm{off}}} t = TM\frac{\mu_{\mathrm{on}}}{\mu} t.$$

Table 4.7 Limits for superposition of cumulative renewal reward processes—tails of interarrival times vs. tails of rewards. The tail parameters $\alpha \in (1,2)$, $\beta \in (1,2)$

Renewal reward processes		
	Rewards	
	$E[Y_1^2] < \infty$	$RV_{-\beta}$, $\beta \in (1,2)$
Interarrival times $E[X_1^2] < \infty$	$\lim_{M\to\infty}\lim_{T\to\infty} = \text{Bm}$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{Bm}$	$\lim_{M\to\infty}\lim_{T\to\infty} = Z_\beta$ $\lim_{T\to\infty}\lim_{M\to\infty} = Z_\beta$
Interarrival times $RV_{-\alpha}$, $\alpha \in (1,2)$	$\lim_{M\to\infty}\lim_{T\to\infty} = Z_\alpha$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{fBm}$	$\alpha < \beta$ $\lim_{M\to\infty}\lim_{T\to\infty} = Z_\alpha$ $\lim_{T\to\infty}\lim_{M\to\infty} = Z_\beta^*$ $\beta < \alpha$ $\lim_{M\to\infty}\lim_{T\to\infty} = Z_\beta$ $\lim_{T\to\infty}\lim_{M\to\infty} = Z_\beta$

Recall from Lemma 2.7 that the ON–OFF process has long memory (in the sense of Definition 1.4), or $\int_0^t W(u)du$ has long memory (in the sense of Definition 1.5) if the ON (or OFF) periods are heavy tailed. In this case we are interested in limit theorems for the superposition of ON–OFF processes. Such studies were conducted in Taqqu et al. (1997), Mikosch et al. (2002) or Dombry and Kaj (2011). Specifically, the following two theorems were proven in Taqqu et al. (1997).

Theorem 4.42 *Assume that ON and OFF periods satisfy* (2.78) *and* (2.79), *i.e.*

$$\bar{F}_{\text{on}}(x) = C_{\text{on}} x^{-\alpha_{\text{on}}}, \qquad \alpha_1 \in (1,2), \tag{4.193}$$

$$\bar{F}_{\text{off}}(x) = C_{\text{off}} x^{-\alpha_{\text{off}}}, \qquad \alpha_2 \in (1,2), \tag{4.194}$$

with $\alpha_{\text{on}} < \alpha_{\text{off}}$. *Then*,

$$\lim_{T\to\infty}\lim_{M\to\infty} \frac{W^*_{T,M}(t)}{T^H M^{1/2}} \overset{\text{d}}{=} C_{\text{on}}^{1/2}\sigma_{\text{on–off}} B_H(t),$$

where $(B_H(t), t \in (0,1))$ *is a fractional Brownian motion with Hurst parameter* $H = (3-\alpha_{\text{on}})/2$, *and*

$$\sigma^2_{\text{on–off}} = \frac{\mu^2_{\text{on–off}}}{(\alpha_{\text{on}}-1)\mu^3}.$$

Sketch of Proof

Step 1: Since $W^{(m)}(\cdot)$ $(m = 1,\ldots,M)$ are independent identically distributed bounded processes, application of Lemma 4.6 implies

$$\lim_{M\to\infty} \frac{1}{M^{1/2}} \sum_{m=1}^{M} \left\{ W^{(m)}(t) - E\left[W^{(m)}(t)\right] \right\} \Rightarrow G(t),$$

where $G(t)$ ($t \in [0,1]$) is a centred stationary Gaussian process with the covariance function $cov(W(0), W(t))$.

Step 2: Therefore, $\int_0^{Tt} G(t)\,du$ is still a Gaussian process with variance $\mathrm{var}(\int_0^{Tt} W(u)\,du)$. By Lemma 2.7, the variance grows at rate $C_{\mathrm{on}}\sigma_{\mathrm{on-off}}^2(Tt)^{2H}$ as $T \to \infty$, which is the same as for fractional Brownian motion. We conclude

$$\lim_{T\to\infty} \frac{1}{T^H} \int_0^{Tt} G(t)\,du \Rightarrow C_{\mathrm{on}}^{1/2}\sigma_{\mathrm{on-off}} B_H(t).$$

Step 3: Let

$$U(Tt) = \lim_{M\to\infty} \frac{W_{T,M}^*(t)}{T^H M^{1/2}}.$$

The tightness is verified by noting that as $T \to \infty$, for $t_1 < t_2$,

$$\begin{aligned} &E\big[\big(U(Tu_1) - U(Tu_2)\big)^2\big] \\ &\quad = T^{-2H}\,\mathrm{var}\left(\int_0^{T(t_2-t_1)} W(u)\,du\right) \sim C_1\sigma_{\mathrm{on-off}}^2(t_2-t_1)^{2H} \end{aligned}$$

and $2H > 1$. The tightness is verified by applying Lemma 4.5. □

However, similarly to the case of superposition of renewal processes, different orders of taking limits yield completely different limiting processes.

Theorem 4.43 *Assume that ON and OFF periods satisfy* (4.193) *and* (4.194) *with* $\alpha_{\mathrm{on}} < \alpha_{\mathrm{off}}$ *and* $\alpha_{\mathrm{on}} \in (1,2)$. *Then*

$$\lim_{M\to\infty}\lim_{T\to\infty} (MT)^{-1/\alpha} \int_0^{Tt} \left(\sum_{m=1}^{M} \big(W^{(m)}(u) - E\big[W^{(m)}(u)\big]\big)\right) du \stackrel{\mathrm{d}}{=} C_0 Z_\alpha(t), \tag{4.195}$$

where $Z_\alpha(t) \stackrel{\mathrm{d}}{=} t^{1/\alpha} S_\alpha(1,1,0)$ *is a Lévy process, and* $C_0 = (\frac{\mu_{\mathrm{off}}}{\mu^{1+1/\alpha}}) C_{\mathrm{on}}^{1/\alpha} C_\alpha^{-1/\alpha}$.

Sketch of Proof

Step 1: First, we show that for each $m = 1, \ldots, M$,

$$\lim_{T\to\infty} T^{-1/\alpha} \int_0^{Tt} \big\{W^{(m)}(u) - E\big[W^{(m)}(u)\big]\big\}\,du \stackrel{\mathrm{d}}{=} \left(\frac{\mu_{\mathrm{off}}}{\mu^{1+1/\alpha}}\right) C_{\mathrm{on}}^{1/\alpha} C_\alpha^{-1/\alpha} Z_\alpha^{(m)}(t), \tag{4.196}$$

where $Z_\alpha^{(m)}(t) \stackrel{\mathrm{d}}{=} t^{1/\alpha} S_\alpha(1,1,0)$ are independent Lévy processes.

If $Tt \le \tau_0$, then there are three scenarios possible: either, at 0, the process is ON, then $\int_0^{Tt} W(u)\,du = \min(Tt, X_{0,\mathrm{on}})$; or at 0, the process is OFF, and $X_{0,\mathrm{off}} > Tt$, then $\int_0^{Tt} W(u)\,du = 0$; or at 0, the process is OFF, and $X_{0,\mathrm{off}} < Tt$, then $\int_0^{Tt} W(u)\,du = Tt - X_{0,\mathrm{off}} \le \tau_0 - X_{0,\mathrm{off}} = X_{0,\mathrm{on}}$ (this last situation is shown on Fig. 4.7). In either case, $\int_0^{Tt} W(u)\,du \le X_{0,\mathrm{on}}$. Since $X_{0,\mathrm{on}}$ is a random variable with a finite mean, we conclude that $X_{0,\mathrm{on}}/T^{1/\alpha} \to 0$ in probability as $T \to \infty$.

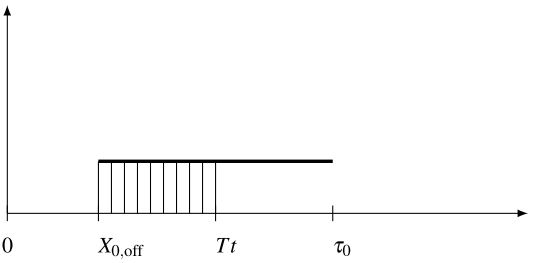

Fig. 4.7 ON–OFF process: The 0th interval starts with OFF period. The *marked area* shows $\int_0^{Tt} W(u)\,du$

If $Tt > \tau_0$, then

$$\int_0^{Tt} W(u)\,du = X_{0,\text{on}} + \sum_{j=1}^{N(Tt)} X_{j,\text{on}} - U,$$

where $U \le X_{N(Tt)+1,\text{on}}$. The first two terms represent the sum of all ON intervals that are at least partially included in $[0, Tt]$. For example, if $\tau_0 < Tt < \tau_1$, then $N(Tt) = 1$ and $\sum_{j=1}^{N(Tt)} X_{j,\text{on}} = X_{1,\text{on}}$; thus, both $X_{0,\text{on}}$ and $X_{1,\text{on}}$ are counted as fully included in $[0, Tt]$. Now, assume that the renewal intervals X_t start with ON periods. It may happen that either $\tau_0 + X_{1,\text{on}} = \tau_0 + X_{N(Tt),\text{on}} < Tt$, and then $U = 0$, or $\tau_0 + X_{1,\text{on}} > Tt$, and in the latter case we have to subtract a portion $(\tau_0 + X_{1,\text{on}} - Tt) \le X_{2,\text{on}}$ that is not included $[0, Tt]$. A similar consideration is valid if the renewal intervals X_t start with OFF periods.

We conclude that the only term that contributes to the limiting behaviour of $\int_0^{Tt} W(u)\,du$ is the sum $\sum_{j=1}^{N(Tt)} X_{j,\text{on}}$. In the same spirit,

$$Tt = X_{0,\text{on}} + X_{0,\text{off}} + \sum_{j=1}^{N(Tt)} X_{j,\text{on}} + \sum_{j=1}^{N(Tt)} X_{j,\text{off}} - Y,$$

where $Y \le X_{N(Tt)+1,\text{on}}$. Thus, informally,

$$\int_0^{Tt} E\big[W(u)\big]\,du = \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} Tt \approx \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} \left(\sum_{j=1}^{N(Tt)} X_{j,\text{on}} + \sum_{j=1}^{N(Tt)} X_{j,\text{off}} \right).$$

Consequently, the limiting behaviour of $T^{-1/\alpha} \int_0^{Tt} \{W(u) - E[W(u)]\}\,du$ is determined by

$$\frac{1}{T^{1/\alpha}} \sum_{j=1}^{N(Tt)} \big(J_j - E[J_j]\big),$$

where after some simple algebra

$$\begin{aligned} J_j &= X_{j,\text{on}} - \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} (X_{j,\text{on}} + X_{j,\text{off}}) \\ &= \frac{\mu_{\text{off}}}{\mu_{\text{on}} + \mu_{\text{off}}} \big(X_{j,\text{on}} - E[X_{j,\text{on}}]\big) - \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} \big(X_{j,\text{off}} - E[X_{j,\text{off}}]\big). \end{aligned}$$

Table 4.8 Limits for superposition of ON–OFF processes

Superposition of ON–OFF processes	
ON times with finite variance	$\lim_{M\to\infty}\lim_{T\to\infty} = \text{Bm}$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{Bm}$
ON times with infinite variance	$\lim_{M\to\infty}\lim_{T\to\infty} = \text{Lévy}$ $\lim_{T\to\infty}\lim_{M\to\infty} = \text{fBm}$

We thus have

$$\frac{1}{T^{1/\alpha}}\sum_{j=1}^{N(Tt)}\big(J_j - E[J_j]\big) = \left(\frac{N(Tt)}{T}\right)^{1/\alpha}\frac{1}{(N(Tt))^{1/\alpha}}\sum_{j=1}^{N(Tt)}\big(J_j - E[J_j]\big).$$

Recall that for a stationary renewal process $N(Tt)/T \to E[N(t)] = (\mu_{\text{on}} + \mu_{\text{off}})^{-1}\mu^{-1}t$ as $T\to\infty$. Therefore, the limiting behaviour of sum is the same as that of

$$\frac{t^{1/\alpha}}{\mu^{1/\alpha}}\frac{1}{(Tt)^{1/\alpha}}\sum_{j=1}^{Tt}\big(J_j - E[J_j]\big).$$

We note that, as $x\to\infty$,

$$P(J_1 > x) \sim \left(\frac{\mu_{\text{off}}}{\mu}\right)^{\alpha_{\text{on}}} C_{\text{on}}x^{-\alpha_{\text{on}}}, \qquad P(J_1 < -x) \sim \left(\frac{\mu_{\text{on}}}{\mu}\right)^{\alpha_{\text{off}}} C_{\text{off}}x^{-\alpha_{\text{off}}}.$$

Since $\alpha = \alpha_{\text{on}} < \alpha_{\text{off}}$, application of (4.80) yields

$$T^{-1/\alpha}\sum_{j=1}^{Tt}\big(J_j - E[J_j]\big) \Rightarrow \left(\frac{\mu_{\text{off}}}{\mu}\right)C_{\text{on}}^{1/\alpha}C_\alpha^{-1/\alpha}Z_\alpha(u),$$

where $Z_\alpha(t) \stackrel{\text{d}}{=} t^{1/\alpha}S_\alpha(1,1,0)$ is a Lévy process. We conclude that (4.196) holds.
Step 2: Since the Lévy processes $Z^{(m)}(t)$ are independent, the result follows.

□

If the ON and OFF times have a finite variance, similar arguments lead to a Brownian motion as a limit for both limiting regimes. We summarize our observations in Table 4.8.

Similar results as for renewal reward and ON–OFF hold for the Infinite Poisson source model, see Konstantopoulos and Lin (1998), Mikosch et al. (2002).

4.9.6 Simultaneous Limits and Further Extensions

What happens when T and M go to infinity simultaneously? The techniques described above fail. Following Mikosch et al. (2002), one can consider the parameter

M as an increasing function of T, i.e. $M = M(T)$. Alternatively, see Mikosch and Samorodnitsky (2007), one can consider the intensity of the point process τ_j to depend on a number of sources M. Consequently, following Mikosch and Samorodnitsky (2007), we consider the process

$$W^*_{\lambda_M,M}(t) = \sum_{m=1}^{M} W^{(m)*}(\lambda_M t) = \sum_{m=1}^{M} \int_0^{\lambda_M t} W^{(m)}(u)\,du,$$

where the $W(\cdot)$, $W^{(m)}(\cdot)$ $(m \geq 1)$ are independent copies of either a renewal reward, an ON–OFF or an $M/G/\infty$ process. We observe that an increase in the intensity can be interpreted as an increase in time in our original cumulative process $W^*_{T,M}(t)$.

Define also a scaling sequence

$$a_M = \sqrt{M \operatorname{var}\left(\int_0^{\lambda_M} W(u)\,du\right)}.$$

In the examples considered above (i.e. renewal reward, ON–OFF, $M/G/\infty$) we have

$$\operatorname{var}\left(\int_0^{\lambda_M} W(u)\,du\right) \sim C\lambda_M^{3-\alpha} L(\lambda_M).$$

For fixed t, convergence of $a_M^{-1} W^*_{\lambda_M,M}(t)$ follows from a classical limit theorem for i.i.d. arrays. Indeed, for some $\delta > 0$, using Hölder's inequality and stationarity of $W(u)$,

$$E\left[\left|W^*_{\lambda_M,M}(t)\right|^{2+\delta}\right] \leq (\lambda_M t)^{1+\delta} \int_0^{\lambda_M t} E\left[\left|W(u) - E\left[W(u)\right]\right|^{2+\delta}\right] du \leq C(\lambda_M t)^{2+\delta}$$

as long as $E[|W(0)|^{2+\delta}] < \infty$. In particular, this is fulfilled for the ON–OFF model and both, renewal reward and $M/G/\infty$, as long as $E[Y_1^{2+\delta}] < \infty$.

If this is the case, we conclude that

$$M^{-\delta/2} \frac{E[|W^*_{\lambda_M,M}(t)|^{2+\delta}]}{(\operatorname{var}(\int_0^{\lambda_M} W(u)\,du))^{1+\delta/2}} \sim M^{-\delta/2} \frac{(\lambda_M t)^{2+\delta}}{\lambda_M^{(3-\alpha)(1+\delta/2)} L^{1+\delta/2}(\lambda_M)}.$$

For each t, the last expression converges to 0 as long as

$$\lambda_M = o\left(M^{1/(\alpha-1+\delta)}\right) \tag{4.197}$$

for some $\delta > 0$.

For each t, we conclude the convergence of $a_M^{-1} W^*_{\lambda_M,M}(t)$ to a normal distribution. The tightness follows clearly from

$$\operatorname{var}\left(a_M^{-1} W^*_{\lambda_M,M}(t-s)\right) = a_M^{-2} M \operatorname{var}\left(\int_0^{\lambda_M(t-s)} W(u)\,du\right) \leq C(t-s)^{3-\alpha}.$$

Therefore, under the fast growth condition (4.197), we conclude the convergence to an fBm. Of course, if we set $\lambda_M = T$, then, as $M \to \infty$, condition (4.197) is clearly fulfilled, and we may recover the convergence in the $\lim_{T\to\infty} \lim_{M\to\infty}$ scheme.

Condition (4.197) is called a *fast growth condition*. Indeed, it means that the number M of sources grows faster than the intensity λ_M, which as mentioned above, can be interpreted as time.

It should be mentioned that in the original paper, Mikosch et al. (2002), the *fast growth* for an $M/G/\infty$ process is defined as

$$\lim_{T\to\infty} \lambda_T T^{1-\alpha} = \infty. \tag{4.198}$$

On the other hand, the *slow growth* is defined as

$$\lim_{T\to\infty} \lambda_T T^{1-\alpha} = 0. \tag{4.199}$$

Similar conditions are imposed in the ON–OFF (Mikosch et al. 2002) or renewal reward context (Taqqu 2002, Pipiras et al. 2004). Roughly speaking, fast growth corresponds to convergence to an fBm, whereas slow growth is responsible for a stable convergence.

Furthermore, similar results to those presented here can be obtained for very general Poisson shot-noise and cluster processes; see Klüppelberg et al. (2003), Klüppelberg and Kühn (2004), Faÿ et al. (2006), Rolls (2010).

However, the picture may change if we consider more complicated models. In particular, we may obtain an fBm limit even in a slow growth regime (see Mikosch and Samorodnitsky 2007, Fasen and Samorodnitsky 2009).

Furthermore, if the limit in (4.199) is a finite, nonnegative constant, then the limiting process is a fractional Poisson process, see Dombry and Kaj (2011).

4.10 Limit Theorems for Extremes

In this section we study the limiting behaviour of partial maxima based on a stationary sequence X_t ($t \in \mathbb{Z}$). We start by recalling some basic results for i.i.d. sequences and illustrating Fréchet and Gumbel domains of attraction. Then, for long-memory sequences, we separate our discussion into the Gumbel and the Fréchet case. A primary example for the first situation is a stationary Gaussian sequence. We argue that there is no influence of dependence (in particular, of long memory) on the limiting behaviour of maxima (Berman 1964, 1971; Leadbetter et al. 1978, 1983; Buchmann and Klüppelberg 2005, 2006). On the other hand, there is no available theory for general linear processes with long memory in the Gumbel case. Furthermore, Breidt and Davis (1998) argue that maxima of Gaussian-based stochastic volatility models (with possible long memory) behave as if the random variables were independent.

Next, we turn our attention to the Fréchet domain of attraction. There, the main tool is point process convergence studied in Sect. 4.3. As we will see, the rate of convergence of maxima of linear processes (weakly or strongly dependent) is the same

as for i.i.d. sequences, however, dependence implies that the so-called extremal index is smaller than one (Davis and Resnick 1985). On the other hand, extremes of heavy-tailed stochastic volatility models (with possible long memory) behave again like independent random variables (Davis and Mikosch 2001; Kulik and Soulier 2012, 2013).

These considerations in the Gumbel and Fréchet case may suggest that *long memory does not play any role in the limiting behaviour of maxima*. However, the picture is much more complicated. This will be illustrated by looking at the extremal behaviour of general stationary stable processes in Sect. 4.10.3. That theory was developed in Samorodnitsky (2004, 2006) and Resnick and Samorodnitsky (2004).

We start our discussion with a sequence X_t $(t \in \mathbb{Z})$ of i.i.d. random variables with common distribution function F. Define partial maxima by $M_n = \max\{X_1, \ldots, X_n\}$. The classical Fisher–Tippett theorem identifies three possible limits for M_n. We refer to Chap. 3 in Embrechts et al. (1997) for further details and examples.

Theorem 4.44 *Assume that X_t $(t \in \mathbb{Z})$ is a sequence of i.i.d. random variables. If there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ and a non-degenerate distribution function Λ such that*

$$c_n^{-1}\big(\max\{X_1, \ldots, X_n\} - d_n\big) \xrightarrow{d} \Lambda,$$

then Λ is one of the following distributions: *Fréchet*, *Weibull or Gumbel*, *defined by the cumulative distribution functions*

$$\begin{aligned}
\Lambda_{\text{Frechet}}(x) &= \exp\big(-x^{-\alpha}\big) \quad (x > 0, \alpha > 0),\\
\Lambda_{\text{Weibull}}(x) &= \exp\big(-(-x)^{-\alpha}\big) \quad (x < 0, \alpha > 0),\\
\Lambda_{\text{Gumbel}}(x) &= \exp\big(-\exp(-x)\big) \quad (x > 0).
\end{aligned}$$

Example 4.28 Assume that X_t $(t \in \mathbb{N})$ are standard normal. Choose $c_n = (2\ln n)^{-1/2}$ and

$$d_n = \frac{1}{2^{1/2}}\left\{2(\log n)^{1/2} - \frac{\log\log n + \log(4\pi)}{2\sqrt{\log n}}\right\}.$$

Then the limiting distribution is Gumbel.

Example 4.29 Assume that X_t $(t \in \mathbb{N})$ fulfill

$$P(X_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(X_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha}. \tag{4.200}$$

(The left-tail behaviour is not needed here, however, we include it for completeness.) Let $A_\beta = A\frac{1+\beta}{2}$. Then,

$$P\big((A_\beta n)^{-1/\alpha}\max\{X_1, \ldots, X_n\} \le x\big) = F^n\big(A_\beta^{1/\alpha} x n^{1/\alpha}\big) = \big(1 - \bar{F}\big(A_\beta^{1/\alpha} x n^{1/\alpha}\big)\big)^n,$$

where $\bar{F}(x) = 1 - F(x)$. Hence, for n large enough,

$$P\big((A_\beta n)^{-1/\alpha} \max\{X_1, \ldots, X_n\} \le x\big) = \left(1 - \frac{x^{-\alpha}}{n}\right)^n \to \exp\big(-x^{-\alpha}\big)$$

as $n \to \infty$. In this case $d_n = 0$, $c_n = (A_\beta n)^{1/\alpha}$, and the limiting law is Fréchet.

These examples identify two main classes of distributions and their corresponding extreme value behaviour: (a) the class of regularly varying distributions, that is $\bar{F}(x) = x^{-\alpha} L(x)$ as $x \to \infty$, where L is a slowly varying function; then the limit is Fréchet; and (b) a class of (informally speaking) light-tailed distributions with unbounded support, like normal, log-normal or Gamma; then the limit is Gumbel. The first class is called the *domain of attraction of the Fréchet law*, and the second one the *domain of attraction of the Gumbel law*. The third type, Weibull, appears when the distribution has a bounded support, with a regularly varying behaviour at a boundary. This case will not be discussed here.

In the context of the examples above, a natural question is what happens if we drop the i.i.d. assumption. We will discuss this problem separately for the Fréchet and Gumbel domains of attraction respectively.

4.10.1 Gumbel Domain of Attraction

It turns out that maxima of a (possibly LRD) Gaussian sequence X_t ($t \in \mathbb{N}$) behaves as if the random variables X_t ($t \in \mathbb{N}$) were independent.

Theorem 4.45 *Let X_t ($t \in \mathbb{N}$) be a stationary Gaussian process with covariance function $\gamma(k)$ such that Berman's condition holds*:

$$\lim_{k \to \infty} \log(k)\gamma(k) = 0. \tag{4.201}$$

Then

$$c_n^{-1}\big(\max(X_1, \ldots, X_n) - d_n\big) \xrightarrow{d} \Lambda_{\text{Gumbel}},$$

where $c_n = (2 \log n)^{-1/2}$, and

$$d_n = \frac{1}{2^{1/2}} \left\{ 2(\log n)^{1/2} - \frac{\log\log n + \log(4\pi)}{2\sqrt{\log n}} \right\},$$

cf. Example 4.28.

Proof The proof is only sketched here; some additional technical details can be found in Berman (1964) or Leadbetter et al. (1978, 1983).

We start with the following special version of the normal comparison lemma (see Lemma 3.2 in Leadbetter et al. 1983). For each y,

$$\left| P\big(\max\{X_1,\dots,X_n\} \le y\big) - \prod_{t=1}^{n} P(X_t \le y) \right|$$
$$\le Cn \sum_{k=1}^{n} \big|\gamma_X(k)\big| \exp\big(-y^2/\big(1+\big|\gamma_X(k)\big|\big)\big).$$

Next, let us fix x and define $u_n = c_n x + d_n$. Then, since $c_n \to 0$ and $d_n \to \infty$, $u_n \sim d_n$ as $n \to \infty$. Furthermore,

$$d_n^2 = 2\log n + \frac{1}{8}\frac{(\log\log n + \log(4\pi))^2}{\log n} - \log\log n \sim 2\log n - \log\log n.$$

Hence,

$$\exp\big(-u_n^2/2\big) \sim \exp\big(-d_n^2/2\big) \sim n^{-1}\sqrt{\log n} \sim \frac{u_n}{\sqrt{2n}}.$$

We may write

$$n\big|\gamma_X(k)\big| \exp\left(-\frac{u_n^2}{(1+|\gamma_X(k)|)}\right) = n\big|\gamma_X(k)\big| \exp\big(-u_n^2\big) \exp\left(-\frac{u_n^2|\gamma_X(k)|}{(1+|\gamma_X(k)|)}\right).$$

Let $\beta > 0$ and $k > n^\beta$. Define $v_n = \sup_{k \ge n^\beta} |\gamma_X(k)|$. Note that

$$v_n u_n^2 \sim 2v_n \log(n) 2\frac{\log n}{\log n^\beta} v_n \log n^\beta = \frac{2}{\beta} v_n \log n^\beta \to 0$$

as $\gamma(n)\log(n) \to 0$. We note that this is exactly the place that Breiman's condition plays a role. Therefore,

$$n \sum_{k=n^\beta}^{n} \big|\gamma_X(k)\big| \exp\left(-\frac{u_n^2}{(1+|\gamma_X(k)|)}\right)$$
$$\le n \exp\big(-u_n^2\big) v_n \sum_{k=n^\beta}^{n} \exp\left(\frac{u_n^2|\gamma_X(k)|}{(1+|\gamma_X(k)|)}\right)$$
$$\le n^2 \exp\big(-u_n^2\big) v_n \exp\big(u_n^2 v_n\big) \le C v_n u_n^2 \exp\big(u_n^2 v_n\big) \to 0.$$

On the other hand, there exists $\delta > 0$ such that $1 + |\gamma_X(k)| < 2 - \delta$. Then

$$
\begin{aligned}
& n \sum_{k \le n^\beta} \left|\gamma_X(k)\right| \exp\left(-\frac{u_n^2}{1+|\gamma_X(k)|}\right) \\
& \quad \le n \sum_{k \le n^\beta} \left|\gamma_X(k)\right| \exp\left(-\frac{u_n^2}{2-\delta}\right) \\
& \quad \sim n n^{-2/(2-\delta)} (\log n)^{1/(2-\delta)} \sum_{k \le n^\beta} \left|\gamma_X(k)\right| \le C n^{1+\beta} n^{-2/(2-\delta)} (\log n)^{1/(2-\delta)}
\end{aligned}
$$

since we may assume without loss of generality that $|\gamma_X(k)| \le 1$. The bound converges to 0 when $\beta < \delta/(2+\delta)$. This finishes the proof. □

In Theorem 4.45 we considered a discrete-time process X_t ($t \in \mathbb{Z}$). The result can be extended to general continuous-time Gaussian processes, in particular to fractional Brownian motion $B_H(u)$; see Berman (1971). Furthermore, the result extends to stochastic differential equations driven by fBm. To illustrate this, we consider a continuous-time process $Y(u)$ ($u \in \mathbb{R}$) that solves

$$
Y(v) - Y(u) = \int_u^v \mu\big(Y(s)\big)\, ds + \int_u^v \sigma\big(Y(s)\big)\, dB_H(s) \quad (u < v), \tag{4.202}
$$

where $\mu(\cdot)$ and $\sigma(\cdot) > 0$ are deterministic functions. We recall from Sect. 2.2.5.2 that if $\mu(x) = \mu < 0$, $\sigma(x) = \sigma$, then the solution is a fractional Ornstein–Uhlenbeck process

$$
Y(u) = \mathrm{FOU}(u) = \sigma \int_{-\infty}^u \exp\big(\mu(u-v)\big)\, dB_H(v).
$$

The general Berman theory applies and

$$
c_T^{-1}\Big(\max_{0 \le u \le T} \mathrm{FOU}(u) - d_T\Big) \xrightarrow{d} \Lambda_{\mathrm{Gumbel}},
$$

where

$$
\begin{aligned}
c_T &= \sigma(-\mu)^{-H} \sqrt{\Gamma(H+1/2)} (2\log T)^{-1/2}, \\
d_T &= \frac{(\Gamma(H+1/2))^{1/2}}{2^{1/2}(-\mu)^H} \left\{ 2(\log n)^{1/2} + \frac{1-H}{2H} \frac{\log\log T}{(\log T)^{1/2}} + \frac{C_0}{(\log T)^{1/2}} \right\}
\end{aligned}
$$

with a constant C_0. We note that the rate of convergence does not depend on the Hurst parameter H. This convergence can be treated as the counterpart to the discrete-time situation in Theorem 4.45.

More generally, Buchmann and Klüppelberg (2005, 2006) study processes of the form $Y_\psi(u) = \psi(\mathrm{FOU}(u))$, where $\mathrm{FOU}(u)$ is a fractional Ornstein–Uhlenbeck process, and ψ is a function. Under general conditions established in those papers,

$Y_\psi(u)$ solves (4.202), and the inverse function ψ^{-1} of ψ fulfills

$$\psi^{-1}(u) = \int_{\psi(0)}^{u} \frac{ds}{\sigma(s)}.$$

Furthermore, the authors give general conditions that guarantee

$$\left(c_T^*\right)^{-1}\left(\max_{0\le u\le T} Y_\psi(u) - \psi(d_T)\right) \xrightarrow{d} \Lambda_{\text{Gumbel}}, \tag{4.203}$$

where c_T^* is possibly different than c_T. The form of c_T^* depends on assumptions on ψ. For example, if

$$\lim_{y\to\infty} \frac{\psi(y + x/y) - \psi(y)}{\psi(y + 1/y) - \psi(y)} = x,$$

then

$$c_T^* = \frac{2^{1/2}(-\mu)^{2H}}{\Gamma(2H+1)}\left\{\psi\left(d_T + \frac{1}{d_T}\right) - \psi(d_T)\right\}.$$

In particular, we can choose $\psi(x) = \exp(x^q)$, $q \in (0, 2)$. Then (4.203) holds with c_T^* as above. We note further that this is not applicable when $q = 2$. Then the limiting distribution is Gumbel. Indeed, note that when Z is standard normal, then e^{Z^2} has a regularly varying tail and hence cannot belong to the Gumbel domain of attraction. We refer to Buchmann and Klüppelberg (2005, 2006) for further results.

A natural question arises. Can we generalize the theorem above to linear processes $X_t = \sum_{k=0}^{\infty} a_k \varepsilon_{t-k}$, where ε_t $(t \in \mathbb{Z})$ belong to the domain of attraction of the Gumbel law? The answer is affirmative for weakly dependent sequences. Davis and Resnick (1988, p. 61; see also Rootzén 1986) show that if

$$P\left(c_n^{-1}\left(\max\{\varepsilon_1, \ldots, \varepsilon_n\} - d_n\right) < x\right) \to_d \Lambda(x),$$

then for the partial maxima of the linear process, we have

$$P\left(c_n^{-1}\left(\max\{X_1, \ldots, X_n\} - d_n\right) < x\right) \to_d \Lambda^\theta(x)$$

with some $\theta \in (0, 1)$. The parameter θ is called the *extremal index* and describes the contribution of dependence to the limiting law (see Embrechts et al. 1997 for more details). However, the authors assumed, in particular, that $\sum_{k=0}^{\infty} |a_k| < \infty$, so that long memory is excluded. At the moment there do not seem to be any results for linear processes in the case of long memory.

Breidt and Davis (1998) study stochastic volatility models

$$X_t = \xi_t \sigma_t = \xi_t \exp(\eta_t/2),$$

where ξ_t $(t \in \mathbb{N})$ is an i.i.d. standard normal sequence, independent of the stationary zero-mean Gaussian sequences η_t. After log-transformation, the sequence

$$Y_t := \log X_t^2 = \eta_t + \log \xi_t^2$$

is represented as the sum of a stationary Gaussian sequence and the log of a χ_1^2 random variables. The tail of Y_t has a complicated form, nevertheless it belongs to the domain of attraction of the Gumbel law. A modification of the normal comparison lemma allows us to prove the following result.

Theorem 4.46 *Let X_t $(t \in \mathbb{N})$ be a stochastic volatility model*

$$X_t = \xi_t \exp(\eta_t/2),$$

where ξ_t $(t \in \mathbb{N})$ is an i.i.d. standard normal sequence, independent of the stationary zero-mean Gaussian sequence η_t. Assume that the covariance function of η_t satisfies Berman's condition (4.201), *and let $Y_t = \log X_t^2$. Then*

$$c_n^{-1}\big(\max(Y_1, \dots, Y_n) - d_n\big) \xrightarrow{d} \Lambda_{\text{Gumbel}},$$

where $c_n = (2\log n)^{-1/2}$,

$$d_n \sim 2\psi_1(\log n)^{1/2} + \psi_2 \log\big((2\log n)^{1/2}\big) - \psi_3(2\log n)^{-1/2}(\log\log n + \psi_4) + \psi_5,$$

where $\psi_1, \psi_2, \psi_3, \psi_4$ are positive constants, and $c_5 \in \mathbb{R}$.

We observe no influence of possible long memory in volatility on the limiting behaviour of maxima. As for Gaussian sequences considered in Theorem 4.45, the only difference appears in the form of the centering constants d_n.

4.10.2 Fréchet Domain of Attraction

Recall Example 4.29. If the random variables are i.i.d. such that (4.200) holds, then the limiting distribution is Fréchet. This result can also be obtained using point processes. We recall from Sect. 4.3, Theorem 4.13, that

$$N_n := \sum_{t=1}^{n} \delta_{\tilde{c}_n^{-1} X_t} \Rightarrow \sum_{l=1}^{\infty} \delta_{j_l} =: N,$$

where j_l are points of a Poisson process with intensity measure

$$d\lambda(x) = \alpha\left[\frac{1+\beta}{2} x^{-(\alpha+1)} 1\{0 < x < \infty\} + \frac{1-\beta}{2}(-x)^{-(\alpha+1)} 1\{-\infty < x < 0\}\right] dx, \tag{4.204}$$

and $\tilde{c}_n$ is such that $P(|X_1| > \tilde{c}_n) \sim n^{-1}$, that is $\tilde{c}_n \sim A^{1/\alpha} n^{1/\alpha}$. We note that the event $\{\max\{X_1, \dots, X_n\} \le x\}$ is equivalent to {no points of N_n in (x, ∞)}. Hence,

for $x > 0$,

$$P\big(\tilde{c}_n^{-1} \max\{X_1, \dots, X_n\} \le x\big) = P\big(N_n(x, \infty) = 0\big) \to P\big(N(x, \infty) = 0\big)$$
$$= \exp\bigg(-\int_x^\infty d\lambda(u)\bigg) = \exp\bigg(-\frac{1+\beta}{2} x^{-\alpha}\bigg).$$

Changing the scaling from $\tilde{c}_n$ to $c_n = (A_\beta n)^{1/\alpha}$, we immediately conclude

$$P\big(c_n^{-1} \max\{X_1, \dots, X_n\} \le x\big) \to \exp\big(-x^{-\alpha}\big) = \Lambda_{\text{Frechet}}(x).$$

This approach to extremes via point processes can be generalized to dependent sequences, including series with long memory.

We start with linear processes. As in Sect. 4.3, we assume that $X_t = \sum_{k=0}^\infty a_k \varepsilon_{t-k}$, where the random variables ε_t are i.i.d. with a regularly varying distribution, that is

$$P(\varepsilon_1 > x) \sim A \frac{1+\beta}{2} x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A \frac{1-\beta}{2} x^{-\alpha}. \tag{4.205}$$

If $\alpha \in (1, 2)$, we assume also that $E(\varepsilon_1) = 0$. Of course, since ε_t are i.i.d.,

$$P\big(c_n^{-1} \max\{\varepsilon_1, \dots, \varepsilon_n\} \le x\big) \to \exp\big(-x^{-\alpha}\big) = \Lambda_{\text{Frechet}}(x),$$

where $c_n = (A_\beta n)^{1/\alpha}$.

We saw in Sect. 4.3 that

$$P(X_1 > x) \sim D_\alpha P(\varepsilon_1 > x), \qquad P(X_1 < -x) \sim D_\alpha P(\varepsilon_1 < -x),$$

where the constant $D_\alpha = \sum_{j=0}^\infty |a_j|^\alpha$ is assumed to be finite. Hence, if X_t^* ($t \in \mathbb{Z}$) is an i.i.d. sequence with the same marginal distribution as X_t, then with the same $c_n = (A_\beta n)^{1/\alpha}$,

$$P\big(c_n^{-1} \max\{X_1^*, \dots, X_n^*\} \le x\big) \to \exp\big(-D_\alpha x^{-\alpha}\big). \tag{4.206}$$

We note that the constant D_α does not play the role of the extremal index (for the definition see e.g. Embrechts et al. 1997) because the i.i.d. random variables X_t^* have the tail $P(X_1 > x) \sim D_\alpha P(\varepsilon_1 > x)$. The limiting distribution above will serve as a benchmark for comparison with dependent linear processes X_t that have the same marginal distribution as X_t^*. To do this, we will assume without loss of generality that $D_\alpha = 1$.

In Theorem 4.14 we showed, in particular, the following convergence of point processes:

$$\sum_{t=1}^n \delta_{\tilde{c}_n^{-1} X_t} \Rightarrow \sum_{l=1}^\infty \sum_{r=0}^\infty \delta_{j_l a_r},$$

where $\tilde{c}_n \sim A^{1/\alpha} n^{1/\alpha}$. Let us also assume for simplicity that all coefficients a_j are nonnegative. When restricted to $(0, \infty)$, the limiting Poisson process has the inten-

sity measure (cf. Davis and Resnick 1985)

$$\alpha \frac{1+\beta}{2} a_+^{\alpha} x^{-(\alpha+1)} dx,$$

where $a_+ = \max_j a_j$. The same argument as described above for the i.i.d. case leads to the following result on sample extremes for heavy-tailed processes with possible long memory. Limiting behaviour of extremes follows directly from Lemma 4.19 and Theorem 4.14, under the assumptions therein.

Theorem 4.47 *Let X_t $(t \in \mathbb{Z})$ be a linear process where the innovations ε_t $(t \in \mathbb{Z})$ are i.i.d. random variables such that (4.205) holds and $E(\varepsilon_1) = 0$ if $\alpha \in (1,2)$. Suppose that either for some $\delta < \alpha$,*

$$\sum_{j=0}^{\infty} |a_j| + \sum_{j=0}^{\infty} |a_j|^{\delta} < \infty,$$

or $a_j \sim c_a j^{d-1}$, $d \in (0, 1 - 1/\alpha)$, and ε_t $(t \in \mathbb{Z})$ are symmetric with $\alpha \in (1,2)$. Moreover, assume that $D_\alpha = 1$ and $a_j \geq 0$. Then with $c_n = (A_\beta n)^{1/\alpha}$,

$$P\big(c_n^{-1} \max\{X_1, \ldots, X_n\} \leq x\big) \to \exp\big(-a_+ x^{-\alpha}\big).$$

This result should be compared with the expression (4.206) for $X_1^*, \ldots, X_n^*$ (with $D_\alpha = 1$). The additional term $\theta := a_+ \in (0,1]$ in the limiting distribution in Theorem 4.47 is the extremal index and describes the effect of dependence on the limiting behaviour of extremes. Since the coefficients a_j are positive, extreme values of the sequence X_t are generated by large positive values of the sequence ε_t. If some of the coefficients are negative, large positive values of X_t are possibly due to large negative values of the innovations, and hence the extremal index will change:

$$\theta = a_+ + a_- \frac{1-\beta}{1+\beta},$$

where $a_- = \max\{\max(-a_j), 0\}$. We refer to Davis and Resnick (1985) and Embrechts et al. (1997) for more details.

We continue our discussion with heavy-tailed stochastic volatility models, as studied in Sect. 4.3.4. We assume that $X_t = \xi_t \sigma_t$, where ξ_t are i.i.d. such that

$$P(\xi_1 > x) \sim A \frac{1+\beta}{2} x^{-\alpha}, \qquad P(\xi_1 < -x) \sim A \frac{1-\beta}{2} x^{-\alpha}. \tag{4.207}$$

We will assume also for simplicity that the sequences σ_t and ξ_t are independent from each other. Then, $P(X_1 > x) \sim A E(\sigma_1^{\alpha}) \frac{1+\beta}{2} x^{-\alpha}$. Hence, if $X_1^*, \ldots, X_n^*$ are independent copies of X_1, then with $c_n = (A_\beta n)^{1/\alpha}$,

$$P\big(c_n^{-1} \max\{X_1^*, \ldots, X_n^*\} \leq x\big) \to \exp\big(-E\big(\sigma_1^{\alpha}\big) x^{-\alpha}\big).$$

Again, the constant $E(\sigma_1^\alpha)$ is related the to marginal behaviour of X_t, not to the dependence structure. In Theorem 4.18 we concluded that the point process based on $X_1, \ldots, X_n$ has the same limit as for the corresponding i.i.d. copies $X_1^*, \ldots, X_n^*$. Directly from Theorem 4.18 we conclude that the limiting behaviour maxima associated with heavy-tailed stochastic volatility models is the same as in the i.i.d. case. There is no influence of any dependence in volatility.

Theorem 4.48 *Consider the LMSV model* $X_t = \xi_t \sigma_t$ $(t \in \mathbb{N})$ *such that* (4.207), *the Breiman condition* (4.94) *and* $E(\sigma_1^{\alpha+\varepsilon}) < \infty$ *with some* $\varepsilon > 0$ *hold. Also, assume that* σ_t $(t \in \mathbb{N})$ *is ergodic. Then*

$$P\big(c_n^{-1} \max\{X_1, \ldots, X_n\} \le x\big) \to \exp\big(-E\big(\sigma_1^\alpha\big)x^{-\alpha}\big).$$

4.10.3 Stationary Stable Processes

Samorodnitsky (2004, 2006) considers a general stationary symmetric α-stable (SαS) process X_t that can be represented by $X_t = \int g_t(s)\, dM(s)$, where M is an SαS random measure. As mentioned in Sect. 1.3.6.3, such processes can be decomposed into a dissipative and a conservative part. As we will indicate below, the dissipative part has no influence on the limiting behaviour of maxima, whereas the conservative part does.

Rosiński (1995) argues that the class of ergodic SαS processes that are generated by the dissipative flow coincides with the class of moving averages $X_t = \int g_t(s)\, dM(s) = \int g(t-s)\, dM(s)$. In particular, consider a Linear Fractional Stable Motion

$$Z_{H,\alpha}(u) = \int_{-\infty}^{\infty} Q_{u,1}(x; H, \alpha)\, dZ_\alpha(x), \tag{4.208}$$

where $Z_\alpha(\cdot)$ is a symmetric α-stable (SαS) Lévy process,

$$Q_{u,1}(x; H, \alpha) = c_1\big[(u-x)_+^{H-1/\alpha} - (-x)_+^{H-1/\alpha}\big] + c_2\big[(u-x)_-^{H-1/\alpha} - (-x)_-^{H-1/\alpha}\big], \tag{4.209}$$

and $H > 1/\alpha$. Let $X_t = Z_{H,\alpha}(t) - Z_{H,\alpha}(t-1)$. Samorodnitsky (2004) proves that in this case

$$P\big(n^{-1/\alpha} \max\{X_1, \ldots, X_n\} \le x\big) \to \exp\big(-Cx^{-\alpha}\big),$$

where C is a positive constant. Hence, the rate of growth of maxima is the same as in the i.i.d. case. We observed this already in the case of moving averages considered in Theorem 4.47.

In contrast, a simple (non-ergodic) example of an SαS process generated by the conservative flow is given by $X_t = Z^{1/\beta}\varepsilon_t$, $(t \in \mathbb{N})$, where Z is a strictly positive α/β-stable random variable, and ε_t is a sequence of i.i.d. symmetric $S_\beta(1, 0, 0)$ random variables, independent of Z, and $0 < \alpha < \beta < 2$. Then, marginally, the random variables X_t are α-stable.

We recall that the β-stability and symmetry of random variables ε_t yield

$$P(\varepsilon_1 > x) \sim \frac{1}{2} C_\beta x^{-\beta},$$

cf. (4.75). Choosing $c_n = (C_\beta/2)^{1/\beta} n^{1/\beta}$, we have

$$\begin{aligned} P\big(c_n^{-1} \max\{X_1, \ldots, X_n\} \le x\big) &= E\big[P\big(c_n^{-1} \max\{\varepsilon_1, \ldots, \varepsilon_n\} \le Z^{-1/\beta} x | Z\big)\big] \\ &\to E\big[\exp\big(-x^{-\alpha} Z^{\alpha/\beta}\big)\big]. \end{aligned}$$

Hence, even though the random variables X_t are α-stable, the scaling involves β, not α. In other words, maxima grow slower than in the i.i.d. case. This is a general pattern for stable processes generated by a dissipative flow. We refer to Samorodnitsky (2004, 2006) and Resnick and Samorodnitsky (2004) for further details.

Chapter 5
Statistical Inference for Stationary Processes

5.1 Introduction

This chapter deals with statistical inference for long-range dependent linear and subordinated processes. Some of the tools will also be used in Chaps. 6 and 7 when we shall consider corresponding problems for nonlinear and nonstationary long-memory time series.

The first step in a statistical analysis is usually the estimation of location and scale parameters. Therefore, Sects. 5.2 and 5.3 are devoted to location and scale estimation, respectively. Suppose we observe $Y_t = \mu + X_t$ $(t = 1, 2, \ldots, n)$ where X_t $(t \in \mathbb{Z})$ is a strictly stationary process with $E(X_t) = 0$. The question is how far the dependence in X_t influences statistical inference about the location parameter μ. We discuss estimation of μ by the sample mean, using limit theorems established in Sects. 4.2 and 4.3 for finite and infinite variance processes, respectively. Resulting confidence intervals and test statistics involve unknown quantities such as a scale parameter and parameters characterizing the dependence structure. These parameters have to be estimated. In particular, one requires knowledge of the long-memory parameter d (or $H = d + \frac{1}{2}$). Usually, these parameters are estimated and plugged into formulas defining standardized statistics and confidence intervals. This may lead to a loss of accuracy, in particular since the long-memory parameter affects rates of convergence. Some possible improvements that can be applied to models such as a FARIMA(p, d, q) process (see Example 5.1) are discussed. Furthermore, we examine how far the sample mean may lose efficiency under long-range dependence or antipersistence.

Then, we turn our attention to M-estimators of location (Huber 1981). In contrast to the weakly dependent case, for linear long-memory sequences robust M-estimators have the same asymptotic efficiency as the sample mean. This was first stated in Beran (1991) in the context of Gaussian subordination and generalized later by Koul (1992), Koul and Mukherjee (1993), Giraitis et al. (1996a), and Koul and Surgailis (2001) to linear processes with finite or infinite variance, respectively. Proofs of such results rely on the reduction principle for empirical processes (Sect. 4.8) and limiting behaviour of the sample mean. However, it should be pointed

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_5, © Springer-Verlag Berlin Heidelberg 2013

out that the asymptotic equivalence of M-estimators and the sample mean does not hold in general for subordinated processes. Moreover, as we will see below, there is an infinite efficiency loss in the case of antipersistence (see Example 5.3).

In Sect. 5.3, we discuss the estimation of a scale parameter. It is assumed that we observe $Y_t = \mu + \sigma X_t$ $(t = 1, \ldots, n)$ where $\sigma > 0$ has to be estimated. A standard approach is to compute the sample variance s^2. In weakly dependent situations, the limiting behaviour of the sample variance does not change, if μ is replaced by a consistent estimator. This is no longer true for long-memory series, as illustrated in Beran and Ghosh (1991)) and Dehling and Taqqu (1991). Furthermore, the sample variance is not the best choice under long-range dependence. For $d > \frac{1}{4}$, the rate of convergence of s^2 is $O_p(n^{2d-1})$ which is slower than for parametric estimators that exploit the relation $\sigma^2 = \int_{-\pi}^{\pi} f_X(\lambda)\, d\lambda$ with f_X denoting the spectral density of X_t. This will be discussed later in Sect. 5.5. Also, the limiting distribution (for $d > \frac{1}{4}$) is quite complicated because it is of the Hermite–Rosenblatt type. Finally, under long memory the sample variance tends to underestimate the true variance.

Besides standard estimators such as the sample mean and the sample variance, or more generally, M-estimators, other methods have been discussed in the literature. For example, Mukherjee (1999) and Sibbertsen (2001) studied L- and S-estimators, respectively. These methods will not be discussed here since they have quite similar properties as M-estimators.

Rates of convergence of the sample mean, or M-estimators of location, involve the long-memory parameter d. Thus, to construct confidence intervals or statistical tests, one needs to estimate d (together with a scale parameter). In Sect. 5.4, we review some heuristic and/or graphical methods commonly used for long-memory identification, including the original R/S method proposed by Hurst (1951) that was studied later by Mandelbrot, or its modified version given in Lo (1991). Other well known approaches include the variance plot, the KPSS statistic, the rescaled variance method, detrended fluctuation analysis (DFA), and temporal aggregation. Some of these methods have been introduced briefly already in Sect. 1.2. Although they are easy to implement and may serve as descriptive tools and a first heuristic check, there are many reasons for using more sophisticated methods when it comes to actual statistical inference. First of all, the methods involve tuning (or cut-off) parameters that are usually based on a subjective visual impression. Depending on the choice of these parameters, one may arrive at completely different conclusions for the same data set. In principle, objective mathematical rules for selecting the cut-off parameters could be worked out under suitable conditions. However, these methods have other properties that restrict their applicability. For instance, they are not robust against departures from stationarity. In particular, trends can be interpreted incorrectly as long memory. Furthermore, even if appropriate cut-off parameters are used and the assumptions of stationarity and long memory are correct, the statistics used in the heuristic methods have poor convergence properties.

Hence, refined estimation procedures are needed. In Sect. 5.5, we start with parametric methods. First, it is assumed that we observe a Gaussian series and we consider maximum likelihood estimation (MLE). The resulting estimator converges with the rate $\sqrt{n}$ and is asymptotically normal (Yajima 1985; Dahlhaus

1989; Hosoya 1997). This result can be generalized to linear processes. The *exact* MLE presents some computational challenges. As an alternative, one therefore considers various approximate versions of the MLE, including Whittle's method (Sect. 5.5.2) and an approach based on the infinite autoregressive representation of X_t. The Whittle estimator and other approximate maximum likelihood methods are consistent (Hannan 1973) and asymptotically equivalent to the MLE (Fox and Taqqu 1986; Beran 1986, 1995; Yajima 1985; Dahlhaus 1989; Giraitis and Surgailis 1990; Horváth and Shao 1999). The main tool to study the asymptotic distribution of the Whittle estimator is the limiting behaviour of quadratic forms, considered in Sect. 4.5. For the autoregressive method, a central limit theorem for martingales can be used. Both estimators are attractive from the theoretical and practical point of view. First of all, proofs are easier than for the exact MLE. In particular, it is clearly visible why long memory does not influence the asymptotic behaviour. Whittle's estimator is obtained by minimizing the normalized periodogram $I_{n,X}(\lambda)/f_X(\lambda)$, where f_X is the spectral density of X_t ($t \in \mathbb{Z}$). If $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ is a linear process with i.i.d. innovations ε_t, then the normalized periodogram can be approximated by $2\pi I_{n,\varepsilon}$ and hence long memory disappears. This approximation is not valid, however, for subordinated sequences and hence the $\sqrt{n}$-rate of convergence no longer holds (Giraitis and Taqqu 1999a, 1999b). The approximate MLE based on the infinite autoregressive representation $X_t - \sum_{j=1}^{\infty} b_j X_{t-j} = \varepsilon_t$ is defined by minimizing the residual sum of squares $\sum \varepsilon_t^2(\eta)$ with respect to a parameter η vector that includes d. Denoting by η^0 the true value of η and by $\dot{\varepsilon}_t$ the derivative of ε_t with respect to η, the asymptotic distribution of $\hat{\eta}$ then essentially follows from the asymptotic distribution of $\sum \dot{\varepsilon}_t(\eta^0)\varepsilon_t(\eta^0)$. Since $\dot{\varepsilon}_t(\eta^0)\varepsilon_t(\eta^0)$ is a martingale difference, a central limit theorem with $\sqrt{n}$-rate of convergence (i.e. $O_p(n^{-\frac{1}{2}})$) follows. From the computational point of view, the Whittle estimator and the AR-estimator are considerably faster than the exact MLE: For the Whittle estimator, one can use a further approximation based on a Riemann sum with Fourier frequencies only. This is particularly attractive because it is shift-invariant (i.e. centring by an estimator of the mean has no effect). This is very important in a long-memory setting since the sample mean has a slow rate of convergence. Moreover, the Fast Fourier Transform (FFT) can be used which makes computations very fast. The approach based on the AR-representation only requires the calculation of parameter-dependent residuals. This can be done using efficient algorithms for linear filters.

The fast rate of convergence of (approximate) maximum likelihood estimators is in particular due to the assumption that a specific parametric model is correct. In practice, this may be rather an optimistic assumption. There is no guarantee that we are able to pick the correct model a priori. In fact, as George Box has once remarked, "no model is correct, but some are useful." In this spirit, there are essentially two approaches to estimating d without knowledge of the "true" model. The first approach is to combine parametric models with a model selection criterion. The best known method is Akaike's information criterion (AIC; Akaike 1973, 1974; Shibata 1976) and related approaches such as BIC or HIC (Schwarz 1978; Hannan and Quinn 1979). This will be discussed in Sect. 5.5.6. In particular, it turns out that not only the asymptotic distribution of the MLE is of the same form as for short-memory

sequences but also the AIC can be derived the same way, and consistency of the BIC holds (Beran et al. 1998). Another approach is semiparametric estimation. In semiparametric estimation, one exploits the simple form of the pole (or root) the spectral density has at zero, namely $f_X(\lambda) \sim L(\lambda)|\lambda|^{-2d}$ (as $|\lambda| \to 0$). This expression is simple in the sense that near the origin only the slowly varying function L and the long-memory parameter d determine the value of f_X. If $L(\lambda)$ converges to a constant c_f, then we have a simple linear relationship $\log f_X(\lambda) \approx \beta_0 + \beta_1 u(\lambda)$ with $\beta_0 = \log c_f$, $u(\lambda) = \log \lambda$ and $\beta_1 = -2d$. Thus, the model one uses is

$$f_X(\lambda) = |\lambda|^{-2d} f_*(\lambda) \quad \big(\lambda \in [-\pi, \pi]\big)$$

where $f_*(\lambda)$ is an (essentially) arbitrary integrable nonnegative function except that for $\lambda \to 0$ we have $f_*(\lambda) \sim L(\lambda)$ as $\lambda \to 0$ and $c_f = \lim_{\lambda \to 0} f_*(\lambda)$ exists and is finite. The model is semiparametric because no assumptions on the shape of $f_*(\lambda)$ outside an arbitrarily small neighbourhood of the origin are made. Semiparametric methods are consistent without specifying a particular model and the asymptotic distribution does not depend on unknown parameters. The price one pays for this generality is a slower rate of convergence than for the MLE and other parametric methods. It should be pointed out, however, that, strictly speaking, in a general setting where models not included in the parametric class are "allowed", a parametric approach combined with a model selection criterion that does not restrict the number of parameters to a finite set also leads to a rate of convergence that is slower than $O_p(n^{-\frac{1}{2}})$. Such an approach is, in fact, closely related to the so-called semiparametric broadband methods where the whole spectral density is estimated asymptotically by increasing the number of parameters in a parametric fit. The best known examples are fractional autoregressive modelling with a growing AR-order $p = p_n$ (Bhansali et al. 2006) and broadband estimation based on FEXP-models considered in Moulines and Soulier (1999, 2000), Hurvich (2001), Hurvich and Brodsky (2001), Hurvich et al. (2002) and Narukawa and Matsuda (2011) (also see Beran 1993 and Robinson (1994a) for the definition of FEXP models). The best rate one can achieve this way is $O_p(n^{-\frac{1}{2}}\sqrt{\log n})$ instead of $O_p(n^{-\frac{1}{2}})$. In this sense, the parametric and semiparametric approach are not as far apart as it may seem.

As already indicated, semiparametric estimators can be divided into the so-called *narrowband* (local) and *broadband* (global) methods. The first type focusses on estimation of d, using frequencies in an asymptotically shrinking neighbourhood of the origin only. In other words, one uses the m lowest Fourier frequencies where $m = m_n = o(n)$. The two main estimators are log-periodogram regression (also called narrowband least squares or Geweke and Porter-Hudak estimator, GPH) studied in Sect. 5.6.2 and the local Whittle estimator (also called narrowband Whittle estimator or Gaussian semiparametric estimator), studied in Sect. 5.6.3. The first one is given in an explicit from, whereas the second is defined implicitly as solution of a nonlinear equation. This makes the GPH estimator attractive for applications and, in its simplicity, similar to heuristic methods. However, it is asymptotically less efficient than a local Whittle estimator. The asymptotic theory for both estimators was

originally established in Robinson (1995a, 1995b). A corresponding asymptotic theory based on wavelets instead of the periodogram was suggested in Abry and Veitch (1998) and Veitch and Abry (1999), with mathematical results derived in Bardet et al. (2000), Moulines et al. (2007b, 2007a, 2008). The second class of semiparametric estimators, the so-called broadband methods, are based on all frequencies in $[-\pi, \pi]$ and provide consistent estimates of the entire spectral density. Knowing $f_X(\lambda)$ for all frequencies is important in many situations. For instance, forecasts require no only an estimate of d but also of $\sigma_\varepsilon^2 = \exp((2\pi)^{-1} \int \log f_X(\lambda)\, d\lambda)$ and all autocovariances $\gamma_X(k) = \int \exp(ik\lambda) f_X(\lambda)\, d\lambda$.

Further topics discussed in this chapter are estimation of d for panel data (Beran et al. 2010), identification of periodicities (see, e.g. Beran and Ghosh 2000; Hosking 1981; Gray et al. 1989, 1994; Giraitis et al. 2001), quantile estimation (see, e.g. Dehling and Taqqu 1989b; Ho and Hsing 1996; Wu 2005; Csörgő et al. 2006; Youndjé and Vieu 2006; Csörgő and Kulik 2008a, 2008b; Coeurjolly 2008a, 2008b; Ghosh et al. 1997; Ghosh and Draghicescu 2002a, 2002b; Draghicescu and Ghosh 2003), density estimation (e.g. Wu and Mielniczuk 2002; Cheng and Robinson 1991; Csörgő and Mielniczuk 1995a; Honda 2000; Kulik 2008b, 2008a), tail index estimation for heavy tailed linear processes with long memory (Beran et al. 2012) and goodness-of-fit tests (see, e.g. Beran and Ghosh 1990, 1991; Ho 2002; Kulik 2008b, 2009; Beran 1992; Deo and Chen 2000; Faÿ and Philippe 2002; Dette and Sen 2010).

5.2 Location Estimation

Suppose we observe

$$Y_t = \mu + X_t \tag{5.1}$$

$(t = 1, 2, \ldots, n)$ where X_t $(t \in \mathbb{N})$ is a stationary process with $E(X_t) = 0$. Our goal is statistical inference for the location parameter μ.

5.2.1 Tests and Confidence Intervals Based on the Sample Mean

The simplest, but not necessarily most efficient, estimator is the sample mean $\hat{\mu} = \bar{y} = n^{-1} \sum_{t=1}^{n} Y_t$. To obtain valid tests and confidence intervals for μ, we need to know how to standardize $\bar{y}$ and what the asymptotic distribution of the standardized sample mean is. This question has been answered already in Sects. 4.2 and 4.3. Assuming that $\mathrm{var}(X_t) < \infty$ and the spectral density of X_t (and thus of Y_t) is of the form $f_X(\lambda) \sim c_f |\lambda|^{-2d}$ as $\lambda \to 0$ (for some $d \in (-1/2, 1/2)$), Corollary 1.2 implies

$$\mathrm{var}(\bar{y}) \sim v(d) f_X\big(n^{-1}\big) n^{-1} \sim v(d) c_f n^{2d-1},$$

where $\nu(0) = \lim_{d\to 0} \nu(d) = 2\pi$ and

$$\nu(d) = \frac{2\sin \pi d}{d(2d+1)} \Gamma(1-2d) \quad (d \neq 0).$$

A natural statistic for obtaining asymptotically correct tests and confidence intervals for μ is therefore

$$T_n = n^{\frac{1}{2}-d} \frac{\bar{y}-\mu}{\sqrt{\nu(d)c_f}} \approx \frac{\sqrt{n}(\bar{y}-\mu)}{\sqrt{\nu(d)\, f_X(n^{-1})}}. \tag{5.2}$$

For example, if X_t is a linear process with short-range dependence ($d = 0$), antipersistence ($d < 0$) or long-range dependence ($d > 0$), then T_n converges in distribution to a standard normal random variable (see in particular Theorem 4.5 for short memory and Theorem 4.6 for long memory). Therefore, tests and confidence intervals for μ can be based on the approximation

$$P(T_n \leq z) \approx \Phi(z) \tag{5.3}$$

with Φ denoting the cumulative standard normal distribution. For instance, critical regions for testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ at a level of significance α are given by $K_\alpha = \{T_n > z_{1-\alpha}\}$ with $z_{1-\alpha}$ denoting the $(1-\alpha)$-quantile of the standard normal distribution. Similarly, a $(1-\alpha)$-confidence interval for μ is given by

$$\bar{y} \pm z_{1-\alpha/2} \sqrt{\nu(d)\, f_X\big(n^{-1}\big)} n^{-\frac{1}{2}}$$

for any $d \in (-\frac{1}{2}, \frac{1}{2})$. For short memory with $d = 0$, this can also be approximated by

$$\bar{y} \pm z_{1-\alpha/2} \sqrt{2\pi f_X(0)} n^{-\frac{1}{2}} \tag{5.4}$$

whereas for $d \neq 0$ (antipersistence or long memory) we can write this interval approximately as

$$\bar{y} \pm z_{1-\alpha/2} \sqrt{\nu(d) c_f} n^{d-\frac{1}{2}}. \tag{5.5}$$

For short-memory processes, the asymptotic confidence interval (5.4) is also valid if instead of a linear process X_t the errors in (5.1) are subordinated to a linear process X_t, i.e. if $Y_t = \mu + G(X_t)$ where G is a suitable function. This is the case, even if X_t ($t \in \mathbb{Z}$) itself has infinite second moments, as long as $\operatorname{var}(G(X_t)) < \infty$. To be more exact, usually some mild additional assumptions on the coefficients of the linear process or some mixing properties are required.

In the case of subordination to a linear process with long-range dependence, the situation is much more complicated. First of all, only the subordinated $Y_t = \mu + G(X_t)$ is observed, whereas X_t is an unobservable latent process. Moreover, the transformation G and its Appell (Hermite, power) rank is not known. If the random variables X_t have finite second moments and a spectral density $f_X(\lambda) \sim c_f \lambda^{-2d}$

with $d \in (0, \frac{1}{2})$, and G has a unique Appell polynomial (Hermite, power) expansion, then the asymptotic distribution of T_n is normal only if either the Appell (Hermite, power) rank m is 1 or if $m > (1 - 2d)^{-1}$ (see Theorem 4.8 or Corollary 4.3). These conditions pose considerable difficulties. Although it is possible to estimate the long-memory parameter $d = d_G$ of the observed process $G(X_t)$ and a correct standardization of $\bar{y}$ consistently (for instance, by applying a semiparametric method to the observed periodogram), one cannot say which distribution to use for tests and confidence intervals. In fact, even if d_G were known exactly, we would need to know the Appell (Hermite, power) rank m. The reason is that the same value of $d_G > 0$ can be obtained by transformations with different Appell (Hermite, power) ranks, and these imply different asymptotic distributions. For instance, suppose that X_t is a Gaussian process with long-memory parameter $d_X = \frac{2}{5}$ and $G(x) = x^2 - 1 = H_2(x)$. Then the long-memory parameter of $G(X_t) = X_t^2 - 1$ is equal to $d_G = 2d_X - \frac{1}{2} = \frac{3}{10}$ and the asymptotic distribution of the standardized sample mean is given by the distribution of an Hermite–Rosenblatt process at time 1, which is completely different than the standard normal distribution. On the other hand, if X_t were a Gaussian process with $d_X = \frac{3}{10}$ and $G(x) = x = H_1(x)$, then $G(X_t) = X_t$ would also have $d_G = \frac{3}{10}$, but the standardized sample mean would be normal. To date no statistical method is known for identifying m from an observed subordinated process. A possible way out is a bootstrap procedure that is also valid for subordinated processes. One such method is the so-called sampling window bootstrap discussed in Sect. 10.5 (Hall et al. 1998). Note that for $m = 1$, the inequality $m < (1 - 2d)^{-1}$ holds for all $d \in (0, \frac{1}{2})$, whereas this is no longer the case for $m \geq 2$. If $m < (1 - 2d)^{-1}$ and $m \geq 2$ then the standardized statistic T_n converges to a non-normal random variable (see, e.g. Corollary 4.3). An even more disturbing problem is that, although the variance of Y_t is finite, the variance of the underlying linear process X_t need not be. Theorem 4.17 then implies that asymptotically T_n has a stable distribution. Thus, using normal quantiles as in (5.5) would lead to completely wrong confidence intervals and tests. Unfortunately, the observed data Y_t give very little indication of this problem, since their variance is finite. In summary, in the long-memory case, inference about μ is very uncertain unless we are willing to accept more specific assumptions.

To avoid the above mentioned problems, one of the specific assumptions used usually in practice is that $Y_t = \mu + G(X_t)$, where X_t is a linear process with long memory and finite variance, and the Appell (Hermite, power) rank m of G is equal to one. In that case, the confidence intervals and tests for μ as given in (5.5) are asymptotically correct. In practice, the nuisance parameters d and c_f are unknown so that T_n is replaced by

$$T_n^* = n^{\frac{1}{2}-\hat{d}} \frac{\bar{y} - \mu}{\sqrt{\nu(\hat{d})\hat{c}_f}} \tag{5.6}$$

where $\hat{d}$ and $\hat{c}_f$ are consistent estimates. Estimation of these parameters is discussed later in this chapter. As we will see, consistent estimation of d and c_f is possible

without detailed knowledge of the process Y_t. Approximate confidence intervals are then of the form

$$\bar{y} \pm z_{1-\alpha/2}\sqrt{v(\hat{d})\hat{c}_f}\, n^{\hat{d}-\frac{1}{2}}. \tag{5.7}$$

For small sample sizes, the normal approximation may not be very accurate, since it does not take into account that the parameters c_f and d are estimated. This is comparable to the situation of i.i.d. data where $T_n = \sqrt{n}(\bar{y} - \mu)/\sigma$ is replaced by the t-statistic $\sqrt{n}(\bar{y} - \mu)/s$ with

$$s = \sqrt{(n-1)^{-1}\sum (Y_i - \bar{y})^2}.$$

However, here the additional variability induced by estimating d is expected to be even more noticeable because it affects the rate of convergence. Focussing on $\hat{d}$, a better approximation can be obtained as follows. Suppose that $\hat{d}$ is asymptotically normal such that, for some $\kappa > 0$,

$$\hat{d} = d + \sigma_d n^{-\kappa}\zeta_2 + o_p\big(n^{-\kappa}\big)$$

where $\zeta_2 \sim N(0,1)$ and σ_d^2 is the asymptotic variance of $\hat{d}$. Ignoring uncertainty about $v(d)$ and c_f, and taking into account the symmetry of the standard normal distribution, we then have (in probability)

$$\begin{aligned} T_n^* &= n^{\frac{1}{2}-\hat{d}}\frac{\bar{y}-\mu}{\sqrt{v(\hat{d})\hat{c}_f}} \\ &= n^{\frac{1}{2}-d}\frac{\bar{y}-\mu}{\sqrt{v(d)c_f}} \cdot \exp\left(\frac{\sigma_d \log n}{n^\kappa}\zeta_2\right) + o_p\big(n^{\frac{1}{2}-d}\big) \\ &= \zeta_1 \exp\left(\frac{\sigma_d \log n}{n^\kappa}\zeta_2\right) + o_p\big(n^{\frac{1}{2}-d}\big) \end{aligned}$$

where ζ_1 and ζ_2 are both standard normal variables. If it can be assumed that ζ_1 and ζ_2 are independent (see, e.g. Nourdin and Rosinski 2012 for results pointing in this direction), then the distribution of T_n^* can be approximated by

$$P\big(T_n^* \le x\big) \approx \int_{-\infty}^{\infty} \Phi\left(x \cdot \exp\left(-\frac{\sigma_d \log n}{n^\kappa}u\right)\right)\phi(u)\,du \tag{5.8}$$

with ϕ denoting the $N(0,1)$-density function. Note that independence of ζ_1 and ζ_2 has been conjectured in Beran (1989). However, no formal proof exists in the literature. Recent results that may be useful for a proof can be found in Nourdin and Rosinski (2012).

Example 5.1 Let X_t ($t \in \mathbb{Z}$) be a FARIMA$(0, d, 0)$ process with innovation variance σ_ε^2 and fractional differencing parameter d. It is shown in Sect. 5.5 that for the

Table 5.1 Comparison of standard normal quantiles z_α (for $\alpha = 0.9, 0.95, 0.975, 0.99, 0.995$) with quantiles q_α^* obtained using approximation (5.9)

n	$\alpha = 0.90$	0.95	0.975	0.99	0.995
Standard normal quantiles z_α					
	1.28	1.65	1.96	2.33	2.58
Corrected quantiles q_α^*					
50	1.38	1.92	2.48	3.25	3.87
100	1.35	1.84	2.33	2.98	3.49
200	1.32	1.77	2.20	2.76	3.18
400	1.31	1.73	2.12	2.61	2.97
1000	1.30	1.69	2.05	2.48	2.78
Ratio of the two quantiles, q_α^*/z_α					
50	1.08	1.17	1.27	1.40	1.50
100	1.05	1.12	1.19	1.28	1.35
200	1.03	1.08	1.12	1.19	1.24
400	1.02	1.05	1.08	1.12	1.15
1000	1.01	1.03	1.05	1.06	1.08

maximum likelihood estimator of d we have $\kappa = \frac{1}{2}$ and $\sigma_d^2 = 6/\pi^2$. Therefore,

$$P\left(T_n^* \le x\right) \approx \int_{-\infty}^{\infty} \Phi\left(x \cdot \exp\left(-\frac{\sqrt{6}}{\pi}\frac{\log n}{\sqrt{n}}u\right)\right)\phi(u)\,du$$
$$= E\left[\Phi\left(x \cdot \exp\left(-\frac{\sqrt{6}}{\pi}\frac{\log n}{\sqrt{n}}\zeta\right)\right)\right] \tag{5.9}$$

where $\zeta \sim N(0,1)$. This expression is easy to calculate, for instance, by Monte Carlo simulation of the expected value. A comparison of quantiles obtained from (5.9) and standard normal quantiles is given in Table 5.1.

5.2.2 *Efficiency of the Sample Mean*

So far we have considered the sample mean only. However, since we have dependent observations, one may ask the question in how far it may be possible to obtain more efficient linear estimators of location by taking into account the dependence structure. Thus, we need to compare the variance of $\bar{y}$ with the variance of the best linear unbiased estimator (BLUE) $\hat{\mu}_{\mathrm{BLUE}} = \sum_{t=1}^{n} w_t Y_t$. The weights w_t are chosen such that $\mathrm{var}(\sum w_t Y_t)$ is minimal under the side condition $\sum w_t = 1$ (which is needed for unbiasedness). Using the notation $w(n) = (w_1, \ldots, w_n)^T$, $Y(n) = (Y_1, \ldots, Y_n)^T$, $X(n) = (X_1, \ldots, X_n)^T$ (where $X_t = Y_t - \mu$) and Σ_n for the covariance matrix of

$Y(n)$, this leads to the formulas

$$w = \frac{\Sigma_n^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma_n^{-1}\mathbf{1}},$$

$$\hat{\mu}_{\text{BLUE}} = \frac{\mathbf{1}^T\Sigma_n^{-1}X(n)}{\mathbf{1}^T\Sigma_n^{-1}\mathbf{1}}$$

and

$$\text{var}(\hat{\mu}_{\text{BLUE}}) = \left(\mathbf{1}^T\Sigma_n^{-1}\mathbf{1}\right)^{-1}.$$

On the other hand, for the sample mean we have

$$\text{var}(\bar{y}) = n^{-1}\sum_{k=-(n-1)}^{n-1}\left(1-\frac{|k|}{n}\right)\gamma_X(k).$$

Suppose now that X_t is either a linear or a subordinated process such that, after appropriate standardization, the sample mean and the BLUE are asymptotically normal. Then the asymptotic efficiency is defined as the ratio of the corresponding asymptotic variances. Since we are in a regular case where the asymptotic variance is the same as the limit of the (standardized) variance, we may calculate the asymptotic efficiency of $\bar{y}$ also by taking the limit of the finite sample efficiency. Thus,

$$\begin{aligned}\text{as.eff}(\bar{y},\hat{\mu}_{\text{BLUE}}) &= \lim_{n\to\infty}\frac{(n^{-1}\mathbf{1}^T\Sigma_n^{-1}\mathbf{1})^{-1}}{\sum_{k=-(n-1)}^{n-1}(1-\frac{|k|}{n})\gamma_X(k)}\\ &= \lim_{n\to\infty}\frac{(n^{-1}\mathbf{1}^T\Sigma_n^{-1}\mathbf{1})^{-1}}{v(d)\,f(n^{-1})} = \lim_{n\to\infty}\frac{(n^{-1}\mathbf{1}^T\Sigma_n^{-1}\mathbf{1})^{-1}}{v(d)c_f n^{2d}}. \qquad (5.10)\end{aligned}$$

In the case of short memory ($d = 0$), this can also be written as

$$\text{as.eff}(\bar{y},\hat{\mu}_{\text{BLUE}}) = \frac{\lim_{n\to\infty}(n^{-1}\mathbf{1}^T\Sigma_n^{-1}\mathbf{1})^{-1}}{2\pi f(0)}.$$

Thus, one needs to investigate the limit of $n^{-1}\mathbf{1}^T\Sigma_n^{-1}\mathbf{1}$. The only difficulty is to characterize the elements of the inverse covariance matrix. In the case of short memory, the key is an approximation of Σ_n^{-1} by a circulant $n\times n$ matrix S. Circulant matrices are defined as follows (see, e.g. Brockwell and Davis 1991):

Definition 5.1 Let

$$S = \left[s(i-j)\right]_{i,j=1,\ldots,n}$$

be a Toeplitz matrix with $s:\mathbb{Z}\to\mathbb{C}$ a periodic function with period n. Then S is called a circulant matrix.

For instance, if $n = 2m+1$ (i.e. n is odd), then we may choose the approximation of Σ_n^{-1} in form of the circulant matrix $S = [s(i-j)]$ with

$$s(k) = \gamma_X(k) \quad \big(0 \le |k| \le m\big)$$

and

$$s(k) = \gamma_X(n-k) \quad \big(m+1 \le |k| \le n-1\big).$$

The advantage of a circulant matrix is that one has explicit expressions for the eigenvalues and eigenvectors. For S one obtains eigenvalues $\alpha_0, \alpha_1, \ldots, \alpha_{2m}$ given by

$$\alpha_o = \sum_{k=-m}^{m} \gamma_X(k),$$

$$\alpha_j = \sum_{k=-m}^{m} \gamma_X(k) \cos k\lambda_j \quad (j = 1, 2, \ldots, m)$$

and

$$\alpha_{n-j} = \alpha_j,$$

with $\lambda_j = 2\pi j/n$ denoting Fourier frequencies. The corresponding orthonormal eigenvectors are given by

$$v_o = n^{-\frac{1}{2}}(1, 1, \ldots, 1)^T = n^{-\frac{1}{2}}\mathbf{1},$$

and, for $j = 1, 2, \ldots, m$,

$$v_j = \sqrt{\frac{2}{n}}\big\{1, \sin\lambda_j, \sin 2\lambda_j, \ldots, \sin\big[(n-1)\lambda_j\big]\big\}^T$$

and

$$v_{n-j} = \sqrt{\frac{2}{n}}\big\{1, \cos\lambda_j, \cos 2\lambda_j, \ldots, \cos\big[(n-1)\lambda_j\big]\big\}^T.$$

This leads to a representation of S and S^{-1} as

$$S = P\Lambda P^T,$$

$$S^{-1} = P\Lambda^{-1}P^T$$

where

$$P = (v_o, v_1, \ldots, v_{n-1})$$

and

$$\Lambda = \begin{pmatrix} \alpha_o & 0 & \cdots & 0 \\ 0 & \alpha_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha_{n-1} \end{pmatrix}.$$

From the very special form of the eigenvalues and eigenvectors, one can then see that $\alpha_j \approx 2\pi f_X(\lambda_j)$ for n large enough and

$$P^T S P \approx D,$$
$$P^T S^{-1} P \approx D^{-1}$$

where

$$D = \begin{pmatrix} 2\pi f(0) & 0 & \cdots & 0 \\ 0 & 2\pi f(\lambda_1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2\pi f(\lambda_{n-1}) \end{pmatrix}.$$

Note that in an exact argument one has to take into account that the dimension of the matrices is increasing. Under suitable assumptions, it can indeed be shown that the approximation of the elements of $P^T S^{-1} P$ is uniform. This then leads to

$$P^T \Sigma_n^{-1} P \approx P^T S^{-1} P \approx D^{-1}$$

and in particular

$$\left[P^T \Sigma_n^{-1} P\right]_{11} = n^{-\frac{1}{2}} \mathbf{1} \Sigma_n^{-1} \mathbf{1} n^{-\frac{1}{2}} = n^{-1} \mathbf{1} \Sigma_n^{-1} \mathbf{1} \approx \big(2\pi f(0)\big)^{-1},$$

which implies

$$\text{as.eff}(\bar{y}, \hat{\mu}_{\text{BLUE}}) = \frac{\lim_{n\to\infty} (n^{-1} \mathbf{1}^T \Sigma_n^{-1} \mathbf{1})^{-1}}{2\pi f(0)} = \frac{2\pi f(0)}{2\pi f(0)} = 1. \tag{5.11}$$

Thus, under some regularity conditions ignored in the heuristic arguments here, we may conclude that for short-memory processes the sample mean is asymptotically efficient. It is therefore not worth going through the complication of calculating $\hat{\mu}_{\text{BLUE}}$ (which would even require estimation of all autocovariances). This is a famous result by Grenander (1954) who derived it in a more general regression context (see Sect. 7.1.2 for further discussion). Assuming f_X to be finite, piecewise continuous and bounded away from zero for all $\lambda \in [-\pi, \pi]$, it is sufficient to derive (5.11).

For $d \neq 0$, the derivation above cannot be applied because for $d > 0$ we have $f_X(0) = \infty$ and for $d < 0$ we would divide by $f_X(0) = 0$. We know, however, that

$$\frac{(n^{-1} \mathbf{1}^T \Sigma_n^{-1} \mathbf{1})^{-1}}{v(d) f(n^{-1})} \sim \frac{(n^{2d-1} \mathbf{1}^T \Sigma_n^{-1} \mathbf{1})^{-1}}{v(d) c_f}$$

where $0 < v(d), c_f < \infty$ are fixed constants. Thus, the quantity one has to deal with is

$$n^{2d-1}\mathbf{1}^T \Sigma_n^{-1}\mathbf{1} = n^{2d-1} \sum_{j,l=1}^{n} c(j-l) \tag{5.12}$$

where $\Sigma_n^{-1} = [c(j-l)]_{j,l=1,2,\ldots,n}$. This matrix is more delicate to deal with than in the short-memory case. A heuristic argument can be given as follows. The asymptotic behaviour of (5.12) is determined by the behaviour $f_X(\lambda) \sim c_f|\lambda|^{-2d}$ at the origin only. Moreover, since we are comparing with the variance of the sample mean, the relative efficiency does not depend on c_f. Thus, the formula we are looking for depends on nothing else than d. In other words, the same formula applies to all processes that have the same value of d. It is therefore sufficient to calculate (5.10) for one particular process because this formula then applies generally. Now, for a FARIMA$(0, d, 0)$ process autocovariances are given by an explicit formula (see Sect. 2.1.1.4). One can then verify by direct calculation that the optimal weights are of the form

$$w_t = w_{t,n} = \binom{n-1}{t-1} \frac{B(t-d, n-t+1-d)}{B(1-d, 1-d)}$$

where $B(\cdot,\cdot)$ is the Beta function. After some calculation one then obtains an explicit formula for the variance of the BLUE and hence a formula for the asymptotic efficiency of the sample mean. The formula, first derived by Adenstedt (1974), is given by

$$\text{as.eff}(\bar{y}, \hat{\mu}_{\text{BLUE}}) = \frac{(2d+1)\Gamma(d+1)\Gamma(2-2d)}{\Gamma(1-d)}$$

for any $d \in (-\frac{1}{2}, \frac{1}{2})$ (also see Samarov and Taqqu 1988; Beran and Künsch 1985; Dahlhaus 1995). For $d = 0$ we obtain the value of 1 as seen before. For all other values of d, the asymptotic efficiency is below one. However, there is a distinct difference between long memory and antipersistence. For $d > 0$, the efficiency loss does not exceed 2 %. This means that under long-range dependence there is no need to abandon the sample mean. In contrast, for $d < 0$, the asymptotic efficiency of $\bar{y}$ can be arbitrarily close to zero if d is close enough to the left border of $-\frac{1}{2}$. This is shown in Fig. 5.1. Thus, one may conclude that for an antipersistent series it may be worth the effort to use the BLUE or a similar improved estimator.

5.2.3 M-Estimation

The sample mean is a special example of an M-estimator of μ defined as the solution of

$$\sum_{t=1}^{n} \psi(Y_t - \hat{\mu}) = 0 \tag{5.13}$$

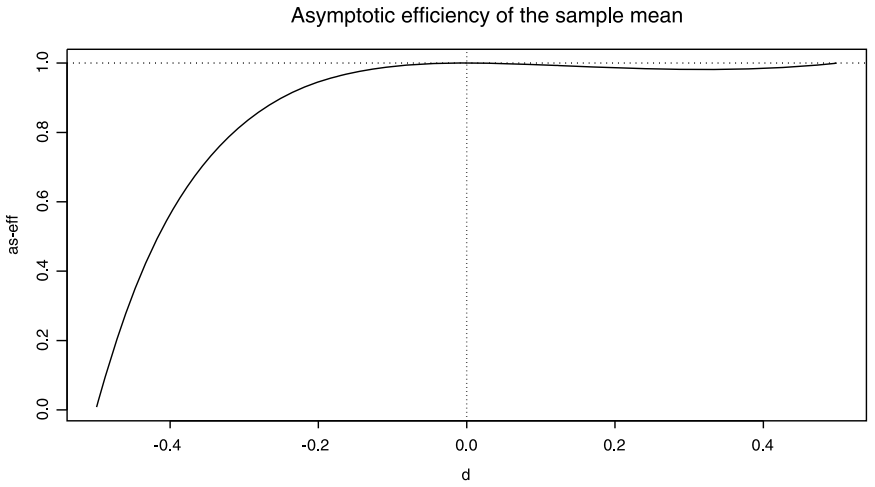

Fig. 5.1 Asymptotic efficiency of the sample mean as a function of d

where ψ is a deterministic function such that $E[\psi(Y-\mu^*)]=0$ if and only if $\mu^*=\mu$. For the sample mean, we have $\psi(x)=x$. Other examples are the median with $\psi(x)=\operatorname{sign}(x)$ and the Huber estimator with

$$\psi(x)=x\cdot 1\{|x|\le c\}+c\cdot\operatorname{sign}(x)\cdot 1\{|x|>c\},$$

where $c>0$ is a suitably chosen tuning parameter. M-estimators have become popular in the context of robust statistics because their robustness is directly related to the ψ-function (Huber 1981; Hampel et al. 1986). Specifically, writing μ as a functional $\mu=T(F)=\int x\,dF(x)$ of the underlying (marginal) distribution F of Y, the influence function defined for an infinitesimal contamination by a point mass at x is defined by

$$IF(x)=\lim_{\varepsilon\to 0}\frac{T((1-\varepsilon)F+\varepsilon\delta_x)-T(F)}{\varepsilon},$$

where δ_x is the Dirac measure at point x. For M-estimators the influence function is proportional to $\psi(x)$. The intuitive interpretation of $IF(x)$ is that it characterizes in how far an infinitesimal contamination of the distribution F by a point mass δ_x influences the estimator. Heuristically, one may write for ε small enough, $T((1-\varepsilon)F+\varepsilon\delta_x)\approx T(F)+\varepsilon IF(x)$. Therefore, $\hat{\mu}$ is said to be robust, if ψ is bounded on $\mathbb{R}$. This is, for instance, the case for the median and the Huber estimator with $c<\infty$, however, not for the sample mean. The reason is that an outlier can change the sample mean to an arbitrary value, whereas this is not the case, if $\psi(x)$ is bounded. Thus, if one wants to guard against outliers, then bounded ψ-functions are useful. However, robustness usually comes at a price, in the sense that robust estimators tend to be less efficient under the ideal uncontaminated model F. For instance, suppose for a moment that X_t ($t\in\mathbb{Z}$) are i.i.d. with density function $p_X(x)$, so that $Y_t=\mu+X_t$ have the common density function $p_X(x-\mu)$. Then, under

some mild conditions on ψ, a Taylor expansion of (5.13) leads to the central limit theorem

$$\sqrt{n}(\hat{\mu} - \mu) \underset{d}{\to} N\left(0, \sigma_\psi^2\right)$$

where

$$\sigma_\psi^2 = \frac{E[\psi^2(Y - \mu)]}{E^2[\psi'(Y - \mu)]}.$$

(Note that this formula is applicable only if $E[\psi'(Y - \mu)]$ is not zero.) The smallest value of σ_ψ^2 is achieved for $\psi(x)$ proportional to the score function

$$s(x, \mu) = \frac{\partial}{\partial \mu} \log p_X(x - \mu).$$

For instance, for the normal distribution $s(x, \mu)$ is proportional to $\psi(x) = x$ and the corresponding asymptotic variance is $\sigma_\psi^2 = \sigma^2 = \text{var}(X)$. For all other ψ-functions, σ_ψ^2 is larger than σ^2. Thus, for i.i.d. observations all robust M-estimators lose efficiency compared to the sample mean. Analogous results also hold, if the assumption of independence is replaced by short memory.

It therefore came as a surprise, when it was discovered that for Gaussian long-memory processes robust M-estimators no longer lose efficiency asymptotically. This was first stated in Beran (1991) in the context of Gaussian subordination and generalized later by Giraitis and Surgailis (1999) and Koul and Surgailis (2001) to linear processes with finite and infinite variance, respectively (also see Wu 2003). Proofs essentially follow from:

- Limit theorems for empirical processes;
- Asymptotic behaviour of the sample mean.

Below, we formulate the result for linear processes with a finite variance (Beran 1991; Giraitis and Surgailis 1999). Extensions, e.g. to stochastic volatility models, require proving an appropriate functional central limit theorem for the corresponding empirical processes or related results (see, e.g. Beran 2006, 2007a; Beran and Schützner 2008, also see Beran and Feng 2007).

Recall that ψ is a function of bounded variation if

$$\sup \sum \left|\psi(x_i) - \psi(x_{i-1})\right| < \infty,$$

where the supremum is taken over all possible partitions of $\mathbb{R}$.

Theorem 5.1 *Let $Y_t = \mu + X_t$, where $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ is a linear process with $E(\varepsilon_i) = 0$, $\sigma_\varepsilon^2 = \text{var}(\varepsilon_i) < \infty$ and $a_j \sim c_a j^{d-1}$ as $j \to \infty$ for some $0 < c_a < \infty$, $0 < d < \frac{1}{2}$. Assume that ψ is a function of bounded variation and such that*

$$E\left[\psi\left(Y - \mu^*\right)\right] = 0$$

if and only if $\mu = \mu^*$. *Moreover, assume that the technical conditions of Theorem* 4.33 *hold and*

$$\int p_X(x)\, d\psi(x) \neq 0. \tag{5.14}$$

Then

$$n^{\frac{1}{2}-d}(\hat{\mu} - \mu) \underset{d}{\to} N\big(0, \sigma_M^2\big),$$

where

$$\sigma_M^2 = \frac{\sigma_\varepsilon^2 c_a^2}{d(2d+1)} \int_0^\infty u^{d-1}(u+1)^{d-1}\, du.$$

Moreover,

$$n^{\frac{1}{2}-d}(\hat{\mu} - \bar{y}) \underset{d}{\to} 0.$$

Proof As mentioned above, the proof follows from the limit theorem for empirical processes (Theorem 4.33), together with the asymptotic behaviour of the sample mean; see Theorem 4.6. Let $F_Y(y)$ and $p_Y(y) = p_X(y - \mu)$ be the marginal distribution and the marginal density function of Y_t. Recall that $F_{n,Y}$ is the empirical distribution function associated with $Y_1, \ldots, Y_n$. By the definition of $\hat{\mu}$, we have

$$\begin{aligned} 0 &= \int \psi(y - \hat{\mu})\, dF_{n,Y}(y) \\ &= \underbrace{\int \psi(y - \hat{\mu})\, d\big[F_{n,Y}(y) - F_Y(y)\big]}_{I_{n,1}} + \underbrace{\int \psi(y - \hat{\mu})\, dF_Y(y)}_{I_{n,2}}. \end{aligned}$$

Recall the reduction principle (4.159). Formally,

$$F_{n,X}(x - \mu) - F_X(x - \mu) + p_X(x - \mu)\bar{x} = o_p\big(n^{d-\frac{1}{2}}\big),$$

uniformly in x. This can be restated as

$$F_{n,Y}(x) - F_Y(x) + p_Y(x)\bar{x} = o_p\big(n^{d-\frac{1}{2}}\big).$$

Thus, using a change of variables and partial integration (which is possible due to bounded variation of ψ), the first term $I_{n,1}$ can be approximated as follows:

$$\begin{aligned} I_{n,1} &= -\int \big[F_{n,Y}(y + \hat{\mu}) - F_Y(y + \hat{\mu})\big]\, d\psi(y) \\ &= \bar{x} \int p_Y(y + \hat{\mu})\, d\psi(y) + o_p\big(n^{d-\frac{1}{2}}\big) \end{aligned}$$

$$= \bar{x} \int p_Y(y+\mu)\, d\psi(y) + (\mu - \hat{\mu}) \cdot \bar{x}_n \int p'_Y(y+\mu)\, d\psi(y) + o_p\big(n^{d-\frac{1}{2}}\big)$$
$$= \bar{x} \int p_X(y)\, d\psi(y) + o_p\big(n^{d-\frac{1}{2}}\big),$$

where the reduction principle was used in the second equality. The last approximation follows from $\hat{\mu} - \mu = o_p(1)$ and $\bar{x}_n = O_p(n^{d-\frac{1}{2}})$.

For the second term $I_{n,2}$, recall that

$$E\big[\psi(Y-\mu)\big] = \int \psi(y-\mu) p_Y(y)\, dy = 0,$$

so that this term can be added to $I_{n,2}$. By a change of variables and Taylor expansion of p_Y,

$$\begin{aligned} I_{n,2} &= \int \big[\psi(y-\hat{\mu}) - \psi(y-\mu)\big] p_Y(y)\, dy \\ &= \int \psi(y) \big[p_Y(y+\hat{\mu}) - p_Y(y+\mu)\big]\, dy \\ &= (\hat{\mu}-\mu) \int \psi(y) p'_Y(y+\mu)\, dy + o_p(\hat{\mu}-\mu) \\ &= (\hat{\mu}-\mu) \int \psi(x) p'_X(x)\, dx + o_p(\hat{\mu}-\mu) \\ &= -(\hat{\mu}-\mu) \int p_X(x)\, d\psi(x) + o_p(\hat{\mu}-\mu). \end{aligned}$$

Thus, bearing in mind (5.14), overall we obtain

$$\hat{\mu} - \mu = \bar{x} + o_p(\hat{\mu}-\mu) + o_p\big(n^{d-\frac{1}{2}}\big).$$

The asymptotic distribution of $\hat{\mu}$ then follows from the asymptotic behaviour of $\bar{x}$ as described in Theorem 4.6. □

Theorem 5.1 is very general. For example, it does not require the existence of a complete Appell polynomial expansion. On the other hand, as stated here, it is not directly applicable to the median (and other ψ-functions with similar properties) because for $\psi(x) = \operatorname{sign}(x)$ the integral $\int p_X(x)\, d\psi(x)$ vanishes. However, by a slight modification of the proof, condition (5.14) can be replaced by

$$\int \psi(x) p'_X(x)\, dx \neq 0. \tag{5.15}$$

Heuristically the result in Theorem 5.1 then follows from the approximation (see Corollary 4.3)

$$n^{-1}\sum_{t=1}^{n}\psi(X_t) \approx J_1\bar{X}_n,$$

where

$$J_k = J_k(\psi) = (-1)^k \int \psi(x)p_X^{(k)}(x)\,dx,$$

and in particular

$$J_1 = -\int \psi(x)p_X'(x)\,dx.$$

In other words, we use the Appell polynomial expansion

$$\psi(X) = \sum_{j=1}^{\infty} \frac{a_{j,\mathrm{app}}}{j!} A_j(X)$$

with

$$a_j = (-1)^j \int \psi(x)p_X^{(j)}(x)\,dx,$$

assuming that a unique Appell polynomial expansion exists. Recall that this also means that we need to impose additional conditions on the marginal density functions. Sufficient conditions include, for instance, $p_X \in C^\infty(\mathbb{R})$ and $\int (p_X^{(j)}/p_X)^2 p_X\,dx < \infty$ (see Giraitis 1985). By definition (and arguments as before), we have

$$\begin{aligned} 0 &\approx n^{-1}\sum_{t=1}^{n}\psi(X_t) + (\hat{\mu}-\mu)\int \psi(x)p_X'(x)\,dx \\ &\approx J_1\bar{x} - J_1(\hat{\mu}-\mu). \end{aligned}$$

Thus, the result of Theorem 5.1 follows, if $J_1 \neq 0$. Note also that for a Gaussian process X_t with $\mathrm{var}(X_t) = 1$ we have $p_X'(x) = -xp_X(x)$ so that condition (5.15) is the same as saying that the Hermite rank of ψ is one.

Example 5.2 Suppose that μ is the median of Y_t. Note that, since $E(X_t)$ was assumed to be zero, this means that the median is equal to the expected value. Then we have for $\psi(x) = \mathrm{sign}(x)$,

$$\begin{aligned} J_1 &= -\int \psi(x)p_X'(x)\,dx \\ &= -\left[-\int_{-\infty}^{0} p_X'(x)\,dx + \int_0^{\infty} p_X'(x)\,dx\right] = 2p_X(0). \end{aligned}$$

Thus, the theorem is applicable, if (apart from the other regularity conditions) we have $p_X(0) \neq 0$. It is interesting to note that this condition is also important in the case of i.i.d. data because there the asymptotic distribution of the sample median is then given by

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow[d]{} N\left(0, \frac{1}{4p_Y^2(\mu)}\right) = N\left(0, \frac{1}{4p_X^2(0)}\right) = N\big(0, J_1^{-2}\big).$$

However, in the case of long memory, the condition $p_X(0) \neq 0$ is only needed to make sure that the asymptotic distribution is valid. The actual value of $p_X(0)$ does not appear in the limiting distribution. This is very much in contrast to the i.i.d. and short memory case.

It should be emphasized that the assumption in the example was that $X_t = Y_t - \mu$ where μ is the median. In other words, the median of Y_t is equal to the expected value. This is, of course, not the case in general. If $\mu = E(Y_t)$ is not identical with the median (say μ_{med}), and our aim is to estimate μ_{med}, then we still have $\psi(u) = \text{sign}(u)$, but $\mu_{\text{med}} = \mu + \Delta_{\text{med}}$ with $\Delta_{\text{med}} \neq 0$. The asymptotic distribution of $\hat{\mu}_{\text{med}}$ is therefore determined by

$$\sum \psi(Y_t - \mu_{\text{med}}) = \sum \psi(X_t - \Delta_{\text{med}}).$$

The Appell polynomial expansion is then

$$\psi(X - \Delta_{\text{med}}) = \sum_{j=1}^{\infty} \frac{a_{j,\text{app}}}{j!} A_j(X)$$

with

$$\begin{aligned} a_{j,\text{app}} &= (-1)^j \int \psi(x - \Delta_{\text{med}}) p_X^{(j)}(x)\, dx \\ &= (-1)^j \int \psi(x) p_X^{(j)}(x + \Delta_{\text{med}})\, dx. \end{aligned}$$

In particular,

$$\begin{aligned} a_{1,\text{app}} = J_1 &= -\int \psi(x) p_X'(x + \Delta_{\text{med}})\, dx \\ &= -\int_{-\infty}^{0} p_X'(x + \Delta_{\text{med}})\, dx + \int_0^{\infty} p_X'(x + \Delta_{\text{med}})\, dx \\ &= -2p_X'(\Delta_{\text{med}}). \end{aligned}$$

The result of Theorem 5.1 is in sharp contrast to the i.i.d. and weakly dependent case where robustness comes at the cost of losing efficiency. It is also worth noting that an analogous asymptotic limit theorem for M-estimators can be obtained

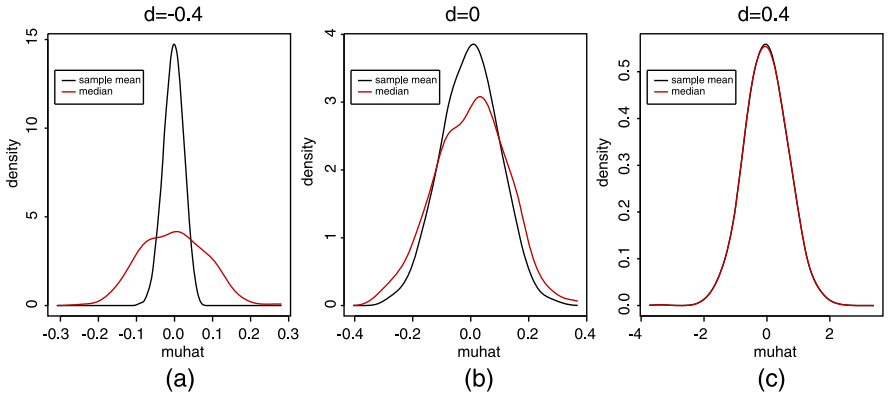

Fig. 5.2 Simulated distributions of the sample mean and the sample median for a FARIMA$(0, d, 0)$ series of length $n = 1000$, with (**a**) $d = -0.4$, (**b**) $d = 0$ and (**c**) $d = 0.4$, respectively

for subordinated processes $Y_t = \mu + G(X_t)$. However, in general, the asymptotic equivalence between the sample mean and other M-estimators is lost.

On the other hand, if X_t $(t \in \mathbb{N})$ is antipersistent, then the opposite happens. Antipersistence implies that $n^{\frac{1}{2}-d}(\bar{y} - \mu)$ is asymptotically normally distributed. Since d is negative, the rate $n^{d-\frac{1}{2}}$ converges faster to zero than $n^{-\frac{1}{2}}$. However, for an M-estimator with a *nonlinear* ψ-function antipersistence is lost so that the rate becomes $n^{-\frac{1}{2}}$ again. Thus, there is an infinite efficiency loss due to robust estimation! This is illustrated numerically in the following example.

Example 5.3 Let $Y_t = \mu + X_t$ where X_t is a fractional ARIMA$(0, d, 0)$ process $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = (1 - B)^{-d} \varepsilon_t$ with standard normal innovations ε_t. Recall that the coefficients a_j behave like $a_j \sim c_a j^{d-1} = \frac{1}{\Gamma(d)} j^{d-1}$. Then $\bar{y}$ is exactly normally distributed with expected value μ and the variance of $\bar{y}$ is approximately equal to

$$\begin{aligned} \operatorname{var}(\bar{y}) \sim v(d) c_f n^{2d-1} &= \frac{\sin \pi d}{\pi d(2d+1)} \Gamma(1-2d) n^{2d-1} \\ &= \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d) d(2d+1)} n^{2d-1}. \end{aligned}$$

We compare this to the median $\hat{\mu}$. If $d > 0$, then $\hat{\mu}$ has the same distribution asymptotically. For $d = 0$, $\sqrt{n}(\hat{\mu} - \mu)$ converges to a normal variable with zero mean and variance $E[\psi^2(X)]/E^2[\psi'(X)] = \pi/2 \approx 1.57$. The relative asymptotic efficiency of $\hat{\mu}$ is therefore $2/\pi \approx 0.64$. On the other hand, for $d < 0$, the relative asymptotic efficiency of $\hat{\mu}$ is zero. This is illustrated in Fig. 5.2 using 1000 Monte Carlo repetitions and a sample size of $n = 1000$. For $d = -0.4$, the distribution of the median is much wider, whereas for $d = 0.4$ the two distributions are hardly distinguishable.

A result analogous to Theorem 5.1 holds for linear long-memory processes with an infinite variance (Koul and Surgailis 2001). As in Theorem 4.17, we consider the linear process $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ such that ε_t are i.i.d. in the domain of attraction of a stable law, i.e. as $x \to \infty$,

$$P(\varepsilon_1 > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\varepsilon_1 < -x) \sim A\frac{1-\beta}{2}x^{-\alpha} \tag{5.16}$$

for some $1 < \alpha < 2$. We also assume that $E(\varepsilon_t) = 0$. Moreover, $a_j \sim c_a j^{d-1}$ and $\sum_{j=0}^{\infty} |a_j|^{\alpha} < \infty$ and hence $0 < d < 1 - \alpha^{-1}$. Also, assume that ψ fulfills the conditions of Theorem 5.1. Then, under further regularity conditions on F_ε (such as those in Theorem 4.35; sufficient is, for instance, if ε_t is $S\alpha S$) we have

$$n^{1-1/\alpha-d}(\hat{\mu} - \bar{Y}_n) \underset{d}{\to} 0.$$

Thus, using Theorem 4.15, applied to the sample mean, we conclude that

$$n^{1-1/\alpha-d}(\hat{\mu} - \mu) \underset{d}{\to} A^{1/\alpha} C_\alpha^{-1/\alpha} \frac{B}{d} \tilde{Z}_{H,\alpha}(1),$$

where $H = d + \alpha^{-1}$, and $\tilde{Z}_{H,\alpha}(1)$ is a symmetric α-stable random variable with scale

$$\eta^{\alpha} = \int_{-\infty}^{1} \left\{ \int_0^1 (u-v)_+^{d-1} \, du \right\}^{\alpha} dv = d^{-1} \left(\int_{-\infty}^{1} \left\{ (1-u)^d - (-u)_+^d \right\}^{\alpha} du \right).$$

5.3 Scale Estimation

Suppose now that we observe $Y_t = \mu + \sigma X_t$ where X_t is a linear process with unit variance. The usual estimator of σ^2 is the sample variance

$$s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (Y_t - \bar{y})^2.$$

For i.i.d. observations the asymptotic distribution of $\sqrt{n}(s^2 - \sigma^2)$ is the same as that of $\sqrt{n}(s_0^2 - \sigma^2)$, where $s_0^2 = (n-1)^{-1} \sum_{t=1}^{n} (Y_t - \mu)^2$. Indeed, we have

$$n^{-1} \sum_{t=1}^{n} (Y_t - \mu)^2 - \sigma^2 = \frac{\sigma^2}{n} \sum_{t=1}^{n} \left(X_t^2 - 1 \right)$$

and

$$n^{-1} \sum_{t=1}^{n} (Y_t - \bar{y})^2 = \frac{\sigma^2}{n} \sum_{t=1}^{n} \left(X_t^2 - 1 \right) - \sigma^2 \bar{x}^2.$$

Now, $\sqrt{n}\sigma^2 n^{-1}\sum_{t=1}^{n}(X_t^2-1)$ converges to a normal distribution with variance $\sigma^4 E[(X^2-1)^2]=2\sigma^2$, whereas $\sqrt{n}\bar{x}^2=o_p(1)$. Therefore, $\sqrt{n}(s^2-\sigma^2)$ and $\sqrt{n}(s_0^2-\sigma^2)$ have the same asymptotic distribution. This asymptotic equivalence no longer holds for strongly dependent data. For instance, if X_t ($t\in\mathbb{Z}$) is a stationary Gaussian process with coefficients $a_j\sim c_a j^{d-1}$ in the Wold decomposition and $1/4<d<1/2$, then (see Theorem 4.3)

$$n^{1-2d}\frac{\sigma^2}{n}\sum_{t=1}^{n}\big(X_t^2-1\big)\underset{d}{\to}\sigma^2 v Z_{2,H}(1),$$

where $Z_{2,H}(\cdot)$ is an Hermite–Rosenblatt process, $H=d+1/2$ and

$$\begin{aligned}v^2&=\frac{\sigma_\varepsilon^4 c_a^4}{d(4d-1)}\left(\int_0^\infty u^{d-1}(u+1)^{d-1}\,du\right)^2\\&=\frac{\sigma_\varepsilon^4 c_a^4}{d(4d-1)}B^2(1-2d,d).\end{aligned}$$

On the other hand,

$$\sigma^2 n^{1-2d}\bar{X}_n^2=\sigma^2\big(n^{1/2-d}\bar{X}_n\big)^2\underset{d}{\to}\sigma^2 v^2 Z^2$$

where $Z\sim N(0,1)$. Thus, the limiting distribution of the appropriately normalized s_0^2 is of the Hermite–Rosenblatt type, whereas for s^2 it is defined by a linear combination of an Hermite–Rosenblatt and a χ_1^2 variable (Dehling and Taqqu 1991). We illustrate this difference in the following example.

A further problem with the sample variance is that, under long memory, it tends to underestimate σ^2 for small sample sizes. The intuitive reason is that sample paths tend to stay at a similar level for a prolonged time and hence vary less than they would over a longer time period. A heuristic argument can be given as follows:

$$\begin{aligned}E\big(s^2\big)&=(n-1)^{-1}\sigma^2\sum_{t=1}^{n}E\big[(X_t-\bar{x}_n)^2\big]\\&=(n-1)^{-1}n\big[\sigma^2-\operatorname{var}(\bar{x}_n)\big]\approx\sigma^2\frac{1-\sigma^{-1}v(d)c_f n^{2d-1}}{1-n^{-1}}\\&=\sigma^2\frac{1-c\cdot n^{2d-1}}{1-o(n^{2d-1})}=\sigma^2\big(1-cn^{2d-1}+o\big(n^{2d-1}\big)\big)\end{aligned}$$

with $c>0$. As d approaches $\frac{1}{2}$, the exponent $2d-1$ approaches zero so that the bias becomes increasingly serious even for relatively large sample sizes. This is illustrated by the following examples.

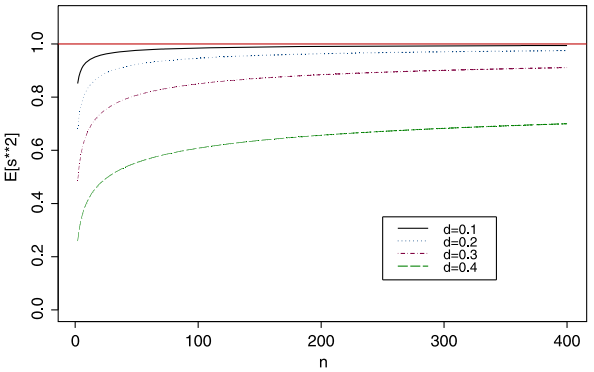

Fig. 5.3 Expected value of s^2 for increments of a self-similar process with variance 1

Example 5.4 For increments $X_t = U_t - U_{t-1}$ of a self-similar process U_t with self-similarity parameter $H = d + \frac{1}{2}$, we have the exact equality $E[s^2] = \sigma^2 \cdot \kappa$ with

$$\kappa = \frac{1 - n^{2d-1}}{1 - n^{-1}}.$$

Figure 5.3 displays κ for n between 2 and 400. One can see that even for $n = 400$ there is a considerable bias with $c = 0.7$, unless long memory is very weak.

Example 5.5 Figure 5.4(a) displays boxplots of simulated sample variances s^2 for a Gaussian FARIMA$(0, d, 0)$ model with $d = 0.4$ and $d = -0.4$, respectively, with $\sigma^2 = 1$. As expected, for $d = -0.4$ no relevant bias appears to be present whereas for $d = 0.4$ most sample variances are far below $\sigma^2 = 1$ even for $n = 1000$.

A further illustration which consequences long-range dependence and the negative bias of s^2 may have is given in the following example.

Example 5.6 Figure 5.5 shows a simulated sample path (of length $n = 2000$) of a Gaussian FARIMA$(2, d, 0)$ process with $d = 0.4$, $\varphi_1 = 0.5$ and $\varphi_2 = 0.4$. Based on the first 1000 observations the sample mean $\bar{x}$ and the sample variance s^2 are calculated. A standard rule of thumb for normally distributed observations is that $\bar{x} \pm 2s$ covers about 95 % of the population distribution. Figure 5.5 illustrates that for data with long memory this rule is not reliable even for large sample sizes. Here, $\bar{x} \pm 2s$ obtained from the first 1000 observations does not even include the expected value $\mu = 0$. The reason is that, due to long memory which is even accompanied by additional strong short memory (due to φ_1, φ_2), almost all of the first 1000 observations are above the expected value and their variability is (therefore) relatively small. Due to stationarity, the process drops below the expected value afterwards but that is too late, if the observer gets to see the first 1000 values only. In contrast, the interval $\bar{x} \pm 2\sqrt{\text{var}(X_t)}$ provides a much more reliable interval. Even though it is also affected by the high value of the sample mean, it is large enough—due to

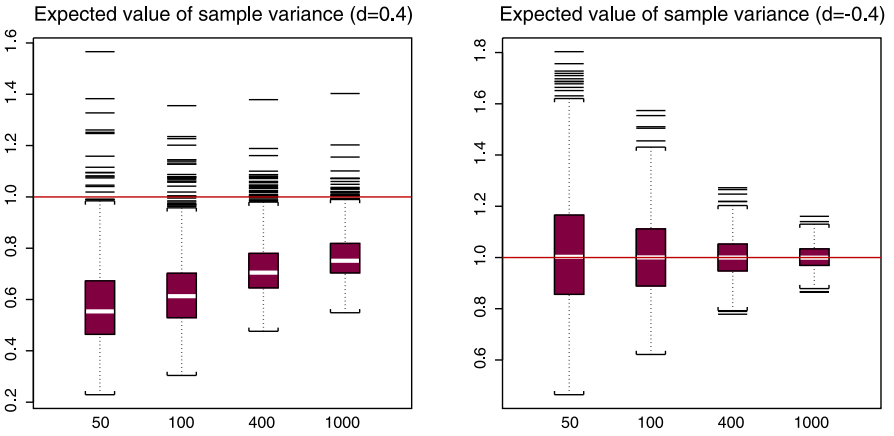

Fig. 5.4 Boxplots of simulated sample variances for a FARIMA$(0, d, 0)$ process with $d = 0.4$ and -0.4, respectively

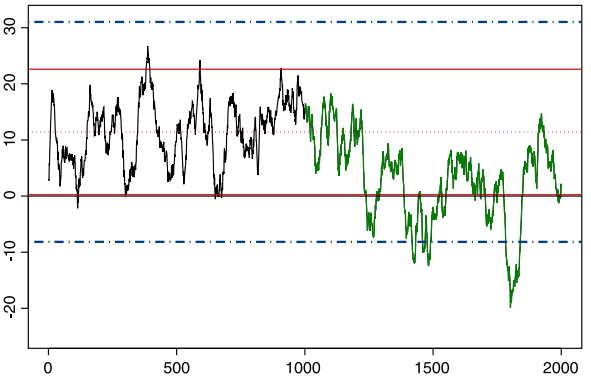

Fig. 5.5 Simulated sample path of a FARIMA$(0, 0.4, 0)$ process of length 2000. The *horizontal lines* correspond to the sample mean (*dotted line in the middle*) and the region $\bar{x} \pm 2s$ based on the first 1000 observations (*full lines*). The *outer dotted lines* correspond to $\bar{x} \pm 2\sqrt{\mathrm{var}(X_t)}$

the *correct* standard deviation—to include at least most of the next 1000 observations.

In summary, the sample variance is biased, and its asymptotic distribution is complicated and depends on specific assumptions that cannot be verified. A further problem is its slow rate of convergence. An alternative approach is to set

$$\hat{\sigma}^2 = \int_{-\pi}^{\pi} \hat{f}_X(\lambda)\, d\lambda$$

where $\hat{f}_X$ is an estimate of the spectral density of X_t ($t \in \mathbb{N}$). In particular, for parametric models we have $\hat{f}_X(\lambda) = f_X(\lambda; \hat{\theta})$ where $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ. Thus, the rate of convergence is better than for the sample variance (see

Sect. 5.5), and we obtain the same asymptotic normal distribution under rather general assumptions.

5.4 Heuristic Estimation of Long Memory

Let X_t ($t \in \mathbb{Z}$) be a stationary linear process with long memory. In this section, we review several heuristic methods for long memory identification (or more generally, the distinction between short memory, long memory or antipersistence):

- Variance plot;
- Rescaled range method (R/S);
- KPSS statistic;
- Rescaled variance method (V/S);
- Detrended Fluctuation Analysis (DFA);
- Temporal Aggregation.

For reasons outlined in the introduction, these methods are mainly useful for descriptive purposes rather than concrete statistical inference or model building.

5.4.1 Variance Plot

Recall that for long-range dependent linear processes

$$\operatorname{var}(\bar{x}) \sim C n^{2d-1},$$

where C is a constant (for simplicity of presentation, we omit more general slowly varying functions here). Consequently,

$$\log\big(\operatorname{var}(\bar{x})\big) \approx \log C + (2d-1)\log n.$$

This suggests applying linear regression with $\log n$ as predictor. For instance, one may define the following procedure:

1. Divide $X_1, \ldots, X_n$ into m non-overlapping, adjacent blocks of length k such that $n = mk$ (or $[mk]$).
2. Compute the sample mean for each block, i.e.

$$\bar{x}_k(j) = \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} X_t \quad (j = 1, \ldots, m).$$

3. Compute the overall mean

$$\bar{x} = \bar{x}_n = \frac{1}{m} \sum_{j=1}^{m} \bar{x}_k(j).$$

4. Compute an estimate of the variance of a sample based on k observations by

$$s^2(k) = \frac{1}{m-1} \sum_{j=1}^{m} \big(\bar{x}_k(j) - \bar{x}\big)^2.$$

5. Heuristically, when k grows, $\operatorname{var}(\bar{x}_k(j)) \sim Ck^{2d-1}$ for each $j = 1, \ldots, m$. Thus, $s^2(k)$ grows approximately at the rate k^{2d-1}. We therefore carry out steps 1–4 for $k = 2, \ldots, n/2$ and plot $\log s^2(k)$ against $\log k$. The slope should be approximately equal to $(2d-1)$. The estimator $\hat{d}_{\mathrm{VAR}}$ of d is then obtained from a least squares fit.

5.4.2 Rescaled Range Method

The rescaled range statistics R/S was introduced by Hurst (1951). Hurst's original definition motivated by the calculation of the minimal capacity of a dam is given in Sect. 1.2. When dealing with stationary processes, one often uses instead a simpler expression of the form

$$R_n = \max_{1 \le k \le n} \sum_{t=1}^{k} (X_t - \bar{x}_n) - \min_{1 \le k \le n} \sum_{t=1}^{k} (X_t - \bar{x}_n)$$

and $S_n^2 = \frac{1}{n-1} \sum_{t=1}^{n} (X_t - \bar{x}_n)^2$ (which is the same as the sample variance s^2). If X_t is second-order stationary, then S_n^2 converges in probability to $\sigma_X^2 = \operatorname{var}(X_t)$. Limiting properties of the R/S statistics were investigated by Mandelbrot (1975). An attractive feature is that the method is robust in the sense that under weak dependence the rate of convergence is $n^{-\frac{1}{2}}$ even if the variance of X_t is infinite (Mandelbrot and Wallis 1969a, 1969b, 1969c; Mandelbrot and Taqqu 1979). In other words, under mild regularity conditions, we have

$$n^{-1/2} \frac{R_n}{S_n} \xrightarrow[d]{} Q$$

where Q is a nondegenerate random variable (Feller 1951; Annis and Lloyd 1976; Mandelbrot and Wallis 1969a, 1969b, 1969c; Mandelbrot and Taqqu 1979). Another rate is obtained under long-range dependence or antipersistence. Suppose, for instance, that $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-k}$ is a linear process with $a_j \sim c_a j^{d-1}$ $(0 < d < \frac{1}{2})$ and $\operatorname{var}(X_t) = \sigma_X^2 < \infty$. Since $S_n^2 \to \sigma_X^2$ in probability, Theorem 4.6 implies

$$\begin{aligned} \sum_{t=1}^{k} (X_t - \bar{x}) &= \sum_{t=1}^{k} X_t - \frac{k}{n} \sum_{t=1}^{n} X_t \\ &\sim C(d) n^H \left[B_H\left(\frac{k}{n}\right) - \frac{k}{n} B_H(1) \right] \end{aligned}$$

and

$$n^{-H}\frac{R_n}{S_n} \underset{d}{\to} C(d)\Big[\sup_{u\in[0,1]} \tilde{B}_H(u) - \inf_{u\in[0,1]} \tilde{B}_H(u)\Big] =: \tilde{Z}_H \tag{5.17}$$

where

$$C^2(d) = \frac{\sigma_\varepsilon^2 c_a^2}{d(2d+1)} \int_0^\infty v^{d-1}(1+v)^{d-1}\, dv \tag{5.18}$$

and $\tilde{B}_H(u) = B_H(u) - uB_H(1)$ $(u \in [0,1])$ is a fractional Brownian bridge with Hurst parameter $H = d + \frac{1}{2}$. Note the similarity to CUSUM statistics used in the context of change point detection, see Sect. 7.9. Also note that, under additional uniform integrability conditions, (5.17) implies

$$E\left[\left(\frac{R_n}{S_n}\right)\right] \sim \text{const} \cdot n^{d+\frac{1}{2}} = \text{const} \cdot n^H. \tag{5.19}$$

This, or (5.17) (and in particular the difference between $H = \frac{1}{2}$ and $H \neq \frac{1}{2}$), is generally known as *Hurst effect*, and motivated Mandelbrot and co-workers to develop graphical techniques for estimating H based on the R/S statistic (Mandelbrot and Wallis 1969a, 1969b, 1969c; Mandelbrot and Taqqu 1979, also see Bassingthwaighte and Raymond 1994; Teverovsky et al. 1999 and references therein). Taking the logarithm on both sides of (5.17), we obtain

$$\begin{aligned} \log(R_n/S_n) &\approx H \log n + \log \tilde{Z}_H \\ &= \beta_0 + \beta_1 \log n + e_H(n) \end{aligned} \tag{5.20}$$

where $\beta_1 = H$, $\beta_0 = E[\log \tilde{Z}_H]$ and

$$e_H(n) \underset{d}{=} \log \tilde{Z}_H - E[\log \tilde{Z}_H].$$

This means that H can be interpreted as the slope of a regression line of $\log(R_n/S_n)$ against $\log n$ with intercept β_0 and random errors $e_H(n)$. The usual R/S-estimate of d is therefore defined by

$$\hat{d}_{R/S} = \hat{H}_{R/S} - \frac{1}{2} = \hat{\beta}_1 - \frac{1}{2}$$

where $\hat{\beta}_1$ is the least squares estimate of β_1. The plot of $\log(R/S)$ against $\log n$ (or originally rather the more complex version defined in Sect. 1.2) is also known as 'pox plot'.

A modified R/S statistic with a limiting distribution that does not depend on short-memory parameters was proposed by Lo (1991), with the purpose of testing

the null hypothesis $H_0 : d = 0$ (no long memory) against $H_1 : d > 0$. The idea is to replace S_n^2 by

$$S_{n,m}^2 = \sum_{k=-(m-1)}^{m-1} \left(1 - \frac{|k|}{m}\right) \hat{\gamma}_X(k) = S_n^2 + 2 \sum_{k=1}^{m-1} \left(1 - \frac{k}{m}\right) \hat{\gamma}_X(k)$$

where $\hat{\gamma}_X(k)$ are sample covariances of X_t ($t \in \mathbb{N}$). Under H_0, $S_{n,m}^2$ converges to $2\pi f_X(0)$, provided that m tends to infinity (in this sense, $(2\pi)^{-1} S_{n,m}^2$ is a non-parametric estimator of $f_X(0)$). This is the correct asymptotic standardization of $\sum_{t=1}^n X_t$ to obtain a standard Brownian bridge $n^{-\frac{1}{2}} R_n / S_{n,m}$ in the limit. Moreover, as shown in Giraitis et al. (2003), the standardization is also correct for $d \neq 0$ except for the factor m^{-2d}. More specifically one has

$$m^{-2d} S_{n,m}^2 \xrightarrow[p]{} \frac{c_\gamma}{d(2d+1)}.$$

The general statement is then

$$\left(\frac{m}{n}\right)^d n^{-d-\frac{1}{2}} R_n / S_{n,m} \xrightarrow[d]{} \sup_{u \in [0,1]} \tilde{B}_H(u) - \inf_{u \in [0,1]} \tilde{B}_H(u).$$

In particular, the test of H_0 is consistent because we use the standardized statistic

$$T_n = n^{-\frac{1}{2}} \frac{R_n}{S_{n,m}}$$

so that under H_1 we have $T_n \to \infty$. A practical problem is, however, an appropriate choice of m for a given data set where we do not know the short-memory structure. Moreover, for large lags (large compared to n) sample covariances do not yield reliable estimates of $\gamma_X(k)$. The testing procedure therefore tends to be quite volatile (see, e.g. the discussion in Teverovsky et al. 1999).

A further practical problem with the R/S approach is that it is not robust against departures from stationarity. Specifically, Bhattacharya et al. (1983) considered the model $X_t = \mu(t) + \varepsilon_t$, where ε_t are i.i.d. random variables and $\mu(t)$ is a hyperbolically decaying trend function. The estimate of d based on the R/S statistic then converges to a value of d larger than $\frac{1}{2}$. Thus, the estimator suggests long memory that is not really present in the data.

The most serious problem with the R/S method is, however, that it is not clear how to choose the cut-off point n_0 after which the linear approximation (5.20) is sufficiently accurate. Typically, there is a transient stretch where short-range correlations play a role and then the plot levels off (for $n \geq n_0$) to fluctuate around a straight line with slope $H = d + \frac{1}{2}$. Depending on how n_0 is selected, the estimated slopes may be quite different. Usually, n_0 is chosen by visual inspection. The accuracy of such a subjective method is, of course, hard to quantify, and different analysts may arrive at different conclusions. It should be noted, however, that this

comment is not specific to the R/S approach but rather applies to all heuristic and graphical methods considered in this section.

5.4.3 KPSS Statistic

An alternative to the R/S method is the so-called *KPSS* statistic proposed by Kwiatkowski et al. (1992) and its modification analogous to Lo's correction (Giraitis et al. 2003). There, the range is replaced by a second moment. The modified statistic is of the form

$$T_{\mathrm{KPSS}} = \frac{M_n}{S^2_{n,m}}$$

with $S^2_{n,m}$ as before (with $m \to \infty$, $m/n \to 0$) and

$$\begin{aligned} M_n &= \frac{1}{n^2} \sum_{k=1}^{n} \left\{ \sum_{t=1}^{k} (X_t - \bar{x}_n) \right\}^2 \\ &= \frac{1}{n^2} \sum_{k=1}^{n} \left\{ \sum_{t=1}^{k} X_t - \frac{k}{n} \sum_{t=1}^{n} X_t \right\}^2 . \end{aligned}$$

Again, using Theorem 4.6, we have

$$\begin{aligned} M_n &\approx C^2(d) n^{-2} \sum_{k=1}^{n} \left\{ B_H(k) - \frac{k}{n} B_H(n) \right\}^2 \\ &\underset{d}{=} C^2(d) n^{2H-1} \sum_{k=1}^{n} \left\{ B_H\left(\frac{k}{n}\right) - \frac{k}{n} B_H(1) \right\}^2 \frac{1}{n} \end{aligned}$$

and conclude

$$n^{-2d} M_n \underset{d}{\to} C^2(d) \int_0^1 \tilde{B}_H^2(u)^2 \, du,$$

where $\tilde{B}_H$ is a fractional Brownian bridge with Hurst parameter $H = d + 1/2$ and $C(d)$ is given in (5.18). For T_{KPSS} similar arguments lead to (see Giraitis et al. 2003)

$$\left(\frac{m}{n}\right)^{2d} T_{\mathrm{KPSS}} \underset{d}{\to} \int_0^1 \tilde{B}_H^2(u)^2 \, du$$

under fairly general assumptions including fourth order stationarity. The estimator of d is therefore defined as $\hat{d}_{\mathrm{KPSS}} = \frac{1}{2}\hat{\beta}_1$ where $\hat{\beta}_1$ is the least squares estimate of the slope when regressing $\log T_{\mathrm{KPSS}}$ on $\log(n/m)$. Finally, note that the *KPSS* approach has also been suggested for testing stationarity against the alternative of a fractionally integrated process (Lee and Schmidt 1996).

5.4.4 Rescaled Variance Method

In the rescaled variance (or V/S) method, Giraitis et al. (2003) propose replacing M_n by

$$V_n = \frac{1}{n^2}\left\{\sum_{k=1}^{n}\left[\sum_{t=1}^{k}(X_t - \bar{x}_n)\right]^2 - \frac{1}{n}\left[\sum_{k=1}^{n}\sum_{t=1}^{k}(X_t - \bar{x}_n)\right]^2\right\}.$$

For the so-called V/S statistic

$$T_{V/S} = \frac{V_n}{S_{n,m}^2},$$

we then have, again due to Theorem 4.6,

$$\left(\frac{m}{n}\right)^{2d} T_{V/S} \xrightarrow[d]{} \int_0^1 \tilde{B}_H^2(u)\,du - \left(\int_0^1 \tilde{B}_H(u)\,du\right)^2.$$

The estimator of d is therefore defined as $\hat{d}_{V/S} = \frac{1}{2}\hat{\beta}_1$ where $\hat{\beta}_1$ is the least squares estimate of the slope when regressing $\log T_{V/S}$ on $\log(n/m)$.

5.4.5 Detrended Fluctuation Analysis (DFA)

DFA was introduced in Peng et al. (1994) to provide some evidence of long memory in DNA sequences (also see Taqqu et al. 1995). The procedure works as follows:

1. Divide $X_1, \ldots, X_n$ into m nonoverlapping and adjacent blocks of size k such that $n = mk$ (or $n = [mk]$).
2. Within each of the m blocks, we regress $T_l = \sum_{t=1}^{l} X_t$ against l and estimate the variance of the residuals by

$$S_k^2(j) = \frac{1}{k}\sum_{l=(j-1)k+1}^{jk}(T_l - \hat{\beta}_{0,j} - \hat{\beta}_{1,j}l)^2 \quad (j = 1, \ldots, m).$$

   Here $\hat{\beta}_{0,j}$ and $\hat{\beta}_{1,j}$ are least squares regression estimates based on the jth block.
3. Compute

$$F^2(k) = \frac{1}{m}\sum_{j=1}^{m} S_k^2(j).$$

4. Heuristically, as the length k of the blocks grows, $S_k^2(j)$ grows at the rate $k^{2H} = k^{2d+1}$ for each j. Thus, $F^2(k)$ grows at the rate k^{2H}. Hence, after carrying out

steps 1–3 for $k = 2, \ldots, n/2$, we plot $\log F(k)$ against $\log k$ to obtain

$$\log F(k) \approx \log C + H \log k = \beta_0 + \beta_1 \log k.$$

The estimator of d is then obtained from the least squares estimate $\hat{\beta}_1$ by $\hat{d}_{\mathrm{DFA}} = \hat{H}_{\mathrm{DFA}} - \frac{1}{2} = \hat{\beta}_1 - \frac{1}{2}$.

This method is quite similar to the variance plot. The main difference is that instead of assuming stationarity a priori, a fitted linear trend function is subtracted first in each block. The method is therefore less sensitive to trends in the data than, for instance, the R/S approach.

5.4.6 Temporal Aggregation

Another idea of estimating the long-memory parameter d is based on the results in Sect. 2.2.1. If X_i $(i = 1, 2, \ldots, n)$ are generated by a second order stationary process with zero mean, then Theorem 2.67 implies that after sufficient aggregation the autocovariances and the spectral density of

$$Y_{i,M} = \sum_{j=(i-1)s+1}^{iM} X_j$$

are close to the spectral density of fractional Gaussian noise. After (sufficient) aggregation, we therefore can apply one of the Gaussian (quasi) maximum likelihood methods discussed in the next section, using fractional Gaussian noise as the parametric model. As for the previous heuristic methods, a problem with this approach is that it is difficult to decide how much aggregation is needed to come sufficiently close to the asymptotic spectral shape (2.71). An advantage is, however, that tests and confidence intervals for d may be based on asymptotic results for the MLE (see Theorem 5.2). In particular, we have $\hat{d} - d = O_p(n^{-\frac{1}{2}})$, though it should be said that, strictly speaking, Theorem 5.2 does not apply exactly because the additional uncertainty introduced by the preceding decision on the degree of aggregation is not taken into account. A practical limitation of the method is that aggregating blocks of length M reduces the data size by the factor $1/M$. One therefore needs a relatively long series to obtain reasonable results. On the other hand, even if the original series is far from Gaussian, aggregation often leads to an almost Gaussian process. An MLE approach based on the assumption of normality is then likely to be quite efficient (except for the reduced sample size due to aggregation). In particular, temporal aggregation is often advantageous when the original observations X_i are very discrete or if the series contains a large portion of zeroes. Data of this type are, for instance, often encountered in telecommunications and computer network engineering, or in climatological data such as ice thickness or precipitation. For a typical application, see, e.g. Beran et al. (1995).

5.4.7 Comments

Although the heuristic estimators discussed above are very easy to implement, they are generally not the best choice when it comes to reliable statistical inference. The main problem is that a suitable cut-off point has to be chosen after which the asymptotic approximation is good enough or, in the aggregation approach, one has to choose a suitable degree of aggregation. The choice is usually based on visual inspection. It is not clear how to quantify uncertainty of such estimates, and how to prevent results being guided by "wishful thinking". Other problems, such as lack of robustness against deterministic trends, are easier to handle and could be amended by suitable modifications. For instance, the DFA method removes linear trends a priori, in contrast to the related variance plot. For some detailed comments on R/S and DFA, also see Mielniczuk and Wojdyłło (2007a, 2007b). In spite of their limitations, the methods described in this section are often useful for a quick check. In particular, one may assess whether there may any chance of finding long memory in the data. A confirmation of the conjecture and concrete mathematical models have to be carried out using more sophisticated methods, some of which will be described in the following sections. Moreover, all heuristic methods described here focus on the long-memory parameter d and ignore any other aspect of the data. Also in this sense, they can be seen as complementary rather than competitors to more elaborate techniques, such as parametric or broadband estimation, where the complete dependence structure is modelled.

5.5 Gaussian Maximum Likelihood and Whittle Estimation

5.5.1 Exact Gaussian or Quasi-maximum Likelihood Estimation

Consider a linear process

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = A(B)\varepsilon_t \tag{5.21}$$

where $a_0 = 1$, $\sum a_j^2 < \infty$, $A(e^{-i\lambda}) = \sum a_j \exp(-ij\lambda)$, B is the backshift operator and ε_t are i.i.d. zero mean random variables with finite variance $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t)$. Suppose that the autocovariance function $\gamma_X(k)$ and the spectral density $f_X(\lambda)$ of X_t are known except for a $(p+1)$-dimensional parameter vector $\vartheta = (\sigma_\varepsilon^2, \theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^p$ characterizing the linear dependence structure. We will use the notation ϑ^0, $\sigma_{\varepsilon,0}^2$ and θ^0 whenever it needs to be emphasized that we are dealing with the true parameter values. Derivatives with respect to θ will be denoted by a dot, e.g.

$$\frac{\partial}{\partial\theta} f_X = \dot{f}_X, \qquad \frac{\partial^2}{\partial\theta^2} f_X = \ddot{f}_X, \ldots.$$

Note that, since X_t is a *linear* process, its distribution is fully specified by the distribution of the innovations ε_t and the parameter θ. In particular, if X_t ($t \in \mathbb{Z}$) is a (zero mean) Gaussian process, then its distribution is fully specified by the parameters σ_ε^2 (scale) and θ (linear dependence). As will be seen below, for Gaussian processes the maximum likelihood estimator (MLE) $\hat{\theta}_{\mathrm{MLE}}$ of θ is defined as the minimizer of a certain quadratic form, and the MLE of σ_ε^2 is obtained by evaluating a related quadratic form at $\theta = \hat{\theta}_{\mathrm{MLE}}$. The asymptotic distribution of the MLE therefore essentially follows from limit theorems for quadratic forms of Gaussian processes discussed in Sect. 4.5.1. If X_t is *not* Gaussian, then minimizing the Gaussian likelihood still leads to a consistent estimator of ϑ under fairly general conditions. In this case, one also speaks of a pseudo- or *quasi*-maximum likelihood estimator (QMLE). The asymptotic distribution then again follows from the corresponding limit theorems for quadratic forms. For instance, for linear processes suitable limit theorems are given in Sect. 4.5.2.

Let us now start with a zero mean Gaussian process X_t ($t \in \mathbb{Z}$) and data consisting of n observed values $X(n) = (X_1, \ldots, X_n)^T$. The joint probability density function of the random vector $X(n)$ is given by

$$p(x;\vartheta) = (2\pi)^{-\frac{n}{2}} |\Sigma_n|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^T \Sigma_n^{-1} x\right) \quad \left(x \in \mathbb{R}^n\right)$$

where

$$\Sigma_n = \Sigma_n(\vartheta) = \left[\gamma_X(r-s;\vartheta)\right]_{r,s=1,\ldots,n}$$

is the $n \times n$ covariance matrix of $X(n)$ and $|\cdot|$ denotes the determinant. The log-likelihood function can therefore be written as

$$\log p(x;\vartheta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma_n(\vartheta)\right| - \frac{1}{2} x^T \Sigma_n^{-1}(\vartheta) x.$$

Multiplying by $-2/n$ and ignoring the constant leads to the "log-likelihood function"

$$\mathscr{L}_{n,\mathrm{exact}}(\vartheta) = \frac{1}{n}\log\left|\Sigma_n(\vartheta)\right| + \frac{1}{n} x' \Sigma_n^{-1}(\vartheta) x. \tag{5.22}$$

The maximum likelihood estimator of ϑ is defined as

$$\hat{\vartheta}_{n,\mathrm{MLE}} = \operatorname{argmin} \mathscr{L}_{n,\mathrm{exact}}(\vartheta). \tag{5.23}$$

The estimator is asymptotically normal as stated in the following theorem. The result was proven by Yajima (1985) for FARIMA$(0, d, 0)$ models, whereas Dahlhaus (1989) considered general Gaussian processes with a possibly unknown mean. Hosoya (1997) considered extensions of the results by Dahlhaus to a bivariate setting.

Theorem 5.2 *Assume that* $X_t = \sum a_j \varepsilon_{t-j}$ ($t \in \mathbb{Z}$) *is a linear process with spectral density*

$$f_X(\lambda) \sim c_f |\lambda|^{-2d} \quad (\text{as } \lambda \to 0)$$

for some $d \in (-\frac{1}{2}, \frac{1}{2})$. *Then, under suitable regularity conditions, we have*

$$\sqrt{n}\big(\hat{\vartheta}_{n,\mathrm{MLE}} - \vartheta^0\big) \underset{d}{\to} N(0, \Sigma_{\mathrm{MLE}}) \tag{5.24}$$

with

$$\Sigma_{\mathrm{MLE}} = 4\pi C_{\mathrm{MLE}}^{-1} + \kappa_4 \sigma_\varepsilon^4 I_{\mathrm{var}},$$

where $C_{\mathrm{MLE}} = [c_{rs}]_{r,s=1,\ldots,p+1}$ *is a matrix with entries*

$$c_{rs} = \int_{-\pi}^{\pi} \left(\frac{\partial}{\partial \vartheta_r} \log f_X(\lambda;\vartheta)\right)\left(\frac{\partial}{\partial \vartheta_s} \log f_X(\lambda;\vartheta)\right) d\lambda \Big|_{\vartheta=\vartheta^0}$$

and all entries in the $(p+1)\times(p+1)$ *matrix* I_{var} *are equal to zero except for the upper left corner with* $[I_{\mathrm{var}}]_{11} = 1$.

Remark 5.1 For $d > 0$, "suitable regularity conditions" are in particular:

- The function $\lambda \to f_X(\lambda;\vartheta)$ is continuous except at $\lambda = 0$;
- The function $\lambda \to 1/f_X(\lambda;\vartheta)$ is continuous;
- The function $\vartheta \to \int_{-\pi}^{\pi} \log f_X(\lambda;\vartheta)\, d\lambda$ is differentiable (twice) under the integral sign.

These conditions guarantee the validity of several approximation arguments, as well as the interchange of differentiation with integration. For $d \in (-\frac{1}{2}, 0]$, conditions have to be adapted accordingly.

If $\kappa_4 = 0$ then the integral involved in the definition of C_{MLE} resembles Fisher's information matrix. In fact, it can be shown that under the assumption of Gaussianity (and some regularity conditions), Σ_{MLE} is indeed the inverse of the limit of Fisher's information matrix (Dahlhaus 1989). Finally note that, in general, Σ_{MLE} depends on the unknown parameter vector ϑ (though for θ the corresponding submatrix does not depend on σ_ε^2). A notable exception is the FARIMA$(0,d,0)$ model (see the discussion below in the context of the Whittle estimator).

The exact MLE, however, is difficult to deal with both theoretically and numerically. To simplify the exact log-likelihood in (5.22), note first that $f_X(\lambda)$ can be written as

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\big|A\big(e^{-i\lambda}\big)\big|^2 = \frac{\sigma_\varepsilon^2}{2\pi} h_X(\lambda),$$

where $h_X(\lambda) = (2\pi/\sigma_\varepsilon^2) f_X(\lambda)$ depends on θ only and not on σ_ε^2. The autocovariance function can be written as

$$\gamma_X(k;\vartheta) = \int_{-\pi}^{\pi} e^{ik\lambda} f_X(\lambda;\vartheta)\, d\lambda = \frac{\sigma_\varepsilon^2}{2\pi}\int_{-\pi}^{\pi} e^{ik\lambda} h_X(\lambda;\theta)\, d\lambda$$

so that

$$\Sigma_n(\vartheta) = \frac{\sigma_\varepsilon^2}{2\pi}\Sigma_{h,n}(\theta)$$

with

$$\Sigma_{h,n} = \Sigma_{h,n}(\theta) = \left[\int_{-\pi}^{\pi} e^{i(r-s)\lambda} h_X(\lambda;\theta)\, d\lambda\right]_{r,s=1,\ldots,n} = 2\pi[\hat{h}_{r-s}]_{r,s=1,\ldots,n}.$$

Here, $\hat{h}_{r-s}$ denotes the $(r-s)$th Fourier coefficient of $h_X(\lambda;\theta)$. Hence

$$\left|\Sigma_n(\vartheta)\right| = \left(\frac{\sigma_\varepsilon^2}{2\pi}\right)^n \left|\Sigma_{h,n}(\theta)\right|.$$

This leads to the alternative formula for the log-likelihood function

$$\mathscr{L}_{n,\text{exact}}\left(\sigma_\varepsilon^2,\theta\right) = \log\sigma_\varepsilon^2 + \frac{1}{n}\log\left|\Sigma_{h,n}(\theta)\right| + \left(\frac{2\pi}{\sigma_\varepsilon^2}\right)\frac{1}{n}x'\Sigma_{h,n}^{-1}(\theta)x. \tag{5.25}$$

In this set-up, a first simplification of $\mathscr{L}_{n,\text{exact}}$ can be made by noting that (see Grenander and Szegö 1958 and (4.8))

$$\lim_{n\to\infty}\frac{1}{n}\log\left|\Sigma_{h,n}(\theta)\right| = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log h_X(\lambda;\theta)\, d\lambda = 0. \tag{5.26}$$

(In an exact proof, convergence has to be shown to be uniform in θ in a suitable way). Note that $\int \log h_X(\lambda;\theta)\, d\lambda$ being zero follows directly from the Wiener–Kolmogorov formula for the one-step prediction error

$$\begin{aligned}\sigma_\varepsilon^2 &= 2\pi\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\log f_X(\lambda)\, d\lambda\right)\\ &= 2\pi\exp\left(\log\frac{\sigma_\varepsilon^2}{2\pi} + \frac{1}{2\pi}\int_{-\pi}^{\pi}\log h_X(\lambda)\, d\lambda\right)\\ &= \sigma_\varepsilon^2\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\log h_X(\lambda)\, d\lambda\right)\end{aligned} \tag{5.27}$$

(see, e.g. Priestley 1981; also see Chap. 8). Thus, using $h_X(\lambda;\theta)$ allows us to separate estimation of σ_ε^2 and θ, and, due to (5.26), to ignore one complicated term in the likelihood function. We therefore replace $\mathscr{L}_{n,\text{exact}}$ by

$$\mathscr{L}_{n,\text{approx}}\left(\sigma_\varepsilon^2,\theta\right) = \int_{-\pi}^{\pi}\log f_X(\lambda)\, d\lambda + \left(\frac{2\pi}{\sigma_\varepsilon^2}\right)\frac{1}{n}x'\Sigma_{h,n}^{-1}(\theta)x, \tag{5.28}$$

or equivalently by

$$\mathscr{L}_{n,\text{approx}}\left(\sigma_\varepsilon^2,\theta\right) = \log\sigma_\varepsilon^2 + \left(\frac{2\pi}{\sigma_\varepsilon^2}\right)\frac{1}{n}x'\Sigma_{h,n}^{-1}(\theta)x. \tag{5.29}$$

In this form, minimization with respect to $\vartheta = (\sigma_\varepsilon^2,\theta)$ can be done for θ and σ_ε^2 separately. First, one minimizes $\mathscr{L}_{n,\text{approx}}(\sigma_\varepsilon^2,\theta)$ with respect to θ to obtain

$$\hat{\theta} = \hat{\theta}_{n,\text{approx MLE}} = \text{argmin}_\theta\mathscr{L}_{n,\text{approx}}\left(\sigma_\varepsilon^2,\theta\right). \tag{5.30}$$

Then $\hat{\sigma}_\varepsilon^2$ is obtained by plugging $\hat{\theta}$ into the log-likelihood expression (5.29) and minimizing with respect to σ_ε^2 while keeping $\hat{\theta}$ fixed. This leads to the explicit solution

$$\hat{\sigma}_\varepsilon^2 = \frac{2\pi}{n} x' \Sigma_{h,n}^{-1}(\hat{\theta}) x. \tag{5.31}$$

If X_t ($t \in \mathbb{Z}$) is not Gaussian, then these equations can still be used for defining $\hat{\vartheta}$, but the estimate is no longer an approximate MLE.

5.5.2 Whittle Estimation

In order to find the approximate MLE in (5.29), one needs to invert the $n \times n$ covariance matrix Σ_n. For large values of n, this is not a pleasant task numerically, in particular since for values of d close $\frac{1}{2}$ the covariance matrix may be close to singularity. Note that, even if the true value of d is not close to $\frac{1}{2}$, trial values of d close to the border may occur frequently during numerical optimization. The problem with the inverse covariance matrix is illustrated in Fig. 5.6 where, for the case of a FARIMA$(0, d, 0)$ process and $n = 1000$, the condition number κ of Σ_n (defined as the ratio of the maximal and minimal eigenvalues) is plotted as a function of d. For $d = 0$, we have a diagonal matrix and hence (with $\kappa = 1$) no problem with the numerical precision of the inverse. However, κ increases quite rapidly for positive values of d which means that the numerical precision when calculating Σ_n^{-1} deteriorates substantially. The problem is less severe for $d < 0$ although very close to the left border of $-\frac{1}{2}$ the condition number is quite large as well. A further problem with handling the inverse covariance matrix directly is that for asymptotic considerations one would need to study the second-order derivatives of $\Sigma_{h,n}^{-1}(\theta)$.

This calls for a further simplification. The following very elegant solution is due to Whittle (1953). To simplify the presentation, we consider the case where θ is one-dimensional, i.e. $\theta = d$. The derivation can easily be carried over to multivariate parameters $\theta \in \mathbb{R}^p$. Thus, let $\theta = d$. We approximate the matrix $\Sigma_{h,n}^{-1}(\theta)$ by $W_n(\theta)$, where

$$W_n(\theta) = \big[w_{r-s}(\theta)\big]_{r,s=1,\dots,n} = \left[\frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} e^{i(r-s)\lambda} \frac{1}{h_X(\lambda;\theta)}\, d\lambda\right]_{r,s=1,\dots,n}. \tag{5.32}$$

Formally, the doubly-infinite matrix $W_\infty(\theta)$ is the inverse of $\Sigma_{h,\infty}(\theta)$. To see this, we note that the entries of the matrix $W_n(\theta)$ are Fourier coefficients of $1/h$ multiplied by $(2\pi)^{-1}$. We compute, for example, the entry (r, r) of the product $W_\infty(\theta)\Sigma_{h,\infty}(\theta)$, and obtain

$$\sum_{l=-\infty}^{\infty} (2\pi)^{-1} \widehat{(1/h)}_{r-l} (2\pi) \hat{h}_{l-r} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{h_X(\lambda)} h_X(\lambda)\, d\lambda = 1$$

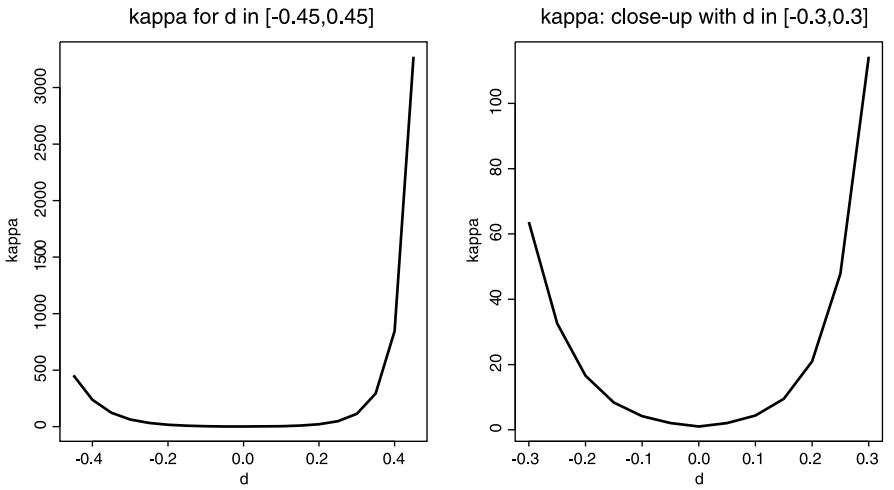

Fig. 5.6 Condition number κ (= ratio of the largest and the smallest eigenvalue) of the covariance matrix Σ_n ($n = 1000$) for a FARIMA$(0, d, 0)$ process, plotted as a function of d. The *right panel* is a zoomed picture of the left one

where we used the Parseval's identity. More generally, we note that

$$\sum_{k=-\infty}^{\infty} e^{-iku} = 2\pi\,\delta(u)$$

where $\delta(\cdot)$ is the Dirac (generalized) function with the property $\int \delta(u)g(u)\,du = g(0)$ (for sufficiently regular functions g). Then, assuming summation and integration to be interchangeable, we have

$$\begin{aligned}\big[W_\infty(\theta)\,\Sigma_{h,\infty}(\theta)\big]_{rs} &= \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{i(r\lambda - s\nu)} \underbrace{\sum_{j=-\infty}^{\infty} e^{-ij(\lambda-\nu)}}_{2\pi\,\delta(\lambda-\nu)} \frac{h_X(\nu;\theta)}{h_X(\lambda;\theta)}\,d\lambda\,d\nu \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(r-s)\lambda}\,d\lambda = \delta_{rs}\end{aligned}$$

with δ_{rs} denoting the Kronecker delta (i.e. $\delta_{rr} = 1$, and $\delta_{rs} = 0$ for $r \neq s$).

Thus, instead of $\mathscr{L}_{n,\text{approx}}$ in (5.29) we consider

$$\begin{aligned}\mathscr{L}_{n,\text{Whittle}}\big(\sigma_\varepsilon^2, \theta\big) &= \log\sigma_\varepsilon^2 + \left(\frac{2\pi}{\sigma_\varepsilon^2}\right)\frac{1}{n}\sum_{t,s=1}^{n} w_{t-s}X_tX_s \\ &= \log\sigma_\varepsilon^2 + \left(\frac{2\pi}{\sigma_\varepsilon^2}\right)\frac{1}{n}x^T W_n(\theta)x. \qquad (5.33)\end{aligned}$$

Now, minimizing $\mathscr{L}_{n,\text{Whittle}}$ with respect to θ yields

$$Q_{n,\text{Whittle}}(\hat{\theta}) = x^T(n)\frac{\partial}{\partial\theta}W_n(\hat{\theta})x = 0. \tag{5.34}$$

The estimator $\hat{\theta}$ that minimizes $\mathscr{L}_{n,\text{Whittle}}$ is often referred to as Whittle estimator. For the particular case with $\theta = d$, the Whittle estimator is

$$\hat{d}_{\text{Whittle}} := \text{argmin}_d \frac{1}{n} x' W_n(d) x.$$

The asymptotic distribution of $\hat{\theta}_{\text{Whittle}}$ therefore essentially follows from limit theorems for quadratic forms, as studied in Sect. 4.5.

The estimator is asymptotically normal as described in the next theorem, and, in fact, turns out to have the same asymptotic distribution as the exact MLE. The result was proven in Fox and Taqqu (1986) in the Gaussian case (also see Beran 1986) and in Giraitis and Surgailis (1990) and Horváth and Shao (1999) for linear processes. We note that consistency of the Whittle estimator was proven in Hannan (1973) for a general class of ergodic sequences. We state the theorem first for the simplest case where d is the only unknown parameter:

Theorem 5.3 *Assume that X_t ($t \in \mathbb{Z}$) is a linear process with spectral density*

$$f_X(\lambda) \sim c_f |\lambda|^{-2d} \quad (as\ \lambda \to 0)$$

for some $-\frac{1}{2} < d < \frac{1}{2}$. Assume that $\theta = d$. Then, under appropriate regularity conditions,

$$\sqrt{n}\big(\hat{d}_{\text{Whittle}} - d^0\big) \to N\big(0, 4\pi V^{-1}\big),$$

where

$$V = \int_{-\pi}^{\pi} \left(\frac{\dot{f}_X(\lambda)}{f_X(\lambda)}\right)^2 d\lambda.$$

Proof We sketch a longish proof here for $d > 0$, postponing details to the end of the section. Recall (5.34). Suitable regularity conditions enable us to apply a Taylor expansion of the form

$$0 = Q_{n,\text{Whittle}}(\hat{\theta}) = Q_{n,\text{Whittle}}\big(\theta^0\big) + \dot{Q}_{n,\text{Whittle}}(\tilde{\theta})\big(\hat{\theta} - \theta^0\big)$$

where $|\tilde{\theta} - \theta^0| \le |\hat{\theta} - \theta^0|$, and thus

$$\hat{\theta} - \theta^0 \approx M^{-1} \cdot n^{-1} Q_{n,\text{Whittle}}\big(\theta^0\big)$$

where

$$M = \lim_{n\to\infty} n^{-1} E_{\theta^0}\big[\dot{Q}_{n,\text{Whittle}}\big(\theta^0\big)\big].$$

Consequently, in view of (5.34), the main ingredient of the proof of asymptotic normality of the Whittle estimator is the limiting behaviour of the quadratic form

$$Q_{n,\text{Whittle}}(\theta) = \sum_{t,s=1}^{n} \dot{w}_{i-j}(\theta) X_t X_s \tag{5.35}$$

where $\dot{w}_k = \dot{w}_k(\theta) = \frac{\partial}{\partial\theta} w_k(\theta)$, i.e. (cf. (5.32))

$$\dot{w}_k = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} e^{ik\lambda} \frac{\partial}{\partial\theta} \frac{1}{h_X(\lambda;\theta)}\, d\lambda.$$

Defining

$$\tilde{g}(\lambda;\theta) := \frac{1}{(2\pi)^2} \frac{\partial}{\partial\theta} \frac{1}{h_X(\lambda;\theta)} = -\frac{1}{(2\pi)^2} \frac{\dot{h}_X}{h_X^2},$$

we have

$$\dot{w}_k = \int_{-\pi}^{\pi} e^{ik\lambda} \tilde{g}(\lambda;\theta)\, d\lambda.$$

First of all, it can be established that the quadratic form is centred, i.e.

$$E\big[Q_{n;\text{Whittle}}(\theta)\big] \approx 0 \tag{5.36}$$

(see technical notes at the end of this section). If now $f_X(\lambda) \sim c_f |\lambda|^{-2d}$ $(\lambda \to 0)$, then $h_X^{-1}(\theta) \sim c_f^{-1} |\lambda|^{2d}$ and differentiating w.r.t. $\theta = d$ yields

$$\frac{\partial}{\partial\theta} h_X^{-1}(\lambda;\theta) \sim 2c_f^{-1} \log\lambda \cdot \lambda^{2d} \quad (\lambda \to 0).$$

Thus, Theorem 4.30 and its multivariate extension is applicable to $Q_{n,\text{Whittle}}$ with $\gamma = -2d$, and arbitrary long-memory parameter $d \in (0, \frac{1}{2})$. In other words, for the Whittle estimator condition (4.131) of Theorem 4.30 is always fulfilled, i.e. the long-memory contribution is neutralized by considering the reciprocal of the spectral density in the definition of the quadratic form $Q_{n,\text{Whittle}}$. The central limit theorem is then of the form

$$n^{-\frac{1}{2}} Q_{n,\text{Whittle}}\big(\theta^0\big) \underset{d}{\to} \sigma_Q N(0,1) \tag{5.37}$$

with

$$\sigma_Q^2 = 16\pi^3 \int_{-\pi}^{\pi} \big[f_X(\lambda)\tilde{g}(\lambda)\big]^2\, d\lambda + \kappa_4 \left(2\pi \int_{-\pi}^{\pi} f_X(\lambda)\tilde{g}(\lambda)\, d\lambda\right)^2.$$

However, here the term with κ_4 disappears since

$$\int_{-\pi}^{\pi} \big[f_X(\lambda)\tilde{g}(\lambda)\big]\, d\lambda = -\frac{1}{(2\pi)^2} \left(\frac{\sigma_\varepsilon^2}{2\pi}\right) \frac{\partial}{\partial\theta} \int_{-\pi}^{\pi} \log h_X(\lambda)\, d\lambda = 0.$$

Furthermore, one can see that

$$\sigma_Q^2 = \left(\frac{\sigma_\varepsilon^2}{2\pi}\right)^2 \frac{1}{\pi} V$$

and

$$M = \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{2\pi} V$$

(see technical details below). Combining the arguments together and denoting by Z a standard normal variable, we obtain

$$\begin{aligned}\sqrt{n}\big(\hat{\theta} - \theta^0\big) &\approx \big[n^{-1}\dot{Q}_{n,\text{Whittle}}\big(\theta^0\big)\big]^{-1} \cdot \frac{1}{\sqrt{n}} Q_n\big(\theta^0\big) \\ &\xrightarrow[d]{} \left(\frac{\sigma_\varepsilon^2}{2\pi} \cdot \frac{1}{2\pi} V\right)^{-1} \left(\frac{\sigma_\varepsilon^2}{2\pi} \cdot \frac{1}{\sqrt{\pi}} V^{\frac{1}{2}}\right) Z \stackrel{d}{=} N(0, \Sigma_\theta)\end{aligned}$$

with

$$\Sigma_\theta = 4\pi V^{-1}.$$

□

The result can be extended to simultaneous estimation of σ_ε^2 and θ, as well as to multivariate parameter vectors θ. The scale estimator

$$\hat{\sigma}_\varepsilon^2 = \frac{2\pi}{n} x^T \Sigma_{h,n}^{-1}(\hat{\theta}) x$$

can be approximated by

$$\hat{\sigma}_\varepsilon^2 = \frac{2\pi}{n} \sum_{t,s=1}^{n} w_{t-s} X_t X_s$$

where w_k are as before, i.e.

$$w_k = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} e^{ik\lambda} \frac{1}{h_X(\lambda;\theta)}\, d\lambda = \int_{-\pi}^{\pi} e^{ik\lambda} \tilde{g}(\lambda;\theta)\, d\lambda.$$

The asymptotic distribution of $\hat{\sigma}_\varepsilon^2$ therefore directly follows from Theorem 4.30. In particular, the asymptotic variance of $\hat{\sigma}_\varepsilon^2$ is equal to

$$\begin{aligned}\sigma_{\text{var}}^2 &= (2\pi)^2 \left[16\pi^3 \int_{-\pi}^{\pi} \big(f_X(\lambda)\tilde{g}(\lambda)\big)^2 d\lambda + \kappa_4 \left(2\pi \int_{-\pi}^{\pi} f_X(\lambda)\tilde{g}(\lambda)\, d\lambda\right)^2\right] \\ &= (2+\kappa_4)\sigma_\varepsilon^4\end{aligned}$$

where $\kappa_4 = E(\varepsilon_t^4) - 3$ is the kurtosis of ε_t and we have a function $\tilde{g}$ that is different than for the other parameters, namely $\tilde{g}(\lambda) = (2\pi)^{-2}(1/h_X(\lambda))$. In the Gaussian case, one has $\kappa_4 = 0$ so that $\sigma_{\text{var}}^2 = 2\sigma_\varepsilon^4$.

Comparing these results for the Whittle estimator $\hat{\vartheta} = (\hat{\sigma}_\varepsilon^2, \hat{\theta}_{n,\text{Whittle}})$ with those for the exact (Gaussian or quasi) MLE (Theorem 5.2), we see that the asymptotic distribution is the same, namely

$$\sqrt{n}\big(\hat{\vartheta}_{\text{Whittle}} - \vartheta^0\big) \underset{d}{\to} N(0, \Sigma_{\text{MLE}}) \tag{5.38}$$

with

$$\Sigma_{\text{MLE}} = 4\pi C_{\text{MLE}}^{-1} + \kappa_4 \sigma_\varepsilon^4 I_{\text{var}}$$

where C_{MLE} is the matrix with entries

$$c_{rs} = \int_{-\pi}^{\pi} \left(\frac{\partial}{\partial \vartheta_r} \log f_X(\lambda; \vartheta)\right)\left(\frac{\partial}{\partial \vartheta_s} \log f_X(\lambda; \vartheta)\right) d\lambda \Big|_{\vartheta=\vartheta^0}.$$

In particular, in the Gaussian case, $\hat{\vartheta}_{\text{MLE}}$ is indeed the actual maximum likelihood estimator so that we can say that $\hat{\vartheta}_{\text{Whittle}}$ is asymptotically equivalent to the exact MLE, and is hence asymptotically efficient (see Dahlhaus 1989 for an exact proof of the asymptotic efficiency of the MLE in the Gaussian case).

It is interesting to look at Σ_{MLE} more closely. First of all, the element c_{11} is equal to

$$c_{11} = 2\pi \sigma_\varepsilon^{-4}$$

so that

$$\sigma_{\text{var}}^2 = [\Sigma_{\text{MLE}}]_{11} = (2 + \kappa_4)\sigma_\varepsilon^4.$$

For $r = 1$ and $s > 1$, we obtain

$$c_{rs} = c_{sr} = \sigma_\varepsilon^{-2} \int_{-\pi}^{\pi} \frac{\partial}{\partial \vartheta_s} \log h_X(\lambda; \vartheta)\, d\lambda \Big|_{\vartheta=\vartheta^0} = 0.$$

The asymptotic covariance matrix therefore simplifies to

$$\Sigma_{\text{MLE}} = \begin{pmatrix} \sigma_{\text{var}}^2 & 0 \\ 0 & \Sigma_\theta \end{pmatrix} = \begin{pmatrix} (2+\kappa_4)\sigma_\varepsilon^4 & 0 \\ 0 & \Sigma_\theta \end{pmatrix} = \begin{pmatrix} (2+\kappa_4)\sigma_\varepsilon^4 & 0 \\ 0 & 4\pi V^{-1} \end{pmatrix} \tag{5.39}$$

where $4\pi V^{-1} = [\Sigma_{\text{MLE}}]_{r,s=2,\ldots,p+1}$. This means that the scale estimator $\hat{\sigma}_\varepsilon^2$ is asymptotically independent of the other parameter estimates. Moreover, the asymptotic distribution of $\hat{\theta}$ does not depend on σ_ε^2 and is the same as if σ_ε^2 were known. Note, however, that σ_ε^2 is the *innovation* variance. Estimating the variance $\sigma_X^2 = \text{var}(X_t)$ of the *process* itself and estimating θ *cannot* be done independently because the variance $\sigma_X^2 = \sigma_\varepsilon^2 \int h_X(\lambda)\, d\lambda$ depends on both parameters, σ_ε^2 *and* θ. Furthermore, note that, in general, the asymptotic covariance matrix of $\hat{\theta}$ depends on the unknown parameters θ. There are, however, models where this is not the case because the derivative of $\log f_X$ (with respect to θ) does not depend on θ. An important example is the fractional ARIMA$(0, d, 0)$ model.

Example 5.7 Let $\theta = d$. The spectral density of a fractional ARIMA$(0, d, 0)$ process is given by

$$f_X(\lambda; d) = \frac{\sigma_\varepsilon^2}{2\pi}(2 - 2\cos\lambda)^{-d}.$$

Therefore,

$$\frac{\partial}{\partial\theta}\log f_X(\lambda; d) \equiv -\log(2 - 2\cos\lambda)$$

does not depend on d. Moreover,

$$(4\pi)^{-1}\int_{-\pi}^{\pi}\left[\log(2 - 2\cos\lambda)\right]^2 d\lambda = \frac{\pi^2}{6}$$

so that the asymptotic variance of $\hat{d}_{\text{Whittle}}$ is equal to $6/\pi^2 \approx 0.608$. More generally, the asymptotic variance of $\hat{d}_{\text{Whittle}}$ (and $\hat{d}_{\text{MLE}}$) is nuisance parameter free for all models with

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}L(\lambda)|\lambda|^{-2d}\exp\left(\sum_{j=2}^{p}\theta_j g_j(\lambda)\right)$$

where $L(\lambda)$, $g_j(\lambda)$ are functions that do not depend on any parameters, $L(\lambda)$ is slowly varying at zero, and $g_j(\lambda)$ are bounded. For $L(\lambda) \equiv 1$ this is the definition of an FEXP model of order $\tilde{p} = p - 1$ (Beran 1993; Robinson 1994a; see Sect. 2.1.1.5).

Finally, it should be noted that the extension to subordinated processes is not straightforward. If instead of a linear process X_t we consider Gaussian subordination, i.e. $G(X_t)$ with X_t Gaussian but G nonlinear, then Giraitis and Taqqu (1999a) showed that the results of Theorem 5.3 are no longer valid. Depending on detailed conditions, one can obtain asymptotic normality with a rate of convergence slower than $n^{-\frac{1}{2}}$ or even a non-normal limiting distribution.

5.5.3 Further Comments on the Whittle Estimator

Here we give a heuristic explanation why long memory cancels out in the Whittle estimator so that we obtain $\sqrt{n}$-convergence. Recall the approximate log-likelihood equation (5.33):

$$\mathscr{L}_{n,\text{Whittle}}\left(\sigma_\varepsilon^2, \theta\right) = \log\sigma_\varepsilon^2 + \left(\frac{2\pi}{\sigma_\varepsilon^2}\right)\frac{1}{n}\sum_{t,s=1}^{n} w_{t-s}X_tX_s.$$

This can also be written in terms of the periodogram as

$$\mathscr{L}_{n,\text{Whittle}}\left(\sigma_\varepsilon^2, \theta\right) = \log\sigma_\varepsilon^2 + \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\frac{1}{n}\sum_{t,s=1}^{n} X_tX_s e^{i(t-s)\lambda}\frac{1}{f_X(\lambda;\theta)}\,d\lambda$$

$$= \log \sigma_\varepsilon^2 + \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_{n,X}(\lambda)}{f_X(\lambda;\theta)}\, d\lambda. \tag{5.40}$$

If $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$, then

$$\frac{1}{2\pi} \frac{I_{n,X}(\lambda)}{f_X(\lambda)} \approx I_{n,\varepsilon}(\lambda),$$

where $I_{n,\varepsilon}(\lambda)$ is the periodogram of the underlying i.i.d. sequence. Therefore, the long-memory effect vanishes. On the other hand, such a simplification is not valid for subordinated processes (see Giraitis and Taqqu 1999a).

The Whittle estimator can also be considered in terms of a smoothed periodogram. Heyde and Gay (1989) considered a general class of smoothed periodograms

$$G(\vartheta) := \int_{-\pi}^{\pi} \psi(\lambda;\vartheta)\big\{I_{n,X}(\lambda) - E\big[I_{n,X}(\lambda)\big]\big\}\, d\lambda$$

with a suitably chosen function $\psi(\cdot;\theta)$ and looked for solutions of $G(\vartheta) = 0$. Note that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log f_X(\lambda;\theta)\, d\lambda = \log \sigma_\varepsilon^2 - \log 2\pi.$$

Thus, ignoring the $\log 2\pi$ term, we can consider minimization of

$$\mathscr{L}_{n,\text{Whittle}}\big(\sigma_\varepsilon^2, \theta\big) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \log f_X(\lambda;\theta) + \frac{I_{n,X}(\lambda)}{f_X(\lambda;\theta)} \right\} d\lambda.$$

This amounts to solving

$$\dot{\mathscr{L}}_{n,\text{Whittle}}\big(\sigma_\varepsilon^2, \theta\big) = \frac{1}{2\pi} \int \frac{\dot{f}_X(\lambda;\vartheta)}{f_X(\lambda;\vartheta)} \left\{ \frac{I_{n,X}(\lambda)}{f_X(\lambda;\vartheta)} - 1 \right\} d\lambda = 0 \tag{5.41}$$

where

$$\dot{f}_X(\lambda;\vartheta) = \frac{\partial}{\partial \vartheta} f_X(\lambda;\vartheta).$$

Hence, we are in the Heyde and Gay (1989) set-up by choosing

$$\psi(\lambda;\vartheta) = \left(\frac{\dot{f}_X(\lambda;\vartheta)}{f_X(\lambda;\vartheta)} \right) \frac{1}{f_X(\lambda;\vartheta)},$$

and

$$\begin{aligned} G(\vartheta) &:= \int_{-\pi}^{\pi} \psi(\lambda;\vartheta)\big\{I_{n,X}(\lambda) - E\big[I_{n,X}(\lambda)\big]\big\}\, d\lambda \\ &= \int \frac{\dot{f}_X(\lambda;\vartheta)}{f_X(\lambda;\vartheta)} \left\{ \frac{I_{n,X}(\lambda)}{f_X(\lambda;\vartheta)} - E\left[\frac{I_{n,X}(\lambda)}{f_X(\lambda;\vartheta)} \right] \right\} d\lambda = 0, \end{aligned}$$

and arguing that $E[I_{n,X}(\lambda;\vartheta)]$ can be replaced by $f_X(\lambda;\vartheta)$ (Rosenblatt 1985). Heyde and Gay (1993) extended the results to random fields. Furthermore, Heyde and Dai (1996) considered a possible effect of nonstationarity by defining the model $Y_t = X_t + z_t$ where z_t is a deterministic function such that $\sum_{t=1}^{n} z_t = 0$. If X_t is weakly dependent, then this has no effect on the Whittle estimator. However, if X_t has long memory, then z_t must decay fast enough to avoid any effect.

The Whittle estimator is also very attractive from a computational point of view, in particular if one uses a further approximation of (5.40). Replacing the integral by a Riemann sum leads to the Whittle approximation

$$\mathscr{L}_{n,\mathrm{Whittle}}\big(\sigma_\varepsilon^2,\theta\big) \approx \log\sigma_\varepsilon^2 + \frac{2}{n}\sum_{j=1}^{N_n}\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j;\vartheta)} =: \mathscr{L}_{n,\mathrm{WR}}\big(\sigma_\varepsilon^2,\theta\big), \tag{5.42}$$

where $N_n = [(n-1)/2]$ and $\lambda_j = 2\pi j/n$ are Fourier frequencies. This approximation is computationally fast due to the Fast Fourier Transform (FFT) (Cooley and Tukey 1965). Furthermore, we note that for Fourier frequencies $\lambda_j = 2\pi j/n$ we have

$$\frac{1}{2\pi n}\left|\sum_{t=1}^{N_n}(X_t-\bar{X}_n)e^{it\lambda_j}\right|^2 = \frac{1}{2\pi n}\left|\sum_{t=1}^{N_n}X_t e^{it\lambda_j}\right|^2 = I_{n,X}(\lambda_j).$$

In other words, estimation of the mean does not affect the periodogram computed at Fourier frequencies. This has important implications on the numerical performance of the estimator, as we shall indicate below at the end of the section. Recalling (5.28), an alternative to approximation (5.42) is

$$\mathscr{L}_{n,\mathrm{Whittle}}\big(\sigma_\varepsilon^2,\theta\big) \approx \frac{2}{n}\sum_{j=1}^{N_n}\log f_X(\lambda_j;\vartheta) + \frac{2}{n}\sum_{j=1}^{N_n}\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j;\vartheta)}. \tag{5.43}$$

Note also that Nordman and Lahiri (2006) proposed general maximum likelihood estimation in the spectral domain of which Whittle estimation is a special case.

It is also worth mentioning that the asymptotic distribution of $\hat{\vartheta}_{\mathrm{Whittle}}$ (and $\hat{\vartheta}_{\mathrm{MLE}}$) is the same as it would be in the following idealized setting. Let $\zeta_j = I_{n,X}(\lambda_j)$ and assume that ζ_j $(j = 1,2,\ldots,N_n)$ are independent exponential variables with expected values $f_X(\lambda_j;\vartheta^0)$ $(j = 1,2,\ldots,N_n)$. Thus, for $z \geq 0$, ζ_j has the probability density function

$$p_j(z) = \alpha_j \exp(-\alpha_j z)$$

where $\alpha_j = \alpha_j(\vartheta) = 1/f_X(\lambda_j;\vartheta^0)$. Given the observations ζ_j $(j = 1,2,\ldots,N_n)$, we would like to estimate $\vartheta = (\sigma_\varepsilon^2,\theta)$ by the maximum likelihood method. The log-likelihood function is given by

$$\mathscr{L}_\zeta\big(\sigma_\varepsilon^2,\theta\big) = \sum_{j=1}^{N_n}\log\alpha_j - \sum_{j=1}^{N_n}\alpha_j\zeta_j$$

so that $\hat{\vartheta}$ is the solution of $\dot{\mathscr{L}}_\zeta(\hat{\vartheta}) = 0$ where

$$\begin{aligned}\dot{\mathscr{L}}_\zeta\left(\sigma_\varepsilon^2, \theta\right) &= \sum_{j=1}^{N_n} \dot{\alpha}_j \left(\frac{1}{\alpha_j} - \zeta_j\right) \\ &= -\sum_{j=1}^{N_n} \frac{\dot{f}_X(\lambda_j; \vartheta)}{f_X(\lambda_j; \vartheta)} \left\{1 - \frac{I_{n,X}(\lambda_j)}{f_X(\lambda; \vartheta)}\right\} \\ &= -\mathscr{L}_{n,\mathrm{WR}}\left(\sigma_\varepsilon^2, \theta\right).\end{aligned}$$

Thus, we obtain the Whittle estimator, and hence also the same asymptotic distribution. The usefulness of this insight is a purely practical one. For FEXP models, the idealized setting with the variables ζ_j can be understood as a generalized linear model with expected value $\mu_j = E(\zeta_j)$ and link function $\eta(\mu) = \log \mu$ (Beran 1993). Recall that FEXP models are defined by a spectral density of the form

$$\begin{aligned}f_X(\lambda) &= \frac{\sigma_\varepsilon^2}{2\pi} |\lambda|^{-2d} \exp\left(\sum_{j=2}^{p} \theta_j g_j(\lambda)\right) \\ &= \exp\left(\sum_{j=0}^{p} \vartheta_{j+1} g_j(\lambda)\right)\end{aligned}$$

with $g_0(\lambda) = 1$, $g_1(\lambda) = -2\log|\lambda|$, g_j $(j \geq 2)$ bounded, $\vartheta_1 = \log \sigma_\varepsilon^2 - \log 2\pi$, $\vartheta_2 = d$, etc. Then

$$\eta(\mu) = \sum_{j=0}^{p} \vartheta_{j+1} g_j(\lambda)$$

is linear in ϑ so that we have indeed a generalized linear model (see McCullagh and Nelder 1989 for an introduction to generalized linear models). Therefore, Whittle's estimator can be calculated using programs for generalized linear models (GLM). As a cautionary remark it should be said that the derivation of the asymptotic distribution of the Whittle estimator using the assumption of i.i.d. variables $\zeta_j / f_X(\lambda_j; \vartheta^0)$ would *not* be correct. The periodogram exhibits a different asymptotic behaviour at frequencies tending to zero (and the number of Fourier frequencies where this is the case tends to infinity). This has been discussed in Sect. 4.6. The conclusion that the nonstandard behaviour of the periodogram near the origin is asymptotically negligible had to be shown by detailed arguments. Thus, only *in retrospect* can we conclude that the GLM methodology with independent observations ζ_j can be used for the calculation of the Whittle estimator.

5.5.4 Some Technical Details for the Whittle Estimator

Here, we provide some technical details that were omitted in the proof of Theorem 5.3. Recall that $\hat{\theta}_{\text{Whittle}}$ is obtained by setting the quadratic form

$$Q_{n,\text{Whittle}}(\theta) = \sum_{t,s=1}^{n} \dot{w}_{t-s}(\theta) X_t X_s$$

with

$$\dot{w}_k = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} e^{ik\lambda} \frac{\partial}{\partial\theta} \frac{1}{h_X(\lambda;\theta)}\, d\lambda$$

equal to zero. As before we use the notation

$$\tilde{g}(\lambda;\theta) := \frac{1}{(2\pi)^2} \frac{\partial}{\partial\theta} \frac{1}{h_X(\lambda;\theta)} = -\frac{1}{(2\pi)^2} \frac{\dot{h}_X}{h_X^2}.$$

We evaluate first (5.36), i.e. we would like to show that the quadratic form is centred. Using summability of $\dot{w}_k \gamma_X(k)$, we have

$$\begin{aligned}
E\big[Q_{n,\text{Whittle}}(\theta)\big] &= \sum_{k=-(n-1)}^{n-1} \big(n - |k|\big) \dot{w}_k \gamma_X(k) \sim n \sum_{k=-\infty}^{\infty} \dot{w}_k \gamma_X(k) \\
&= n \int \left(\sum_{k=-\infty}^{\infty} \dot{w}_k e^{ik\lambda} \right) f_X(\lambda;\theta)\, d\lambda \\
&= 2\pi n \int \tilde{g}(\lambda;\theta) f_X(\lambda;\theta)\, d\lambda \\
&= -(2\pi)^{-1} \left(\frac{\sigma_\varepsilon^2}{2\pi} \right) n \int_{-\pi}^{\pi} \frac{\dot{h}_X(\lambda;\theta)}{h_X(\lambda;\theta)}\, d\lambda \\
&= -(2\pi)^{-1} \left(\frac{\sigma_\varepsilon^2}{2\pi} \right) n \frac{\partial}{\partial\theta} \int_{-\pi}^{\pi} \log h_X(\lambda;\theta)\, d\lambda = 0.
\end{aligned}$$

In the last equation, we assumed that the derivative can be interchanged with integration.

Next, we evaluate expressions for the asymptotic variance σ_Q^2:

$$\begin{aligned}
\sigma_Q^2 &= 16\pi^3 \int_{-\pi}^{\pi} \big[f_X(\lambda) \tilde{g}(\lambda) \big]^2 d\lambda = 16\pi^3 \int_{-\pi}^{\pi} \left[f_X(\lambda) \frac{1}{(2\pi)^2} \frac{\dot{h}_X}{h_X^2} \right]^2 d\lambda \\
&= \left(\frac{\sigma_\varepsilon^2}{2\pi} \right)^2 \cdot \frac{1}{\pi} \int_{-\pi}^{\pi} \left(\frac{\dot{h}_X}{h_X} \right)^2 d\lambda = \left(\frac{\sigma_\varepsilon^2}{2\pi} \right)^2 \cdot \frac{1}{\pi} \int_{-\pi}^{\pi} \left(\frac{\dot{f}_X}{f_X} \right)^2 d\lambda
\end{aligned}$$

$$= \left(\frac{\sigma_\varepsilon^2}{2\pi}\right)^2 \cdot \frac{1}{\pi} V.$$

To conclude, we need to obtain the asymptotic quantity M. Recall that M is the limiting expected value of

$$\dot{Q}_{n,\text{Whittle}}(\theta) = \sum_{t,s=1}^{n} \ddot{w}_{t-s}(\theta) X_t X_s.$$

We assume that also the second derivative and integration can be interchanged so that

$$0 = \frac{\partial^2}{\partial\theta^2} \int_{-\pi}^{\pi} \log h_X \, d\lambda = -\int_{-\pi}^{\pi} \left[\left(\frac{\dot{h}_X}{h_X}\right)^2 - \frac{\ddot{h}_X}{h_X} \right] d\lambda.$$

Then

$$\begin{aligned} M &= \sum_{k=-\infty}^{\infty} \ddot{w}_k \gamma_X(k) = \int_{-\pi}^{\pi} \left[\sum_{k=-\infty}^{\infty} e^{ik\lambda} \ddot{w}_k \right] f_X(\lambda) \, d\lambda \\ &= 2\pi \int \frac{\partial}{\partial\theta} \tilde{g}(\lambda;\theta) f_X(\lambda) \, d\lambda = \left(\frac{\sigma_\varepsilon^2}{2\pi}\right) \int_{-\pi}^{\pi} \frac{1}{2\pi} \left[2\frac{\dot{h}_X^2}{h_X^3} - \frac{\ddot{h}_X}{h_X^2} \right] h_X(\lambda) \, d\lambda \\ &= \frac{1}{2\pi} \left(\frac{\sigma_\varepsilon^2}{2\pi}\right) \int_{-\pi}^{\pi} \left(\frac{\dot{h}_X}{h_X}\right)^2 d\lambda = \left(\frac{\sigma_\varepsilon^2}{2\pi}\right) \cdot \frac{1}{2\pi} V. \end{aligned}$$

5.5.5 *Further Approximation Methods for the MLE*

Sometimes it is of interest to obtain estimates (say $\hat{\vartheta}^{(j)}$) of ϑ for disjoint blocks of observations, $X_{1+(j-1)l}, \ldots, X_{n+(j-1)l}$ ($j = 1, \ldots, m$; $m = [n/l]$). For instance, if one has to handle very long time series, computations based on smaller blocks can be faster. In the context of change point detection, this is useful for detecting possible changes in the parameter (see Sect. 7.9). If $l \to \infty$, $l/n \to p \in (0, 1]$, and X_t is a stationary linear process, then it can be shown that the averaged value

$$\bar{\vartheta} = m^{-1} \sum_{j=1}^{m} \hat{\vartheta}^{(j)}$$

has the same asymptotic distribution as the MLE based on the whole series (Beran and Terrin 1994).

Apart from different versions of the Whittle estimator, there is another useful approximation that avoids inversion of the covariance matrix and is computationally fast. The idea is to use the infinite autoregressive representation of invertible linear processes. Suppose that $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ is a second-order linear process with spectral density $f_X(\lambda) \sim c_f |\lambda|^{-2d}$ for some $-\frac{1}{2} < d < \frac{1}{2}$. As before, we assume

that f_X is characterized by a parameter vector $\vartheta = (\sigma_\varepsilon^2, \theta)$ with $\theta = (d, \ldots)$. If X_t is invertible, then ε_t can be represented by past values of the process as

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \tag{5.44}$$

with $\pi_0 = 1$ and π_j obtained from

$$\sum_{k=0}^{\infty} \pi_j z^j = \frac{1}{A(z)} = \left(\sum_{j=0}^{\infty} a_j z^j \right)^{-1}$$

($|z| \le 1$, $z \neq 1$). In other words, we have an autoregressive representation

$$X_t = \sum_{j=1}^{\infty} b_j X_{t-j} + \varepsilon_t \tag{5.45}$$

with $b_j = -\pi_j$ ($j \ge 1$). In the case of a Gaussian process, the log-likelihood of $X_1, \ldots, X_n$ may be expressed in terms of $\log \sigma_\varepsilon^2$ and a sum of ε_{t-j}^2. This leads to the idea of estimating θ^0 by minimizing the residual sum of squares

$$S(\theta) = \sum \varepsilon_t^2(\theta)$$

with respect to θ and then setting

$$\hat{\sigma}_\varepsilon^2 = n^{-1} S(\hat{\theta}) = n^{-1} \sum \varepsilon_t^2(\hat{\theta}).$$

Taking derivatives with respect to θ, we obtain $\hat{\theta} = \hat{\theta}_{\mathrm{AR}}$ as the solution of

$$\dot{S}(\hat{\theta}_{\mathrm{AR}}) = \sum \dot{\varepsilon}_t(\hat{\theta}_{\mathrm{AR}}) \varepsilon_t(\hat{\theta}_{\mathrm{AR}}) = 0.$$

One difficulty that has to be addressed is that (5.44) includes the infinite past X_t ($t \le n$), whereas only a finite number of observations X_t ($1 \le t \le n$) are available. The simplest solution is truncation, which amounts to setting all unobserved values equal to zero. Thus, for $t = 2, \ldots, n$ one defines

$$e_t(\theta) = \sum_{j=0}^{t-1} \pi_j(\theta) X_{t-j}$$

to obtain $\hat{\theta}_{\mathrm{AR}}$ as the solution of

$$\sum_{j=2}^{n} \dot{e}_t(\hat{\theta}_{\mathrm{AR}}) e_t(\hat{\theta}_{\mathrm{AR}}) = 0 \tag{5.46}$$

(Beran 1995). In the previous terminology, this means that we are maximizing an approximate (Gaussian or quasi) log-likelihood function

$$\mathscr{L}_{\mathrm{AR}} = -\frac{n}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{t=2}^{n} e_t^2(\theta). \tag{5.47}$$

Under suitable regularity conditions, the asymptotic distribution of $\hat{\theta}_{\mathrm{AR}}$ and $\hat{\sigma}_\varepsilon^2$ turns out to be the same as for the other approximate methods considered above (Beran 1995), i.e.

$$\sqrt{n}\big(\hat{\vartheta}_{\mathrm{AR}} - \vartheta^0\big) \underset{d}{\to} N(0, \Sigma_{\mathrm{MLE}}).$$

The essential reason is that $\dot{\varepsilon}_t(\theta^0)\varepsilon_t(\theta^0)$ is a martingale difference so that a central limit theorem applies.

An advantage of the autoregressive approach is that it can be generalized to integrated processes. Suppose that we observe an integrated process Y_t as follows. Let $m \in \{0, 1, 2, \ldots\}$ be the smallest integer such that the mth difference, $(1-B)^m Y_t$, is stationary, and such that

$$(1-B)^m Y_t = X_t$$

with X_t as before, characterized by $\vartheta = (\sigma_\varepsilon^2, \theta)$. If m were known, then all values of X_t $(2 \le t \le n)$ could be recovered by differencing. It is therefore possible to express e_t defined above as a linear combination of $Y_1, \ldots, Y_n$. In other words, we can define e_t as a function of m and ϑ and minimize

$$S(m, \vartheta) = \sum e_t^2(m, \vartheta)$$

with respect to m and ϑ. Under suitable regularity conditions, we have, as $n \to \infty$, $P(\hat{m} = m^0) \to 1$ and $\hat{\vartheta}$ has the same asymptotic distribution as in the case where m is known (for heuristic arguments see Beran 1995, also see Velasco and Robinson 2000 for similar results in the semiparametric context, Robinson 1994b for nonstationarity tests, and Ling and Li 1997 for an extension to GARCH innovations).

Example 5.8 Consider an ARFIMA$(0, d, 0)$ process X_t with $d = d^0 \in (-\frac{1}{2}, \frac{1}{2})$. Recall that

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

with

$$\pi_j = \pi_j(d) = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \sim \frac{1}{\Gamma(-d)} j^{-(d+1)}.$$

Then $\hat{d}_{\mathrm{AR}}$ is defined by minimizing $S(d) = \sum e_t^2(d)$ with

$$e_t(d) = X_t + \frac{\Gamma(1-d)}{\Gamma(2)\Gamma(-d)} X_{t-1} + \cdots + \frac{\Gamma(t-1-d)}{\Gamma(t)\Gamma(-d)} X_1.$$

Example 5.9 Let X_t be defined as in the previous example, but we observe instead an integrated process

$$Y_1 = X_1, \qquad Y_2 = X_1 + X_2, \qquad \ldots, \qquad Y_n = X_1 + \cdots + X_n.$$

Then, for $2 \le t \le n$, we have

$$X_t = Y_t - Y_{t-1}.$$

Since the true integer differencing order $m^0 = 1$ is unknown, we need to minimize $S(m, \vartheta) = \sum e_t^2(m, \vartheta)$ with respect to ϑ for a set of integer values $0 \le m \le m_{\max}$. Usually one does not go beyond $m_{\max} = 2$. Once we have minimal values $S(0, \hat{\vartheta}_{\mathrm{AR},1})$, $S(1, \hat{\vartheta}_{\mathrm{AR},2}), \ldots, S(m_{\max}, \hat{\vartheta}_{\mathrm{AR},m_{\max}})$ (with $\hat{\vartheta}_{\mathrm{AR,i}}$ denoting the corresponding estimates of ϑ), we set

$$\hat{m} = \arg\min_m S(m, \hat{\vartheta}_{\mathrm{AR},m}).$$

The estimate of $\vartheta = (\sigma_\varepsilon^2, d)$ is then equal to

$$\hat{\vartheta}_{\mathrm{AR}} = \big(\hat{\sigma}_\varepsilon^2, \hat{d}_{\mathrm{AR}}\big) := \hat{\vartheta}_{\mathrm{AR},\hat{m}}.$$

A $(1-\alpha)$-level confidence interval for $d \in (-\frac{1}{2}, \frac{1}{2})$ is of the form

$$\hat{d}_{\mathrm{AR}} \pm z_{1-\frac{\alpha}{2}} \sqrt{[\Sigma_{\mathrm{MLE}}]_{22}} n^{-\frac{1}{2}} = \hat{d}_{\mathrm{AR}} \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{6}}{\pi} n^{-\frac{1}{2}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. At the same time, one can also provide a $(1-\alpha)$-level confidence interval for the total differencing parameter $d_{\mathrm{total}} = m + d$ by

$$\hat{m} + \hat{d}_{\mathrm{AR}} \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{6}}{\pi} n^{-\frac{1}{2}}.$$

For instance, $S(0, \vartheta) = \sum e_t^2(0, \vartheta)$ with

$$\begin{aligned} e_t(0, d) &= e_t(d) \\ &= X_t + \frac{\Gamma(1-d)}{\Gamma(2)\Gamma(-d)} X_{t-1} + \cdots + \frac{\Gamma(t-1-d)}{\Gamma(t)\Gamma(-d)} X_1 \end{aligned}$$

and $S(1, \vartheta) = \sum e_t^2(1, \vartheta)$ with

$$\begin{aligned} e_t(1, d) &= \sum_{j=0}^{t-1} \pi_j (Y_{t-j} - Y_{t-j-1}) \\ &= (Y_t - Y_{t-1}) + \frac{\Gamma(1-d)}{\Gamma(2)\Gamma(-d)} (Y_{t-1} - Y_{t-2}) + \cdots \end{aligned}$$

$$+\frac{\Gamma(t-1-d)}{\Gamma(t)\Gamma(-d)}(Y_1 - Y_0).$$

Finally, note that a problem not considered here is maximum likelihood estimation of ϑ for time series with missing values. An approach that has been used successfully for short-memory time series is based on the state space representation. For Markov processes this essentially amounts to representing X_t in vector form as a multivariate AR(1) model. Long-memory processes do not have the Markov property so that an exact state space representation is not possible. Nevertheless, Chan and Palma (1998) could show that, under suitable regularity conditions, an approximate state space representation with an asymptotically increasing dimension can be obtained and used to define an approximate MLE.

5.5.6 Model Choice

The asymptotic results obtained above are valid under the assumption that the assumed parametric model is correct. If the model is misspecified, then this means that the true spectral density f_X is not in the parametric class of spectral densities specified a priori. The estimates therefore converge to a value of ϑ^* that minimizes the discrepancy between the true spectral density f_X and a member of the parametric family. The convergence to the "pseudo value" ϑ^* may be slower than $O_p(n^{-\frac{1}{2}})$ (see, e.g. Yajima 1993; Chen and Deo 2006). In order to be applicable in practice, parametric models have to be combined with an appropriate model choice criterion. The best known method is Akaike's information criterion (AIC; see Akaike 1973, 1974; Shibata 1976, 1980) and related methods such as the BIC and HIC (Schwarz 1978; Hannan and Quinn 1979). The reason for the success of the AIC is its simplicity and its fundamental justification in terms of information theory. The foundation in information theory makes the AIC potentially applicable to most situations encountered in statistics. However, the exact form of the AIC relies on approximations that may depend on the particular setup.

Suppose, for instance, that X_t is generated by a FARIMA$(p_0, d, 0)$ process with unknown order $p_0 \geq 0$ and parameter vector $\vartheta^0 = (\sigma^2_{\varepsilon,0}, d^0, \varphi^0_1, \ldots, \varphi^0_p)$ (with $\vartheta^0 = (\sigma^2_{\varepsilon,0}, d^0)$ for $p^0 = 0$). Following the same line of thought as in Akaike (1973) (also see Bhansali 1986 and references therein), we then consider a process Y_t, independent of X_t, but being generated by the same model as X_t. Denoting by $\hat{\vartheta}_{\mathrm{MLE}}(X, p)$ the MLE of $\vartheta(p)$ (based on the observations $X(n) = (X_1, \ldots, X_n)$) for a FARIMA$(p, d, 0)$ process with $p \geq p_0$, the quality of the fit using this (possibly overparametrized) model is measured by the loss function

$$L\big(p; \hat{\vartheta}_{\mathrm{MLE}}\big(X(n), p\big)\big) = -2E_{y,\vartheta^0}\big[\mathscr{L}\big(Y(n), \hat{\vartheta}_{\mathrm{MLE}}(X, p)\big)\big]$$

where $Y(n) = (Y_1, \ldots, Y_n)$, $\mathscr{L}(Y(n), \hat{\vartheta}_{\mathrm{MLE}}(X, p))$ is the log-likelihood function evaluated at $Y(n)$ and the (now fixed) parameter $\hat{\vartheta}_{\mathrm{MLE}}(X, p)$, and the expected value

$E_{y,\vartheta^0}[\cdot]$ is taken with respect to the distribution of $Y(n)$ (which is specified by the correct parameter ϑ^0). The corresponding risk function of the MLE based on the order p is then

$$R(p) = E_{x,\theta_0}\big[L\big(p; \hat{\vartheta}_{\mathrm{MLE}}\big(X(n), p\big)\big)\big]$$

where the expectation $E_{x,\vartheta^0}[\cdot]$ is taken with respect to the distribution of $X(n)$. If instead of the exact MLE, an approximate MLE is used, then $\mathscr{L}$ in the definition of L is replaced by the corresponding approximate log-likelihood function. For instance, for the autoregressive approach above (Beran 1995), $\mathscr{L}$ is replaced by $\mathscr{L}_{\mathrm{AR}}$ in (5.47). Note that for $Y_1, \ldots, Y_n$,

$$\begin{aligned} e_{t,Y}(\theta) &:= \sum_{j=0}^{t-1} \pi_j(\theta) Y_{t-j} \\ &= \int_{-\pi}^{\pi} \sum_{j=0}^{t-1} \pi_j e^{i(t-j)\lambda}\, dM(\lambda) \\ &= \int_{-\pi}^{\pi} e^{it\lambda} \Pi_t\big(e^{-i\lambda}\big)\, dM(\lambda) \end{aligned}$$

where

$$Y_t = \int_{-\pi}^{\pi} e^{it\lambda}\, dM(\lambda) \underset{d}{=} X_t$$

is the spectral representation of Y_t (and hence also of X_t) and

$$\Pi_t\big(e^{-i\lambda}, \theta\big) = \sum_{j=0}^{t-1} \pi_j(\theta) e^{-ij\lambda}.$$

Since $E_{y,\vartheta^0}[dM(\lambda)\overline{dM(\nu)}] = 0$ $(\lambda \neq \nu)$ and $E_{y,\vartheta^0}[dM(\lambda)\overline{dM(\lambda)}] = f(\lambda)$, we then have

$$E_{y,\vartheta^0}\big[e_{t,Y}^2(\theta)\big] = \int_{-\pi}^{\pi} \big|\Pi_t\big(e^{-i\lambda}\big)\big|^2 f(\lambda)\, d\lambda.$$

Therefore, for a fixed $\hat{\vartheta}_{\mathrm{AR}}$,

$$\begin{aligned} L\big(p; \hat{\vartheta}_{\mathrm{AR}}\big(X(n), p\big)\big) &= -2E_{y,\vartheta^0}\big[\mathscr{L}_{\mathrm{AR}}\big(Y(n); \hat{\vartheta}\big)\big] \\ &= n \log \hat{\sigma}_{\varepsilon}^2 + \hat{\sigma}_{\varepsilon,\mathrm{AR}}^{-2} \sum_{t=2}^{n} \int_{-\pi}^{\pi} \big|\Pi_t\big(e^{-i\lambda}, \hat{\theta}_{\mathrm{AR}}\big)\big|^2 f(\lambda)\, d\lambda \end{aligned}$$

and

$$R(p) = nE_{x,\theta^0}\big(\log \hat{\sigma}_{\varepsilon,\mathrm{AR}}^2\big) + \sum_{t=2}^{n} E_{x,\theta^0}\left[\hat{\sigma}_{\varepsilon,\mathrm{AR}}^{-2} \int_{-\pi}^{\pi} \big|\Pi_t\big(e^{-i\lambda}, \hat{\theta}_{\mathrm{AR}}\big)\big|^2 f(\lambda)\, d\lambda\right].$$

(Note that there is a typo in Beran et al. 1998, equation (12), since the sum over t is missing.) The main difficulty at this stage is to show that $|\Pi_t(e^{-i\lambda}, \hat{\theta}_{\mathrm{AR}})|^2$ may be replaced by $|\Pi_\infty(e^{-i\lambda}, \hat{\theta}_{\mathrm{AR}})|^2$. This does not follow from corresponding results for short-memory processes because here the decay of the coefficients π_j is hyperbolic instead of exponential. The second approximation to be established is

$$E_{x,\theta^0}\big(\log \hat{\sigma}^2_{\varepsilon,\mathrm{AR}}\big) = \log \sigma^2_\varepsilon - \frac{p+2}{n} + o\big(n^{-1}\big).$$

Once these two facts are established (see Beran et al. 1998 for details), one can obtain—up to a constant that does not depend on p—an asymptotically unbiased estimator of $R(p)$ in quite the same way as for short-memory processes, namely

$$\hat{R}(p) = n \log \hat{\sigma}^2_{\varepsilon,\mathrm{AR}} + 2(p+2) + C\big(n, \vartheta^0\big) + o(1)$$

with C being fixed for a given sample size n and true parameter ϑ^0. Thus, comparing models with different orders p can be done in a first approximation by comparing the values of

$$AIC(p) = n \log \hat{\sigma}^2_{\varepsilon,\mathrm{AR}} + 2(p+2).$$

The reason is that

$$E_{x,\vartheta^0}\big[AIC(p_1) - AIC(p_2)\big] = R(p_1) - R(p_2) + o(1).$$

The order p_0 is therefore estimated by

$$\hat{p} = \hat{p}_{\mathrm{AIC}} = \arg \min_{0 \le p \le p_{\max}} AIC(p)$$

where $p_{\max}$ is a certain maximal order up to which one is willing to compare models. It can be shown that asymptotically, $\hat{p}_{\mathrm{AIC}}$ does not underestimate p_0, i.e.

$$P(\hat{p}_{\mathrm{AIC}} < p_0) \to 0.$$

On the other hand, as for short-memory series, also in the more general setting with $d \in (-\frac{1}{2}, \frac{1}{2})$, the probability of overestimation is not zero, i.e.

$$P(\hat{p}_{\mathrm{AIC}} > p_0) \to c \in (0, 1).$$

Consistency of $\hat{p}$ can be established by using a stronger penalty for the number of parameters. For instance, minimizing the BIC criterion

$$BIC(p) = n \log \hat{\sigma}^2_{\varepsilon,\mathrm{AR}} + 2 \log n \cdot (p+2)$$

or the HIC defined by

$$HIC(p) = n \log \hat{\sigma}^2_{\varepsilon,\mathrm{AR}} + 2c \log \log n \cdot (p+2)$$

with $c > 1$, one obtains

$$P(\hat{p} = p_0) \to 1$$

(Beran et al. 1998, Theorem 2). This is analogous to the case of short-memory series where one assumes d to be equal to zero a priori (see, e.g. Shibata 1976; Schwarz 1978; Hannan and Quinn 1979). Moreover, the results also hold in the more general case where we may observe an integrated fractional process Y_t with $(1-B)^m Y_t = X_t$ and m unknown (see the discussion above, and Beran et al. 1998). For other results on model choice for fractional processes, see, e.g. Crato and Ray (1996) and Baillie et al. (2012).

It should be noted that the consistency results for $\hat{p}$ make sense only if there is actually a finite p_0 such that the true spectral density falls into the corresponding parametric family. In a more general setting, this may not be the case. In principle, one may therefore increase $p_{\max}$ with increasing sample size. This implies, however, that in the general case where $p_0 = \infty$ the convergence results in Theorem 5.2, and in particular $\sqrt{n}$-convergence, no longer apply. One may conjecture, however, that the rate $O_p(n^{-\frac{1}{2}})$ may at most deteriorate by a logarithmic factor. This conjecture is supported by results for broadband estimators discussed below (see Sect. 5.9). For instance, Bhansali et al. (2006) consider the model

$$(1-B)^d X_t = Y_t,$$

where Y_t is a weakly dependent infinite order moving average defined by

$$\sum_{j=0}^{\infty} a_{\text{short},j} Y_{t-j} = \varepsilon_t,$$

with $\sum_j |a_{\text{short},j}| < \infty$ and ε_t i.i.d. centred random variables with a finite fourth moment. The spectral density is then

$$f_X(\lambda) = \frac{\sigma_\xi^2}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d} \left|A_{\text{short}}\left(e^{-i\lambda}\right)\right|^{-2},$$

where $A_{\text{short}}(z) = \sum_{j=0}^{\infty} a_{\text{short},j} z^j$. The estimator is obtained by minimizing the objective function

$$\int_{-\pi}^{\pi} \frac{I_{n,X}(\lambda)}{f_{X,p}(\lambda;\theta)}\, d\lambda,$$

where

$$f_{X,p}(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d} \left|A_p(\lambda)\right|^{-2} = \frac{\sigma_\varepsilon^2}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d} \left|\sum_{j=0}^{p} a_{\text{short},j} e^{i\lambda j}\right|^{-2}.$$

In other words, the approximation of the likelihood function is in the spectral domain rather than the time domain, similar to Beran (1993). The authors show that

$\sqrt{n/p}(\hat{d} - d)$ converges to a normal random variable if $p = p_n$ tends to infinity at an appropriate rate. For example, one can choose $p_n = \log n$ so that there is very little loss in terms of the convergence rate, as compared to the MLE. Related results can also be found in Poskitt (2007a). Other broadband methods (see Sect. 5.9) are based on FEXP models (Beran 1993; Robinson 1994a; Moulines and Soulier 1999, 2000; Hurvich 2001; Hurvich and Brodsky 2001; Hurvich et al. 2002; Narukawa and Matsuda 2011). The essential feature of broadband methods is that consistency and the rate $O_p(\sqrt{n^{-1}\log n})$ can be achieved under rather general conditions, just by letting the number of parameters tend to infinity in a suitable way. This is where the parametric and the semiparametric worlds meet. In the end, both methodologies may lead to similar or even identical results, either formally or when in the hands of experienced data analysts.

5.5.7 *Comments on Finite Sample Properties and Further Extensions*

Sowell (1992) studied numerical properties of the MLE and Whittle estimator in the case of a known mean. Although both estimators have the same limiting variance, the MLE appears to have better finite sample properties (also see, e.g. Hauser 1999 for an extended simulation study). However, as indicated by Cheung and Diebold (1994), this effect disappears when $\mu = E(X_t)$ is replaced by an estimator. The mean squared error of the MLE is increased due to a bias induced when replacing μ by $\hat{\mu}$. In contrast, the Whittle estimator based on the Riemann approximation (5.42) does not depend on the location parameter. Also, not quite surprisingly (see, e.g. Huber 1981; Hampel et al. 1986), the MLE performs poorly when data are contaminated (Haldrup and Nielsen 2007). For robust versions of the MLE, see, e.g. Beran (1994a, 1994b). Also note that theoretical results on finite sample properties and the possibility of applying bootstrap methods for inference in the context of MLE and Whittle estimation can be found in Lieberman et al. (2000, 2003) and Andrews et al. (2006). The bias and methods for bias reduction of the MLE and approximate MLE is considered, for instance, in Cheung and Diebold (1994), Smith et al. (1997), Hauser (1999), Lieberman (2001), and Doornik and Ooms (2003). Extensions to spatial processes are discussed, for instance, in Heyde and Gay (1989), Boissy et al. (2005), Leonenko and Sakhno (2006), Beran et al. (2009) (also see Angulo et al. 2000; Guo et al. 2009). For space-time processes, see Haslett and Raftery (1989). For multivariate extensions, see Luceno (1996), Morana (2007), Tsay (2000).

In summary, the MLE and the different approximations discussed above have the same asymptotic distribution given in Theorem 5.2. From the numerical point of view, the approximate Whittle and the AR-based estimator are more tractable. Moreover, the Whittle estimator based on Fourier frequencies is free of the effect of centring.

5.6 Semiparametric Narrowband Methods in the Fourier Domain

5.6.1 Introduction

In this section, we describe semiparametric methods for estimating d in the spectral domain. We start with narrowband methods which include log-periodogram regression (5.6.2) and local Whittle estimation (5.6.3).

The semiparametric approach to long-memory estimation was initiated in Geweke and Porter-Hudak (1983), the first mathematical results are due to Robinson (1995a, 1995b). The two papers by Robinson form the basis for many subsequent theoretical studies. The first estimator was suggested by Geweke and Porter-Hudak (1983) and is therefore known as the GPH estimator. Its asymptotic distribution was established in Robinson (1995a), detailed studies of the mean squared error and the question of bandwidth choice can be found in Hurvich and Beltrao (1994a, 1994b), Hurvich et al. (1998), Hurvich and Deo (1999), Andrews and Sun (2004), Andrews and Guggenberger (2003), Robinson and Henry (2003), among others. The local Whittle estimator was suggested by Künsch (1987), its asymptotic normality was derived by Robinson (1995b). Since then, there has been an enormous number of papers dealing with various aspects of narrowband estimation—to name a few (in alphabetical order): Abadir et al. (2007), Andrews and Sun (2004), Arteche (2004, 2006), Chen and Hurvich (2003a, 2003b), Christensen and Nielsen (2006), Frederiksen et al. (2012), Hassler et al. (2006), Henry (2007), Henry and Robinson (1996), Hurvich et al. (2005a), Hurvich and Ray (1995), Lobato (1995, 1997, 1999), Lobato and Robinson (1996), Marinucci (2000), Nielsen (2004a), Nielsen and Frederiksen (2011), Phillips and Shimotsu (2004), Phillips (2007), Poskitt (2007a, 2007b), Robinson (1994c, 1995a, 1995b, 2005), Robinson and Marinucci (2001, 2003), Robinson and Yajima (2002), Shimotsu (2012), Shimotsu and Phillips (2006), Souza (2007), Velasco (1999a, 1999b, 2000).

Both estimators (local periodogram regression, local Whittle) can be considered in terms of sums of weighted periodogram ordinates. This approach appeared in a sequence of papers, including Moulines and Soulier (1999, 2003), Faÿ and Soulier (2001), Hurvich et al. (2002, 2005a). We present this approach as well as technical details for asymptotic normality and asymptotic expressions for the mean squared error in Sect. 5.6.4.

An important question for semiparametric estimators is the optimal rate of convergence. Such results are established in Giraitis et al. (1997) and Soulier (2010). This is discussed in Sect. 5.8. In contrast to a fully parametric setting, semiparametric estimation suffers from a potentially serious bias and loss of efficiency. To overcome such problems, various bias and variance reduction techniques are proposed in the literature, including tapering and pooling (Hannan 1970; Hurvich and Beltrao 1993; Hurvich and Chen 2000; Robinson 1995a; Hurvich et al. 2002). These methods extend also the applicability of semiparametric estimators to nonstationary models (Hurvich and Ray 1995; Kim and Phillips 1999; Velasco 1999a, 1999b; Hurvich et al. 2005a; Arteche and Velasco 2005; Hidalgo 2005). This is presented in Sect. 5.6.5.

5.6.2 Log-periodogram Regression—Narrowband LSE

Suppose that X_t is a stationary process with spectral density

$$f_X(\lambda) = \left|1 - \exp(-i\lambda)\right|^{-2d} f_*(\lambda) \sim |\lambda|^{-2d} f_*(\lambda) \sim c_f |\lambda|^{-2d}, \tag{5.48}$$

as $\lambda \to 0$ where $-\frac{1}{2} < d < \frac{1}{2}$ and f_* is a function such that $f_*(0) = c_f \neq 0$. Recall that the empirical analog to the spectral density is the periodogram

$$I_{n,X}(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t e^{-it\lambda}\right|^2,$$

and, for Fourier frequencies $\lambda_j = 2\pi j/n$ ($j = 1, \ldots, N_n = [(n-1)/2]$), $I_{n,X}(\lambda)$ is not affected by centring of the time series. From Sect. 4.6 we recall that for a weakly dependent time series the following asymptotics holds:

$$\left(\frac{I_{n,X}(\lambda_{j_1})}{f_X(\lambda_{j_1})}, \ldots, \frac{I_{n,X}(\lambda_{j_k})}{f_X(\lambda_{j_k})}\right) \underset{d}{\to} (Z_1, \ldots, Z_k) \tag{5.49}$$

where $Z_1, \ldots, Z_k$ are i.i.d. standard exponential random variables and $\lambda_{j_1}, \ldots, \lambda_{j_k}$ are distinct Fourier frequencies. This property facilitates the derivation of asymptotic results for functionals of the periodogram $I_{n,X}(\lambda)$ ($\lambda \in [-\pi, \pi]$), provided that the functional can be approximated by using Fourier frequencies only. Together with (5.48) this motivated the following definition of a least squares regression estimator of d by Geweke and Porter-Hudak (1983). First of all, since

$$\log f_X(\lambda) \sim \log c_f + d \cdot b(\lambda) \tag{5.50}$$

with $b(\lambda) = -2\log(\lambda)$, a natural idea is to estimate c_f and d by linear regression techniques. Replacing f_X by the periodogram and assuming that $\log I_{n,X}(\lambda)$ is approximately equal to $\log f_X(\lambda) + \log Z$, where Z is a standard exponential variable, we have

$$E\left[\log I_{n,X}(\lambda)\right] \approx \int_0^\infty e^{-x} \log x \, dx + \log f_X(\lambda) = -\eta + \log f_X(\lambda),$$

where $\eta \approx 0.57722$ is the Euler constant. Thus, combining this with (5.50) yields, for $\lambda \to 0$,

$$E\left[\log I_{n,X}(\lambda)\right] \approx \underbrace{(\log c_f - \eta)}_{\beta_0} + \underbrace{d}_{\beta_1} \cdot b(\lambda) = \beta_0 + \beta_1 b(\lambda)$$

and the corresponding regression equation is

$$\log I_{n,X}(\lambda) = \log f_X(\lambda) + \log Z = \beta_0 + \beta_1 b(\lambda) + e(\lambda) \tag{5.51}$$

with

$$e(\lambda) = \log Z + \eta.$$

Motivated by (5.51), Geweke and Porter-Hudak (1983) suggested the least squares estimator (also called the GPH-estimator)

$$\hat{d}_{\mathrm{GPH}} = \frac{\sum_{j=1}^{m}(b_j - \bar{b}) \log I_{n,X}(\lambda_j)}{\sum_{j=1}^{m}(b_j - \bar{b})^2} \tag{5.52}$$

where

$$b_j = -2\log(\lambda_j), \qquad \bar{b} = \frac{1}{m}\sum_{j=1}^{m} b_j$$

and

$$\lambda_j = \frac{2\pi j}{n} \quad (j = 1, 2, \ldots, m)$$

are the m smallest Fourier frequencies. The number m is called the *bandwidth parameter*. To obtain a consistent estimator that is not disturbed by deviations from (5.50) outside an infinitesimal neighbourhood of the origin, m is chosen such that $m/n \to 0$. At the same time, the variance needs to converge to zero so that we need $m \to \infty$. The balance between these two conditions calls for a compromise between bias and variance, comparable to other situations in nonparametric statistics. In general, this is a difficult empirical optimization problem.

Avoiding complications at the moment, simple heuristics can be used to get an idea about the possible asymptotic distribution of $\hat{d}_{\mathrm{GPH}}$. If we assume $Z_j = I_{n,X}(\lambda_j)/f_X(\lambda_j)$ $(j = 1, \ldots, m)$ to be exactly independent standard exponential random variables, then

$$\mathrm{var}(\hat{d}_{\mathrm{GPH}}) = \frac{\mathrm{var}(\log I_{n,X}(\lambda))}{\sum_{j=1}^{m}(b_j - \bar{b})^2} = \frac{\mathrm{var}(\log(Z))}{\sum_{j=1}^{m}(b_j - \bar{b})^2}$$

and $\mathrm{var}(\log Z) = \pi^2/6$ (cf. (4.151)). Moreover, straightforward calculation yields

$$\begin{aligned}\sum_{j=1}^{m}(b_j - \bar{b})^2 &= 4\left\{\sum_{j=1}^{m}(\log j)^2 - \frac{1}{m}\left(\sum_{j=1}^{m}\log j\right)^2\right\} \\ &\sim 4m\left\{\int_0^1 (\log\lambda)^2\, d\lambda - \left(\int_0^1 \log\lambda\, d\lambda\right)^2\right\} = 4m. \end{aligned} \tag{5.53}$$

Thus,

$$\mathrm{var}(\hat{d}_{\mathrm{GPH}}) \sim \frac{\pi^2}{24m}.$$

Consequently, if the bias (caused by treating Z_j as standard exponential random variables) is of smaller order than $m^{-\frac{1}{2}}$, then we may expect to obtain a central limit theorem of the form

$$\sqrt{m}(\hat{d}_{\mathrm{GPH}} - d) \underset{d}{\to} N\left(0, \frac{\pi^2}{24}\right). \tag{5.54}$$

Although (5.54) is generally correct, an exact mathematical derivation turned out to be much more difficult. The main reason is that under long memory, property (5.49) no longer holds for Fourier frequencies that are very close to the origin (see Sect. 4.6). Since the behaviour of f_X at the origin is exactly what we are interested in, this is rather troublesome. Moreover, in the derivation of (5.54) it has to be taken into account that it is not sufficient to derive a limit theorem for $I_{n,X}(\lambda)$ and the discrete Fourier transform $d_{n,X}(\lambda) = (2\pi n)^{-1/2} \sum_{t=1}^{n} X_t e^{-it\lambda}$ at a finite number of fixed non-zero frequencies and then "plug" this into a functional that depends on an increasing number of Fourier frequencies (this was done in the incorrect heuristic "derivation" above). By applying a refined analysis, it is, however, possible to salvage the result.

The first proof of (5.54) in a long-memory case is due to Robinson (1995a). He uses the following assumptions:

- (GPH1) X_t is a stationary Gaussian process.
- (GPH2)

  $$f_X(\lambda) = |\lambda|^{-2d} f_*(\lambda) = |\lambda|^{-2d}\big(f_*(0) + O\big(\lambda^{\rho}\big)\big)$$

  for some $0 < \rho \leq 2$ and $-\frac{1}{2} < d < \frac{1}{2}$. The parameter ρ controls the smoothness of f_X away from the origin and plays a crucial role when determining the optimal number of Fourier frequencies m. Furthermore, f_X is assumed to be differentiable at $\lambda \in (0, \varepsilon)$ with $\varepsilon > 0$ a suitable constant, and also

  $$f_X'(\lambda) = O\big(\lambda^{-2d-1}\big) \quad (\lambda \to 0).$$

- (GPH3)

  $$m \to \infty, \qquad m = o\big(n^{\frac{2\rho}{2\rho+1}}\big).$$

These assumptions respectively describe the type of model considered, smoothness of the spectral density $f_X(\lambda)$ (specifically of the short memory part f_*), and the choice of the number of the Fourier frequencies. In particular, the condition $m = o(n^{\frac{2\rho}{2\rho+1}})$ implies that the bias of the GPH estimator is negligible compared to the variance. Indeed, we will argue later (see, e.g. (5.69)) that for long-memory sequences

$$\begin{aligned} E\big[(\hat{d}_{\mathrm{GPH}} - d)^2\big] &= \mathrm{Bias}^2 + \mathrm{Variance} \\ &\approx \frac{\pi^2}{24m} + \mathrm{const} \cdot \frac{m^4}{n^4}. \end{aligned}$$

Therefore, for m proportional to $n^{\frac{2\rho}{2\rho+1}}$, the squared bias is of a smaller order than the variance as long as $\rho < 2$.

Robinson's (1995a) method of proof requires also trimming of some low frequencies to remove the effect of the asymptotic dependence of periodogram ordinates. Let $l > 0$ be an integer and define

$$\bar{b}(l) = \frac{1}{m-l+1}\sum_{j=l}^{m} b_j, \qquad \bar{y}(l) = \frac{1}{m-l+1}\sum_{j=l}^{m} \log I_{n,X}(\lambda_j)$$

and

$$\hat{d}(l) = \hat{d}_{\mathrm{GPH}}(l) = \frac{\sum_{j=l}^{m}(b_j - \bar{b}(l))\log I_{n,X}(\lambda_j)}{\sum_{j=l}^{m}(b_j - \bar{b}(l))^2}, \tag{5.55}$$

$$\hat{\beta}_0 = \bar{y}(l) - \hat{d}(l)\bar{b}(l). \tag{5.56}$$

Robinson's (1995a) result reads as follows.

Theorem 5.4 *Under assumptions* (GPH1)–(GPH3) *and*

$$\frac{\sqrt{m}\log m}{l} = o(1), \qquad l = o\left(\frac{m}{(\log n)^2}\right), \tag{5.57}$$

we have

$$\sqrt{m}\left(\frac{1}{2\log m}(\hat{\beta}_0 - \beta_0), \hat{d}_{\mathrm{GPH}}(l) - d\right) \underset{d}{\to} N(0, V), \tag{5.58}$$

where

$$V = \frac{\pi^2}{24}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

We note that the rate of convergence $\sqrt{m}$ is slower than $\sqrt{n}$. In fact, the best rate is achieved when $\rho = 2$ in assumption (GPH2). Then, assumption (GPH3) yields that m cannot grow faster than $m = n^{4/5}$ and hence the rate of convergence of $\hat{d}$ cannot be faster than $O_p(n^{-2/5})$.

A remarkable feature of (5.58) is that $\hat{\beta}_0$ and $\hat{d}_{\mathrm{GPH}}(l)$ are asymptotically perfectly negatively correlated. Thus, whenever $\hat{d}_{\mathrm{GPH}}(l)$ underestimates d, the intercept (and hence the scale parameter) is overestimated and vice versa.

Robinson's (1995a) ideas were further exploited leading to improvement of his results. In particular, the assumption of Gaussianity (condition (GPH1)) is not necessary, nor is it required that l tends to infinity. Moulines and Soulier (2003, Theorems 6.2 and 6.3) state an alternative condition under which asymptotic normality of $\hat{d}_{\mathrm{GPH}}$ holds. Their assumption (GPH1′) is more general in the sense that the process need not be Gaussian; though, on the other hand, it is more restrictive with respect to the shape of the spectral density:

- (GPH1′) $X_t = (1-B)^{-d}\varepsilon_t$ is a FARIMA(p,d,q) process with $d \in (-\frac{1}{2}, \frac{1}{2})$, and ε_t an i.i.d. sequence with finite fourth moment and such that $\int |E[e^{it\varepsilon_0}]|^r \, dt$ is finite for some $r \geq 1$.

5.6.3 Local Whittle Estimation—Narrowband Whittle Estimation

Let us recall the discrete Whittle approximation of the Gaussian likelihood (see Sect. 5.5, (5.43))

$$\mathscr{L}_{n,\text{Whittle}}\big(\sigma_\varepsilon^2, \theta\big) \approx \frac{2}{n} \sum_{j=1}^{N_n} \bigg(\log f_X(\lambda_j; \vartheta) + \frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j; \vartheta)} \bigg).$$

The idea of local Whittle estimation is to use the lowest m frequencies only (Künsch 1987). This leads to

$$\mathscr{L}_{m,\text{Whittle}}\big(\sigma_\varepsilon^2, \theta\big) = \frac{2}{m} \sum_{j=1}^{m} \bigg(\log f_X(\lambda_j; \theta) + \frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j; \theta)} \bigg),$$

where $\lambda_j = 2\pi j/n$, $j = 1, \ldots, m$. Assuming $f_X(\lambda) \sim c_f \lambda^{-2d}$ and $m/n \to 0$, minimization of $\mathscr{L}_{m,\text{Whittle}}$ can be replaced by the minimization of

$$Q(c_f, d) = \frac{1}{m} \sum_{j=1}^{m} \bigg[\log\big(c_f \lambda_j^{-2d}\big) + \frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d}} \bigg]. \tag{5.59}$$

The partial derivative with respect to c_f is

$$\frac{\partial}{\partial c_f} Q(c_f, d) = \frac{1}{m} \sum_{j=1}^{m} \bigg[c_f^{-1} - \frac{I_{n,X}(\lambda_j)}{c_f^2 \lambda_j^{-2d}} \bigg].$$

Thus, for any d, setting this expression to zero yields the explicit expression

$$\hat{c}_f = G_m(d) = \frac{1}{m} \sum_{j=1}^{m} \frac{I_{n,X}(\lambda_j)}{\lambda_j^{-2d}}. \tag{5.60}$$

Plugging $\hat{c}_f$ into (5.59) leads to

$$Q(\hat{c}_f, d) = \frac{1}{m} \sum_{j=1}^{m} \log \hat{c}_f \lambda_j^{-2d} + \frac{\hat{c}_f}{\hat{c}_f} = \log \hat{c}_f - d \frac{2}{m} \sum_{j=1}^{m} \log \lambda_j + 1.$$

Thus, given a permissible range $d \in \Theta \subseteq (-\frac{1}{2}, \frac{1}{2})$, the local Whittle estimator of d is defined by

$$\hat{d}_{\text{LW}} = \arg\min_{d \in \Theta} K_m(d) \tag{5.61}$$

where

$$K_m(d) = \log G_m(d) - d\left(\frac{2}{m}\sum_{j=1}^{m}\log\lambda_j\right). \tag{5.62}$$

This estimator is also called the Gaussian semiparametric estimator (GSE).

Robinson (1995b) derives the asymptotic distribution of $\hat{d}_{\mathrm{LW}}$ under assumptions that mimic those for the GPH estimator. In particular:

- (LW1) X_t is a second order stationary process with Wold representation $X_t = \mu_0 + \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$, $\sum a_j^2 < \infty$ and $E[\varepsilon^4] < \infty$.
- (LW2)

$$f_X(\lambda) = |\lambda|^{-2d} f_*(\lambda) = |\lambda|^{-2d}\big(f_*(0) + O\big(\lambda^{\rho}\big)\big) = |\lambda|^{-2d}\big(c_f + O\big(\lambda^{\rho}\big)\big)$$

  for some $0 < \rho \le 2$ and $-\frac{1}{2} < d < \frac{1}{2}$.
- (LW3)

$$m \to \infty, \qquad \frac{(\log m)^2 m^{1+2\rho}}{n^{2\rho}} \to 0.$$

These are sufficient assumptions for deriving the asymptotic distribution of $\hat{d}_{\mathrm{LW}}$. Note in particular that the innovations ε_t do not need to be independent. Thus, this includes also some nonlinear processes. Slightly weaker conditions can be used to prove weak consistency only. The asymptotic distribution is given as follows:

Theorem 5.5 *Under the assumptions* (LW1)–(LW3),

$$\sqrt{m}(\hat{d}_{\mathrm{LW}} - d_0) \underset{d}{\to} N\left(0, \frac{1}{4}\right).$$

As in the case of the log-periodogram estimator, we will give a sketch of the proof here, deferring details to Sect. 5.6.4. Unlike in the GPH case, it is more difficult to establish consistency of the local Whittle estimator because it is not given in an explicit form.

Proof The standard approach to obtain the asymptotic distribution of $\hat{d}_{\mathrm{LW}}$ is a Taylor expansion. Without loss of generality, we may assume $\sigma_\varepsilon^2 = 1$ (since it cancels out). Using the notation

$$\dot{K}(d) = \dot{K}_m(d) = \frac{\partial}{\partial d} K_m(d), \qquad \ddot{K}(d) = \ddot{K}_m(d) = \frac{\partial^2}{\partial d^2} K_m(d),$$

we may write heuristically

$$\begin{aligned} 0 &= \dot{K}(\hat{d}_{\mathrm{LW}}) = \dot{K}(d_0) + \ddot{K}(\tilde{d})(\hat{d}_{\mathrm{LW}} - d_0) \\ &= \dot{K}(d_0) + \big[\ddot{K}(d_0) + o_p(1)\big](\hat{d}_{\mathrm{LW}} - d_0) \end{aligned}$$

where $\tilde{d}$ is a suitable (random) intermediate value with $|\tilde{d} - d_0| \le |\hat{d}_{\mathrm{LW}} - d_0|$. Specifically,

$$\dot{K}(d_0) = \left(\frac{2}{m} \sum_{j=1}^{m} \log \lambda_j \cdot \frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}} \right) \left(\frac{1}{m} \sum_{j=1}^{m} \frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}} \right)^{-1} - \frac{2}{m} \sum_{j=1}^{m} \log \lambda_j .$$

Now, the procedure is very similar as in the case of the GPH estimator. We handle the denominator

$$\left(\frac{1}{m} \sum_{j=1}^{m} \frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}} \right)$$

by applying Bartlett's approximation

$$\frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}} \approx \frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} \approx 2\pi I_{n,\varepsilon}(\lambda_j),$$

where $I_{n,\varepsilon}(\lambda)$ is the periodogram of the innovation process ε_t. Heuristically, we then have

$$\frac{1}{m} \sum_{j=1}^{m} \frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}} \approx \frac{2\pi}{m} \sum_{j=1}^{m} I_{n,\varepsilon}(\lambda_j) \approx 2\pi E\big[I_{n,\varepsilon}(\lambda_j)\big] = \sigma_\varepsilon^2 = 1.$$

Therefore, the limiting behaviour of $\dot{K}(d_0)$ is the same as that of

$$\frac{2}{m} \sum_{j=1}^{m} \log \lambda_j \left[\frac{I_{n,X}(\lambda_j)}{c_f \lambda^{-2d_0}} - 1 \right].$$

Thus, using the notation $b_j = -2 \log \lambda_j$ and $\bar{b} = m^{-1} \sum_{j=1}^{m} b_j$, we conclude that the asymptotic behaviour of $\dot{K}(d_0)$ is the same as that of

$$\dot{K}_m^*(d_0) := -\frac{1}{m} \sum_{j=1}^{m} (b_j - \bar{b}) \left[\frac{I_{n,X}(\lambda_j)}{c_f \lambda^{-2d_0}} - 1 \right]. \tag{5.63}$$

Using again Bartlett's approximation, we deduce that $\dot{K}_m^*(d_0)$ has the same asymptotic behaviour as

$$\dot{K}_m^{**}(d_0) := \frac{1}{m} \sum_{j=1}^{m} (b_j - \bar{b}) \big[2\pi I_{n,\varepsilon}(\lambda_j) - 1 \big].$$

We will argue below (see Sect. 5.6.4) that $m \cdot \mathrm{var}(\dot{K}_m^{**}(d_0)) \to 4$ as $m \to \infty$ and also

$$\sqrt{m}\, \dot{K}_m^{**}(d_0) \underset{d}{\to} N(0, 4).$$

By similar arguments, one obtains

$$\ddot{K}_m(d_0) \xrightarrow[p]{} 4.$$

Thus

$$\sqrt{m}(\hat{d}_{\mathrm{LW}} - d_0) = -\frac{\sqrt{m}\dot{K}_m(d_0)}{\ddot{K}_m(d_0)} + o_p(1),$$

which leads to the asymptotic normality with variance $4/4^2 = \frac{1}{4}$. □

5.6.4 Technical Details for Semiparametric Estimators in the Fourier Domain

In this section, we provide some technical tools that are used to prove asymptotic normality of the GPH and local Whittle estimators in Theorems 5.4 and 5.5. We start with asymptotic normality of a weighted sum of periodogram ordinates for i.i.d. sequences. This is an important step in proving asymptotic normality under long memory.

We start with some formulas for the variance of $\hat{d}_{\mathrm{GPH}}$ that explain in particular the low trimming condition used by Robinson (1995a). Then, we will discuss the proof of asymptotic normality following steps suggested in Moulines and Soulier (1999, 2003), Hurvich et al. (2002), Lang and Soulier (2000) and Hurvich et al. (2005a).

5.6.4.1 Decomposition of the GPH Estimator

Recall that $b_j = -2\log\lambda_j$, $\lambda_j = 2\pi j/n$ and

$$\hat{d}_{\mathrm{GPH}} = \frac{\sum_{j=1}^m (b_j - \bar{b})\log I_{n,X}(\lambda_j)}{\sum_{j=1}^m (b_j - \bar{b})^2}.$$

Also, under the specification $f_X(\lambda) = |\lambda|^{-2d} f_*(\lambda)$ a straightforward computation yields

$$\frac{\sum_{j=1}^m (b_j - \bar{b})\log f_X(\lambda_j)}{\sum_{j=1}^m (b_j - \bar{b})^2} = d + \frac{\sum_{j=1}^m (b_j - \bar{b})\log f_*(\lambda_j)}{\sum_{j=1}^m (b_j - \bar{b})^2}.$$

Thus, we can decompose $\hat{d}_{\mathrm{GPH}} - d$ as

$$\hat{d}_{\mathrm{GPH}} - d = \sum_{j=1}^m b^*_{j,m}\left\{\log\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} + \eta\right\} + \sum_{j=1}^m b^*_{j,m}\log f_*(\lambda_j), \tag{5.64}$$

where

$$b^*_{j,m} = \frac{b_j - \bar{b}}{\sum_{k=1}^{m}(b_k - \bar{b})^2}$$

and η is the Euler constant. Note that the constant η can be added since $\sum_{j=1}^{m}(b_j - \bar{b}) = 0$. A similar decomposition is applied to $\sqrt{m}(\hat{d}_{\text{GPH}} - d)$ and the resulting stochastic term is written as the weighted sum

$$S_{m,X}(\log) := \sum_{j=1}^{m} b_{j,m}\left\{\log \frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} + \eta\right\}$$

where

$$b_{j,m} = \sqrt{m}b^*_{j,m} = \sqrt{m}\frac{b_j - \bar{b}}{\sum_{k=1}^{m}(b_k - \bar{b})^2}. \tag{5.65}$$

5.6.4.2 Decomposition of the Local Whittle Estimator

Also, in the case of the local Whittle estimator we assume $f_X(\lambda) \sim f_*(0)\lambda^{-2d}$. Defining $g_X(\lambda) = f_*(0)\lambda^{-2d}$, we have to study (cf. (5.63))

$$\sqrt{m}\dot{K}^*_m(d) = \sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,X}(\lambda_j)}{g_X(\lambda_j)} - 1\right]$$

where now

$$b_{j,m} = \frac{b_j - \bar{b}}{\sqrt{m}}. \tag{5.66}$$

Then $\sqrt{m}\dot{K}^*_m(d)$ is decomposed into a sum of two terms,

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} - 1\right] + \sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,X}(\lambda_j)}{g_X(\lambda_j)} - \frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)}\right]. \tag{5.67}$$

We denote this weighted sum of periodogram ordinates as

$$S_{m,X}(\text{linear}) = \sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} - 1\right].$$

5.6.4.3 Bias of the GPH Estimator

We apply the expected value to the decomposition (5.64) to obtain

$$\text{Bias}_{\text{GPH}} = E[\hat{d}_{\text{GPH}} - d] = \sum_{j=1}^{m} b_{j,m} E[Y_j] + \sum_{j=1}^{m} b^*_{j,m} \log f_*(\lambda_j),$$

where

$$Y_j = \log \frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} + \eta.$$

Let us note that when the normalized periodogram ordinates $I_{n,X}(\lambda_j)/f_X(\lambda_j)$ are exactly standard exponential (like in the case of i.i.d. random variables), then the first part disappears. In other words, in the long-memory case there are two sources of bias. The first, deterministic part, that arises from treating the spectral density $f_X(\lambda)$ as being *exactly* equal to $f_*(\lambda)|\lambda|^{-2d}$. The second, stochastic part, comes from treating the normalized periodogram ordinates as standard exponential random variables.

Let us deal with the deterministic part first (cf. Hurvich et al. 1998). Under the assumptions $f'_*(0) = 0$ and $f''_*(0) < \infty$, we have

$$\sum_{j=1}^{m} b^*_{j,m} \log f_*(\lambda_j) \approx \frac{2\pi^2}{9} \frac{f''_*(0)}{f_*(0)} \frac{m^2}{n^2}.$$

This can be obtained by expanding

$$\log f_*(\lambda_j) \approx \log f_*(0) + \frac{f''_*(0)}{f_*(0)} \frac{\lambda_j^2}{2},$$

which leads to

$$\sum_{j=1}^{m} b^*_{j,m} \log f_*(\lambda_j) \approx 2\pi^2 \frac{f''_*(0)}{f_*(0)} \frac{1}{n^2} \sum_{j=1}^{m} (b_j - \bar{b}) j^2,$$

and the behaviour of the deterministic part follows from a careful study of the sum in the latter expression. Furthermore, the stochastic term is negligible (see, e.g. Hurvich et al. 1998). Summarizing, under the assumption of the existence of the second order derivative of f_*, the bias of the GPH estimator is given by

$$\text{Bias}_{\text{GPH}} = \frac{2\pi^2}{9} \frac{f''_*(0)}{f_*(0)} \left(\frac{m}{n}\right)^2 + o\big((m/n)^2\big). \tag{5.68}$$

In general, if we assume (GPH3) only, the bias is of the order $(m/n)^\rho$.

5.6.4.4 Variance of GPH Estimator

In the decomposition (5.64) of the GPH estimator only the first part is stochastic. We compute

$$\operatorname{var}\left(\sum_{j=1}^{m} b_{j,m}^{*} \log I_{n,X}(\lambda_j)\right) = \sum_{j=1}^{m} \left(b_{j,m}^{*}\right)^2 \operatorname{var}\left(\log I_{n,X}(\lambda_j)\right) + \sum_{\substack{k,j=1 \\ k\neq j}}^{m} b_{k,m}^{*} b_{j,m}^{*} cov\left(\log I_{n,X}(\lambda_k), \log I_{n,X}(\lambda_j)\right).$$

Theorem 4.32 can be used to remove a sufficiently large number l of low frequencies in the GPH estimator so that the covariance between $Y_j = \log I_{n,X}(\lambda_j)$ at different frequencies (with $j \geq l+1$) is negligible and the second term in the expression for the variance disappears asymptotically (see condition (5.57)). Note, however, that a refined analysis shows that this trimming is actually not necessary asymptotically.

5.6.4.5 Mean Squared Error and Optimal Bandwidth for the GPH Estimator

Combining (5.68) with the computation of the variance above, we obtain

$$MSE(\hat{d}_{\mathrm{GPH}}) = E\left[(\hat{d}_{\mathrm{GPH}} - d)^2\right] \approx \left(\frac{2\pi^2}{9} \frac{f_*''(0)}{f_*(0)} \frac{m^2}{n^2}\right)^2 + \frac{\pi^2}{24m}. \tag{5.69}$$

Minimizing the MSE, we obtain the optimal value of m as

$$m_{\mathrm{opt}} = C n^{\frac{4}{5}}$$

with

$$C = \left(\frac{27}{129\pi^2}\right)^{\frac{1}{5}} \left(\frac{f_*(0)}{f_*''(0)}\right)^{\frac{2}{5}} \tag{5.70}$$

and the optimal rate of the *MSE*,

$$MSE_{\mathrm{opt}} = O\left(n^{-\frac{4}{5}}\right).$$

It is interesting to note that this rate is the same as encountered, for example, in nonparametric regression models with i.i.d. or weakly correlated residuals. The intuitive explanation is that most of the residuals in the GPH regression are approximately uncorrelated, and this turns out to be enough to obtain a result analogous to an i.i.d. situation. We note further that for the optimal choice of m the contribution of the bias is of the same order as the variance. On the other hand, the bias is negligible, if $m = o(n^{4/5})$, or more generally, if $m = o(n^{\frac{2\rho}{2\rho+1}})$, i.e. when (GPH3) and (LW3) hold.

The optimal bandwidth derived from formula (5.70) involves unknown quantities. Hurvich and Beltrao (1994a) therefore consider bandwidth choice by cross-validation in the spectral domain. Hurvich and Deo (1999) estimate the constant C and consider a plug-in bandwidth.

The local Whittle estimator is not given in an explicit form. Therefore, the evaluation of the variance and the bias is more complicated. Bounds for the second part in (5.67) are discussed in Robinson (1995b), Lang and Soulier (2000) and Hurvich et al. (2005a). Henry and Robinson (1996) consider plug-in bandwidth selection for the local Whittle estimator.

5.6.4.6 Asymptotic Normality

The procedure for establishing asymptotic normality of the GPH or the local Whittle estimator runs as follows:

- Use Bartlett's decomposition for linear processes:

$$I_{n,X}(\lambda) = 2\pi f_X(\lambda) I_{n,\varepsilon}(\lambda) + R_n(\lambda),$$

  where $R_n(\cdot)$ is a remainder.
- This suggests a decomposition of the stochastic terms $S_{m,X}(\log)$ and $S_{m,X}(\text{linear})$ respectively as:

$$\underbrace{\sum_{j=1}^{m} b_{j,m} \log\bigl(2\pi I_{n,\varepsilon}(\lambda_j)\bigr)}_{=:S_{m,\varepsilon}(\log)} + \underbrace{\sum_{j=1}^{m} b_{j,m} \log\left(\frac{I_{n,X}(\lambda_j)}{2\pi f_X(\lambda_j) I_{n,\varepsilon}(\lambda_j)}\right)}_{=R_{m,\varepsilon}(\log)}$$

  and

$$\underbrace{\sum_{j=1}^{m} b_{j,m}\bigl[2\pi I_{n,\varepsilon}(\lambda_j) - 1\bigr]}_{S_{m,\varepsilon}(\text{linear})} + \underbrace{\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,X}(\lambda_j)}{f_X(\lambda_j)} - 2\pi I_{n,\varepsilon}(\lambda_j)\right]}_{R_{m,\varepsilon}(\text{linear})},$$

  where $b_{j,m}$'s are given in (5.65) and (5.66) for the GPH and local Whittle estimator, respectively.
- Establish convergence of $S_{m,\varepsilon}(\log)$ and $S_{m,\varepsilon}(\text{linear})$, the sums of periodogram ordinates based on the i.i.d. sequence ε_t.
- Show that the remainder terms $R_{m,\varepsilon}(\log)$ and $R_{m,\varepsilon}(\text{linear})$ are negligible.
- The steps above establish convergence of $S_{m,X}(\log)$ and $S_{m,X}(\text{linear})$. Combine this with conditions on bias negligibility discussed above: $m = o(n^{\frac{2\rho}{2\rho+1}})$.

5.6.4.7 Periodogram for i.i.d. Sequences

Let ε_t be an i.i.d. sequence. Consider a weighted sum

$$S_{m,\varepsilon}(\phi) = \sum_{j=1}^{m} b_{j,m}\phi\big(2\pi I_{n,\varepsilon}(\lambda_j)\big).$$

Here, ϕ is a deterministic function, $b_{j,m}$ are deterministic constants and $m \to \infty$ as $n \to \infty$. Faÿ and Soulier (2001) (see also Theorem 9.4 in Moulines and Soulier 2003) give general conditions under which $S_{n,\varepsilon}(\phi)$ converges:

- (FS1) ε_t is an i.i.d. sequence with finite fourth moment and such that $\int |E[e^{it\varepsilon_0}]|^r\, dt$ is finite for some $r \geq 1$.
- (FS2) $\sum_{j=1}^{m} b_{j,m} = 0$ (so that centring in $S_{m,\varepsilon}(\phi)$ is not needed), $\sum_{j=1}^{m} b_{j,m}^2 = 1$, $\lim_{n\to\infty}\{\sum_{j=1}^{m} |b_{j,m} - b_{j+1,m}| + |b_{m,m}|\}\log^2(m) = 0$;
- (FS3) The Lindeberg condition

$$\lim_{n\to\infty} \max_{1\leq j\leq m} |b_{j,m}| = 0.$$

- (FS4)

$$\lim_{n\to\infty} \sum_{j=1}^{m} b_{j,m}^2 \operatorname{var}\big(\phi\big(2\pi I_{n,\varepsilon}(\lambda_j)\big)\big) = \sigma_0^2.$$

The results can be formulated as follows (see Theorem 9.4 in Moulines and Soulier 2003 or Faÿ and Soulier 2001).

Theorem 5.6 *Under the conditions* (FS1)–(FS4) *we have*

$$S_{m,\varepsilon}(\phi) \underset{d}{\to} N\big(0, \sigma_0^2\big).$$

We note in passing that if $\sum_{j=1}^{m} b_{j,m} \neq 0$, then the asymptotic variance has to be changed. It will involve κ_4, the fourth cumulant of ε_1; see Faÿ and Soulier (2001), Moulines and Soulier (2003) as well as Example 5.11 below.

We will shed some light on this theorem in the following examples that apply directly to the GPH and local Whittle estimator.

Example 5.10 Let $\phi(x) = \log(x)$. Set

$$b_{j,m} = \sqrt{m}\frac{b_j - \bar{b}}{\sum_{k=1}^{m}(b_k - \bar{b})^2}, \quad b_j = -2\log\lambda_j, \quad \lambda_j = \frac{2\pi j}{n}.$$

Note that (cf. (5.53)) $\sum_{j=1}^{m} b_{j,m}^2 \sim 1/4$ as $m \to \infty$. Now assume that ε_t are standard normal. Then the normalized periodogram ordinates $I_{n,\varepsilon}(\lambda_k)/f_\varepsilon(\lambda_k)$ are independent, with the same standard exponential distribution. Furthermore, $\operatorname{var}(\log(Z)) =$

$\pi^2/6$, where Z is standard exponential. Thus

$$\operatorname{var}\left(\sum_{j=1}^{m} b_{j,m}\phi\big(2\pi I_{n,\varepsilon}(\lambda_j)\big)\right) = \sum_{j=1}^{m} b_{j,m}^2 \operatorname{var}\big(\phi\big(2\pi I_{n,\varepsilon}(\lambda_j)\big)\big) = \frac{\pi^2}{6}\sum_{j=1}^{m} b_{j,m}^2 \sim \frac{\pi^2}{24},$$

and

$$S_{m,\varepsilon}(\log) \underset{d}{\to} N\big(0, \pi^2/24\big).$$

Example 5.11 Let $\phi(x) = x$ and

$$b_{j,m} = \frac{b_j - \bar{b}}{\sqrt{m}}, \quad b_j = -2\log\lambda_j, \quad \lambda_j = \frac{2\pi j}{n}.$$

We have (cf. (5.53)) $\sum_{j=1}^{m} b_{j,m}^2 \sim 4$. Assume that ε_t are standard normal. Then the periodogram ordinates $2\pi I_{n,\varepsilon}(\lambda_k)$ are independent, with the same standard exponential distribution and hence with unit variance. Thus

$$\operatorname{var}\left(\sum_{j=1}^{m} b_{j,m}\phi\big(2\pi I_{n,\varepsilon}(\lambda_j)\big)\right) = \sum_{j=1}^{m} b_{j,m}^2 \operatorname{var}\big(2\pi I_{n,\varepsilon}(\lambda_j)\big) = \sum_{j=1}^{m} b_{j,m}^2 \sim 4.$$

Hence,

$$S_{m,\varepsilon}(\phi) \underset{d}{\to} N(0,4). \tag{5.71}$$

If ε_t are not Gaussian, then (cf. (4.142))

$$Cov\big(I_{n,\varepsilon}(\lambda_k), I_{n,\varepsilon}(\lambda_l)\big) = \frac{\kappa_4}{4\pi^2 n} \quad (j \neq k),$$

where κ_4 is the fourth cumulant. Therefore,

$$\sum_{\substack{j,k=1\\ j\neq k}}^{m} b_{j,m}b_{k,m}Cov\big(2\pi I_{n,\varepsilon}(\lambda_j), 2\pi I_{n,\varepsilon}(\lambda_k)\big) = \frac{\kappa_4}{n}\sum_{\substack{j,k=1\\ j\neq k}}^{m} b_{j,m}b_{k,m} = -\frac{\kappa_4}{n}\sum_{j=1}^{m} b_{j,m}^2,$$

since $\sum_{j=1}^{m} b_{j,m} = 0$. Hence, the covariance term is negligible and (5.71) is valid in the non-Gaussian case even though the periodogram ordinates are dependent.

The idea behind the proof of Theorem 5.6 can be illustrated for linear functionals $\phi(x) = x$. Indeed, it follows from

$$\sum_{j=1}^{m} b_{j,m}\big(2\pi I_{n,\varepsilon}(\lambda_j) - 1\big) = \sum_{j=1}^{m} b_{j,m}\left[\frac{1}{n}\sum_{t,s=1}^{n} \varepsilon_t\varepsilon_s e^{-i(t-s)\lambda_j} - 1\right]$$

$$= \sum_{t=1}^{n} \left[\sum_{s=1}^{t-1} \varepsilon_s \frac{1}{n} \sum_{j=1}^{m} b_{j,m} \underbrace{\left(e^{-i(t-s)\lambda_j} + e^{i(t-s)\lambda_j} \right)}_{2\cos(t-s)\lambda_j} \right]$$

$$= \sum_{t=1}^{n} \varepsilon_t \sum_{s=1}^{t-1} \varepsilon_s \cdot \underbrace{\frac{2}{n} \sum_{j=1}^{m} b_{j,m} \cos(t-s)\lambda_j}_{c_{t-s}(n,m)}$$

$$= \sum_{t=1}^{n} \varepsilon_t \sum_{s=1}^{t-1} \varepsilon_s c_{t-s}(n,m) = 2 \sum_{t=1}^{n} z_t(n,m)$$

where

$$z_1(1,m)$$
$$z_1(2,m), z_2(2,m)$$
$$\vdots$$
$$z_1(n,m), \ldots, z_n(n,m)$$

is a martingale difference array. Once it is shown that

$$\sum_{t=1}^{n} E\left[z_t^2(n,m) \mid \mathscr{F}_{t-1} \right] - 1 \underset{p}{\to} 0$$

(with $\mathscr{F}_t$ denoting the σ-algebra generated by ε_s, $s \le t$), and

$$\sum_{t=1}^{n} E\left[z_t^2(n,m) I\left\{ \left| z_t(n,m) \right| > \delta \right\} \right] \to 0$$

for all $\delta > 0$, asymptotic normality follows from a standard martingale central limit theorem.

5.6.4.8 Periodogram for Long-Memory Sequences

Finally, we show that the remainder terms are negligible. We sketch a proof for the GPH estimator only, referring for technical details and the local Whittle estimator to Hurvich et al. (2002) or Robinson (1995b).

Lemma 5.1 *Assume* (GPH1′), (FS2) *and* (FS3). *Then*

$$\sum_{j=1}^{m} b_{j,m} \log\left(\frac{I_{n,X}(\lambda_j)}{2\pi f_X(\lambda_j) I_{n,\varepsilon}(\lambda_j)} \right) = o_P(1).$$

Proof The details can be found in Hurvich et al. (2002). The sum is split into two parts, $\sum_{j=1}^{m_0}$ and $\sum_{j=m_0+1}^{m}$. The first part is treated easily since m_0 is treated as fixed. First, by the continuous mapping theorem, (4.138) and (4.150),

$$\sum_{j=1}^{m_0} \log\bigl(2\pi I_{n,\varepsilon}(\lambda_j)\bigr) \underset{d}{\to} V_2,$$

$$\sum_{j=1}^{m_0} \log\bigl(I_{n,X}(\lambda_j)/f_X(\lambda_j)\bigr) \underset{d}{\to} V_1,$$

where V_1 and V_2 are finite random variables. Thus,

$$\begin{aligned}
&\sum_{j=1}^{m_0} b_{j,m} \log\left(\frac{I_{n,X}(\lambda_j)}{2\pi f_X(\lambda_j) I_{n,\varepsilon}(\lambda_j)}\right) \\
&\quad \le \max_{1\le j\le m} b_{j,m} \left\{ \left| \sum_{j=1}^{m_0} \log\bigl(I_{n,X}(\lambda_j)/f_X(\lambda_j)\bigr) \right| + \left| \sum_{j=1}^{m_0} \log\bigl(2\pi I_{n,\varepsilon}(\lambda_j)\bigr) \right| \right\} \\
&\quad = o(1) O_P(1) = o_P(1).
\end{aligned}$$

The second part is more technical, and we refer to Hurvich et al. (2002). □

The combination of Theorem 5.6 and Lemma 5.1 implies that the asymptotic behaviour of $\sum_{j=1}^{m} b_{j,m} \log(I_{n,X}(\lambda_j))$ is the same as that of $\sum_{j=1}^{m} b_{j,m} \log(2\pi \times I_{n,\varepsilon}(\lambda_j))$.

5.6.4.9 Consistency of the Local Whittle Estimator

For consistency of the local Whittle estimator, one considers a δ-neighbourhood of d_0 ($0 < \delta < \frac{1}{2}$),

$$N_\delta = \bigl\{ d : |d - d_0| < \delta \bigr\},$$

and the probability

$$p_n = P\bigl(|\hat{d} - d_0| > \delta\bigr) = P\bigl(\hat{d} \in N_\delta^c \cap \Theta\bigr) = P\Bigl(\inf_{N_\delta^c \cap \Theta} K_m(d) \le \inf_{N_\delta \cap \Theta} K_m(d)\Bigr),$$

where $\Theta = (-1/2, 1/2)$. Since $d_0 \in N_\delta \cap \Theta$, we have

$$\inf_{N_\delta \cap \Theta} K_m(d) \le K_m(d_0).$$

Hence, using the notation

$$S_m(d; d_0) = K_m(d) - K_m(d_0),$$

one obtains

$$p_n \le P\Big(\inf_{N_\delta^c \cap \Theta} S_m(d; d_0) \le 0\Big).$$

Note that

$$S_m(d; d_0) = \log \frac{G_m(d)}{G_m(d_0)} - (d - d_0)\left(\frac{2}{m}\sum_{j=1}^{m} \log \lambda_j\right).$$

Intuitively, one may use the approximation

$$\begin{aligned} \log \frac{G_m(d)}{G_m(d_0)} &= \log\left[1 + \frac{G_m(d) - G_m(d_0)}{G_m(d_0)}\right] \\ &= \frac{G_m(d) - G_m(d_0)}{G_m(d_0)} + o_p(1); \end{aligned}$$

however, a detailed argument must use a uniform approximation (in probability). Moreover,

$$\begin{aligned} G_m(d) &= \frac{1}{m} c_f \sum_{j=1}^{m} \lambda_j^{2(d-d_0)} \left[\frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}}\right] \\ &= c_f \left\{\sum_{j=1}^{m} \lambda_j^{2(d-d_0)} \frac{1}{m} + \sum_{j=1}^{m} \lambda_j^{2(d-d_0)} \left[\frac{I_{n,X}(\lambda_j)}{c_f \lambda_j^{-2d_0}} - 1\right] \frac{1}{m}\right\}. \end{aligned}$$

If $d_0 < d + \frac{1}{2}$, then the first sum is a Riemann sum that converges to the corresponding integral. For instance, if $\Theta \subset (0, \frac{1}{2})$, then this is always the case and convergence is uniform in d. For $d + \frac{1}{2} < d_0$, a different approximation has to be used. The second sum is stochastic with expected value approaching zero. Again a uniformity argument must be used. Careful analysis along this line finally yields

$$\lim_{n \to \infty} P\Big(\inf_{N_\delta^c \cap \Theta} S_m(d; d_0) \le 0\Big) = 0$$

so that $\hat{d}_{\mathrm{LW}}$ converges to d_0 in probability. For details see Robinson (1995b).

5.6.5 *Comparison and Modifications of Semiparametric Estimators in the Fourier Domain*

Let us summarize the theory of semiparametric estimators in the Fourier domain:

Estimator	Linear
GPH (Theorem 5.4)	$\sqrt{m}(\hat{d}_{\mathrm{GPH}} - d) \underset{d}{\to} N(0, \mathrm{var})$ $\mathrm{var} = \pi^2/24$
Local Whittle (Theorem 5.5)	$\sqrt{m}(\hat{d}_{\mathrm{LW}} - d) \underset{d}{\to} N(0, \mathrm{var})$ $\mathrm{var} = 1/4$

Although there is no closed form formula for the local Whittle estimator, the proof of asymptotic normality is easier than for the GPH estimator. The reason is that long memory is filtered out automatically due to the division of $I_{n,X}(\lambda)$ by $f_X(\lambda)$, so that (apart from some details regarding the accuracy of the approximations) standard results for the periodogram of i.i.d. observations can be applied. From the applied point of view, the GPH estimator is easier to calculate since simple least squares regression (together with the correction by the Euler constant) can be applied. However, the asymptotic loss of efficiency compared to $\hat{d}_{\mathrm{LW}}$ is considerable, namely

$$\mathrm{as.eff}(\hat{d}_{\mathrm{GPH}}, \hat{d}_{\mathrm{LW}}) = \frac{1/4}{\pi^2/24} = \frac{6}{\pi^2} \approx 0.61. \tag{5.72}$$

On the other hand, the GPH estimator can be modified to account for this loss of efficiency, though the charm of simplicity is lost by the modifications. This will be discussed below, along with other extensions and modifications.

5.6.5.1 Bias Reduction—Tapering

Semiparametric estimators—though asymptotically unbiased—can have a considerable finite sample bias. For example, the bias of the GPH estimator is of the order m^2/n^2. One way to avoid or reduce the finite sample bias is to consider a tapered periodogram in the definition of the GPH or the local Whittle estimator. A tapered periodogram is defined as

$$I^T_{n,X}(\lambda) = \frac{|\sum_{t=1}^n w_{n,t} X_t e^{-it\lambda}|^2}{2\pi \sum_{t=1}^n |w_{n,t}|^2}, \tag{5.73}$$

where $w_j \in \mathbb{C}$ is an appropriate "taper". A classical choice is a cosine bell taper (Tukey 1967) defined by

$$w_{n,t} = 0.5\left\{1 - \cos\left(\frac{2\pi(t - 1/2)}{n}\right)\right\}.$$

We note that this taper is shift invariant which means that the resulting tapered periodogram is shift invariant as well. Recall that this property is important in order to avoid the problem that replacing the mean by its estimate deteriorates the performance. Hurvich and Beltrao (1993) indicated by means of simulation that tapering

reduces the bias of the GPH estimator. Hurvich and Chen (2000) also suggested a modified cosine bell taper,

$$w_{n,t} = 0.5\{1 - e^{i\frac{2\pi(t-1/2)}{n}}\},$$

and showed asymptotic normality of the resulting local Whittle estimator:

$$\sqrt{m}(\hat{d} - d) \underset{d}{\to} N\left(0, \frac{3}{8}\right).$$

We note that the asymptotic variance is still smaller than for the GPH, but larger than for the *untapered* local Whittle estimator. To explain the increase in the variance (and hence loss of efficiency), let us recall that for i.i.d. data the values of the discrete Fourier transform (DFT) at Fourier frequencies are uncorrelated. For the tapered DFT, we have

$$E\left[d_{n,X}(\lambda_j)\,\overline{d_{n,X}(\lambda_k)}\right] \neq 0,$$

for $|j - k| \leq p$. Thus, while tapering reduces the bias, it introduces some dependence.

A different approach to bias reduction is taken in Andrews and Guggenberger (2003). The authors consider the GPH estimator, but instead of Fourier frequencies λ_j $(j = 1, \ldots, m)$ they consider λ_j^{2r} for some integer r. The idea comes from a local polynomial regression that reduces the bias. The estimator of Andrews and Guggenbauer has a bias of the order $(m/n)^{2+2r}$ which is smaller than for the original GPH method.

5.6.5.2 Improved Efficiency—Pooling

The idea of pooling dates back to Hannan (1970). As mentioned above, the GPH method is less efficient than the local Whittle estimator. To address this problem, Robinson (1995a) considers modified GPH estimators $\hat{d}_J$ based on averages of the periodogram over disjoint blocks of adjacent frequencies, each of length J. The asymptotic variance of $\hat{d}_J$ (say $V_{11}(J)$) decreases monotonically in J, with limit

$$\lim_{J\to\infty} V_{11}(J) = \frac{1}{4} < \frac{\pi^2}{24}.$$

Thus, in the limit, the pooled log-periodogram estimator has the same efficiency as the local Whittle approach. However, a practical problem with blockwise averaging is that an observed series may be too short to use a large value of J. Hurvich et al. (2002) suggest using blocks of size 4.

5.6.5.3 Nonstationary and Noninvertible Models

Another direction of research deals with nonstationary processes. Assume, for instance, that the sequence Y_t ($t \in \mathbb{Z}$) is such that

$$Y_t - Y_{t-1} = X_t,$$

where X_t is an ARFIMA$(0, d, 0)$ with $d \in (-\frac{1}{2}, \frac{1}{2})$. The sequence Y_t is fractionally differenced of order $d^* = d + 1$ and hence nonstationary. Another situation is that we observe a noninvertible process with $d < -\frac{1}{2}$.

Hurvich and Chen (2000) considered a tapered local Whittle estimator and allowed $d \in (-\frac{1}{2}, \frac{3}{2})$. Data are differenced so that the memory parameter becomes $d^* = d - 1 \in (-\frac{3}{2}, \frac{1}{2})$ and the sequence may become noninvertible. They showed that the asymptotic variance of the local Whittle estimator of d^* is

$$\frac{\pi \Gamma^2(2p-1)\Gamma^2((4p-3)/2)}{\Gamma^4((2p-1)/2)\Gamma(4p-3)} \approx \left(\frac{p\pi}{2}\right)^{1/2},$$

where p is the parameter that appears in the definition of the tapered periodogram. We observe a loss of efficiency, but the estimator is applicable to noninvertible processes. Velasco (1999a) shows that the local Whittle estimator is consistent when $d \in (-\frac{1}{2}, 1)$ and asymptotically normal when $d \in (-\frac{1}{2}, \frac{3}{4})$. Phillips and Shimotsu (2004) show that the local Whittle is also consistent for $d \in [\frac{3}{4}, 1)$. However, the rate of convergence becomes m^{2-2d} and the limiting distribution is χ^2_1. If $d > 1$, then the local Whittle estimator is not consistent. Abadir et al. (2007) and Hurvich et al. (2005a) allow for deterministic or stochastic trends. Kim and Phillips (1999) argue that the range of consistency and normality is the same for both GPH and local Whittle estimator. Hurvich and Ray (1995) demonstrated that the GPH estimator of d^* (applied to Y_t) does not equal to the GPH estimator of d, increased by one, applied to the differenced data X_t.

5.6.5.4 Other Models and Approaches

Another extension is to time series where the spectral density has a pole at a known location ω that may not necessarily be equal to zero. Arteche and Velasco (2005) assume that

$$f_X(\omega \pm \lambda) \sim C_{\pm}\lambda^{-2d_{\pm}},$$

where $d_{\pm}$ may be different. The authors consider both the GPH and the local Whittle estimator. Hidalgo (2005) considers the case of spectral densities where the locations of poles are unknown.

Both estimators also apply to more complicated stationary models. Robinson and Henry (1996) extend Robinson's (1995b) results to linear processes with dependent innovations. Shao and Wu (2007) consider local Whittle for ARFIMA with

GARCH innovations. Lobato (1999) and Robinson (2008) develop semiparametric estimation for multivariate long-memory time series. Giraitis and Robinson (2003) derive an Edgeworth expansion for the local Whittle estimator and provide improved confidence intervals.

One can also consider different estimators. Robinson (1994a, 1994b), Lobato and Robinson (1996) and Lobato (1997) suggest the averaged periodogram

$$F_{n,X}(\lambda) = \frac{2\pi}{n} \sum_{j=1}^{[n\lambda/2\pi]} I_{n,X}(\lambda_j)$$

to estimate $F_X(\lambda) = \int_0^\lambda f_X(\omega)\,d\omega$. If $f_X(\lambda) \sim C|\lambda|^{-2d}$, then $F_X(\lambda) \sim C\lambda^{-2d+1}/(1-2d)$ and also $F_X(q\lambda)/F_X(\lambda) \to q^{-2d+1}$ as $\lambda \to 0$ for any $q > 0$. The estimator is defined as

$$\hat{d} = \frac{1}{2} - \frac{\log(F_{n,X}(q\lambda_m)/F_{n,X}(\lambda))}{2\log q}$$

where m is a bandwidth. The estimator is both location and scale invariant. However, the asymptotic variance depends on d, and, if $d \in (1/4, 1/2)$, then the limit is non-normal (of Hermite–Rosenblatt type). Thus, the estimator is not particularly useful from a practical point of view. However, the theory developed in Robinson (1994a, 1994b) formed a basis for further considerations in Robinson (1995a, 1995b) and thereafter.

5.7 Semiparametric Narrowband Methods in the Wavelet Domain

5.7.1 Log Wavelet Regression

In this section, we complement semiparametric estimation in the Fourier domain with corresponding methods in the wavelet domain. Estimation of the memory parameter in the wavelet domain originates in the works of Wornell and Oppenheim (1992) and Abry et al. (1995). The log-wavelet estimator we are going to analyse here was investigated in Abry and Veitch (1998) and Veitch and Abry (1999). An asymptotic theory was developed in Bardet et al. (2000), Moulines et al. (2007b, 2008). We refer also to overview articles by Faÿ et al. (2009) and Abry et al. (2003). There is also a corresponding theory for a wavelet version of the local Whittle estimator studied in Sect. 5.6.3. We will not discuss this here, referring the reader to Moulines et al. (2008).

Assume that X_t $(t \in \mathbb{Z})$ is a discrete-time long-memory time series (for instance, a FARIMA(p, d, q)). Assume further that X_t $(t \in \mathbb{Z})$ is centred with covariance function $\gamma_X(k)$. In order to apply the discrete wavelet transform, we replace the

sequence by its continuous-time interpolation $X(u)$ $(u \in \mathbb{R})$ defined as

$$X(u) = \sum_{t \in \mathbb{Z}} X_t \phi(u - t),$$

where $\phi(\cdot)$ is a father wavelet. In particular, if ϕ is the Haar scaling function, then $X(u)$ is just a piecewise constant interpolation of the discrete time sequence X_t $(t \in \mathbb{Z})$. For a finite sample $X_1, \ldots, X_n$, we define

$$X_n(u) = \sum_{t=1}^{n} X_t \phi(u - t).$$

Without loss of generality, we may assume that the support of ϕ and ψ is contained in $[-T, 0]$ and $[0, T]$, respectively, where T is a positive integer. This means that the processes $X(u)$ and $X_n(u)$ agree for all $u \in [0, n - T + 1]$. Furthermore, the support of $\psi_{j,k}$ is contained in $[2^{-j}k, 2^{-j}(k + T)]$. Recall that for instance for ψ-functions whose support has length 1, $\phi_{j,k}$ is defined by $\psi_{j,k} = T^{\frac{1}{2}} 2^{\frac{j}{2}} \psi(2^j N \cdot - k)$. Hence, the wavelet coefficients

$$d_{j,k} := \int_{-\infty}^{\infty} X(u) \psi_{j,k}(u)\, du$$

and

$$d_{j,k}^{(n)} := \int_{-\infty}^{\infty} X_n(u) \psi_{j,k}(u)\, du$$

are the same as long as the support of $\psi_{j,k}$ is contained in $[0, n - T + 1]$. Thus, since we consider $j \leq 0$, the restriction on k is $k \leq n_j - 1$ where

$$n_j := \left[2^j (n - T + 1) - (T - 1)\right],$$

and $[x]$ is the largest integer smaller than x. This motivates the following definition:

$$\mathscr{I}_n := \left\{(j, k) : j \leq 0, 0 \leq k \leq n_j - 1\right\}.$$

In other words, $\mathscr{I}_n$ is the set of indices (j, k) for which we can compute wavelet coefficients $d_{j,k} = d_{j,k}^{(n)}$. For a given resolution level j, we can compute $d_{j,0}, \ldots, d_{j,n_j - 1}$. Often, T is chosen to be equal to 1. This means that $n_j = 2^j n$ and $\mathscr{I}_n := \{(j, k) : j \leq 0, 0 \leq k \leq 2^j n - 1\}$. Also note that by definition we may use the decomposition

$$X(u) = \sum_{j=0}^{-\infty} \sum_{k=0}^{n_j - 1} b_{j,k} \psi_{j,k}(u)$$

for $u \in [0, n - T + 1]$.

Recall now the formula (4.154) for the variance of wavelet coefficients:

$$\sigma_j^2 := \mathrm{var}(d_{j,0}) \approx 2^{-2jd} c_f \int |\lambda|^{-2d} \left|\hat{\psi}(\lambda)\right|^2 d\lambda = 2^{-2jd} c_f \Psi(2d). \tag{5.74}$$

Taking logarithm on both sides, we have

$$\log\big(\mathrm{var}(d_{j,0})\big) \approx \log\big(c_f \Psi(2d)\big) - 2dj \log(2).$$

Since the sequence $d_{j,k}$ ($k \in \mathbb{Z}$) is stationary, we can estimate $\mathrm{var}(d_{j,0})$ by using the sample variance based on $d_{j,0}, \ldots, d_{j,n_j-1}$:

$$\hat{\sigma}_j^2 := \widehat{\mathrm{var}(d_{j,0})} = \frac{1}{n_j} \sum_{k=0}^{n_j-1} d_{j,k}^2.$$

This leads to the following regression problem

$$\log\big(\hat{\sigma}_j^2\big) = \log\big(\widehat{\mathrm{var}(d_{j,0})}\big) = \log\big(c_f \Psi(2d)\big) - 2dj \log(2) + U_j,$$

where $U_j = \log(\widehat{\mathrm{var}(d_{j,0})}/(c_f \Psi(2d) 2^{-2dj}))$.

We note a similarity to the log-periodogram regression set-up considered in Sect. 5.6, see (5.51). However, there is a significant difference between wavelet regression and the corresponding log-periodogram regression. We note that the errors U_j are defined explicitly in terms of d. Hence, unlike in the log-periodogram case, one can expect that the limiting variance of a log-wavelet regression estimator will depend on d.

The log-wavelet regression estimator of d may be obtained by regressing $\log(\hat{\sigma}_j^2)$ on $-2j\log(2)$, where $j = j_0, \ldots, j_1$. Resolution levels j_0 and j_1 have to be chosen by the user. In particular, $j_1 = j_0 - r$, where r is fixed and the choice of the finest resolution level $j_0 = j_0(n)$ depends on the sample size n. The idea is, of course, to let j_0 tend to $-\infty$ because the hyperbolic long-memory decay shows at coarse resolution levels.

Since the variance of $\mathrm{var}(d_{j,0})$ is larger the coarser the resolution level (i.e. the lower j) is, it is recommended to use a weighted linear regression. Specifically, the log-wavelet regression estimator is defined as

$$\hat{d}_{\mathrm{WR}} = \sum_{j=j_0}^{j_1} w_{j_0-j} \log\big(\hat{\sigma}_j^2\big),$$

where the weights w_j have the following properties:

$$\sum_{j=j_0}^{j_0-r} w_{j_0-j} = \sum_{j=0}^{r} w_j = 0 \tag{5.75}$$

and

$$-2\log(2)\sum_{j=0}^{r} j w_j = 1. \tag{5.76}$$

In particular, Moulines et al. (2007b) suggest the following weights:

$$\mathbf{w} = (w_1, \ldots, w_r)^T = DB\left(B^T DB\right)^{-1}\mathbf{b},$$

where D is a positive definite matrix,

$$B = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & r \end{bmatrix}^T,$$

and

$$\mathbf{b} = \left[0, -\left(2\log(2)\right)^{-1}\right]^T.$$

Consider the total number of wavelet coefficients used in the estimation:

$$m := \sum_{j=j_0}^{j_0-r} n_j.$$

The parameter m will play a similar role as the number of Fourier coefficients in case of the log-periodogram estimation. Since $n_j = 2^j(n - T + 1) - (T - 1)$, we have

$$\begin{aligned}\sum_{j=j_0}^{j_1} n_j &= \sum_{j=j_0}^{j_1} 2^j(n - T + 1) - \sum_{j=j_0}^{j_1}(T - 1) \\ &= (n - T + 1)2^{j_0}\left(2 - 2^{-r}\right) - (r + 1)(T - 1).\end{aligned}$$

If r is fixed and $n2^{j_0} = n2^{j_0(n)} \to \infty$ as $n \to \infty$ (note that $j_0(n) \to -\infty$), then

$$m = m(n) = \sum_{j=j_0}^{j_1} n_j \sim n2^{j_0}\left(2 - 2^{-r}\right).$$

Finally, to formulate a central limit theorem for the wavelet estimator, we state the following assumptions:

- (W1) X_t is a stationary Gaussian process;
- (W2)

$$f_X(\lambda) = |\lambda|^{-2d}\left(f_*(0) + O\left(\lambda^\rho\right)\right);$$

- (W3)

$$j_0(n) \to -\infty$$

and

$$\lim_{n\to\infty} n2^{j_0(n)} = \infty, \qquad \lim_{n\to\infty} n2^{(1+2\rho)j_0(n)} = 0; \tag{5.77}$$

- (W4) The wavelet and scaling functions have the following properties:
  - W4(a) ψ and ϕ have support $[0, T]$ and $[-T, 0]$, respectively;
  - W4(b) The wavelet function ψ has M vanishing moments;
  - W4(c) There exists $\beta > 1$ such that $\sup_{\lambda\in\mathbb{R}} |\hat{\psi}(\lambda)|(1 + |\lambda|)^{\beta} < \infty$.

Conditions (W1)–(W3) are almost the same as for the GPH estimator. The additional condition (W4) involves assumptions on wavelets used in the estimation procedure. For example, for Daubechies' wavelets the parameters M and β can be chosen arbitrarily large.

Theorem 5.7 *Under assumptions* (W1)–(W4), *we have*

$$\sqrt{m}(\hat{d}_{WR} - d) \underset{d}{\to} N\big(0, v^2\big),$$

where $m = n2^{j_0(n)}(2 - 2^{-r})$ *and*

$$v^2 = \big(2 - 2^{-r}\big)\frac{2}{\Psi^2(d)} \sum_{j=j_0}^{j_0-r} \sum_{j'=j_0}^{j_0-r} w_{j_0-j} w_{j_0-j'} 2^{(j_0-j)/2} 2^{(j_0-j')/2} \gamma\big(j, j'\big)$$

where the constant $\gamma(j, j')$ *is defined in* (5.80) (*as a function of* j, j').

Note that the asymptotic variance is quite complicated and depends on d. Also note that the second part of condition (5.77) is needed to assure that the bias is negligible. This is similar to what was needed for log-periodogram regression.

5.7.2 Technical Details for Wavelet Estimators

In this section, we present some technicalities for the log-wavelet estimator. Details can be found in Bardet et al. (2000), Moulines et al. (2007a, 2008) and in an overview article by Faÿ et al. (2009).

In what follows, we shall assume that the conditions of Theorem 5.7 are fulfilled.

5.7.2.1 Variance and Covariance of the Wavelet Sample Variance

Recall that $\sigma_j^2 = \mathrm{var}(d_{j,0})$ and $\hat{\sigma}_j^2 = n_j^{-1} \sum_{k=0}^{n_j-1} d_{j,k}^2$. Since the random variables X_t ($t \in \mathbb{Z}$) are normal, the wavelet coefficients $d_{j,k}$ are Gaussian as well. Hence,

$$cov\big(d_{j,k}^2, d_{j',k'}^2\big) = 2cov^2(d_{j,k}, d_{j',k'})$$

and

$$\operatorname{var}\big(\hat{\sigma}_j^2\big) = \frac{1}{n_j^2} \sum_{k,k'=0}^{n_j-1} cov\big(d_{j,k}^2, d_{j,k'}^2\big) = \frac{2}{n_j^2} \sum_{k,k'=0}^{n_j-1} cov^2(d_{j,k}, d_{j,k'})$$

$$= \frac{2}{n_j^2} n_j \sum_{k=-(n_j-1)}^{n_j-1} \left(1 - \frac{k}{|n_j|}\right) cov^2(d_{j,k}, d_{j,k'}).$$

On account of Lemma 4.23, the sequence $d_{j,k}$ $(k \in \mathbb{Z})$ has summable covariances. Hence, as $n_j \to \infty$,

$$\operatorname{var}\big(\hat{\sigma}_j^2\big) \sim \frac{2}{n_j} \sum_{k=-\infty}^{\infty} cov^2(d_{j,k}, d_{j,k'}). \tag{5.78}$$

Furthermore, the weak dependence of the wavelet coefficients implies that

$$cov\big(\hat{\sigma}_j^2, \hat{\sigma}_{j'}^2\big) = \frac{1}{n_j n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} cov\big(d_{j,k}^2, d_{j',k'}^2\big) = \frac{2}{n_j n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} cov^2(d_{j,k}, d_{j',k'})$$

$$\approx 2\big(f_*(0)\big)^2 \frac{2^{-2jd}}{n_j} \frac{2^{-2j'd}}{n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} \Psi_{j,j'}^2\big(k, k'\big), \tag{5.79}$$

where $\Psi_{j,j'}(k, k')$ was defined in (4.153). The weak dependence of the wavelet coefficients then also implies that the limit

$$\gamma\big(j, j'\big) := \lim_{n\to\infty} \sqrt{n_j n_{j'}} \frac{1}{n_j} \frac{1}{n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} \Psi_{j,j'}^2\big(k, k'\big) \tag{5.80}$$

exists and is finite (recall that n_j is proportional to $n2^j$).

5.7.2.2 Bias of the Log-wavelet Estimator

As in the case log-periodogram estimation, we begin with the bias term. In what follows, we will argue that the bias is

$$E[\hat{d}_{\mathrm{WR}}] - d = O\left(m^{-1} + \left(\frac{m}{n}\right)^{\rho}\right), \tag{5.81}$$

where $m = m(n) \sim n2^{j_0(n)}(2 - 2^{-r})$ (as $j_0 \to -\infty$). A precise constant is given in Bardet et al. (2000).

Let us start with the following important inequality. A proof is omitted (see Moulines et al. 2007a).

Lemma 5.2 *Let ξ be a centred Gaussian vector with covariance matrix Σ, and let A be a positive definite matrix. Then*

$$\left|E\left[\log\left(\xi^T A\xi\right)\right] - \log\left(E\left(\xi^T A\xi\right)\right)\right| \le C\left\{1 \wedge \frac{\rho_A^2 \rho_\Sigma^2}{\mathrm{var}(\xi^T A\xi)}\right\}$$

where ρ_A and ρ_Σ denotes the spectral radius of A and Σ, respectively.

Recall that $\sigma_j^2 = \mathrm{var}(d_{j,0})$ and $\hat{\sigma}_j^2 = n_j^{-1}\sum_{k=0}^{n_j-1} d_{j,k}^2$. We split the bias as

$$\begin{aligned} E[\hat{d}_{\mathrm{WR}}] - d &= \sum_{j=j_0}^{j_1} w_{j_0-j} E\left[\log\left(\hat{\sigma}_j^2\right)\right] - d \\ &= \sum_{j=j_0}^{j_0-r} w_{j_0-j} \log\left(\sigma_j^2\right) - d + \sum_{j=j_0}^{j_0-r} w_{j_0-j}\left\{E\left[\log\hat{\sigma}_j^2\right] - \log\left(E\left[\hat{\sigma}_j^2\right]\right)\right\}. \end{aligned}$$

As for the first term, we note that due to (5.75) and (5.76) we have

$$\begin{aligned} \sum_{j=j_0}^{j_0-r} w_{j_0-j} \log\left(c_f \Psi(d) 2^{-2jd}\right) &= \sum_{j=j_0}^{j_0-r} w_{j_0-j} \log\left(c_f \Psi(d)\right) - 2d\log(2) \sum_{j=j_0}^{j_0-r} j w_{j_0-j} \\ &= 0 + d. \end{aligned}$$

Hence, the first term can be written as

$$\sum_{j=j_0}^{j_0-r} w_{j_0-j} \log\left(1 + \frac{\sigma_j^2 - c_f \Psi(d) 2^{-2jd}}{c_f \Psi(d) 2^{-2jd}}\right),$$

and applying Lemma 4.24 yields a bound

$$\sum_{j=j_0}^{j_0-r} w_{j_0-j} \log\left(1 + C 2^{j\rho}\right) \le 2C \sum_{j=j_0}^{j_0-r} w_{j_0-j} 2^{j\rho}$$

(note the inequality $|\log(1+x)| \le 2x$ for $x > 0$). Hence, the first term is bounded by

$$2C 2^{j_0\rho} \sum_{j=0}^{-r} w_j 2^{j\rho} = C\left(\frac{m}{n}\right)^{\rho}. \tag{5.82}$$

As for the second term, we apply Lemma 5.2 to $\xi = (d_{j,0}, \ldots, d_{j,n_j-1})^T$, $\Sigma = \Sigma_j = cov(\xi)$ and $A = \mathrm{diag}(n_j^{-1})$ being an $n_j \times n_j$ diagonal matrix with the same diagonal entries n_j^{-1}, so that $\mathrm{trace}(A) = n_j^{-1}$. Then $\xi^T A\xi = \hat{\sigma}_j^2$ and the weak dependence of the wavelet coefficients yield $\mathrm{var}(\hat{\sigma}_j^2) \sim C n_j^{-1}$, see (5.78). Also,

$\mathrm{Sp}(\Sigma_j) \le 2\pi \sup_x |f_j(\lambda)|$, where f_j is the spectral density of $d_{j,k}$ $(k \in \mathbb{Z})$, see Lemma 4.8. Hence, applying Lemma 5.2, we obtain

$$\left|E\left[\log \hat{\sigma}_j^2\right] - \log\left(E\left[\hat{\sigma}_j^2\right]\right)\right| \le C\left(1 \wedge \frac{\|f_j\|_\infty^2}{n_j^2 \operatorname{var}(\hat{\sigma}_j^2)}\right) \le C n_j^{-1}.$$

By definition $n_j = 2^j(n - T + 1) - (T - 1)$, so that $j \to n_j$ is an increasing function. Hence, for each $j = j_0, \ldots, j_0 - r$ the bound above is at most of order $n_{j_0-r}^{-1}$. Furthermore, since $m \sim n2^{j_0}(2 - 2^{-r})$, the quantity n_{j_0-r} is proportional to m. Consequently, the second term in the bias decomposition is bounded by

$$C n_{j_0-r}^{-1} \sum_{j=j_0}^{j_0-r} w_{j_0-j} \le C n_{j_0-r}^{-1} \sim m^{-1}. \tag{5.83}$$

Consequently, (5.82) and (5.83) yield the bias bound (5.81).

5.7.2.3 Variance of the Log-wavelet Estimator

Next, we find a precise expression for the variance of the log-wavelet estimator. Specifically, we will show that

$$m \cdot \operatorname{var}(\hat{d}_{\mathrm{RW}}) \to v^2$$

as $m = n2^{j_0(n)}(2 - 2^{-r}) \to \infty$ when $n \to \infty$. The constant v^2 is defined in Theorem 5.7.

As we did for the bias, we start with the following inequality. A proof of this inequality is similar to the proof of Lemma 5.2.

Lemma 5.3 *Let ξ and $\tilde{\xi}$ be centred Gaussian vectors with covariance matrix Σ and $\tilde{\Sigma}$, respectively, and let A and $\tilde{A}$ be positive definite matrices. Then*

$$\left|cov\left(\log\left(\xi^T A\xi\right), \log\left(\tilde{\xi}^T \tilde{A}\tilde{\xi}\right)\right) - \frac{cov(\xi^T A\xi, \tilde{\xi}^T \tilde{A}\tilde{\xi})}{E(\xi^T A\xi)E(\tilde{\xi}^T \tilde{A}\tilde{\xi})}\right|$$
$$\le C\left\{\frac{\mathrm{sp}^3(A)\mathrm{sp}^3(\Sigma)}{\mathrm{var}^{3/2}(\xi^T A\xi)} \vee \frac{\mathrm{sp}^3(\tilde{A})\mathrm{sp}^3(\tilde{\Sigma})}{\mathrm{var}^{3/2}(\tilde{\xi}^T \tilde{A}\tilde{\xi})}\right\}.$$

We use this lemma with $\xi = (d_{j,0}, \ldots, d_{j,n_j-1})^T$, $\tilde{\xi} = (d_{k,0}, \ldots, d_{k,n_k-1})^T$, $A = \mathrm{diag}(n_j^{-1})$, $\tilde{A} = \mathrm{diag}(n_k^{-1})$ in order to approximate

$$\operatorname{var}(\hat{d}_{\mathrm{WR}}) = \sum_{j=j_0}^{j_0-r} \sum_{j'=j_0}^{j_0-r} w_{j_0-j} w_{j_0-j'} cov\left(\log\left(\hat{\sigma}_j^2\right), \log\left(\hat{\sigma}_{j'}^2\right)\right)$$

by

$$A(j_0, j_0 - r) := \sum_{j=j_0}^{j_0-r} \sum_{j'=j_0}^{j_0-r} w_{j_0-j} w_{j_0-j'} \frac{cov(\hat{\sigma}_j^2, \hat{\sigma}_{j'}^2)}{E(\hat{\sigma}_j^2) E(\hat{\sigma}_{j'}^2)}.$$

We proceed as we did for the bias: on account of Lemma 5.3, the error of this approximation is controlled by

$$\sum_{j=j_0}^{j_0-r} \sum_{k=j_0}^{j_0-r} w_{j_0-j} w_{j_0-k} \left\{ \frac{\|f_j\|_\infty^2}{n_j^3 \operatorname{var}^{3/2}(\hat{\sigma}_j^2)} \vee \frac{\|f_k\|_\infty^2}{n_k^3 \operatorname{var}^{3/2}(\hat{\sigma}_k^2)} \right\},$$

where f_j and f_k are the spectral densities of the sequences $d_{j,l}$ and $d_{k,l}$ ($l \in \mathbb{Z}$), respectively. As previously for the bias, we argue that the error of the approximation is of the order

$$C \sum_{j=j_0}^{j_0-r} \sum_{k=j_0}^{j_0-r} w_{j_0-j} w_{j_0-k} \left\{ n_j^{-1} \vee n_k^{-1} \right\} = C n_{j_0-r}^{-3/2} = C m^{-3/2} = o\left(m^{-1}\right).$$

Since we will see below that the variance $\operatorname{var}(\hat{d}_{\mathrm{WR}})$ is of the order m^{-1}, this approximation error is negligible.

Hence, it suffices to study the term $A(j_0, j_0 - r)$. Formula (5.79),

$$cov\left(\hat{\sigma}_j^2, \hat{\sigma}_{j'}^2\right) \approx 2 f_*^2(0) \frac{2^{-2jd}}{n_j} \frac{2^{-2j'd}}{n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} \Psi_{j,j'}^2\left(k, k'\right),$$

and (cf. (4.154))

$$E\left(\hat{\sigma}_j^2\right) = E\left(d_{j,0}^2\right) \sim 2^{-2jd} f_*(0) \Psi(d)$$

yield, for $j, j' \to -\infty$,

$$\frac{cov(\hat{\sigma}_j^2, \hat{\sigma}_{j'}^2)}{E(\hat{\sigma}_j^2) E(\hat{\sigma}_{j'}^2)} \sim 2 \frac{1}{\Psi^2(d)} \frac{1}{n_j} \frac{1}{n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} \Psi_{j,j'}^2\left(k, k'\right).$$

Also, the limit

$$\gamma\left(j, j'\right) := \lim_{n \to \infty} \sqrt{n_j n_{j'}} \frac{1}{n_j} \frac{1}{n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} \Psi_{j,j'}^2\left(k, k'\right)$$

exists and is finite (recall that n_j is proportional to $n2^j$).

Hence, $\operatorname{var}(\hat{d}_{\mathrm{WR}})$ behaves asymptotically (as $n \to \infty$, $j_0 \to -\infty$) like

$$\operatorname{var}(\hat{d}_{\mathrm{WR}}) \sim \frac{2}{\Psi^2(d)} \sum_{j=j_0}^{j_0-r} \sum_{j'=j_0}^{j_0-r} w_{j_0-j} w_{j_0-j'} \frac{1}{\sqrt{n_j n_{j'}}} \sqrt{n_j n_{j'}}$$

$$\times \frac{1}{n_j n_{j'}} \sum_{k=0}^{n_j-1} \sum_{k'=0}^{n_{j'}-1} \Psi^2_{j,j'}(k,k')$$

$$\sim \frac{2}{\Psi^2(d)} \sum_{j=j_0}^{j_0-r} \sum_{j'=j_0}^{j_0-r} w_{j_0-j} w_{j_0-j'} \frac{1}{\sqrt{n_j n_{j'}}} \gamma(j,j').$$

Using $n_j \sim 2^j n$, $m \sim n 2^{j_0(n)}(2-2^{-r})$, we have

$$n_j n_{j'} \sim 2^j 2^{j'} n^2 = \left(2^{j_0} n\right)^2 2^{(j-j_0)} 2^{(j'-j_0)} \sim \left(\frac{m}{2-2^{-r}}\right)^2 2^{(j-j_0)} 2^{(j'-j_0)}.$$

We conclude that

$$m \cdot \mathrm{var}(\hat{d}_{\mathrm{WR}}) \sim \left(2-2^{-r}\right) \frac{2}{\Psi^2(d)} \sum_{j=j_0}^{j_0-r} \sum_{j'=j_0}^{j_0-r} w_{j_0-j} w_{j_0-j'} 2^{(j_0-j)/2} 2^{(j_0-j')/2} \gamma(j,j').$$

5.8 Optimal Rate for Narrowband Methods

Both the log-periodogram and the local Whittle estimator are asymptotically normal with variances $\pi^2/24$ and $1/4$, respectively (see Theorems 5.4 and 5.5). It is very useful for applications that in both cases the asymptotic variance does not depend on any unknown parameters. On the other hand, the problem both methods (and local semiparametric methods in general, including the wavelet approach studied in Sect. 5.7) have in common is that, given an observed data set, the choice of the cut-off point m is not really specified. In fact, the choice of a cut-off point is shared by all local methods, including the wavelet approach studied in Sect. 5.7 (there $j_0(n)$ and r have to be chosen). Various solutions to this problem have been suggested in the literature. The essential idea is to choose m such that the mean squared error $MSE = E[(\hat{d}-d)^2]$ is minimized. This will discussed in this section.

Let us recall first that, given a sequence $m = m(n)$, the asymptotic efficiency of the GPH compared to the local Whittle estimator is

$$\mathrm{as.eff}(\hat{d}_{\mathrm{GPH}}, \hat{d}_{\mathrm{LW}}) = \frac{1/4}{\pi^2/24} = \frac{6}{\pi^2} \approx 0.61. \tag{5.84}$$

On the other hand, comparing bandwidth condition (GPH3) with (LW3), we observe that the second condition requires

$$(\log m)^{\frac{2}{1+2\rho}} m = o\left(n^{\frac{2\rho}{1+2\rho}}\right)$$

which is a slightly lower bound for m than imposed by (GPH3) where $m = o(n^{2\rho/(1+2\rho)})$ is sufficient. Thus, the way the results are formulated in Robinson

(1995a, 1995b), a logarithmically faster rate could be used for the GPH estimator so that $\hat{d}_{\mathrm{GPH}}$ would be infinitely more efficient asymptotically than $\hat{d}_W$. However, (GPH3) and (LW3) are merely sufficient conditions so that one may conjecture that this difference is not real but rather due to the specific way the results are formulated and derived. Thus, the fundamental questions in this context are:

1. What is the sharpest lower bound for $O_p(\hat{d} - d)$ among *all* estimators of d?
2. Is this the optimal rate in the sense that there is an estimator that achieves it?

To answer these questions, it has to be decided first which criterion and what kind of situations to consider for measuring the quality of an estimator. For instance, if we were willing to assume a priori that X_t is generated by a known parametric family of linear time series models, then the optimal rate would be $n^{-\frac{1}{2}}$ and it would be achieved, for instance, by a maximum likelihood or Whittle estimator (see Theorems 5.2 and 5.3). However, the point of semiparametric estimation is that the shape of the spectral density is not sufficiently known to be associated with a fixed parametric family. Instead, one needs to consider rate optimality within a sufficiently rich set of spectral functions that are not specified explicitly. Given such a set of functions, one may then consider the minimax risk of $\hat{d}$ over this set or various versions of Bayes risks and so on. A concrete result along this line is derived in Giraitis et al. (1997) as follows. Consider the class $\mathscr{N} = \mathscr{N}(C_0, K_0, \rho)$ of stationary Gaussian processes X_t ($t \in \mathbb{Z}$) with spectral densities f_X such that

$$\begin{aligned} &f_X(\lambda) = c_f |\lambda|^{-2d}\big(1 + \Delta(\lambda)\big) \quad \left(-\frac{1}{2} < d < \frac{1}{2}\right), \\ &0 < c_f < C_0, \qquad \big|\Delta(\lambda)\big| \le K_0 |\lambda|^{\rho} \end{aligned} \tag{5.85}$$

with C_0, K_0 and ρ fixed. In the following, $P_f \in \mathscr{N}$ will denote the probability distribution function of the process X_t for a given spectral density $f = f_X$, $d(f)$ the corresponding value of d and $\mathscr{D}_n$ the set of all estimators of d based on a series of length n. The following result shows that an estimator of d cannot have a better rate of convergence than $n^{-\rho/(2\rho+1)}$ when taking into account the worst possible case within $\mathscr{N}$.

Theorem 5.8 *Assume that* (5.85) *holds. Let*

$$r = \frac{\rho}{2\rho + 1}. \tag{5.86}$$

Then there exists a constant $c > 0$ such that

$$\liminf_n \left\{ \inf_{\hat{d}_n \in \mathscr{D}_n} \left[\sup_{P_f \in \mathscr{N}} P_f\big(n^r \big|\hat{d}_n - d(f)\big| \ge c\big) \right] \right\} > 0. \tag{5.87}$$

Proof Without loss of generality, we will assume $C_0 > 1$. Since the supremum over $\mathscr{N}$ is considered, it is sufficient to find a sequence of spectral densities f_n with

$P_{f_n} \in \mathscr{N}$ such that

$$\liminf_n P_{f_n}\left(n^r \left|\hat{d} - d(f)\right| \geq c\right) > 0$$

for some $c > 0$. Such a sequence can be constructed, for instance, by starting with f_0, $d(f_0) = 0$ and defining a sequence f_n such that $d_n = d(f_n)$ approaches $d(f_0)$ at the rate n^{-r}. Specifically, let

$$\delta_n = n^{-\frac{r}{\rho}} = n^{-\frac{1}{2\rho+1}}, \qquad d_n = d_1 n^{-r}$$

where $0 < d_1 < \frac{1}{2}$ and

$$f_0(\lambda) \equiv 1,$$
$$f_n(\lambda) = \begin{cases} c|\lambda|^{-2d_n} & (0 < |\lambda| \leq \delta_n), \\ 1 & (\delta_n < |\lambda| \leq \pi). \end{cases}$$

Then $d(f_0) = 0$ and, for $n \geq 1$,

$$d(f_n) - d(f_0) = d(n) = d(1)n^{-r}.$$

By detailed calculation, one can show that $P_{f_n} \in \mathscr{N}$. Moreover, P_{f_0} and P_{f_n} are close in the sense that

$$\int_{-\pi}^{\pi} \left[f_n(\lambda) - f_0(\lambda)\right]^2 d\lambda = O\left(n^{-1}\right)$$

(as $n \to \infty$). Consider now the log-likelihood ratio

$$\Lambda_n = \log \frac{L_{f_n}(X_1, \ldots, X_n)}{L_{f_0}(X_1, \ldots, X_n)} = \log \frac{dP_{f_n}(X_1, \ldots, X_n)}{dP_{f_0}(X_1, \ldots, X_n)}.$$

Then there exist finite positive constants K_1, K_2 such that for all n

$$\mu_n = E_{f_n}(\Lambda_n) \leq K_1,$$
$$\sigma_n^2 = E_{f_n}\left[(\Lambda_n - \mu_n)^2\right] \leq K_2$$

and, for all events A and any constant $a > 0$,

$$P_{f_n}(A) \leq e^a P_{f_0}(A) + \frac{M}{a^2}$$

where $M = K_1^2 + K_2$. Now, consider the specific event

$$A = A_n(f) = \left\{n^r \left|\hat{d}_n - d(f)\right| \geq c\right\}.$$

Since for any $0 \le \varepsilon \le 1$ the mixture distribution $\varepsilon P_{f_0} + (1-\varepsilon)P_{f_n}$ is in $\mathcal{N}$, we have the lower bound

$$\begin{aligned}\sup_{P_f \in \mathcal{N}} P_f\big(A_n(f)\big) &\ge \varepsilon P_{f_0}\big(A_n(f_0)\big) + (1-\varepsilon) P_{f_n}\big(A_n(f_n)\big) \\ &\ge \varepsilon\left[P_{f_n}\big(A_n(f_0)\big) - \frac{M}{a^2}\right] e^{-a} + (1-\varepsilon) P_{f_n}\big(A_n(f_n)\big).\end{aligned}$$

However, $d(f_n) - d(f_0) = d_n = d_1 n^{-r}$ so that for $c < \frac{1}{2}d(1)$ at least one of the inequalities

$$\big|\hat{d}_n - d(f_n)\big| \ge cn^{-r}, \qquad \big|\hat{d}_n - d(f_0)\big| \ge cn^{-r}$$

holds. Hence

$$P_{f_n}\big(A_n(f_0)\big) + P_{f_n}\big(A_n(f_n)\big) \ge 1$$

which implies

$$P_{f_n}\big(A_n(f_0)\big) - \frac{M}{a^2} \ge 1 - P_{f_n}\big(A_n(f_n)\big) - \frac{M}{a^2}$$

and

$$\sup_{P_f \in \mathcal{N}} P_f\big(A_n(f)\big) \ge \varepsilon\left(1 - \frac{M}{a^2}\right) e^{-a} + \big(1 - \varepsilon - \varepsilon e^{-a}\big) P_{f_n}\big(A_n(f_n)\big).$$

We may choose ε such that $(1-\varepsilon-\varepsilon e^{-a}) = 0$, namely $\varepsilon = (1+e^{-a})^{-1}$. This yields

$$\sup_{P_f \in \mathcal{N}} P_f\big(A_n(f)\big) \ge \left(1 - \frac{M}{a^2}\right)\frac{e^{-a}}{1+e^{-a}} = c(a, M)$$

which is independent of n and larger than zero for $a > \sqrt{M}$. □

The intuitive meaning of equation (5.87) is as follows. Suppose we use a certain estimator $\hat{d}_n$. If we have no prior knowledge where in $\mathcal{N}$ the true distribution P_f may be, then the probability that $\hat{d}_n$ differs from the true value by at least $\pm cn^{-r}$ can be, in the worst case, larger or equal to c. This probability cannot be made smaller, no matter which estimation procedure is used—at least ultimately, i.e. as n tends to infinity.

Question 1 is now resolved, at least when considering the family of distributions specified by $\mathcal{N}$. The next question is whether the rate n^{-r} can actually be achieved by a concrete estimator. The answer is affirmative. Giraitis et al. (1997) provide a solution based on a suitable modification of the GPH method. (It is to be expected that an analogous method could be constructed using the local Whittle approach, though up to date no concrete results seem to be available in the literature.) The idea is to use the trimmed GPH estimator as described (5.55), using optimal sequences of lower and upper bounds l_n, m_n. The specific conditions proposed in Giraitis et al. (1997) are

- (O1)

$$D_0^{-1}(\log n)^3 \le l \le D_0 \frac{n^{2r}}{(\log n)^3},$$

- (O2)

$$D_0^{-1} n^{2r} \le m \le D_0 n^{2r}$$

where $D_0 > 1$.

Then the following holds.

Theorem 5.9 *Assume that* (5.85) *holds. Let* $r = \rho/(2\rho+1)$. *Define*

$$\mathscr{J}_n = \{(l_n, m_n) : \text{(O1) and (O2) hold}\}$$

and let $\hat{d}_{m_n}(l_n)$ *be the trimmed* GPH *estimator based on frequencies* λ_j $(l_n \le j \le m_n)$. *Then*

$$\limsup_n \Big\{ \max_{(l_n,m_n)\in\mathscr{N}} \Big[\sup_{P_f \in \mathscr{N}} n^{2r} E_f\big[\big(\hat{d}_{m_n}(l_n) - d\big)^2\big] \Big] \Big\} < \infty.$$

The essence of the proof is to show that

$$n^{-2r} E\left[\left(\sum_{j=l_n}^{m_n} (b_j - \bar{b}) Z_j \right) \right] = O(1)$$

uniformly over $(l_n, m_n) \in \mathscr{J}_n$ and $P_f \in \mathscr{N}$ where Z_j are i.i.d. standard exponential random variables and $b_j = -2\log\lambda_j$. This is done by similar arguments as in Robinson (1995a).

The result essentially means that no matter which sequence from $\mathscr{J}_n$ one takes and which distribution from $\mathscr{N}$ is true, we have an upper bound for the worst mean squared error,

$$MSE\big(\hat{d}_{m_n}(l_n)\big) = E\big[\big(\hat{d}_{m_n}(l_n) - d\big)^2\big] = O\big(n^{-2r}\big) = O\big(n^{-\frac{2\rho}{2\rho+1}}\big). \tag{5.88}$$

We recognize the formula for the mean squared error of the GPH estimator (cf. (5.69)). It is interesting to note that the rate $m = O(n^{2r})$ was actually excluded in the original proof by Robinson (1995a), since there $m = o(n^{2r})$ (see (GPH3)). Indeed, if $m \approx n^{2r}$, then the variance and squared bias in the decomposition of the mean squared error are of the same order, whereas for the central limit theorem without bias correction we need the bias to be negligible compared to the square root of the variance. Furthermore, as in the case of asymptotic normality of the GPH estimator, trimming is not needed; see Soulier (2010).

We illustrate Theorems 5.8 and 5.9 for the special case of FARIMA processes.

Example 5.12 To illustrate the meaning of Theorems 5.8 and 5.9 and in particular condition (5.85), consider, for instance, a FARIMA$(0, d, 0)$ process with spectral density

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|1 - e^{-i\lambda}\right|^{-2d}, \quad d \in (-1/2, 1/2).$$

Then

$$f_X(\lambda) = c_f\left(2\sin\frac{|\lambda|}{2}\right)^{-2d} = c_f|\lambda|^{-2d}\big(1 + O\big(\lambda^2\big)\big).$$

Thus, (5.85) holds with $\rho = 2$.

This example shows that the best achievable rate within the class of FARIMA$(0, d, 0)$ processes (and more generally FARIMA(p, d, q)) is $m = n^{4/5}$, i.e. $\hat{d} - d = O_p(n^{-\frac{2}{5}})$, and this rate is indeed achieved by the GPH estimator. On the other hand, if we had chosen $b_j = b(\lambda_j) = 2\sin\frac{1}{2}|\lambda_j|$ in the definition of the GPH estimator, then with respect to the FARIMA$(0, d, 0)$ model, we would have $\rho = \infty$ and hence the parametric rate $m = \sqrt{n}$, $\hat{d} - d = O_p(n^{-\frac{1}{2}})$. The reason is simply that we are assuming the correct model for which $f_X(\lambda) \equiv c_f \exp\{d \cdot b(\lambda)\}$ for all frequencies. However, if we consider the more general class $\mathcal{N}(C_0, K_0, 2)$, then the advantage of using $b(\lambda) = 2\sin\frac{1}{2}|\lambda|$ disappears because $\mathcal{N}$ also includes models that deviate from the FARIMA$(0, d, 0)$ spectrum. Thus, the best achievable rate within $\mathcal{N}(C_0, K_0, 2)$ class is

$$m \sim cn^{\frac{4}{5}}, \qquad MSE(\hat{d}) = O\big(n^{-\frac{4}{5}}\big).$$

In summary, one can say that (5.85) together with Theorems 5.8 and 5.9 shows a dilemma often encountered in statistics. The best possible rate is $m \propto \sqrt{n}$ (and $\hat{d} - d = O_p(n^{-\frac{1}{2}})$) which is obtained whenever a correctly specified parametric model is assumed. On the other hand, the less we are willing to assume a priori, the smaller the value of ρ is. In reality, a compromise between these two extremes needs to be assumed. Often, a "FARIMA-neighbourhood" with $\rho = 2$ is used, since it includes all FARIMA(p, d, q) models (with p, q arbitrary). However, in some situations such an assumption may not be realistic. The first such situation is related to long-memory processes observed with an additive noise (see Example 5.13), the second is related to the case of spectral densities that are slowly varying at 0. In the latter case, the results of Theorems 5.8 and 5.9 are no longer valid. We will illustrate this in Example 5.14 (for a general theory, see Soulier 2010).

Example 5.13 Assume that $Y_t = X_t + Z_t$, where X_t $(t \in \mathbb{Z})$ is a long-memory process with spectral density $f_X(\lambda) = \lambda^{-2d} f_*(\lambda)$, and Z_t $(t \in \mathbb{Z})$ is an i.i.d. sequence with spectral density $\sigma_Z^2/(2\pi)$, independent of the sequence X_t. Then

$$f_Y(\lambda) = f_X(\lambda) + \sigma_Z^2/(2\pi) = \lambda^{-2d} f_*(\lambda) + \sigma_Z^2/(2\pi)$$

$$\approx \lambda^{-2d} f_*(0) + \sigma_Z^2/(2\pi) = \lambda^{-2d} f_*(0)\big(1 + O\big(\lambda^{2d}\big)\big).$$

Thus, (5.85) holds with $\rho = 2d$, yielding the optimal values

$$m \sim n^{\frac{2\rho}{2\rho+1}} = n^{\frac{4d}{4d+1}}, \qquad MSE(\hat{d}) = O\big(n^{-\frac{2\rho}{2\rho+1}}\big) = O\big(n^{-\frac{4d}{4d+1}}\big).$$

Example 5.14 Assume that the spectral density is of the form $f_X(\lambda) = \lambda^{-2d} f_*(\lambda)$ ($d \in (-\frac{1}{2}, \frac{1}{2})$) and f_* can be written as

$$f_*(\lambda) = f_*(\pi) \exp\left\{-\int_\lambda^\pi \frac{\eta(u)}{u}\, du\right\}, \quad \lambda \in (0, \pi),$$

where $\eta(\cdot)$ is regularly varying at 0 with index $\rho \geq 0$, i.e. $\lim_{\lambda \to 0} \eta(c\lambda)/\eta(\lambda) = c^\rho$ for each positive c. Recall that for $\rho = 0$, $\eta(\cdot)$ is said to be slowly varying. We will also assume for simplicity that $\eta(\cdot)$ is non-decreasing on $[0, \pi]$. For example, if $\eta(u) = Cu^\rho$, $\rho > 0$, $C > 0$, then

$$f_*(\lambda) = \text{const} \cdot \exp\big(\lambda^\rho\big) = \text{const} + O\big(\lambda^\rho\big),$$

and hence we are in the situation of (5.85). This situation is referred to as *second order regular variation*. As proven in Theorem 5.8, the rate of convergence is $n^{\rho/(2\rho+1)}$. However, if $\eta(u) = (\log\log(1/u)\log(1/u))^{-1}$, then $f_*(\lambda) = \log\log(1/\lambda)$ is slowly varying and the rate of convergence is $\log(n)\log\log(n)$. Likewise, if $f_X(\lambda) = \log|1 - e^{i\lambda}|^2$, then the spectral density is of the required form with $d = 0$ and so $f_X(\lambda) = f_*(\lambda)$. The spectral density is slowly varying and the rate of convergence is logarithmic. We note that $f_X(\lambda) = \log|1 - e^{i\lambda}|^2$ is the spectral density of a Gaussian sequence with covariance $\gamma_X(k) = 1/(k+1)$. In summary, the results of Theorems 5.8 and 5.9 are valid under the assumption of second order regular variation only.

5.9 Broadband Methods

5.9.1 Broadband LSE for FEXP(∞) Models

So far, we considered local methods (also called narrowband methods) that focus on the behaviour of the spectral density at the origin. Several questions remained unanswered:

1. How should we estimate the complete spectral density $f_X(\lambda)$ ($\lambda \in [-\pi, \pi]$)?
2. The best achievable rate of local methods under reasonable conditions is $\hat{d} - d = O(n^{-r})$ with $r = \frac{2}{5}$. Can one improve this rate, possibly by considering other realistic neighbourhoods?
3. Is the GPH or the local Whittle estimator better?

First, some explanations regarding the second question are needed. Condition (5.85) implies that

$$f_X(\lambda) = c_f|\lambda|^{-2d}\big(1 + \Delta(\lambda)\big) = c_f|\lambda|^{-2d} f_*(\lambda)$$

with

$$\frac{d^k}{d\lambda^k} f_*(0) = 0 \quad \big(1 \le k \le [\rho]\big).$$

Thus, if $\rho = 2$, then the first two derivatives of f_* vanish at $\lambda = 0$ which means that f_* is very flat (and close to 1) around the origin, The higher the value of ρ is, the flatter f_* becomes and the assumption becomes very restrictive. In particular, the assumption that $\rho > 2$ can be rather unrealistic. Therefore, $\rho = 2$ is the highest reasonable value one may be willing to accept. This means that, in practice, the best attainable rate within neighbourhoods of the type described by $\mathcal{N}$ (the set of Gaussian processes with the spectral density given above) is $\hat{d} - d = O_p(n^{-2/5})$. This is rather disappointing and quite far from the parametric rate of $n^{-\frac{1}{2}}$. For instance, for $n = 1000$ we have $n^{-2/5} \approx 0.06$ whereas $n^{-1/2} \approx 0.03$. The conclusion is that it may be perhaps too ambitious to expect that an estimator can perform well in a large neighbourhood of the type described by $\mathcal{N}$. A possible way out is to consider a different type of neighbourhood that may characterize departures from the "ideal" model in another direction. This is the idea of global estimators. As a by-product, global estimators also solve question 1, since the whole spectral density is estimated.

Essentially three types of global methods are discussed in the literature: (i) broadband LSE type methods (also called broadband FEXP); (ii) broadband Whittle estimation; (iii) adaptive fractional autoregressive (FAR(p, d)) fitting. We are not aware of broadband approaches in the wavelet domain. Here, we first consider approach (i). The starting point is the class of fractional exponential models, or FEXP models, as proposed in Beran (1993) and Robinson (1994a) (also see Bloomfield 1973 for EXP models in the context of short-memory time series). An FEXP(p) model has a spectral density of the form

$$\log f_{\mathrm{FEXP}(p)}(\lambda) := -2d \log\big|1 - e^{-i\lambda}\big| + \sum_{j=0}^{p} \vartheta_j h_j(\lambda)$$

where $-\frac{1}{2} < d < \frac{1}{2}$, the functions h_j are bounded, continuous at the origin and not linearly dependent (in the sense of the scalar product $\langle h_j, h_l\rangle = \int h_j(\lambda)h_l(\lambda)\,d\lambda$). The unknown parameter vector is $\theta = (d, \vartheta_0, \ldots, \vartheta_p)'$. Motivated by Fourier analysis, a standard choice for h_j is

$$h_0(\lambda) = \frac{1}{\sqrt{2\pi}}, \qquad h_j(\lambda) = \frac{1}{\sqrt{\pi}} \cos j\lambda \quad (j \ge 1). \tag{5.89}$$

This way one has $\langle h_j, h_l\rangle = \delta_{jl}$. The idea of broadband FEXP estimation (Moulines and Soulier 1999, 2000; Hurvich 2001; Hurvich and Brodsky 2001; Hurvich et al.

2002) is to assume that the spectral density is of the form

$$\begin{aligned}\log f_X(\lambda) &= -2d \log\left|1 - e^{-i\lambda}\right| + \log f_*(\lambda) \\ &=: d \cdot a(\lambda) + L_*(\lambda)\end{aligned}$$

where $L_*(\lambda) = \log f_*(\lambda)$ has the Fourier series representation

$$L_*(\lambda) = \log L(\lambda) = \sum_{j=0}^{\infty} \vartheta_j h_j(\lambda)$$

with h_j defined by (5.89). As n tends to infinity, $L_*(\lambda)$ is approximated by a finite Fourier series with p_n terms and estimated parameters $\hat{\vartheta}_j$ $(j = 0, \ldots, p)$. If p_n tends to infinity, then one obtains a perfect approximation, provided that the estimates $\hat{\vartheta}_j$ converge fast enough to the true values. The latter is guaranteed by preventing that p_n grows too fast, since otherwise there would be too many parameters to estimate. Thus this method can be understood as an empirical Fourier approximation of $\log f$. More specifically, taking into account the Fourier representation

$$-2\log\left|1 - e^{-i\lambda}\right| = \sum_{j=1}^{\infty} c_j \cos j\lambda = \sum_{j=1}^{\infty} \frac{2\sqrt{\pi}}{j} \cos j\lambda, \tag{5.90}$$

we have

$$\log f_X(\lambda) = \frac{\vartheta_0}{\sqrt{2\pi}} + \sum_{j=1}^{\infty} a_j(d, \vartheta_j) \cdot \left(\frac{1}{\sqrt{\pi}} \cos j\lambda\right)$$

with

$$a_j(d, \vartheta_j) = \frac{2\pi}{j} d + \vartheta_j.$$

If the parameters were known, then the difference between $\log f_X$ and the best approximation by an FEXP model of order p would be

$$\log f_X(\lambda) - \log f_{\mathrm{FEXP}(p)}(\lambda) = \sum_{j=p+1}^{\infty} \vartheta_j \cdot \left(\frac{1}{\sqrt{\pi}} \cos j\lambda\right).$$

This bias has to be balanced against the error due to simultaneous estimation of $\vartheta_0, \ldots, \vartheta_p$ and d.

The specific assumptions in Moulines and Soulier (1999) are as follows:

- (F1) X_t is Gaussian;
- (F2) $f_*'(\lambda)$ exists for $\lambda \neq 0$ and there exists a finite constant c such that

$$\left|f_*'(\lambda)\right| \leq \frac{c}{|\lambda|};$$

- (F3)

$$L_*(\lambda) = \log f_*(\lambda) = \sum_{j=0}^{\infty} \vartheta_j h_j(\lambda) \tag{5.91}$$

with

$$|\vartheta_0| + \sum_{j=1}^{\infty} |\vartheta_j| j^{\rho} \le K_0 < \infty \tag{5.92}$$

for some finite ρ, $K_0 > 0$;

- (F4)

$$p_n \to \infty, \qquad p_n^3 \left(\frac{\log n}{n} \right)^2 \to \infty, \qquad n(\log n)^2 p_n^{-1-2\rho} \to 0. \tag{5.93}$$

The assumption of Gaussianity is not really necessary (see, e.g. Hurvich et al. 2002) but simplifies calculations. In the following, the scalar product between infinite dimensional real vectors $x = (x_1, x_2, \ldots)'$, $y = (y_1, y_2, \ldots)'$ and the corresponding norm will be defined by

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \|x\|^2 = \langle x, x \rangle.$$

Furthermore, we will denote by $\theta^0 = (d^0, \vartheta_0^0, \vartheta_1^0, \ldots)'$ the true (infinite dimensional) value of the parameter θ, by $\theta^{(p)} = (d, \vartheta_0, \vartheta_1, \ldots, \vartheta_p)'$ the restriction of θ to the first $p + 2$ components and by η the Euler constant. Similarly, we define the remaining parameter vector $\theta(p) = (\vartheta_{p+1}, \vartheta_{p+2}, \ldots)'$ and also $c(p) = (c_{p+1}, c_{p+2}, \ldots)'$. For a given order $p = p_n$, the FEXP estimator of θ^0 is defined by

$$\hat{\theta}(p) = (0, 0, \ldots)'$$

and

$$\hat{\theta}^{(p)} = \underset{d, \vartheta_0, \ldots, \vartheta_p}{\arg\min} \sum_{j=1}^{[n/2]} \left[\log I_{n,X}(\lambda_j) + \eta - d \cdot a(\lambda_j) - \sum_{k=0}^{p} \vartheta_j \cdot h_j(\lambda_j) \right]^2.$$

Computationally, this is very easy because $\hat{\theta}^{(p)}$ is obtained directly form the least squares estimator in a multiple linear regression of the form

$$y_j = \beta_1 a(\lambda_j) + \beta_2 h_0(\lambda_j) + \cdots + \beta_{p+2} h_p(\lambda_j) + \varepsilon_j$$

with

$$y_j = \log I_{n,X}(\lambda_j) + \eta, \qquad d = \beta_1, \qquad \vartheta_j = \beta_{j+2} \quad (j = 0, \ldots, p).$$

An asymptotic expression for the mean squared error of $\hat{d}$ (which is the first component of $\hat{\theta}$) can be given as follows (Moulines and Soulier 1999):

Theorem 5.10 *Under the assumptions* (F1)–(F3),

$$\begin{aligned} MSE(\hat{d}) &= E\big[(\hat{d}-d)^2\big] = \text{Bias}^2 + \text{Variance} \\ &= \left(\frac{\langle c(p), \theta(p)\rangle}{\|c(p)\|^2}\right)^2 + 4\pi \cdot \frac{\pi^2}{6\|c(p)\|^2} n^{-1} + R_{n,p} \end{aligned} \tag{5.94}$$

where an upper bound for the remainder term $R_{n,p}$ can be given by

$$|R_{n,p}| \le C\frac{p}{n}\left[\frac{(\log n)^3}{p^\rho} + \frac{p(\log n)^6}{n}\right]$$

with a finite constant C.

Under slightly stronger conditions than in Theorem 5.10, asymptotic normality of $\hat{d}$ is derived in Moulines and Soulier (1999).

Theorem 5.11 *Suppose that* (F1), (F2), (F3) *and* (F4) *hold with $\rho > 1/4$. Then*

$$\sqrt{\frac{n}{p_n}}(\hat{d}-d) \underset{d}{\to} N\left(0, \frac{\pi^2}{6}\right).$$

Since $c(p)$ is fully specified by (5.90) and the coefficients ϑ_j satisfy condition (5.92), $\|c(p)\|^2$ can be simplified, as $p \to \infty$, to

$$\big\|c(p)\big\|^2 = 4\pi \sum_{j=p+1}^{\infty} j^{-2} \sim 4\pi p^{-1} \int_1^\infty x^{-2}\,dx = 4\pi p^{-1}. \tag{5.95}$$

Therefore, the variance term can be approximated asymptotically by

$$4\pi \cdot \frac{\pi^2}{6\|c(p)\|^2} n^{-1} \sim \frac{\pi^2}{6}\frac{p}{n}.$$

Moreover, (5.92) implies $\vartheta_j = o(j^{-\rho-1})$ so that

$$\big|\langle c(p), \theta(p)\rangle\big| = \left|2\sqrt{\pi} \sum_{j=p+1}^{\infty} \left(\vartheta_j j^\rho\right) \cdot j^{-1-\rho}\right| \le \text{const} \cdot p^{-1-\rho}.$$

For the bias term, we therefore obtain the upper bound

$$\left(\frac{\langle c(p), \theta(p)\rangle}{\|c(p)\|^2}\right)^2 \le \text{const} \cdot p^{-2\rho}.$$

This means that the bias term converges to zero whenever $p \to \infty$ whereas the variance term converges to zero whenever $pn^{-1} \to 0$. This is a classical situation in nonparametric statistics where a balance between bias and variance has to be found. Also note that the remainder term is always of a smaller order, as long as both conditions hold, and p^ρ grows faster than $(\log n)^3$ and $pn^{-1}(\log n)^3 \to 0$. The MSE can then be approximated by

$$MSE = A_1 p^{-2\rho} + A_2 \frac{p}{n} = A_1 \exp(-2\rho \log p) + A_2 n^{-1} \cdot p \tag{5.96}$$

where A_1, A_2 are suitable constants. The optimal value of p is obtained by

$$\frac{\partial}{\partial p} MSE = -2\rho A_1 \cdot p^{-1-2\rho} + A_2 n^{-1} = 0$$

which yields

$$\begin{aligned} p_{\text{opt}} &= C_{\text{opt}} \cdot n^{\frac{1}{2\rho+1}}, \\ C_{\text{opt}} &= \left(\frac{2\rho A_1}{A_2} \right)^{\frac{1}{2\rho+1}}. \end{aligned} \tag{5.97}$$

As always in nonparametric optimization problems, the optimal choice is such that the contributions of the bias and the variance are of the same order. The corresponding optimal MSE is of the order

$$\begin{aligned} MSE_{\text{opt}} &= O\left(\frac{p_{\text{opt}}}{n} \right) = O\left(n^{-2r}\right), \\ r &= \frac{2\rho}{2\rho+1}. \end{aligned} \tag{5.98}$$

At first sight, (5.98) looks the same as the rate obtained in (5.88) for semiparametric methods. In particular, for $\rho \to \infty$, one approaches the parametric rate n^{-1}. As for (5.88), the limit is never reached unless one is in a parametric setting. Here, this can be seen from (5.97) because p_n becomes $O(1)$, or in other words, one is confined to a model with a finite number of parameters ϑ_j. There is, however, an essential difference between (5.98) and (5.88) because the interpretation of the regularity parameter ρ is completely different. For neighbourhoods $\mathcal{N}(C_0, K_0, \rho)$ used in the context of narrowband estimation, a high value of ρ implies that all derivatives of $f_*(\lambda)$ up to order $k \le \rho$ are zero at the origin. For $\rho > 2$, this is a very strong unrealistic restriction. For the broadband method, ρ only restricts the rate at which Fourier coefficients ϑ_j of $L_*(\lambda) = \log f_*(\lambda)$ converge to zero (see (5.92)), without imposing any specific differentiability properties. Thus, large values of ρ are *not* unrealistic. In fact, for a large class of functions f_*, ϑ_j even decays at an exponential rate so that $\rho = \infty$. Of course, in this case, the heuristic "derivation" of p_{opt} given above cannot be applied directly. To be specific, suppose that

$$|\vartheta_j| \le \text{const} \cdot \varphi^j$$

for some $0 \le \varphi < 1$. Then there exists a $\lambda_0 > 0$ such that

$$\sum_{j=0}^{\infty} e^{j\lambda_0} |\vartheta_j| \le K_0.$$

This implies

$$\begin{aligned} \left|\langle c(p), \theta(p)\rangle\right| &= 2\sqrt{\pi} \sum_{j=p+1}^{\infty} |\vartheta_j| j^{-1} = 2\sqrt{\pi} \sum_{j=p+1}^{\infty} e^{j\lambda_0} |\vartheta_j| \left(j e^{j\lambda_0}\right)^{-1} \\ &\le K_0 p^{-1} e^{-p\lambda_0} \end{aligned}$$

and, due to (5.94),

$$MSE(\hat{d}) \approx A_1 e^{-2p\lambda_0} + A_2 \frac{p}{n}. \tag{5.99}$$

Minimizing with respect to p leads to

$$p_{\text{opt}} \sim C_{\text{opt}} \log n \tag{5.100}$$

(where C_{opt} can be calculated from A_1, A_2 and λ_0) and an optimal MSE of the order

$$MSE_{\text{opt}}(\hat{d}) = O\left(\frac{p_{\text{opt}}}{n}\right) = O\left(\frac{\log n}{n}\right). \tag{5.101}$$

This result can be summarized as follows. We replace assumption (F3) by assumption (F3′),

$$|\vartheta_0| + \sum_{j=1}^{\infty} |\vartheta_j| e^{j\lambda_0} \le K_0 < \infty. \tag{F3′}$$

Then the following holds (Moulines and Soulier 1999).

Theorem 5.12 *Let*

$$p_n \sim \lambda_0^{-1} \log n.$$

Then, under assumptions (F1), (F2), (F3′) *and* (F4), p_n *minimizes the MSE asymptotically and*

$$\lim_{n\to\infty} \frac{n}{\log n} MSE(\hat{d}) = \lambda_0^{-1} \frac{\pi^2}{6}.$$

Thus, up to a logarithmic factor, we obtain the parametric rate. In comparison, the rate obtained here is much better than n^{-r} with $r = 2\rho/(2\rho+1)$ and $\rho \le 2$ for local methods. The reason is that here ρ represents a different aspect of the approximation so that assuming $\rho = \infty$ is not unrealistic. A further remarkable consequence of the theorem is that asymptotically the contribution of the variance is larger than

the one of the bias by the factor $\log n$. Thus, at least in theory, no bias correction is needed when applying the results to confidence intervals or testing. Practically speaking, of course, $\log n$ grows very slowly so that for small to moderate sample sizes the contribution of the bias may not really be negligible. Moreover, it is not quite obvious how to guess the value of λ_0.

In analogy to local methods, rate optimality (in the minimax sense) of the broadband LSE can be derived, though over more general sets of spectral densities. In particular, $O(n^{-1}\log n)$ turns out to be the best attainable rate for the mean squared error of $\hat{d}$. In comparison with local methods where the best rate is $O(n^{-4/5})$, this is a considerable improvement. This, together with Theorem 5.12, provides a strong argument in favour of broadband methods. However, for finite samples, the actual values of p_{opt} and MSE_{opt} very much depend on the constant C_{opt}. Thus, a data-adaptive algorithm would be needed where this constant would be estimated. If we are satisfied with purely asymptotic optimality, then only λ_0 or a lower bound for λ_0 in (F3′) would need to be estimated. Even more pragmatically, to be on the safe side, a relatively small lower bound $\lambda_{\text{low}} \le \lambda_0$ may be used. Such an assumption is more general than a parametric model and thus leads to a realistic asymptotic bound for the mean squared error,

$$\lim_{n\to\infty} \frac{n}{\log n} MSE(\hat{d}) \le \frac{\pi^2}{6}\frac{1}{\lambda_{\text{low}}}.$$

As always, staying on the safe side (or in other words applying a "robust" procedure in the sense specified above), leads to a loss of efficiency of the size $\lambda_{\text{low}}/\lambda_0$. Suggestions for adaptive semiparametric estimation are discussed, for instance, in Moulines and Soulier (2003), Henry and Robinson (1996), Hurvich (2001), Henry (2007). Moulines and Soulier (2000) and Hurvich (2001) propose using a version of Mallows C_p criterion for choosing p. In particular, if the aim is to minimize the integrated mean squared error

$$\int E\big[\big(\log \hat{f}(\lambda) - \log f_X(\lambda)\big)^2\big]\, d\lambda,$$

it is proposed to minimize a C_p-statistic defined by

$$C_p^* = RSS + \frac{\pi^2 p_n}{3} \tag{5.102}$$

with

$$RSS = \sum (y_i - \hat{y}_i)^2,$$

$y_j = \log I_{n,X}(\lambda_j) + \eta$ and $\hat{y}_i = \log f_X(\lambda; \hat{\sigma}_\xi^2, \hat{\theta})$. Optimality properties of this criterion are discussed in Moulines and Soulier (2000) (also see Hurvich and Brodsky 2001 where (5.102) was suggested originally).

5.9.2 Broadband Whittle Estimation for FEXP(∞) Models

In analogy to narrowband estimation, a possibly more efficient alternative to broadband LSE is a Whittle approach. First asymptotic results were derived in Narukawa and Matsuda (2011). As before, it is assumed that the spectral density is of FEXP(∞)-type, i.e.

$$f_X(\lambda) = \left|1 - e^{-i\lambda}\right|^{-2d} f_*(\lambda; \vartheta)$$

where

$$L_*(\lambda; \vartheta) = \log f_*(\lambda; \vartheta) = \sum_{j=0}^{\infty} \vartheta_j \cos j\lambda.$$

Note that f_X can also be written as

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|1 - e^{-i\lambda}\right|^{-2d} k(\lambda; \vartheta)$$

with $\sigma_\varepsilon^2 = 2\pi \exp \vartheta_0$ equal to innovation variance in the Wold decomposition of X_t, and

$$\log k(\lambda; \vartheta) = \sum_{j=1}^{\infty} \vartheta_j \cos j\lambda.$$

Estimation is based on a sequence of FEXP(p_n) models with spectral densities

$$f_{p_n}(\lambda) = \left|1 - e^{-i\lambda}\right|^{-2d} f_{*,p_n}(\lambda; \vartheta) = \frac{\sigma_\varepsilon^2}{2\pi}\left|1 - e^{-i\lambda}\right|^{-2d} k_{p_n}(\lambda; \vartheta)$$

where

$$L_{*,p_n}(\lambda; \vartheta) = \log f_{*,p_n}(\lambda; \vartheta) = \sum_{j=0}^{p_n} \vartheta_j \cos j\lambda,$$

$$\log k_{p_n}(\lambda; \vartheta) = \sum_{j=1}^{p_n} \vartheta_j \cos j\lambda.$$

Given p_n, the Whittle log-likelihood is proportional to

$$\mathscr{L}_{p_n}(d, \vartheta) = \sum_{j=1}^{[n/2]} \left\{ \frac{I_{n,X}(\lambda_j)}{f_{p_n}(\lambda; d, \vartheta)} + \log f_{p_n}(\lambda; d, \vartheta) \right\}$$

and $\theta = (d, \vartheta)$ is estimated by minimizing $\mathscr{L}_{p_n}(d, \vartheta)$. Using the notation

$$\mathbf{a}_p(\lambda) = \left(-2\log\left|1 - e^{-i\lambda}\right|, 1, \ldots, \cos p\lambda\right)'$$

$\mathscr{L}_{p_n}(d, \vartheta)$ can also be written as

$$\mathscr{L}_{p_n}(\theta) = \sum_{j=1}^{[n/2]} \left\{ \frac{I_{n,X}(\lambda_j)}{\exp(\mathbf{a}'_{p_n}(\lambda)\theta)} + \mathbf{a}'_{p_n}(\lambda)\theta \right\}.$$

This is a convex function of θ so that minimization is no problem. To derive the asymptotic distribution of this estimator, Narukawa and Matsuda (2011) used the following conditions:

- (W1) X_t is a second-order linear process, i.e. $X_t = \mu + \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$, ε_t are i.i.d. with $E(\varepsilon_t) = 0$, $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t) = 1$, and have finite fourth moment;
- (W2) The true parameter value θ^0 is in the interior of the parameter space

$$\Theta = \left\{ 0 \le d < \frac{1}{2}, \vartheta_0, \vartheta_1, \ldots \in \mathbb{R}, |\vartheta_j| \le C j^{-\delta} \ (j \ge 1) \right\}$$

  where $0 < C < \infty$ and $\delta > 1$ are some fixed constants. Moreover, the spectral densities corresponding to two different values of d, say $d_1 \ne d_2$, differ for some frequencies $\lambda \in A \subseteq [-\pi, \pi]$ where A (which may depend on the particular parameters) has positive Lebesgue measure.
- (W3) Condition (F3) holds for some $\rho > 9/2$;
- (W4) $p_n \to \infty$, $p_n/n \to 0$ such that

$$\frac{p_n^{2\rho}}{n} \to \infty, \qquad \frac{p_n^9 (\log n)^4}{n} \to 0$$

  for some $\rho > 9/2$.

Note that the assumption that σ_ε^2 is equal to 1 is not really needed because the estimation of d and the asymptotic distribution do not depend on the value of σ_ε^2.

Theorem 5.13 *Under the assumptions* (W1)–(W4), $\hat{d}$ *converges to* d^0 *in probability and*

$$\sqrt{\frac{n}{p_n}}(\hat{\theta} - \theta) \underset{d}{\to} N(0, 1).$$

As expected, the asymptotic variance is smaller than $\pi^2/6 \approx 1.65$ obtained for the broadband LSE in Theorem 5.11. The loss of efficiency of the LSE compared to the Whittle approach is exactly the same as in the comparison of the local methods (see (5.84)),

$$\mathrm{eff}(\mathrm{LSE}, \mathrm{Whittle}) = \frac{6}{\pi^2} \approx 0.61. \tag{5.103}$$

Note, however, that Narukawa and Matsuda (2011) did not consider antipersistent processes ($-\frac{1}{2} < d < 0$) nor did they analyse the case where the Fourier coefficients ϑ_j decay exponentially—though it may be conjectured that the rate $\sqrt{\log n/n}$ can also be achieved for the Whittle broadband approach.

5.9.3 Adaptive Fractional Autoregressive Fitting

The idea of fitting AR(p_n) processes with p_n tending to infinity has been suggested in the context of short-memory time series in early papers by Parzen (1968, 1969, 1974), Akaike (1969) and Berk (1974) (also see Bhansali 1978, 1980, 1993; Shibata (1980, 1981). Following a suggestion by Beran, Bhansali et al. (2006) extended AR-fitting to the cases of long memory and antipersistence (also see Poskitt 2007a). The idea is to fit FAR(p_n, d) models to series of length n, with p_n tending to infinity with increasing sample size. The autoregressive order p_n has to diverge fast enough to avoid an asymptotic bias. On the other hand, the number of estimated parameters (which is equal to $p_n + 2$) should not grow too fast in order to keep the variance under control. Finding an optimal balance between these two requirements is analogous to the question of choosing an appropriate number of Fourier coefficients ϑ_j in the FEXP-approach considered previously. Here, it is assumed that the spectral density of the observed process can be written as

$$f_X\big(\lambda; \sigma_\varepsilon^2, \theta\big) = f_*(\lambda)\big|1 - e^{-i\lambda}\big|^{-2d},$$

with

$$f_*(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\big|\varphi\big(e^{-i\lambda}\big)\big|^{-2} = \frac{\sigma_\varepsilon^2}{2\pi}\left|\sum_{j=0}^{\infty}\varphi_j e^{-ij\lambda}\right|^{-2},$$

where $\varphi_0 = 1$, $\theta = (d, \varphi) = (d, \varphi_1, \varphi_2, \ldots)$ is the interior of the parameter space $\Theta \subset (-\frac{1}{2}, \frac{1}{2}) \times \mathbb{R}^{\mathbb{N}}$ with Θ such that X_t is a stationary causal process and

$$\sum_{j=0}^{\infty}|\varphi_j| < \infty.$$

In other words, X_t is assumed to have the Wold representation

$$X_t = \left(\sum_{j=0}^{\infty}\varphi_j B^j\right)^{-1}(1-B)^{-d}\varepsilon_t = \sum_{j=0}^{\infty}a_j\varepsilon_{t-j}$$

where $a_0 = 1$ and ε_t are uncorrelated identically distributed zero mean random variables with $\mathrm{var}(\varepsilon_t) = \sigma_\varepsilon^2$. Since σ_ε^2 is the one-step mean squared prediction error of the best linear forecast, we have $\int \log|\varphi(e^{-i\lambda})|\,d\lambda = 0$ so that the continuous version of Whittle's likelihood approximation is (up to a constant) of the form (cf. Whittle likelihood in (5.40))

$$\begin{aligned}\mathscr{L}_W^0\big(\sigma_\varepsilon^2, \theta\big) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_{n,X}(\lambda)}{f_X(\lambda;\theta)}\,d\lambda + \log\sigma_\varepsilon^2 \\ &= \sigma_\varepsilon^{-2}\int_{-\pi}^{\pi} I_{n,X}(\lambda)\big|\varphi\big(e^{-i\lambda}\big)\big|^2\big|1-e^{-i\lambda}\big|^{2d}\,d\lambda + \log\sigma_\varepsilon^2.\end{aligned}$$

Minimization with respect to θ is therefore independent of σ_ε^2. To obtain an estimator of θ, the quantity to minimize is

$$\mathscr{L}_W(\theta) = \int_{-\pi}^{\pi} I_{n,X}(\lambda)\left|\varphi\left(e^{-i\lambda}\right)\right|^2 \left|1 - e^{-i\lambda}\right|^{2d} d\lambda.$$

Since θ is infinite-dimensional, one uses instead a sequence of FAR(p_n) models. In analogy to the FEXP approach, the notation will be θ^0 for the true parameter, $\theta^{(p)} = (d, \varphi_1, \ldots, \varphi_p)'$ for the finite part of θ and $\theta(p) = (\varphi_{p+1}, \varphi_{p+2}, \ldots)'$ for the rest. Also, we will use the norms

$$\|\theta\|_2 = \sqrt{d^2 + \sum_{j=1}^{\infty} \varphi_j^2}, \qquad \|\theta\|_1 = |d| + \sum_{j=1}^{\infty} |\varphi_j|.$$

The FAR(p_n)-estimator of θ^0 is defined by

$$\hat{\theta}(p_n) = (0, 0, \ldots)' \tag{5.104}$$

and

$$\hat{\theta}^{(p_n)} = \underset{d,\varphi_1,\ldots,\varphi_{p_n}}{\arg\min}\ \mathscr{L}_W(d, \varphi_1, \ldots, \varphi_p, 0, 0, \ldots). \tag{5.105}$$

Finally, the innovation variance is estimated by

$$\hat{\sigma}_\varepsilon^2 = \mathscr{L}_W(\hat{\theta}) = \int_{-\pi}^{\pi} I_{n,X}(\lambda)\left|\hat{\varphi}\left(e^{-i\lambda}\right)\right|^2 \left|1 - e^{-i\lambda}\right|^{2\hat{d}} d\lambda. \tag{5.106}$$

Computationally, the method of fractional autoregressive fitting is less elegant than the FEXP-approach because $\mathscr{L}_W(\theta)$ is not a convex function. However, $\mathscr{L}_W(\theta)$ is a convex function of $\varphi_1, \varphi_2, \ldots, \varphi_p$, if d is fixed. The simplest way of computing (5.105) is therefore to minimize $\mathscr{L}_W(\theta)$ with respect to $\varphi_1, \varphi_2, \ldots, \varphi_p$ for each fixed d on a fine grid in $(-\frac{1}{2}, \frac{1}{2})$, and then take the solution with the overall smallest value of $\mathscr{L}_W(\theta)$. More specifically, Bhansali et al. (2006) define $\hat{\theta}$ as follows:

- Step 1: An initial consistent estimate $\tilde{d}$ of d is computed such that $\tilde{d} - d^0 = o_p(n^{-r})$ for some $0 < r < 1$. Once $\tilde{d}$ is given, improved estimates of d are searched for in the interval $\tilde{d} \pm C_0 p_n^{-s}$ only, where $C_0 > 0$, and $s > 2$ is such that

$$p_n^s = o\left(\min\left(n^{\frac{1}{4}}, n^r\right)\right).$$

- Step 2: $\hat{\theta}^{(p_n)}$ is defined by (5.104) and (5.105), but minimization is restricted by the condition $d \in [\tilde{d} - C_0 p_n^{-s}, \tilde{d} + C_0 p_n^{-s}]$.

We will use the notation

$$\tau_0 = 1, \qquad \tau_i = \sum_{j=1}^{i} \frac{\varphi_{i-j}}{j} \quad (i \geq 1). \tag{5.107}$$

The results in Bhansali et al. (2006) (Theorems 5.14 and 5.15 below) include not only the derivation of the asymptotic distribution of $\hat{d}$ but also a simultaneous limit theorem for all parameters. The following conditions are used:

- (FAR1)

$$E\left(\varepsilon_t^4\right) < \infty,$$

$\varphi_0 = 1$ and, for some $\varepsilon > 0$,

$$\varphi(z) = \sum_{j=0}^{\infty} \varphi_j z^j \neq 0 \quad \left(|z| < 1 + \varepsilon\right);$$

- (FAR2)

$$\Theta = \left\{ \theta \in \left[-\frac{1}{2}, \frac{1}{2} \right] \times \prod_{j=0}^{\infty} [C_j, D_j] : \|\theta\|_1 < \infty \right\}$$

and $\theta^0 \in \Theta^0$;

- (FAR3) As $n \to \infty$,

$$p_n \to \infty, \qquad p_n = o\left(\min\left(n^{\frac{1}{8}}, \frac{n^{1-2d^0}}{(\log n)^4} \right) \right) \tag{5.108}$$

and

$$\sum_{j=p_n}^{\infty} \left|\varphi_j^0\right| = o\left(n^{-\frac{1}{2}}\right). \tag{5.109}$$

Theorem 5.14 *Under* (FAR1)–(FAR3) *and* $J_n = o(p_n)$,

$$\sqrt{\frac{n}{p_n}} \left(\hat{d} - d^0, \hat{\varphi}_1 - \varphi_1, \ldots, \hat{\varphi}_{J_n} - \varphi_{J_n} \right)' = (\tau_0, \tau_1, \ldots, \tau_{J_n})' Z_n + r_n$$

where the random variable Z_n *is real-valued* (*and one-dimensional*), *and*

$$Z_n \underset{d}{\to} Z \sim N(0, 1), \qquad \|r_n\|_2 = o_p(1).$$

Since J_n may converge to infinity, Theorem 5.14 can be used to obtain simultaneous confidence bands for an increasing number of parameters $d^0, \varphi_1, \ldots, \varphi_{J_n}$. More specifically, we standardize by

$$s_{J_n} = \sqrt{\sum_{i=0}^{J_n} \tau_i^2}$$

or $\hat{s}_{J_n}$ with φ_j in (5.107) replaced by $\hat{\varphi}_j$ to obtain

Theorem 5.15 *Under the same assumptions as above*,

$$\hat{s}_{J_n} - s_{J_n} \underset{p}{\to} 0,$$

and

$$\sqrt{\frac{n}{p_n}}\hat{s}_{J_n}^{-1}\left\|\left(\hat{d} - d^0, \hat{\varphi}_1 - \varphi_1, \ldots, \hat{\varphi}_{J_n} - \varphi_{J_n}\right)\right\|_2 \underset{d}{\to} |Z|.$$

How fast p_n may diverge to infinity can be seen from condition (FAR3). On the one hand, (5.108) sets an upper bound which is required in order that the variance of the parameter estimates does not become too large due to overparametrization. On the other hand, (5.109) makes sure that p_n is large enough to avoid an asymptotic bias. Since the rate of convergence is $\sqrt{p_n/n}$, it is desirable to choose p_n as small as possible while satisfying the other requirements. Condition (FAR1) implies an exponential decay $|\varphi_j| = O(\rho^j)$ for some $0 < \rho < 1$ so that

$$\sum_{j=p_n}^{\infty} \left|\varphi_j^0\right| = O\left(\rho^{p_n}\right). \tag{5.110}$$

In order that $O(\rho^{p_n}) = o(n^{-\frac{1}{2}})$, it is sufficient to have

$$p_n \log \rho^{-2} - \log n \to \infty.$$

Since $\log \rho^{-2} > 1$, this implies that sequences of the order $p_n \sim c \log n$ with $c \geq 1$ are possible. In particular, for $p_n = \log n$ we obtain

$$\hat{d} - d^0 = O_p\left(\sqrt{\frac{\log n}{n}}\right)$$

and even

$$\left\|\left(\hat{d} - d^0, \hat{\varphi}_1 - \varphi_1, \ldots, \hat{\varphi}_{J_n} - \varphi_{J_n}\right)\right\|_2 = O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

In other words, as for the broadband methods above, the parametric rate of $n^{-\frac{1}{2}}$ can be reached up to a logarithmic factor.

5.9.4 General Conclusions on Broadband Estimators

Broadband methods have, at least in theory, two main advantages compared to local methods:

1. An almost parametric rate of convergence, $\hat{d} - d^0 = O_p(n^{-\frac{1}{2}} \log n)$ can be achieved under realistic conditions;

2. An estimate (and confidence intervals) of the complete spectral density is provided.

Compared to parametric estimation, semiparametric methods have the advantage of providing consistent estimates without the necessity of specifying a fixed functional form of the spectral density. This "robustness" comes at the cost of a slower rate of convergence, but in many situations the rate deteriorates by a logarithmic factor only (see point 1 above). Practically speaking, the discrepancy between the parametric and the semiparametric approach is, however, not as vast as it may seem. An experienced data analyst will never fit a parametric model without trying out alternative models with more parameters or even a completely different structure. An essentially objective way of choosing a parametric model is to apply an appropriate criterion such the AIC or BIC (see Sect. 5.5.6). In i.i.d. situations or in the regression context, the AIC can be shown to be closely related to Mallow's C_p. It may be conjectured that a similar result holds for fractional time series models. Thus, we may, for instance, fit FAR$(p, d, 0)$ models by parametric maximum likelihood estimation for $p = 0, 1, \ldots, p_{\max}$ with $p_{\max}$ large but fixed, and choose the best among these models by the AIC. This is, in principle, a parametric fit. On the other hand, we may carry out the same procedure using $p_{\max}(n) = O((\log n)^{1+\delta})$ (for some $\delta > 0$) and choose the model that minimizes C_p^* defined in (5.102). This is then called a semiparametric fit. A third approach is to fit one FAR$(p_n, d, 0)$ with $p_n = \log n$. Again, this is a semiparametric fit, however, applied "mechanically" using the order p_n by default. Now, p_n grows very slowly. Also, the data dependent semiparametric order (with $p_{\max}(n) \to \infty$) will generally grow very slowly. Therefore, the difference between a semiparametric fit with $p_n \to \infty$ and a parametric fit, chosen using a "reasonable" fixed upper bound for p, is likely to be small. One should also bear in mind that applying semiparametric fitting mechanically by letting p tend to infinity without regarding the observed data is likely to be less effective than a data driven approach where the order or even the selection of specific parameters is carried out by a suitable information criterion. The information criterion to be used may depend on the purpose of the analysis. For instance, in the short-memory context, the AIC is known to be suitable for predictions but less for model identification (due to inconsistency, see above). The same may be true for series with long-range dependence.

An additional question is which of the broadband methods is preferable: FEXP or FAR fitting (or possibly semiparametric fitting of another class of nested models)? The answer depends on which approximation we expect to be more parsimonious. For instance, if the true process is a fractional autoregressive process of finite order p^0, then FAR fitting is likely to provide better results. The reason is that ultimately only $p^0 + 2$ parameters are required in the FAR-representation of the spectral density, whereas the FEXP-series is infinite. The opposite applies when X_t is generated by an FEXP-process of finite order.

5.10 Parametric and Semiparametric Estimators—Summary

We give a brief summary of the main estimators considered in this chapter. Consider a second-order stationary time series X_t $(t = 1, 2, \dots)$ with expected value zero, autocovariance function $\gamma_X(k)$ and spectral density

$$f_X(\lambda) = f_*(\lambda)|\lambda|^{-2d} \quad (\lambda \to 0).$$

- Whittle estimator $\hat{d}_{\text{Whittle}}$—parametric (Theorem 5.3):

$$\hat{d}_{\text{Whittle}} = \operatorname{argmin} \frac{1}{n} x^T W_n(\theta) x,$$

  where

$$W_n(\theta) = \left[(2\pi)^{-2} \int_{-\pi}^{\pi} e^{i(r-s)\lambda} \frac{1}{h_X(\lambda;\theta)}\, d\lambda \right]_{r,s=-n,\dots,n},$$

  $h_X(\lambda) = (\sigma_\varepsilon^2/(2\pi))^{-1} f_X(\lambda)$ and $x = X(n) = (X_1, \dots, X_n)^T$.
  - $\sqrt{n}$-rate of convergence for linear processes (Theorem 5.3);
  - $\sqrt{n}$-rate of convergence not valid for subordinated processes;
- Geweke and Porter-Hudak estimator $\hat{d}_{\text{GPH}}$—semiparametric estimator in the Fourier domain (Theorem 5.4):
  - $\sqrt{m}$-rate of convergence, where the best possible value is $m = o(n^{4/5})$;
- Local Whittle estimator $\hat{d}_{\text{LW}}$—semiparametric estimator in the Fourier domain (Theorem 5.5):
  - $\sqrt{m}$-rate of convergence, where the best possible value is $m = o(n^{4/5})$;
  - More efficient than the GPH estimator;
- Log-wavelet regression estimator $\hat{d}_{WR}$—semiparametric estimator in the wavelet domain (Theorem 5.7):
  - $\sqrt{m}$-rate of convergence, where the best possible value is $m = o(n^{4/5})$;
  - Limiting variance has a complicated form and depends on d;
- Broadband estimators (Theorems 5.11, 5.13, 5.14):
  - $\sqrt{n/p_n}$-rate of convergence, where p_n may be chosen as $\log n$.

We summarize the limit theorems in Table 5.2.

5.11 Estimation for Panel Data

As discussed in Sect. 2.2.2, long memory can be generated by aggregation of short-memory processes $X_{i,t}$ $(t \in \mathbb{Z}; i = 1, 2, \dots, N)$ with randomly selected parameters. The particular case we considered was aggregation of $AR(1)$ processes with i.i.d. AR-parameters φ_i and such that φ_i^2 is Beta distributed with parameters $\alpha, \beta > 1$. If N is large, then the aggregated process $X_t^{(N)} = N^{-\frac{1}{2}} \sum_{i=1}^{N} X_{i,t}$ is close to the limiting process (obtained by $N \to \infty$). The long-term dependence structure is therefore characterized by a long-memory parameter d_0 that can be estimated by one of the

Table 5.2 Parametric and semiparametric estimators—asymptotic distributions

Estimator	Limit theorem
Whittle (Theorem 5.3)	$\sqrt{n}(\hat{d}_W - d) \xrightarrow[d]{} N(0, \text{var})$ $\text{var} = 4\pi V^{-1}$ FARIMA$(0, d, 0)$: $\text{var} = 6/\pi^2$
Narrowband LSE (Theorem 5.4)	$\sqrt{m}(\hat{d}_{\text{GPH}} - d) \xrightarrow[d]{} N(0, \text{var})$ $\text{var} = \pi^2/24$
Narrowband Whittle (Theorem 5.5)	$\sqrt{m}(\hat{d}_{\text{LW}} - d) \xrightarrow[d]{} N(0, \text{var})$ $\text{var} = 1/4$
Narrowband LSE (Theorem 5.11)	$\sqrt{m}(\hat{d} - d) \xrightarrow[d]{} N(0, \text{var})$ $\text{var} = \pi^2/24$
Broadband Whittle (Theorem 5.13)	$\sqrt{n/p_n}(\hat{d} - d) \xrightarrow[d]{} N(0, \text{var})$ $\text{var} = 1/4$
Narrowband Wavelet LSE (Theorem 5.7)	$\sqrt{m}(\hat{d}_{WR} - d) \xrightarrow[d]{} N(0, \text{var})$ var complicated and depends on d

maximum likelihood, narrowband or broadband methods described so far in this chapter. Sometimes, however, not only the aggregated but also the individual series are available. This allows for estimation of α and β (or more generally, the distribution φ_i, $i = 1, 2, \ldots$, are generated from). Once these parameters are estimated, one can obtain an alternative estimate of d_0 by plugging in $\hat{\alpha}$ and $\hat{\beta}$ into the aggregation formula (2.72), i.e.

$$\hat{\gamma}(k) = \frac{B(\hat{\alpha} + \frac{k}{2}, \hat{\beta} - 1)}{B(\hat{\alpha}, \hat{\beta})}$$

and thus setting $\hat{d} = 1 - \hat{\beta}/2$.

This approach is considered in Beran et al. (2010). As in Sect. 2.2.2, the squared coefficients φ_i^2 are assumed to be i.i.d. Beta distributed with parameters $\alpha, \beta > 1$. Given $\varphi_1^2, \ldots, \varphi_N^2$, the conditional MLE of $\theta^0 = (\alpha_0, \beta_0)$ is the defined by minimizing

$$\sum_{i=1}^{N} \left\{ \ln\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \right) - (\alpha - 1) \ln \varphi_i^2 - (\beta - 1) \ln\left(1 - \varphi_i^2\right) \right\},$$

or equivalently, by finding the root of the two equations

$$\begin{aligned}\psi(\alpha)-\psi(\alpha+\beta)&=N^{-1}\sum_{i=1}^{N}\ln\varphi_i^2,\\ \psi(\beta)-\psi(\alpha+\beta)&=N^{-1}\sum_{i=1}^{N}\ln\bigl(1-\varphi_i^2\bigr)\end{aligned}\tag{5.111}$$

where $\psi(x)=\frac{d}{dx}\Gamma(x)$ is the digamma function. Since φ_i are not known, the idea is to plug in (approximate) maximum likelihood estimates

$$\hat{\varphi}_i=\hat{\varphi}_{i,n}=\frac{\sum_{t=1}^{n}X_{i,t}X_{i,t-1}}{\sum_{t=1}^{n}X_{i,t}^2}$$

obtained from the individual series. The asymptotic distribution of $\hat{\alpha}$ and $\hat{\beta}$ can then be derived under suitable conditions on N and n. For each individual series, the asymptotic distribution of the Yule–Walker estimator $\hat{\varphi}_{i,n}$ (as $n\to\infty$) is well known. The difficulty in deriving the asymptotic distribution of $\hat{\theta}=(\hat{\alpha},\hat{\beta})$ is, however, that N tends to infinity simultaneously and the randomly generated values of φ_i can get arbitrarily close to the unit root boundary of 1. It can be shown, however, that under the conditions stated in Sect. 2.2.2 there is a uniform bound $E[(\hat{\varphi}_{i,n,h}-\varphi_i)^2]\le cn^{-1}$ where $\hat{\varphi}_{i,n,h}$ is a truncated estimator defined by

$$\hat{\varphi}_{i,n,h}=\min\bigl\{\max\{\hat{\varphi}_{i,n},h\},1-h\bigr\}$$

with $h\to 0$ as $N,n\to\infty$. Plugging this estimator into (5.111), one obtains an estimator $\hat{\theta}_{n,h}$ for which asymptotic normality can be derived (Beran et al. 2010):

Theorem 5.16 *Denote by $\theta_0=(\alpha_0,\beta_0)^T$ the true parameter vector, set $\eta_0=\min\{\alpha_0,\beta_0\}$, and let $N,n\to\infty$ and $h\to 0$ be such that*

$$\log h=o\bigl(N^{\frac{1}{4}}\bigr),\qquad N=o\bigl(h^{-2\eta_0}\bigr),\qquad N=o\bigl(n^2h^4\bigr).$$

Then $\hat{\theta}_{N,n,h}$ converges to θ_0 in probability and

$$\sqrt{N}(\hat{\theta}_{N,n,h}-\theta_0)\xrightarrow{d}N\bigl(0,A^{-1}(\theta_0)\bigr),$$

where

$$\begin{aligned}A(\theta)&=\frac{\partial}{\partial\theta}\begin{pmatrix}\psi(\alpha)-\psi(\alpha+\beta)\\ \psi(\beta)-\psi(\alpha+\beta)\end{pmatrix}\\ &=\begin{pmatrix}\psi_1(\alpha)-\psi_1(\alpha+\beta) & -\psi_1(\alpha+\beta)\\ -\psi_1(\alpha+\beta) & \psi_1(\beta)-\psi_1(\alpha+\beta)\end{pmatrix}\end{aligned}$$

and $\psi_1(x)=\frac{d^2}{dx^2}\ln\Gamma(x)$ denotes the trigamma function.

The conditions on N, n and h essentially imply that the stronger long memory is in the limiting aggregated process, the longer each replicate has to be in comparison with the number of replicates. For example, if we have $1 < \alpha_0, \beta_0 < 2$, then an admissible choice of n and h is

$$n \sim c_1 N^{\frac{1}{2}+\eta_0^{-1}+\delta}, \qquad h \sim c_2 N^{-\eta_0}$$

for some $\delta > 0$. This means, however, that n has to tend to infinity faster than N by the factor N^λ with $\lambda = \eta_0^{-1} + \delta - \frac{1}{2}$. The value of λ is larger the closer β_0 is to the lower limit of 1 which corresponds to $d_0 = 1 - \beta_0/2$ approaching the upper limit of $\frac{1}{2}$. Thus, in the most extreme case, n has to be about $\sqrt{N}$ times larger than N. On the other hand, in the limit towards short memory, i.e. $\beta_0 \to 2$ (and thus $d_0 \to 0$), λ tends to (the arbitrarily small value of) δ so that the length of each series ultimately does not need to be of a (much) larger order than N.

5.12 Estimating Periodicities

One of the standard questions in time series analysis is whether there may be periodicities in the data. In principle, one can distinguish two main types of periodicities: deterministic periodicities and stochastic periodicities. The first type (seasonal trends) may be handled, for instance, by suitable trigonometric regression models. For general results on fixed design regression under long memory, see Sect. 7.1. Here we consider the second type, i.e. stochastic periodicities.

5.12.1 Identifying Local Maxima

We first consider a linear process $X_t = \sum a_j \varepsilon_{t-j}$ with spectral density

$$f_X\big(\lambda; \sigma_\varepsilon^2, \theta\big) = \frac{\sigma_\varepsilon^2}{2\pi} h(\lambda; \theta) = \frac{\sigma_\varepsilon^2}{2\pi} \big|1 - e^{-i\lambda}\big|^{-2d} f_*(\lambda; \theta)$$

belonging to a parametric family such that f_* is twice continuously differentiable (with respect to λ) in $[-\pi, \pi]$, and $\theta = (d, \theta_2, \ldots, \theta_p)$. The true value of θ will be denoted by θ^0. We say that the process has stochastic periodicities if the spectral density has at least one (isolated) local maximum $\lambda_{\max}$. In other words, the set

$$\Lambda = \big\{\lambda \in [-\pi, \pi] : h'\big(\lambda; \theta^0\big) = 0, h''\big(\lambda; \theta^0\big) < 0\big\} \tag{5.112}$$

is not empty (with h' and h'' denoting derivatives with respect to λ). Suppose first that there is just one local maximum. Given an estimator $\hat{\theta}$ of θ^0, a natural estimator of $\lambda_{\max}$ is defined as a solution of

$$h'(\hat{\lambda}_{\max}; \hat{\theta}) = 0.$$

Under some regularity conditions, the asymptotic distribution of $\hat{\lambda}_{\max}$ can then be derived by applying a Taylor expansion with respect to λ and θ. In particular, let $\hat{\theta}$ be one of the approximate Gaussian maximum likelihood estimators (or the exact one) discussed previously. Then under the assumptions of Theorem 5.2, one obtains (Beran and Ghosh 2000)

$$\sqrt{n}(\hat{\lambda}_{\max} - \lambda_{\max}) \underset{d}{\to} N\big(0, \tau^2_{\max}\big) \tag{5.113}$$

with

$$\tau^2_{\max} = \frac{[\dot{h}'(\lambda_{\max}; \theta^0)]^T \, \Sigma_{\mathrm{MLE}} \dot{h}'(\lambda_{\max}; \theta^0)}{[h''(\lambda_{\max}; \theta^0)]^2} \tag{5.114}$$

where Σ_{MLE} is given in Theorem 5.2, and

$$\big[\dot{h}'\big(\lambda_{\max}; \theta^0\big)\big]^T = \left(\frac{\partial}{\partial \theta_1} h'\big(\lambda_{\max}; \theta^0\big), \ldots, \frac{\partial}{\partial \theta_p} h'\big(\lambda_{\max}; \theta^0\big) \right).$$

Note that the variance is inversely proportional to the curvature at the maximal point. This means that sharp peaks are easier to estimate. The same asymptotic result follows if one has more than one local maximum. Moreover, the result also holds for integrated process with unknown integer differencing order m if the approximate MLE defined in (5.46) is applied to estimate m and θ^0 (Beran and Ghosh 2000). The method can also be combined with a consistent model choice criterion and nonparametric trend removal (see Sects. 5.5.6, 7.4). Finally, note that the method and the central limit theorem (5.113) obtained in Beran and Ghosh (2000) is an extension of analogous results for short memory by Newton and Pagano (1983) (also see Diggle 1990).

Example 5.15 We consider the deseasonalized monthly average discharge series (m^3/s) of the Nile river at Dongola (1912–1984) as discussed in Sect. 1.2. In Fig. 5.7, the periodogram is plotted in log–log-coordinates together with a fitted FARIMA$(12, d, 0)$ spectral density obtained after model selection based on the BIC, and 95 %-confidence bands for frequencies where local maxima occur. A Bonferroni correction was applied so that the confidence intervals are simultaneous. The confidence intervals for the five peaks are about $[0.49, 0.53]$, $[0.99, 1.08]$, $[1.50, 1.61]$, $[2.00, 2.18]$ and $[2.42, 2.76]$. Noting that the seasonal frequency corresponding to one year is $\lambda_0 = 2\pi/12 \approx 0.524$ and its first four harmonics $\lambda_j = j\lambda_0$ ($j = 2, 3, 4, 5$) are equal to 1.047, 1.571, 2.094 and 2.618, respectively, we can see that each is in the corresponding confidence interval. This confirms the suspicion that the simple deseasonalizing filter applied in Sect. 1.2 did not remove all stochastic seasonality.

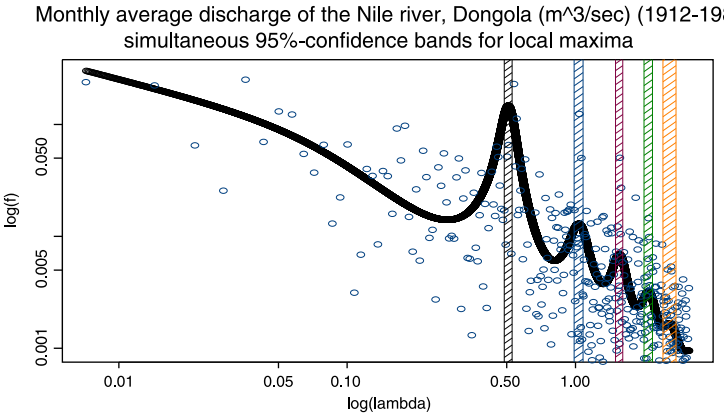

Fig. 5.7 Log–log-periodogram of the deseasonalized monthly average discharge series (m^3/s) of the Nile river at Dongola (1912–1984) as discussed in Sect. 1.2. Also plotted is a FARIMA$(12, d, 0)$ fit obtained after model selection based on the BIC, and 95 %-confidence bands for frequencies where local maxima occur. A Bonferroni correction was applied so that the confidence intervals are simultaneous

5.12.2 Identifying Strong Stochastic Periodicities

An extreme version of a local maximum at a certain frequency $\lambda_{\max}$ is a pole at this frequency. As a generalization of fractional ARIMA processes—following a suggestion by Hosking (1981)—Gray et al. (1989, 1994) studied the so-called GARMA(p, d, q) processes defined as stationary solutions of

$$\varphi(B)X_t = \left(1 - 2uB + B^2\right)^{-\frac{d}{2}} \psi(B)\varepsilon_t \tag{5.115}$$

where $-1 \le u \le 1$, $\varphi(z)$ and $\psi(z)$ are the usual AR- and MA-polynomials of orders p and q and ε_t are i.i.d. zero mean variables with a finite variance σ_ε^2. More generally, one could also consider merely uncorrelated ε_t's, such as martingale differences. For $p = q = 0$, X_t is also called a Gegenbauer process. The reason is that the coefficients in the Wold representation

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$$

are of the form

$$a_j = \frac{1}{\Gamma(\frac{d}{2})} \sum_{s=0}^{[j/2]} (-1)^s \frac{\Gamma(\frac{d}{2} + j - s)(2u)^{j-2s}}{s!(j-2s)!}$$

which turn out to be identical with Gegenbauer polynomials (see Sect. 3.1.4). Given the usual conditions on φ and ψ (i.e. $\varphi(z)$ and $\psi(z)$ have no roots for $|z| \le 1$) a

stationary invertible solution of (5.115) exists if

$$-1 < u, d < 1,$$

or

$$u = \pm 1 \quad \text{and} \quad -\frac{1}{2} < d < \frac{1}{2}.$$

The spectral density is given by

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} 2^{-d} (\cos\lambda - u)^{-d} \left| \frac{\psi(e^{-i\lambda})}{\varphi(e^{-i\lambda})} \right|^2.$$

Thus, for $u = 1$ one obtains a FARIMA(p, d, q) process with the hyperbolic behaviour $f_X(\lambda) \sim c_f |\lambda|^{-2d}$ at the origin $\lambda_0 = \arccos u = 0$. However, the behaviour is quite different for $|u| < 1$ since the pole (for $d > 0$) or zero value (for $d < 0$) occurs elsewhere, namely at the frequency

$$\lambda_0 = \arccos u \neq 0.$$

In particular, for $d > 0$, this means that we have a pole of the form $f_X(\lambda) \sim c_f |\lambda - \lambda_0|^{-2d}$ as λ approaches the non-zero frequency λ_0. This can be interpreted as an extreme version of stochastic periodicity within the realm of stationarity. Note that (5.115) can also be written as

$$\varphi(B)\big(1 - 2uB + B^2\big)^{\frac{d}{2}} X_t = \psi(B)\varepsilon_t.$$

This means that after multiplying the spectral density of X_t by $|1 - 2ue^{-i\lambda} + e^{-i2\lambda}|^d$ one obtains the spectral density of an ARMA process. Considering the spectral representation of X_t one can interpret this as follows. The same way the fractional differencing filter $(1 - \exp(-i\lambda))^d$ replaced the usual (integer) differencing filter $1 - \exp(-i\lambda)$ in the case of FARIMA processes, the fractional filter $(1 - 2ue^{-i\lambda} + e^{-i2\lambda})^{d/2}$ replaces the filter $1 - \exp(-i(2\pi/\lambda_0)\lambda)$ which is often used for removal of periodic components.

The parameters $\vartheta = (\sigma_\varepsilon^2, d, \varphi_1, \ldots, \varphi_p, \psi_1, \ldots, \psi_q)$ and $\lambda_0 = \arccos u$ can be estimated, for example, by Whittle's approximate MLE. The asymptotic distribution of $\hat{\vartheta}_{\text{Whittle}}$ turns out to be of the same form as in Theorem 5.2. This result, established in Giraitis et al. (2001) (for a related heuristic result see Chung 1996a, 1996b), is remarkable because it means that estimation of λ_0 does not change the asymptotic result. Intuitively, the reason is that a pole is ultimately highly visible so that it cannot be missed. Indeed, Giraitis et al. (2001) show that the rate of convergence of $\hat{\lambda}_0$ is very fast, namely $\hat{\lambda}_0 = \lambda_0 + O_p(n^{-1})$. This is in contrast to the estimation of local maxima for differentiable spectra considered in (5.112).

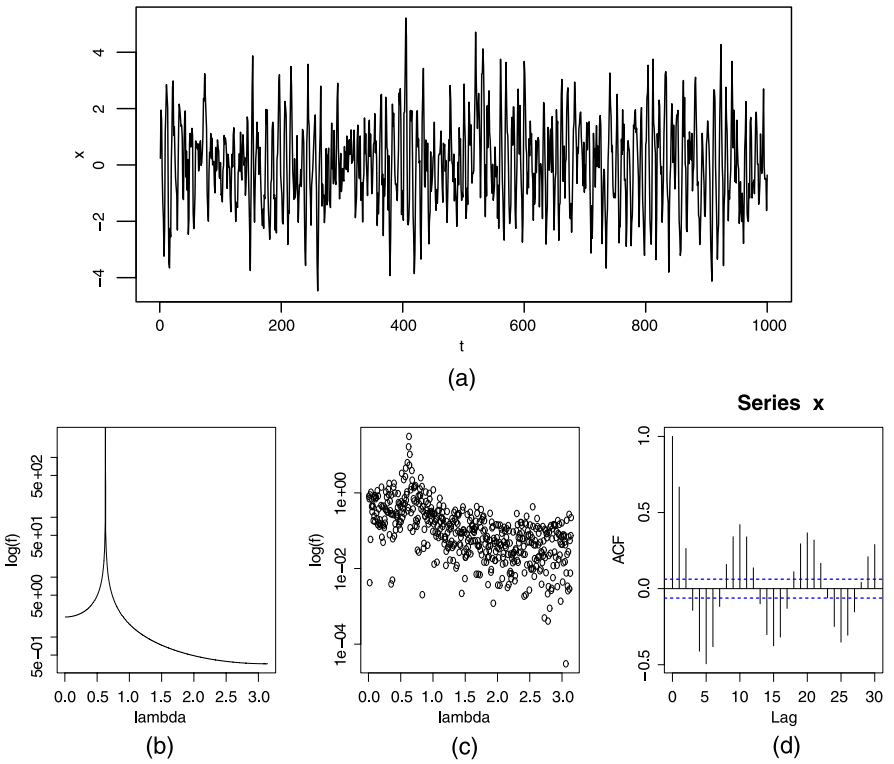

Fig. 5.8 Simulated sample path of a Gegenbauer process with $\lambda_0 = \pi/5$ and $d = 0.4$ (**a**). Also shown are the logarithm of the spectral density (**b**) and of the periodogram (**c**), and the sample autocorrelations (**d**)

For local estimation of ϑ and λ_0, see, e.g. Arteche and Robinson (2000), Hidalgo and Soulier (2004). Note also that instead of one pole one can include an arbitrary finite number of frequencies $0 \leq \lambda_1 < \lambda_2 < \cdots < \pi$ where a pole (or a zero) occurs, by multiplying the right-hand side by the corresponding filters $(1 - 2\cos\lambda_j B + B^2)^{-s/2}$. For further results, see Chung (1996a, 1996b), Ferrara and Guégan (2000, 2001a, 2001b), Giraitis and Leipus (1995), Gray et al. (1994), Guégan (1999, 2000), Olhede et al. (2004), Porter-Hudak (1990), Woodward et al. (1998) and Yajima (1996).

Example 5.16 Figure 5.8(a) shows a simulated sample path of a Gegenbauer process with $\lambda_0 = \pi/5$ and $d = 0.4$. Also shown are the logarithm of the spectral density (Fig. 5.8(b)) and of the periodogram (Fig. 5.8(c)), and the sample autocorrelations (Fig. 5.8(d)). Note that here the logarithm was taken for better visibility; however, a log–log-plot (i.e. taking also the logarithm of λ) is not meaningful because there is no pole at the origin.

5.13 Quantile Estimation

For stationary linear processes with long memory, inference for quantiles follows directly from the corresponding limit theorems discussed in Sect. 4.8. As mentioned there, the main literature includes Dehling and Taqqu (1989b), Ho and Hsing (1996), Wu (2005), Csörgő et al. (2006), Youndjé and Vieu (2006), Csörgő and Kulik (2008a, 2008b), Coeurjolly (2008a, 2008b) among others. Thus, suppose that $X_t = \sum a_j \varepsilon_{t-j}$ is a linear process with $a_j \sim c_a j^{d-1}$ for some $d \in (0, \frac{1}{2})$ and such that the conditions in Theorem 4.33 hold. Denote by F and $p_X = F'$ the marginal distribution and probability density function of X_t, respectively. Also, for a given value of $\alpha \in (0,1)$ we denote by $Q(\alpha) = F^{-1}(\alpha)$ the α-quantile of F and by $Q_{n,X}(\alpha) = F_n^{-1}(\alpha)$ the empirical quantile based on observations $X_1, \ldots, X_n$. As usual, $F_n(x) = n^{-1} \sum 1\{X_t \le x\}$ denotes the empirical distribution function, and the inverse is defined by $F_n^{-1} = \inf\{x : F_n(x) \ge \alpha\}$. As outlined in Sect. 4.8, the empirical quantile is asymptotically equivalent to the sample mean, irrespective of the value of α. Since, for simplicity, we assume the slowly varying function in a_j to be equal to a constant c_a, the spectral density has a pole of the form $f_X(\lambda) \sim c_f |\lambda|^{-2d}$ so that the simplified standardization by $c_f \nu(d)$ applies and we have the functional limiting result

$$n^{\frac{1}{2}-d} \frac{Q_{n,X}(\alpha) - Q(\alpha)}{\sqrt{c_f \nu(d)}} \Longrightarrow Z$$

where convergence is in $D[\alpha_{\text{low}}, \alpha_{\text{up}}]$ equipped with the supremum norm and $0 < \alpha_{\text{low}} < \alpha_{\text{up}} < 1$. This makes inference for quantiles rather simple. Because of convergence in the sup-norm, it is possible to define simultaneous confidence bands for an arbitrary (and even uncountable) number of quantiles. For instance, a 95 % confidence interval for all quantiles between $\alpha_{\text{low}} = 0.005$ and $\alpha_{\text{up}} = 0.995$ can be defined as

$$Q_{n,X}(\alpha) \pm 1.96 \sqrt{c_f \nu(d)} n^{d-\frac{1}{2}}. \tag{5.116}$$

If c_f and d have to be estimated, then exactly the same finite sample corrections as discussed in Sect. 5.2 can be applied, since the standardization is the same as for the sample mean. Formula (5.116) is very much in contrast to the case of i.i.d. observations (and also similar results under short memory) where the asymptotic distribution of $Q_{n,X}(\alpha)$ depends on α and in particular $p_X(\alpha)$. In particular, for i.i.d. observations the asymptotic variance is equal to $\alpha(1-\alpha)/p_X^2(Q(\alpha))$. For instance, if the marginal distribution is standard normal, then under the i.i.d. assumption the asymptotic variance of the median is $\frac{1}{4} 2\pi = \pi/2 \approx 1.57$ whereas for the 5 %-quantile it is about 4.47. In contrast, under long memory the asymptotic variance of both empirical quantiles is the same. It should be noted that the simplicity induced by $Q_{n,X}(\alpha) - Q(\alpha)$ converging to the same random variable is not necessarily good for statistical inference because it also means that for a given data set all quantiles simultaneously either under- or overestimate the corresponding true values. This is illustrated by the following example.

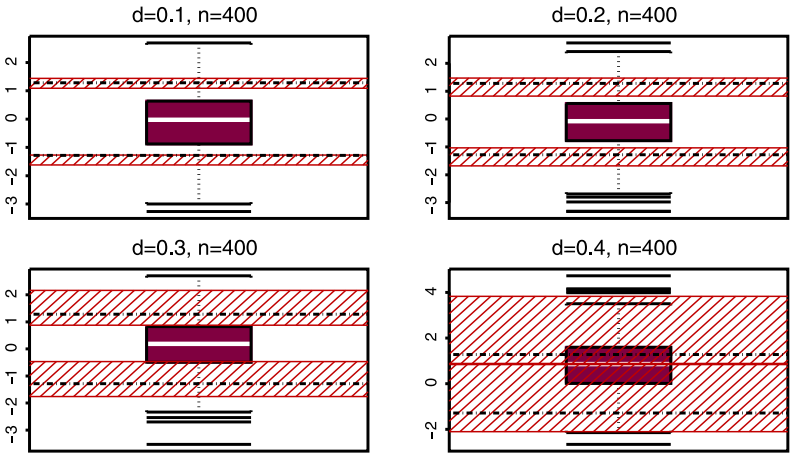

Fig. 5.9 *Boxplots* of simulated observations of a FARIMA$(0, d, 0)$ process of length $n = 400$ with $d = 0.1, 0.2, 0.3$ and 0.4, respectively, together with 95 %-confidence intervals (*shaded areas*) for the 10 %- and 90 %-quantiles. The *dashed lines* represent the true quantiles $Q(0.1)$ and $Q(0.9)$

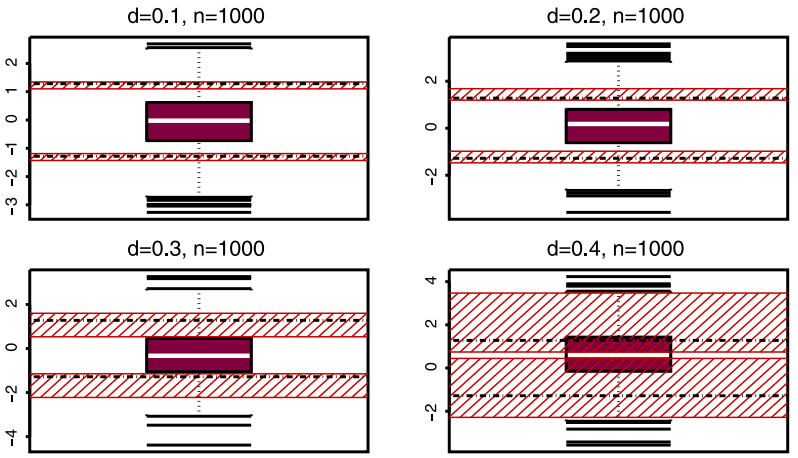

Fig. 5.10 *Boxplots* of observations of a simulated FARIMA$(0, d, 0)$ process of length $n = 1000$ with $d = 0.1, 0.2, 0.3$ and 0.4, respectively, together with 95 %-confidence intervals (*shaded areas*) for the 10 %- and 90 %- quantiles. The *dashed lines* represent the true quantiles $Q(0.1)$ and $Q(0.9)$

Example 5.17 Consider a FARIMA$(0, d, 0)$ process with $\sigma_\varepsilon^2 = 1$, and simultaneous estimation of the 10 %- and 90 %-quantiles. Figures 5.9 and 5.10 display boxplots of observations from one simulated path of length $n = 400$ and 1000, respectively, with $d = 0.1, 0.2, 0.3$ and 0.4. The dashed horizontal lines represent the correct quantiles. The shaded areas are the corresponding 95 % confidence intervals based on the observed path (assuming that c_f and d are known). Note that here the intervals are

of the form

$$Q_{n,X}(\alpha) \pm 1.96\sqrt{\frac{\nu(d)}{2\pi}}n^{d-\frac{1}{2}}$$
$$= Q_{n,X}(\alpha)\frac{1}{2} \pm 1.96\sqrt{\frac{\sin \pi d}{\pi d(2d+1)}\Gamma(1-2d)}n^{d-\frac{1}{2}}.$$

All intervals contain the true values. However, generally there tends to be either over- or underestimation for both quantiles. Moreover, the confidence intervals for $d = 0.4$ are very large, and even overlap for $n = 400$. The reason is the very slow rate of convergence $O_p(n^{-0.1})$.

In applications, quantile estimation is of particular interest when data are not exactly stationary. Extensions of quantile estimation to locally stationary processes were developed, for instance, in Ghosh et al. (1997), Ghosh and Draghicescu (2002a, 2002b), Draghicescu and Ghosh (2003). This will discussed in more detail in Sect. 7.6.

5.14 Density Estimation

5.14.1 Introduction

Here we first recall some standard results for nonparametric density estimation. Suppose we observe $X_1, \ldots, X_n$ generated by a (univariate) stationary process with marginal probability density function p_X. A kernel estimator of p_X at point x_0 is defined by

$$\hat{p}_X(x_0) = \frac{1}{nb}\sum_{t=1}^{n} K\left(\frac{x_0 - X_t}{b}\right) \tag{5.117}$$

where $b > 0$ is the bandwidth and K is a kernel such that $K \geq 0$, $K(u) = K(-u)$ and $\int K(u)\,du = 1$. Often one uses kernels with support $[-1, 1]$. The estimate is simply an average of weights $w_t(x_0) = b^{-1}K((x_0 - X_t)/b)$. For instance, for the rectangular kernel $K(u) = \frac{1}{2}1\{-1 \leq u \leq 1\}$ we consider a neighbourhood $x_0 \pm nb$ of length $2nb$, count the number of observations X_t in this neighbourhood and divide it by the length of the interval. This corresponds to a histogram, with the essential difference that instead of considering disjoint blocks of length $2nb$ we move a window (block) of the same length continuously along the x-axis. To judge the quality of the estimator, we consider the mean squared error

$$MSE(x_0, b) = E\left[\left(\hat{p}_X(x_0) - p_X(x_0)\right)^2\right] = \text{Bias}^2 + \text{Variance}.$$

Alternatively, one may rather look at a global criterion such as the integrated mean squared error

$$IMSE(b) = \int MSE(x, b)\, dx$$

(or a suitably weighted integral). The bias $E[\hat{p}_X(x_0)] - p_X(x_0)$ is smaller the smaller the bandwidth, whereas the variance $\mathrm{var}(\hat{p}_X(x_0))$ decreases for larger values of the bandwidth. This is the standard dilemma in a nonparametric setting. In order to make the bias asymptotically negligible, one needs $b \to 0$ as n tends to infinity. At the same time the variance should tend to zero so that one also needs $nb \to \infty$.

The essential question is how to choose a bandwidth such that it minimizes the MSE or IMSE. In a first step, one derives an asymptotic formula for the MSE. This leads to a formula for an asymptotically optimal bandwidth. Since this formula usually depends on unknown parameters, one finally has to design an adaptive data driven algorithm that estimates the MSE (or IMSE) and the optimal bandwidth.

The bias does not depend on the dependence structure. Suppose that K has support $[-1, 1]$. We write $u_i = (x_0 - X_t)/b$ so that $|u_i| \le 1$ means $x_0 - b \le X_t \le x_0 + b$. An approximate expression for the bias can essentially be derived by a Taylor expansion. Note first that, as $b \to 0$,

$$\begin{aligned} E\big[K(u_i)\big] &= \int K\left(\frac{x_0 - x}{b}\right) p_X(x)\, dx = \int K\left(\frac{x - x_0}{b}\right) p_X(x)\, dx \\ &= b \int K(u)\, p_X(x_0 + bu)\, du \\ &\approx b p_X(x_0) \int K(u)\, du + b^2 p_X'(x_0) \int K(u) u\, du \\ &\quad + \frac{b^3}{2} p_X''(x_0) \int K(u) u^2\, du, \end{aligned}$$

provided that K and p_X are well behaved. By assumptions we have $\int K(u)\, du = 1$, $\int K(u) u\, du = 0$ and $0 < \int K(u) u^2\, du < \infty$. Therefore,

$$\begin{aligned} E\big[\hat{p}_X(x_0)\big] &= \frac{1}{nb} \sum_{t=1}^{n} E\big[K(u_t)\big] = b^{-1} E\big[K(u_1)\big] \\ &\approx p_X(x_0) + \frac{b^2}{2} p_X''(x_0) \int K(u) u^2\, du, \end{aligned}$$

and for the bias we obtain

$$\text{Bias} = \frac{b^2}{2} p_X''(x_0) \int K(u) u^2\, du + o\big(b^2\big).$$

As one can see, the bias is larger in absolute value at points with larger (absolute) curvature. The reason is that at points with high curvature neighbouring values dif-

fer more from the value at x_0 so that averaging over neighbouring values is more harmful.

In contrast to the bias, the variance depends on the autocovariance structure of the process X_t. For uncorrelated data, a similar Taylor expansion leads to the formula

$$\operatorname{Var}\big(\hat{p}_X(x_0)\big) = \frac{1}{nb} p_X(x_0) \int_{-\infty}^{\infty} K^2(u)\,du + o\left(\frac{1}{nb}\right),$$

and hence for the mean squared error we have

$$\begin{aligned} MSE(x_0, b) &= b^4 \left(\frac{1}{2} p_X''(x) \int_{-\infty}^{\infty} u^2 K(u)\,du \right)^2 \\ &\quad + \frac{1}{nb} p_X(x_0) \int_{-\infty}^{\infty} K^2(u)\,du + o\big((nb)^{-1}\big) + o\big(b^2\big) \\ &= \tilde{C}_1(x) b^4 + \tilde{C}_2(x)(nb)^{-1} + o\big((nb)^{-1}\big) + o\big(b^2\big). \end{aligned}$$

Setting the derivative with respect to b equal to zero leads to the asymptotically optimal local bandwidth

$$b_{\text{opt}} = b_{\text{opt}}(x) = C_{\text{opt}} n^{-\frac{1}{5}}$$

with

$$C_{\text{opt}} = C_{\text{opt}}(x) = \left(\frac{1}{4} \tilde{C}_2 / \tilde{C}_1 \right)^{\frac{1}{5}}$$

and an optimal mean squared error of the order $MSE_{\text{opt}} = O(n^{-\frac{4}{5}})$. If, for simplicity, one prefers using a global bandwidth, then one can minimize the integrated MSE,

$$IMSE = \int MSE(x, b) q(x)\,dx = C_1 b^4 + C_2 (nb)^{-1} + o\big((nb)^{-1}\big) + o\big(b^2\big)$$

with $C_i = \int \tilde{C}_i(x) q(x)\,dx$ and q an appropriate weight function. Note that the optimal bandwidth is such that the contributions of the bias and variance to the MSE are of the same order.

The constant in the optimal bandwidth depends on unknown parameters. Therefore, data driven algorithms have been designed to estimate this constant or to obtain a good estimate of the IMSE (see, e.g. Stone 1974; Geisser 1975; Silverman 1986; Bowman 1984). Under short-range dependence and antipersistence, the result remains the same (under suitable regularity conditions), except that C_2 changes and hence the optimal constant C_{opt} is different. Under long memory, the situation changes, however, since the asymptotic behaviour, including the rate of convergence, of the variance depends on whether one uses a sequence of relatively large or small bandwidths. This makes the question of optimal bandwidth choice more complicated and will be discussed in the following section.

An alternative approach that can be used for linear processes is based on the empirical cumulant distribution of the innovation process. This is discussed in Sect. 5.14.3.

5.14.2 Nonparametric Kernel Density Estimation Under LRD

5.14.2.1 Main Results

Here we consider nonparametric kernel density estimation (5.117) under the assumption of long memory. Thus, let $X_1, \ldots, X_n$ be generated by a linear process $X_t = \mu_X + \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ where ε_i are i.i.d. zero mean variables with variance σ_ε^2 and $a_j \sim c_a j^{d-1}$ $(0 < d < \frac{1}{2})$ as $j \to \infty$. Without loss of generality, we will assume $\mu_X = 0$ and $\sigma_X^2 = 1$. The following results can be generalized to the case where c_a is replaced by a slowly varying function L_a. For practical purposes, the case with $L_a \sim c_a$ is usually sufficient.

The first asymptotic results for $\hat{p}_X$ defined in (5.117) were established in Cheng and Robinson (1991) and Csörgő and Mielniczuk (1995a), under the assumption that X_t are subordinated Gaussian variables and for the case of "*large*" bandwidths (see below for a definition). A general result on the asymptotic distribution of $\hat{p}_X(x_0) - p_X(x_0)$ is derived in Wu and Mielniczuk (2002), under the assumption that the bias is asymptotically negligible compared to the variance. Our aim in bandwidth selection is to minimize the mean squared error. The contribution of the optimal bandwidth to the MSE is then automatically of the same order as the variance. We therefore rewrite the limit theorem by considering $\hat{p}_X(x_0) - E[\hat{p}_X(x_0)]$, without any bias condition.

Theorem 5.17 *Let X_t be a linear process with long memory as defined above and denote by $c_X = c_f \nu(d)$ the constant in the asymptotic expression* $\operatorname{var}(\sum X_t) \sim c_X n^{2d+1}$. *Let x_0 be in the interior of the support of p_X, assume that p_X is twice continuously differentiable in a neighbourhood of x_0, denote by Z a standard normal random variable and assume that $b \to 0$ and $nb \to \infty$. Then the following holds*:

- *If $b = o(n^{-2d})$, then*

$$\sqrt{nb}\big(\hat{p}_X(x_0) - E\big[\hat{p}_X(x_0)\big]\big) \xrightarrow{\mathrm{d}} Z\sqrt{p_X(x_0)\int K^2(u)\,du}. \tag{5.118}$$

- *If $b \gg n^{-2d}$ (i.e. $bn^{2d} \to \infty$), then*

$$n^{\frac{1}{2}-d} c_X^{-\frac{1}{2}}\big(\hat{p}_X(x_0) - E\big[\hat{p}_X(x_0)\big]\big) \xrightarrow{\mathrm{d}} p_X'(x_0) Z. \tag{5.119}$$

Proof Let $\mathscr{F}_i = \sigma(\varepsilon_i, \varepsilon_{i-1}, \ldots)$ be the σ-algebra generated by ε_s $(s \le i)$ and denote by p_ε the density of ε_i. We denote by

$$\hat{X}_i = X_i - \varepsilon_i = \sum_{j=1}^{\infty} a_j \varepsilon_{i-j}$$

the best linear forecast of X_i given $\mathscr{F}_{i-1}$ and by $p_{\hat{X}}$ its probability density function. Using the notation

$$\nu_n(X_i) = K\left(\frac{x - X_i}{b}\right),$$

we obtain the decomposition

$$\begin{aligned}
&\hat{p}_X(x_0) - E\big[\hat{p}_X(x_0)\big] \\
&\quad = \frac{1}{nb}\sum_{i=1}^{n}\big\{\nu_n(X_i) - E\big[\nu_n(X_i)|\mathscr{F}_{i-1}\big]\big\} + \frac{1}{nb}\sum_{i=1}^{n} E\big[\nu_n(X_i)|\mathscr{F}_{i-1}\big] - E\big[\hat{p}_X(x_0)\big] \\
&\quad =: R_{n,2} + R_{n,1} \qquad (5.120)
\end{aligned}$$

where

$$E\big[\nu_n(X_i)|\mathscr{F}_{i-1}\big] = \frac{1}{nb}\sum_{i=1}^{n}\int K\left(\frac{x_0 - (\hat{X}_i + u)}{b}\right) p_\varepsilon(u)\, du.$$

We will call (5.120) the M/L-decomposition (marginal/long memory) (see also Sect. 7.2.3). Application of the martingale central limit theorem implies that

$$\sqrt{nb}R_{n,2} \xrightarrow{\mathrm{d}} Z_1\sqrt{p_X(x)\int K^2(u)\, du}. \qquad (5.121)$$

If we assume that X_i are normal, then the term $R_{n,1}$ can be treated using an Hermite polynomial expansion (also called (H)-decomposition in Sect. 7.2.3). However, if we assumed Gaussianity a priori, then there would be no need for nonparametric density estimation. Thus, we apply a different method instead (see Wu and Mielniczuk 2002 for all technical details). We note that $R_{n,1}$ can be written as

$$\zeta_{n,b} := nbR_{n,1} = \sum_{i=1}^{n}\left\{\int K\left(\frac{x_0 - (\hat{X}_i + z)}{b}\right) p_\varepsilon(z)\, dz - \int K\left(\frac{x_0 - x}{b}\right) p_X(x)\, dx\right\}.$$

Applying a change of variables with $w = (x_0 - \hat{X}_i - z)/b$ in the first and $w = (x_0 - x)/b$ in the second integral, we obtain

$$\int K\left(\frac{x_0 - z - \hat{X}_i}{b}\right) p_\varepsilon(z)\, dz - \int K\left(\frac{x_0 - x}{b}\right) p_X(x)\, dx$$

$$= -b \int K(w)\big[p_\varepsilon(x_0 - \hat{X}_i - bw) - p_X(x_0 - bw)\big]\,dw$$

$$= b \int K(w)\big[p_\varepsilon(x_0 - \hat{X}_i + bw) - p_X(x_0 + bw)\big]\,dw$$

and hence

$$\zeta_{n,b} = b \int K(w) S_{n,b}(w)\,dw$$

where

$$S_{n,b}(w) = \sum_{i=1}^{n}\big[p_\varepsilon(x_0 - \hat{X}_i + bw) - p_X(x_0 + bw)\big].$$

Since b tends to zero, a Taylor expansion at $b = 0$ suggests that asymptotically the distribution of $\zeta_{n,b}$ is the same as for

$$\zeta_{n,0} = b \int K(w) S_{n,b}(0)\,dw = b S_{n,b}(0).$$

Now for

$$S_{n,b}(0) = \sum_{i=1}^{n}\big[p_\varepsilon(x_0 - \hat{X}_i) - p_X(x_0)\big]$$

we use the convergence results for empirical processes from Sect. 4.8. Thus, let

$$E_{\hat{X},n}(x) := F_{\hat{X},n}(x) - F_{\hat{X}}(x) = \frac{1}{n}\sum_{i=1}^{n}\big[1\{\hat{X}_i \le x\} - F_{\hat{X}}(x)\big],$$

where $F_{\hat{X}}(x) = P(\hat{X} \le x)$. Then

$$\begin{aligned} S_{n,b}(0) &= \sum_{i=1}^{n}\big[p_\varepsilon(x_0 - \hat{X}_i) - p_X(x_0)\big] \\ &= n \int p_\varepsilon(x_0 - z)\,F_{\hat{X},n}(z) - n p_X(x_0) \\ &= n\left[\int p_\varepsilon(x_0 - z)\,dF_{\hat{X},n}(z) - \int p_\varepsilon(x_0 - z)\,dF_{\hat{X}}(z)\right] \\ &= n \int p_\varepsilon(x_0 - z)\,dE_{\hat{X},n}(z) \\ &= n \int E_{\hat{X},n}(z)\,p'_\varepsilon(x - z)\,dz. \end{aligned}$$

From the reduction principle for empirical processes (4.159) (and $L_X(n) = c_X$), we have

$$nE_{\hat{X},n}(z) = p_{\hat{X}}(z) \sum_{i=1}^{n} \hat{X}_i + o_p\big(n^{\frac{1}{2}+d}\big),$$

uniformly in $z \in \mathbb{R}$. Thus, since p'_ε is integrable,

$$\begin{aligned} S_{n,b}(0) &= \sum_{i=1}^{n} \hat{X}_i \int p_{\hat{X}}(z) p'_\varepsilon(x_0 - z)\, dz + o_p\big(n^{\frac{1}{2}+d}\big) \\ &= p'_X(x_0) \sum_{i=1}^{n} \hat{X}_i + o_p\big(n^{\frac{1}{2}+d}\big) \end{aligned}$$

(note that the slowly varying function is a constant here). Thus, applying the limit Theorem 4.6 for partial sums, we obtain

$$n^{\frac{1}{2}-d_X} c_X^{-\frac{1}{2}} R_{n,1} \xrightarrow{\mathrm{d}} p'_X(x_0) Z. \qquad \square$$

The theorem reveals a "smoothing dichotomy" with a completely different behaviour for "small" and "large" bandwidths, respectively. A similar phenomenon can also be observed in random design nonparametric regression (see Sect. 7.4.8). If b is small then the estimator behaves as if the sequence was i.i.d. For large bandwidths b, long-range dependence influences the variance of the estimator. Note also that the limit in (5.119) degenerates if $p'_X(x) = 0$. If that is the case, then the scaling factor and the limit will change, which leads to a *smoothing trichotomy* (see Wu and Mielniczuk 2002).

To understand the consequences of Theorem 5.17, let us look at the conditions of the two cases more closely. First, note that in (5.118) the rate and even the exact asymptotic distribution of $\sqrt{nb}(\hat{p}_X(x_0) - E[\hat{p}_X(x_0)])$ is *exactly* the same as if the data were i.i.d. Now recall that the mean squared error is the sum of the squared bias (which is always of the order b^4) and the variance. If $b = o(n^{-2d})$, then both terms depend on b and the optimal bandwidth is exactly the same as for i.i.d. data. Thus,

$$b_{\mathrm{opt}} = C_{\mathrm{opt}} n^{-\frac{1}{5}} \tag{5.122}$$

(with C_{opt} as for i.i.d. data) and the optimal MSE of the order $n^{-4/5}$. This is a very nice result because the solution does not depend on the autocorrelation structure (which therefore does not need to be estimated) and the rate is much better than if long memory had played a role. However, it needs to be clarified whether it is always possible to achieve this rate. In other words, if we consider $b = cn^{-\alpha}$ (with $\alpha > 0$), is it possible to choose α such that the bias is of the same order as the variance, under the side condition that $b = o(n^{-2d})$? The side condition implies $\alpha > 2d$. As seen before, the bias–variance condition implies $\alpha = 1/5$. The two conditions together imply $d < 1/10$. Thus, unfortunately the nice result in (5.118) and (5.122) is only

Table 5.3 Overview of rates in kernel density estimation for linear long-memory processes

d	Large b	Optimal b	Small b
$0 \le d < \frac{1}{10}$	$0 < \alpha < \frac{1}{5}$ $\text{Var} \ll \text{Bias}^2$ $MSE = O(n^{-4\alpha})$	$\alpha = \frac{1}{5}$ $\text{Var} \propto \text{Bias}^2$ $MSE = O(n^{-\frac{4}{5}})$	$\alpha > \frac{1}{5}$ $\text{Var} \gg \text{Bias}^2$ $MSE = O(n^{-(1-\alpha)})$
$\frac{1}{10} < d < \frac{1}{2}$	$\alpha < \frac{1}{5} - \frac{1}{2}(d - \frac{1}{10})$ $\text{Var} \ll \text{Bias}^2$ $MSE = O(n^{-4\alpha})$	$\frac{1}{5} - \frac{1}{2}(d - \frac{1}{10}) < \alpha < 2d$ $\text{Var} \propto \text{Bias}^2$ $MSE = O(n^{-(1-2d)})$	$\alpha > 2d$ $\text{Var} \gg \text{Bias}^2$ $MSE = O(n^{-(1-\alpha)})$

applicable for a small range of very weak long-range dependence characterized by $0 < d < \frac{1}{10}$. Note in particular that for $d \uparrow \frac{1}{10}$, the interval $2d < \alpha < \frac{1}{5}$ within which $b^4 = cn^{-4\alpha}$ is of a larger order than the order $1/(nb) = n^{\alpha-1}$ of the variance shrinks to an empty interval. The common feature that is independent of the value of d is that the bandwidth decides whether the bias, the variance or neither of them dominates the mean squared error asymptotically. If the bias is very small, then the MSE is dominated by the variance. If the biased is too large, then the variance does not play any role. However, there is a major difference with respect to the optimal rate of the MSE. For $d < 0.1$ (including short-range dependent and i.i.d. data with $d = 0$), there is *exactly one optimal bandwidth* between the two ranges of "too small" and "too large" bandwidths, namely $b_{\text{opt}} = C_{\text{opt}} n^{-\frac{1}{5}}$. In this optimal case, the contributions of the bias and variance are of the same order. In contrast, for $d > 0.1$, there is a whole *range of bandwidths in between* "too small" and "too large". For all bandwidths in this range, the MSE achieves its optimal rate, and is dominated by the variance. The optimal rate is slower than for $d < 0.1$ and depends on d. However, the good news is that one does not need to estimate one single optimal value of b, since it is sufficient to identify a suitable range from which to choose b. This is illustrated in Fig. 5.14. For $n = 100$ and different values of d, we plot the rate $n^{-\beta}$ of the MSE as a function of $\alpha > 0$. For each value of $d < 0.1$, the curve has a unique minimum at $\alpha = \frac{1}{5}$. For $d > 0.1$, there is an intermediate interval of α-values where $n^{-\beta}$ is constant at its minimal value. This range increases in size as d increases. At the same time, however, the whole curve is shifted upwards for larger values of d because the rate of the optimal MSE deteriorates. An overview of the different rates is also given in Table 5.3. Note in particular that with $d \uparrow \frac{1}{2}$ the range of optimal bandwidths,

$$\frac{1}{4}(1 - 2d) < \alpha < 2d, \tag{5.123}$$

converges to the largest possible range $0 < \alpha < 1$. In other words, the stronger the long memory, the less important the choice of the bandwidth, in the sense that we can choose from a large interval of possible bandwidths without changing the result asymptotically. Loosely speaking, we may also say that under strong long memory the main source of error is the variance. Intuitively, this is due to the fact that under

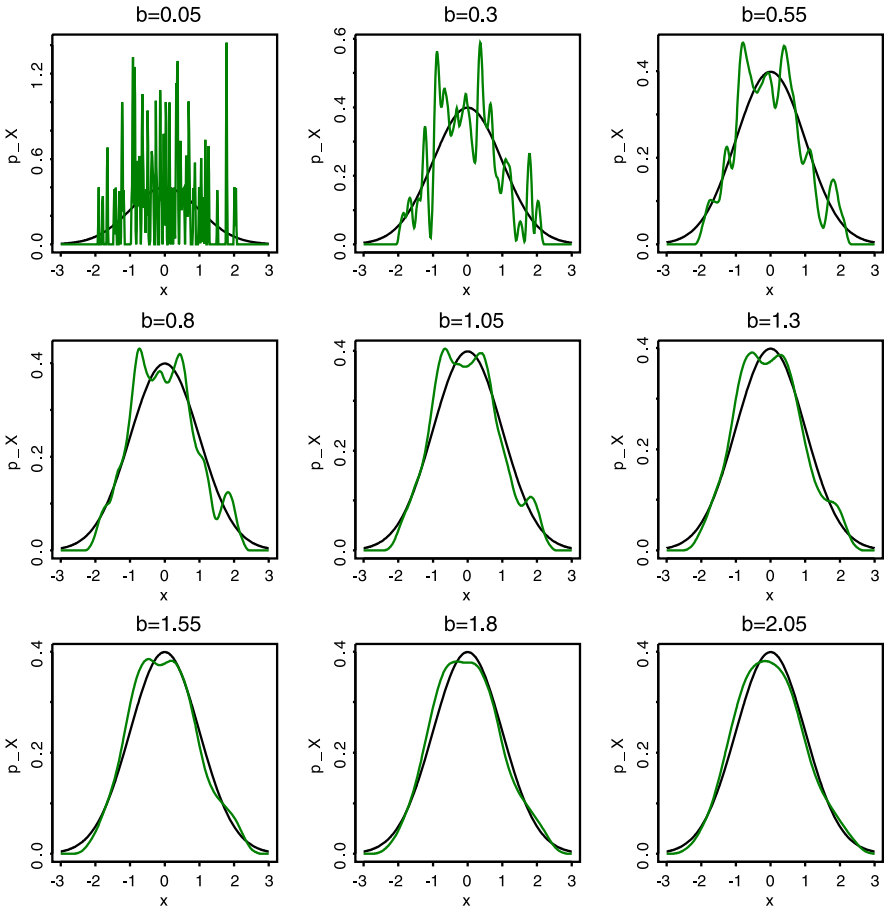

Fig. 5.11 Kernel estimate of the marginal density based on $n = 100$ i.i.d. standard normal variables, together with the true (standard normal) density function and bandwidths b ranging from small (*top left*) to large (*bottom right*)

long memory the whole density tends to be shifted to the left or right randomly because the process is likely to stay on one side of the distribution for extended periods of the time. (Note that this is also related to the notion of "long strange segments" as considered in Sect. 1.3.6.3.) This problem does not show up in the expected value but in the variance. Figures 5.11, 5.12 and 5.13 illustrate this by considering $\hat{p}_X$ for i.i.d. normal observations (Fig. 5.11), and two simulated series of a FARIMA$(0, 0.4, 0)$ process (Figs. 5.12 and 5.13). For the FARIMA process one can observe the typical phenomenon that the whole estimated density is shifted to the left or right, and more concentrated because—due to strong long memory—observations tend stay within a relatively small range for a long time.

The intuitive explanation that under long memory the whole density tends to be shifted to the left or right randomly is supported by a functional limit theorem

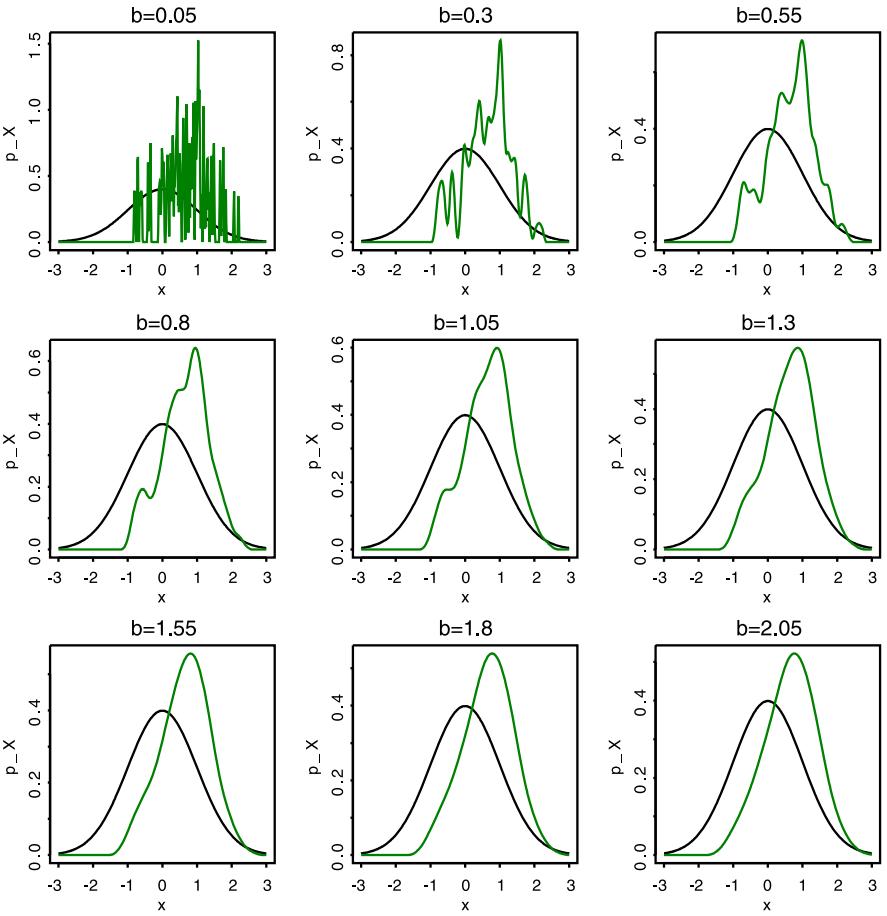

Fig. 5.12 Kernel estimate of the marginal density based on $n = 100$ observations of a standardized FARIMA$(0, 0.4, 0)$ process, together with the true (standard normal) density function and bandwidths b ranging from small (*top left*) to large (*bottom right*)

for large bandwidths first derived in Csörgő and Mielniczuk (1995a). Their result is originally derived under the assumption of Gaussian subordination but can be extended to linear processes. The essential part is that the pointwise convergence in the second part of Theorem 5.17 (i.e. for large densities $b \gg n^{-2d}$) holds uniformly, i.e.

$$\sup_{x_0 \in \mathbb{R}} \left| n^{\frac{1}{2}-d} c_X^{-\frac{1}{2}} \left(\hat{p}_X(x_0) - E\left[\hat{p}_X(x_0)\right] \right) - p'_X(x_0) Z \right| \underset{p}{\to} 0. \tag{5.124}$$

(Under Gaussian subordination such that $1\{X_i \le x\} - F_X(x)$ has Hermite rank $m \ge 2$, Z has to be replaced by an Hermite-variable of rank m and p'_X by another constant obtained from the Hermite expansion.) Csörgő and Mielniczuk call this "globalization" of nonparametric density estimation because the random variable Z

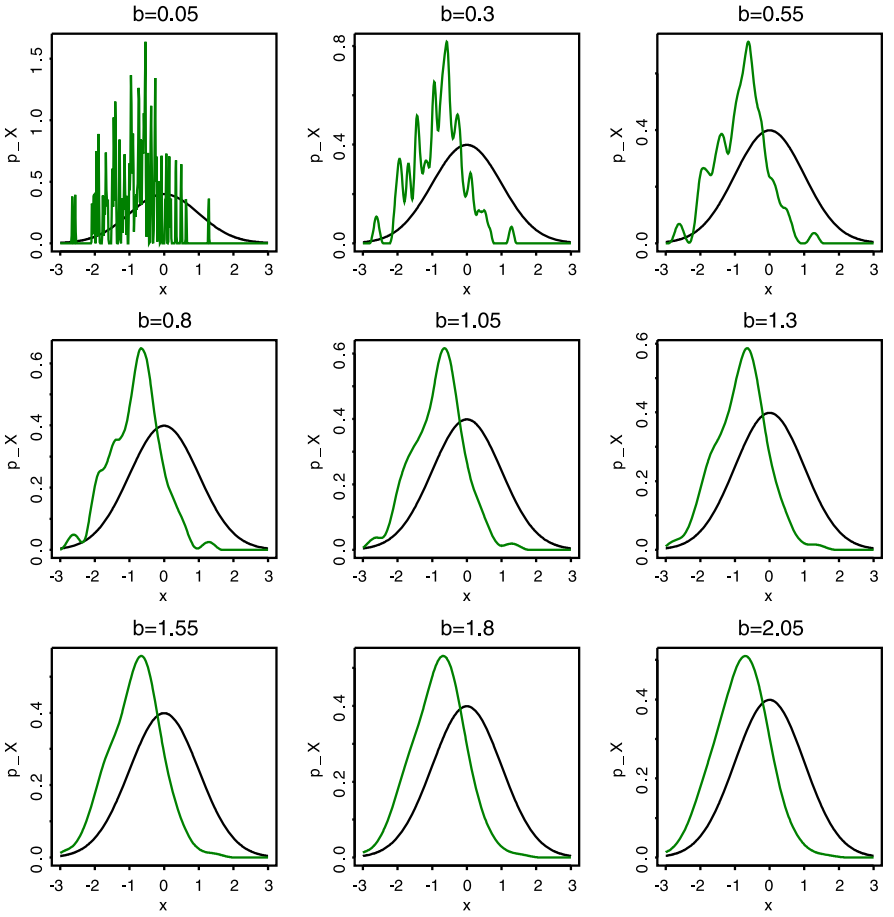

Fig. 5.13 Kernel estimate of the marginal density based on $n = 100$ observations of a standardized FARIMA$(0, 0.4, 0)$ process, together with the true (standard normal) density function and bandwidths b ranging from small (*top left*) to large (*bottom right*)

determines for the whole estimated density function $\hat{p}_X(x_0)$ $(x_o \in \mathbb{R})$ whether and how far it is lower or higher than the whole curve $p_X(x_0)$. More specifically, if $Z > 0$ and b is such that the bias is negligible (see above), then we have asymptotically

$$\hat{p}_X(x_0) > p_X(x_0) \quad \text{for } x_0 \text{ with } p'_X(x_0) > 0$$

and

$$\hat{p}_X(x_0) < p_X(x_0) \quad \text{for } x_0 \text{ with } p'_X(x_0) < 0.$$

For $Z < 0$, the opposite inequalities hold.

Finally, note that a refined optimization of the bandwidth can be considered by taking into account the higher order term $b^2 n^{2d-1} c_X$ from the variance (see

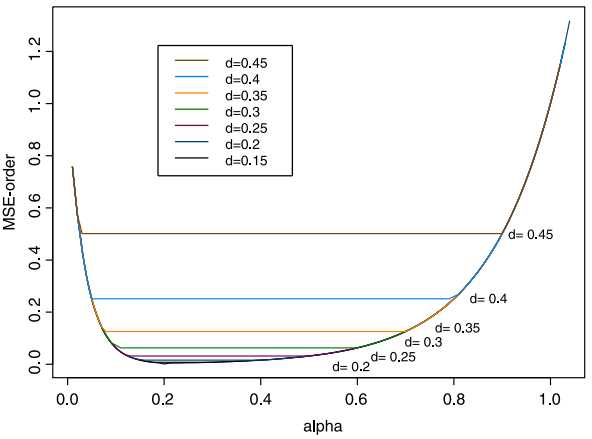

Fig. 5.14 For $n = 100$ and different values of d, the plot shows the rate $n^{-\beta}$ of the MSE as a function of $\alpha > 0$ where $b = cn^{-\alpha}$ is the bandwidth

Claeskens and Hall 2002). The IMSE is then of the asymptotic form

$$IMSE(b) \sim C_1(nb)^{-1} + C_2 b^4 + n^{-(1-2d_X)} c_X + b^2 n^{2d-1} c_X .$$

Although for $d > \frac{1}{10}$ asymptotically negligible, the optimal rate of b may be chosen by minimization of the second-order terms with respect to b. This leads to $b_{\text{opt}} = cn^{-\frac{1}{5}}$ for $d < 0.3$ and $b_{\text{opt}} = cn^{-\frac{2}{3}d}$ for $d > 0.3$. This bandwidth choice involves unknown parameters (in c for $d < 0.3$, and in c *and* the rate for $d > 0.3$), including the unknown density function itself. These have to be replaced by suitable estimates. In particular, a good method for estimating the IMSE has to be applied. This turns out to be quite difficult. The reason can be summarized briefly as follows. Suppose that we are able to calculate (or approximate with high accuracy) the actual error

$$ISE(b) = \int \left(\hat{p}_X(x) - p_X(x)\right)^2 dx$$

as a function of b. Minimizing this quantity yields an estimated optimal bandwidth $\hat{b}_{\text{opt}}$. It turns out, however, that, in spite of the ideal situation with $ISE(b)$ known, $\hat{b}_{\text{opt}}$ converges to b_{opt} in probability only if $d < 0.1$ (see Claeskens and Hall 2002 where asymptotic results are given in a Gaussian context). Therefore, for instance, standard cross-validation cannot be applied to processes with $d > 0.1$.

Example 5.18 We consider the monthly average discharge series of the Danube at Hofkirchen (1901–1984) introduced in Sect. 1.2. Figure 5.15(a) shows a histogram of the series together with a kernel density fit $\hat{p}_X(x)$ and a simultaneous 95 % confidence band (Fig. 5.15(a)), based on the second part of Theorem 5.17 and (5.124). Figure 5.15(b) shows a kernel estimate of the first derivative $p'_X(x)$ used to calculate the confidence band. Note that the width (or rather height) of the band reduces to zero where $p'_X(x) = 0$. This should not be interpreted as absolute certainty about the value of the density at that point but rather as a limitation of the asymptotic approach based on the first order approximation in (5.124).

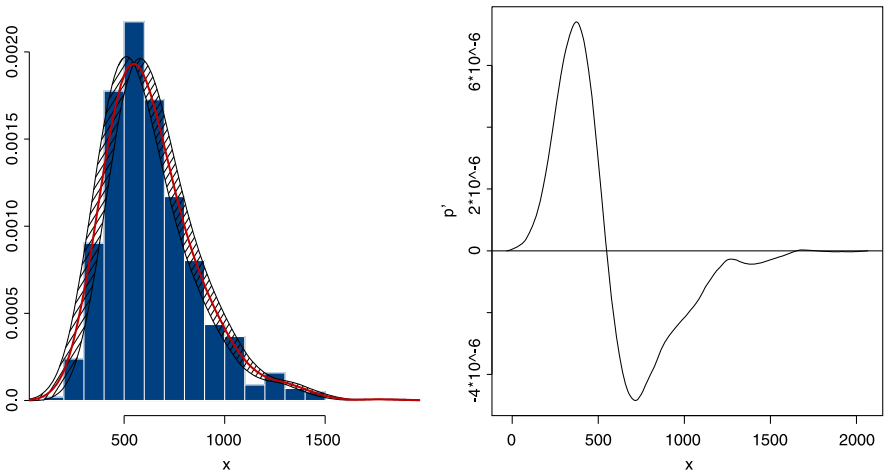

Fig. 5.15 Histogram of the monthly average discharge series of the Danube at Hofkirchen (1901–1984), together with a kernel density fit $\hat{p}_X(x)$ and a simultaneous 95 % confidence band (**a**). Figure (**b**) shows a kernel estimate of the first derivative $p'_X(x)$

We conclude this section with some bibliographical comments. Wu and Mielniczuk (2002) present general limit theorems for $\hat{p}_X(x) - E[\hat{p}_X(x)]$ in the presence of LRD. Previous asymptotic results can be found, for instance, in Cheng and Robinson (1991), Csörgő and Mielniczuk (1995a) (for subordinated processes and large bandwidths). The smoothing dichotomy was shown for the first time in Ho (1996) for subordinated Gaussian and in Honda (2000) for linear processes. See also Hidalgo (1997) and Gajek and Mielniczuk (1999) for multivariate extensions. Asymptotic results for kernel density estimation in an errors-in-variables setting were derived in Kulik (2008b). Hall and Hart (1990a) were the first to establish the formula for the mean squared error of the kernel density estimator; see also Mielniczuk (1997) and Estévez and Vieu (2003). Properties of empirical bandwidth choice were studied in Hall et al. (1995b) and Claeskens and Hall (2002).

5.14.3 Density Estimation Based on the Cumulant Generating Function

An alternative approach to estimating the marginal distribution of a linear process $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ is suggested in Ghosh and Beran (2006). The idea is to exploit linearity directly using the cumulative distribution function. The main advantage of linearity is that, once the coefficients a_j are given, the marginal distribution of X_t is fully determined by the distribution of ε_t. Moreover, from Sects. 5.5 and 5.9 we know that, under some regularity conditions, the coefficients a_j may be estimated with good accuracy. It may therefore be possible to obtain fairly good estimates

$\hat{\varepsilon}_t$ of the innovations ε_t, though a detailed analysis would be required in specific situations.

We consider the case of long memory with $a_j \sim c_a j^{d-1}$ (as $j \to \infty$) for some $0 < d < \frac{1}{2}$, and assume that the moment generating function $m_\varepsilon(u) = E[\exp(u\varepsilon)]$ of ε exists in an open neighbourhood $-\delta < u < \delta$ of the origin. Let $m_X(u)$ be the moment generating function of X_t and denote by $\ell_\varepsilon(u) = \log m_\varepsilon(u)$ and $\ell_X(u) = \log m_X(u)$ the corresponding cumulant generating functions. Then, due to independence of the ε's, we have

$$\ell_X(u) = \sum_{j=0}^{\infty} \ell_\varepsilon(ua_j). \tag{5.125}$$

Suppose now that ε_t $(t = 1, \ldots, n)$ are known (or estimated with sufficient accuracy). Then the moment generating function of ε can be estimated by the empirical moment generating function

$$\hat{m}_\varepsilon(u) = n^{-1} \sum_{t=1}^{n} \exp(u\varepsilon_t).$$

For the cumulant generating function of ε, we then have the estimate

$$\hat{\ell}_\varepsilon(u) = \log \hat{m}_\varepsilon(u).$$

Using (5.125), we then may estimate the corresponding quantities for X_t by

$$\hat{m}_{X,\text{linear}}(u) = \prod_{j=0}^{N_n} \hat{m}_\varepsilon(ua_j), \qquad \hat{\ell}_{X,\text{linear}}(u) = \sum_{j=0}^{N_n} \hat{\ell}_\varepsilon(ua_j) \tag{5.126}$$

where $N_n \to \infty$. (Note that in general, setting $N_n = \infty$ is not a viable option because the variance becomes infinite.) An estimate of the density function p_X can then be obtained by Laplace inversion of $\hat{m}_X$. Note that an analogous approach can be based on the characteristic function which would have the advantage that no moment conditions are required (see, e.g. Feuerverger and Mureika 1977; Csörgő 1981, 1986; Murota and Takeuchi 1981; Ghosh and Ruymgaart 1992; Gürtler and Henze 2000 for asymptotic results and ideas based on the empirical characteristic function in the i.i.d. context). For limit theorems and statistical methods based on the empirical moment generating function in the i.i.d. setting, see, e.g. Csörgő (1982), Epps et al. (1982), Feuerverger (1989), Baringhaus and Henze (1991, 1992), Ghosh (1996), Ghosh and Beran (2000, 2006), Kalliorasa et al. (2006).

To see whether using the linear structure may improve estimation, one can compare the corresponding mean squared errors

$$MSE\big(\hat{\ell}_X(u)\big) = E\big[\big(\hat{\ell}_X(u) - \ell_X(u)\big)^2\big],$$
$$MSE\big(\hat{\ell}_{X,\text{linear}}(u)\big) = E\big[\big(\hat{\ell}_{X,\text{linear}}(u) - \ell_X(u)\big)^2\big]$$

where

$$\hat{\ell}_X(u) = \log \hat{m}_X(u), \quad \hat{m}_X(u) = n^{-1} \sum_{t=1}^{n} \exp(uX_t).$$

For $\hat{\ell}_X(u)$ the asymptotic variance and distribution follows directly from limit theorems for sums discussed in Sect. 4.2. In particular, we have

$$\operatorname{var}\big(\hat{\ell}_X(u)\big) = w(u)n^{2d-1} + o_p\big(n^{2d-1}\big)$$

where

$$w(u) = \frac{\sigma_\varepsilon^2 c_a^2 u^2}{d(d+1)} \left[\int_0^\infty x^{d-1}(1+x)^{d-1}\,dx - \frac{1}{2}(1-2d)^{-1} \right].$$

Since the bias is of the order $O(n^{2d-1})$, its square is asymptotically negligible and

$$MSE\big(\hat{\ell}_X(u)\big) = w(u)n^{2d-1} + o\big(n^{2d-1}\big). \tag{5.127}$$

The bias of $\hat{\ell}_{X,\text{linear}}(u)$ depends on N_n, namely

$$E\big[\hat{\ell}_{X,\text{linear}}(u)\big] - \ell_X(u) = B(u)N_n^{2d-1} + o\big(N_n^{2d-1}\big)$$

with

$$B(u) = -\frac{\sigma_\varepsilon^2 c_a^2 u^2}{2(1-2d)}.$$

The variance is of the form

$$\operatorname{var}\big(\hat{\ell}_{X,\text{linear}}(u)\big) = N_n^{2d} n^{-1} D(u)$$

with

$$D(t) = \frac{\sigma_\varepsilon^2 c_a^2 u^2}{d^2}.$$

Thus, the MSE can be approximated asymptotically by

$$MSE\big(\hat{\ell}_{X,\text{linear}}(u)\big) \approx B^2(u)N_n^{4d-2} + D(u)N_n^{2d}n^{-1}. \tag{5.128}$$

The asymptotically optimal choice of N_n is therefore given by

$$N_n = C_{\text{opt}} n^{\frac{1}{2-2d}}$$

with

$$C_{\text{opt}} = \left(\frac{c_a^2 u^2 d}{4(1-2d)} \right)^{\frac{1}{2-2d}}.$$

The optimal MSE is then of the order

$$MSE_{\text{opt}}\big(\hat{\ell}_{X,\text{linear}}(u)\big) = O\big(n^{\frac{2d-1}{1-d}}\big).$$

Comparing (5.127) with (5.128) and assuming that $N_n = cn^\alpha$ for some $\alpha > 0$, we see that the first MSE is of a larger order than the second one if $\alpha(4d-2) < 2d-1$ and $2\alpha d - 1 < 2d - 1$. This is equivalent to

$$\frac{1}{2} < \alpha < 1.$$

Thus, we have the following result (Ghosh and Beran 2006).

Theorem 5.18 *Let X_t be a linear process as defined above and $N_n = cn^\alpha$ for some $\frac{1}{2} < \alpha < 1$. Then there are constants $0 < \delta, q(u) < \infty$ such that*

$$\lim_{n\to\infty} n^{-\delta}\left\{\frac{MSE(\hat{\ell}_X(u))}{MSE(\hat{\ell}_{X,\text{linear}}(u))}\right\} = q(u).$$

The result means that, as long as N_n tends to infinity at a hyperbolic rate that is faster than $\sqrt{n}$ but slower than n, the estimator $\hat{\ell}_X(u)$ has asymptotically an infinitely larger MSE than $\hat{\ell}_{X,\text{linear}}(u)$. The general reason is that $\hat{\ell}_{X,\text{linear}}(u)$ exploits the additional information of linearity and the good convergence of $\hat{\ell}_\varepsilon$. By limiting the number of terms in the sum to $N_n = o(n)$, the variance is kept low. At the same time, the bias is controlled by the condition $N_n \gg \sqrt{n}$. The optimal choice of N_n balances the bias and variance in the MSE. Note that in the case of short memory with exponentially decaying coefficients a_j, the possibility of balancing bias and variance disappears because the variance dominates asymptotically as long as $N_n \to \infty$. It is therefore no longer possible to improve the rate of convergence by a smart choice of the sequence N_n.

5.15 Tail Index Estimation

Suppose $X_t \in \mathbb{R}$ ($t \in \mathbb{N}$) have a marginal distribution F_X such that

$$\lim_{x\to-\infty} |x|^\alpha F_X(x) = C \tag{5.129}$$

for some finite constant. Consistent estimation of the tail index α is possible under fairly general conditions (see, e.g. Embrechts et al. 1997), for instance, by using the property that equation (5.129) implies conditional Pareto type behaviour in the sense that $P(X > c + x | X > c) \sim x^{-\alpha}$ as $c \to \infty$ (for the Pareto distribution with density $\alpha x^{-\alpha-1}$ ($x > 1$) this relation is true exactly for any $c > 1$). Best known is the classical Hill estimator discussed previously (see, e.g. Hill 1975) that makes use of the Pareto approximation for a certain number k_n of upper order statistics with $k_n \to \infty$

but $k_n/n \to 0$. For the extended literature on such methods, mainly in the i.i.d. or weakly dependent setting, see, e.g. Mason (1982), Hall (1982), Davis and Resnick (1984), Csörgő and Mason (1985), Csörgő et al. (1985), Häusler and Teugels (1985), Deheuvels et al. (1988), Csörgő and Viharos (1997), De Haan and Peng (1998), De Haan and Resnick (1998), Hsing (1991), Resnick and Starica (1995, 1998), Dekkers and de Haan (1989), Rootzén et al. (1998), Drees (2000), Hall and Welsh (1984), Hall (1990), Drees (1998), Beirlant et al. (2006), Resnick (1997, 2007). We are not aware of any results for the Hill estimator in the case of linear long-memory models. Since k_n diverges at a slower rate than the number of observations, consistent estimation of α comes at the price of a slower rate of convergence. For instance, for i.i.d. data, the variance of $\hat{\alpha}$ is asymptotically proportional to k_n^{-1} instead of n^{-1}. A further problem is that for a given data set with a fixed sample size n and unknown marginal distribution, it is difficult to decide which concrete value of k_n to choose (for instance, to minimize the mean squared error of $\hat{\alpha}$). This makes estimation of α quite unreliable for small to moderate sample sizes. As an alternative, Beran and Schell (2010) proposed a method in the spirit of robust statistics (also see Beran 1997). The approach is consistent under an ideal "central" model, and the asymptotic bias is bounded under deviations from this model. At the same time, the variance of $\hat{\alpha}$ achieves the parametric rate. To be more specific, suppose, for instance, that the "ideal model" has the Pareto distribution. If X_t $(t = 1, 2, \ldots, n)$ are i.i.d. and *exactly* Pareto distributed with $\alpha = \alpha_0$, then the derivative of the log-likelihood function with respect to α is equal to $n\alpha^{-1} - \sum_{t=1}^{n} \log X_t$, so that the maximum likelihood estimator of α_0 can be understood as a solution of the equation

$$\alpha \sum_{t=1}^{n} \log X_t - n = \sum_{t=1}^{n} [\alpha \log X_t - 1] = 0. \tag{5.130}$$

Robustness against deviations from the Pareto distribution can be achieved by bounding the score function $\alpha \log X_t - 1$. The simplest version of this idea is to replace $\alpha \log X_t - 1$ in (5.130) by

$$\psi_{v,u}(x, \alpha) = \big[\alpha \log(x) - 1\big]_v^u - \mathrm{E}\big\{\big[\alpha \log(X) - 1\big]_v^u\big\} \tag{5.131}$$

$$= \big[\alpha \log(x) - 1\big]_v^u - C(\alpha; v, u) \tag{5.132}$$

where $[y]_v^u = \max\{v, \min(y, u)\}$ and the expectation is taken with respect to the Pareto distribution. Note that, by subtracting $C(\alpha; u, v)$, we make sure that $\hat{\alpha}$ is consistent under the Pareto assumption. If the true distribution is not exactly Pareto, then $\hat{\alpha}$ generally does not converge to the true value of α_0. However, if $\psi_{v,u}$ is bounded sufficiently in the range of x-values where deviations from the Pareto distribution is most noticeable, then the asymptotic bias of $\hat{\alpha}$ is small (see, e.g. Huber 1981; Hampel et al. 1986). At the same time, the variance of $\hat{\alpha}$ is proportional to n^{-1} because all observed values are used. For small sample sizes, this robust procedure therefore tends to have a smaller mean squared error than Hill type estimators (for illustrative examples, see Beran and Schell 2010). Note that in view of the validity of the Pareto approximation for large conditional quantiles, the most important

deviations occur for small quantiles. The lower truncation parameter v is therefore more important. Moreover, a possible modification of $\psi_{v,u}$ is to use more narrow truncation intervals $[v, u]$ for smaller quantiles.

In the i.i.d. setting, the asymptotic distribution of the estimator based on $\psi_{v,u}$ can be obtained by standard approximations for M-estimators (Beran and Schell 2010; Serfling 1980). Here, we consider the case of a long-memory series with long-tailed marginals. Although the joint log-likelihood of $X_1, \ldots, X_n$ is much more complicated, the M-estimator above can still be applied. Nothing changes with respect to the bias. However, limit theorems for the empirical process summarized in Sect. 4.8.4 imply a completely different asymptotic distribution of $\hat{\alpha} - \mathrm{E}(\hat{\alpha})$. The following results are from Beran et al. (2012).

First of all, the process X_t has to be defined appropriately. The simplest model is a linear process with i.i.d. symmetric innovations ε_t. Thus, let

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{5.133}$$

where ε_t are i.i.d. symmetric (standard) α-stable (SαS) with characteristic function $\varphi_\varepsilon(u) = E[\exp(iu\varepsilon)] = \exp(-|u|^\alpha)$ for some $\alpha = \alpha_0$ $(1 < \alpha_0 < 2)$ and weights a_j such that

$$a_j \sim c_a j^{d-1} \tag{5.134}$$

as $j \to \infty$, with $c_a \neq 0$ and

$$0 < d < 1 - \alpha_0^{-1}. \tag{5.135}$$

Note that X_t inherits the tail index from the innovation process ε_t. Since the distribution F_ε of ε_t is (standard) SαS, we have

$$\lim_{x \to -\infty} |x|^\alpha F_\varepsilon(x) = \lim_{x \to \infty} x^\alpha \big(1 - F_\varepsilon(x)\big) = C_\alpha \tag{5.136}$$

with

$$C_\alpha = \frac{\sin(\frac{\pi\alpha}{2})\Gamma(\alpha)}{\pi} = \frac{1-\alpha}{2\Gamma(2-\alpha)\cos(\frac{\pi\alpha}{2})} \tag{5.137}$$

(see Nolan 2011, Theorem 1.12; Samorodnitsky and Taqqu 1994). The assumption on d implies that $\sum |a_j| = \infty$, $A_\alpha := \sum |a_j|^\alpha < \infty$, and the process X_t is well-defined (in the sense of convergence in probability) with a marginal distribution F_X satisfying the tail condition

$$\lim_{x \to -\infty} |x|^\alpha F_X(x) = \lim_{x \to \infty} x^\alpha \big(1 - F_X(x)\big) = C_\alpha A_\alpha. \tag{5.138}$$

More exactly, F_X is $S\alpha S$ with scale parameter $\gamma = A_\alpha^{1/\alpha}$ (we will use the notation $X_t \sim S_\alpha(0, \gamma, 0)$). This can be seen by considering the characteristic function

$$\begin{aligned}\varphi_X(u) &= E\left[\exp\left(iu\sum_{j=0}^{\infty} a_j\varepsilon_{t-j}\right)\right] = \prod_{j=0}^{\infty} E\big[\exp(iua_j\varepsilon_{t-j})\big] \\ &= \prod_{j=0}^{\infty}\varphi_\varepsilon(a_j u) = \prod_{j=0}^{\infty}\exp\big(-|a_j|^\alpha|u|^\alpha\big) \\ &= \exp\left(-|u|^\alpha\sum_{j=0}^{\infty}|a_j|^\alpha\right) = \exp\big(-\gamma^\alpha|u|^\alpha\big) \\ &= \varphi_{\gamma\varepsilon}(u).\end{aligned}$$

Then, as discussed in detail in Sect. 4.8.4, we have

$$n^{1-H}\big(F_n(x) - F_X(x)\big) \Rightarrow \frac{c_a}{d}p_X(x)\tilde{Z}_{H,\alpha}(1),$$

where $H = d + \alpha^{-1}$, $p_X(x) = F_X'(x)$ and $\tilde{Z}_{H,\alpha}(1)$ is a symmetric α-stable random variable with scale η, where

$$\eta = \left(\int_{-\infty}^{1}\big\{(1-v)^d - (-v)_+^d\big\}^\alpha dv\right)^{1/\alpha}.$$

Here we use the notation $x^+ = \max(0, x)$ and $x^- = -\min(0, x)$, and $\Rightarrow_{D(\overline{\mathbb{R}})}$ for weak convergence of random processes in the space of càdlàg functions on $\overline{\mathbb{R}} = [-\infty, \infty]$ under the sup-norm metric.

Since $X_t \sim S_\alpha(0, A_\alpha^{1/\alpha}, 0)$, the standardized variables $Y_t = X_t/A_\alpha^{1/\alpha}$ are standard SαS (we will write in short $X_t/A_\alpha^{1/\alpha} \sim$ SαS). It is well-known that for a SαS random variable Y the Pareto approximation for the conditional distribution given $Y > c$ holds, i.e. $P(Y > c + x|Y > c) \sim x^{-\alpha}$ $(c \to \infty)$, and the analogous statement can be made for the left tail. The M-estimator introduced above for i.i.d. data can therefore be used for the right and left tail separately as follows. For the right tail, we define, for any $-\infty \le v < u \le \infty$,

$$\begin{aligned}\psi_{v,u}^+(x,\alpha) &= \big[\alpha\log(x) - 1\big]_v^u 1\{x > 0\} - E\big(\big[\alpha\log(X_t) - 1\big]_v^u 1\{X_t > 0\}\big) \quad (5.139)\\ &= \big[\alpha\log(x) - 1\big]_v^u 1\{x > 0\} - C(\alpha, v, u). \quad (5.140)\end{aligned}$$

The analogous function for the left tail is

$$\begin{aligned}\psi_{v,u}^-(x,\alpha) &= \big[\alpha\log(-x) - 1\big]_v^u 1\{x < 0\} - E\big(\big[\alpha\log(-X_t) - 1\big]_v^u 1\{X_t < 0\}\big)\\ &= \big[\alpha\log(-x) - 1\big]_v^u 1\{x < 0\} - C(\alpha, v, u). \quad (5.141)\end{aligned}$$

Note that $\psi_{v,u}^-(x,\alpha) = \psi_{v,u}^+(-x,\alpha)$. An estimator T_n^+ of α can now be defined by setting $T_n^+ = [\tau_0]_1^2 = \max\{1, \min(2, \tau_0)\}$ where τ_0 solves the equation

$$\lambda_{F_n}^+(\tau) = \sum_{t=1}^{n} \psi_{v,u}^+(X_t, \tau) = 0. \tag{5.142}$$

The analogous definition is used for the left-tail estimator T_n^- with $\psi_{v,u}^+$ replaced by $\psi_{v,u}^-$. Note that by definition T_n^+ and T_n^- are functionals of the empirical distribution function F_n. The corresponding functionals of F_X are solutions of the equations

$$\lambda_F^{\pm}(\tau) = \int_{\mathbb{R}} \psi_{v,u}^{\pm}(x,\tau)\, dF_X(x) = 0. \tag{5.143}$$

In general, the constant $C(\alpha, v, u)$ has to be evaluated numerically. In the special case where $v = -\infty$ and $u = \infty$, we have

$$C(\alpha, -\infty, \infty) = \frac{C_e(1-\alpha) + \log A - 1}{2} \tag{5.144}$$

with

$$E\left(\log X_t^+\right) = E\left(\log X_t^-\right) = \frac{C_e}{2}\left(\frac{1}{\alpha} - 1\right) + \frac{1}{2\alpha}\log A_\alpha \tag{5.145}$$

and

$$C_e = \lim_{n\to\infty}\left(\sum_{i=1}^{n}\frac{1}{i} - \log n\right) = 0.577215\ldots \tag{5.146}$$

being the Euler constant (Zolotarev 1986, p. 215). The results in Sect. 4.173 imply

$$n^{-H}\sum_{t=1}^{n}\psi_{v,u}^+(X_t,\alpha) \underset{d}{\to} -c_a h^+ \tilde{Z}_{H,\alpha}(1), \tag{5.147}$$

$$n^{-H}\sum_{t=1}^{n}\psi_{v,u}^-(X_t,\alpha) \underset{d}{\to} -c_a h^- \tilde{Z}_{H,\alpha}(1), \tag{5.148}$$

where $H = d + \frac{1}{\alpha}$ is the self-similarity parameter, $\tilde{Z}_{H,\alpha}(1)$ is in both cases the same standard SαS random variable with the scale η and

$$\begin{aligned} h^{\pm} &= -\int_{\mathbb{R}} \psi_{v,u}^{\pm}(x,\alpha)\, p_X{}'(x)\, dx = \int_{\mathbb{R}} p_X(x)\, d\psi_{v,u}^{\pm}(x,\alpha) \\ &= \int_{\mathbb{R}} p_X(x)\frac{\partial}{\partial x}\psi_{v,u}^{\pm}(x,\alpha)\, dx. \end{aligned} \tag{5.149}$$

The asymptotic distribution of T_n^+ and T_n^- then essentially follows by standard arguments for M-estimators:

$$n^{1-H}\big(T_n^{\pm} - \alpha\big) \underset{d}{\to} \frac{h^{\pm}}{\mu^{\pm}(\alpha)} c_a \tilde{Z}_{H,\alpha}(1)$$

where

$$\mu^{\pm}(\alpha) = E\big[\dot{\psi}_{v,u}^{\pm}(X_t, \alpha)\big] = \int_{\mathbb{R}} \dot{\psi}_{v,u}^{\pm}(x, \alpha) p_X(x)\, dx \tag{5.150}$$

and $\dot{\psi}_{v,u}^{\pm}(x, \alpha) = \partial/\partial\alpha\, \psi_{v,u}^{\pm}(x, \alpha)$. Because of the symmetry of the underlying SαS-distribution, there is a direct relationship between h^+ and h^-, and μ^+ and μ^-, respectively. Note that $\psi_{v,u}^{+}(x, \alpha) = \psi_{v,u}^{-}(-x, \alpha)$, $p_X(x)$ is an even and $p_X{}'(x)$ an odd function. This implies

$$\begin{aligned} h^- &= -\int_{-\infty}^{\infty} \psi_{v,u}^{-}(x, \alpha) p_X{}'(x)\, dx = -\int_{-\infty}^{\infty} \psi_{v,u}^{+}(-x, \alpha) p_X{}'(x)\, dx \\ &= -\int_{-\infty}^{\infty} \psi_{v,u}^{+}(x, \alpha) p_X{}'(-x)\, dx = \int_{-\infty}^{\infty} \psi_{v,u}^{+}(x, \alpha) p_X{}'(x)\, dx \\ &= -h^+ \end{aligned}$$

and

$$\mu^+(\alpha) = E\big[\dot{\psi}_{v,u}^{+}(X_t, \alpha)\big] = \int_0^{\infty} \dot{\psi}_{v,u}^{+}(x, \alpha) p_X(x)\, dx \tag{5.151}$$

$$= \int_{-\infty}^{0} \dot{\psi}_{v,u}^{+}(-x, \alpha) p_X(x)\, dx = \int_{-\infty}^{0} \dot{\psi}_{v,u}^{-}(x, \alpha) p_X(x)\, dx \tag{5.152}$$

$$= \mu^-(\alpha). \tag{5.153}$$

Thus,

$$n^{1-H}\big(T_n^{+} - \alpha\big) \underset{d}{\to} \frac{h^{+}}{\mu^{+}(\alpha)} c_a \tilde{Z}_{H,\alpha}(1)$$

whereas

$$n^{1-H}\big(T_n^{-} - \alpha\big) \underset{d}{\to} -\frac{h^{+}}{\mu^{+}(\alpha)} c_a \tilde{Z}_{H,\alpha}(1)$$

with $\tilde{Z}_{H,\alpha}(1)$ denoting the *same* random variable.

On the average, each of the two estimators uses about half of the observed values only. One may therefore try to obtain an improved estimator by combining T_n^+ and T_n^-. Because of the symmetry, the logical choice would be the average $\bar{T}_n = \frac{1}{2}(T_n^+ + T_n^-)$. However, the result above implies that $n^{1-H}(\bar{T}_n - \alpha)$ converges to zero in distribution. This means that the rate of convergence of $\bar{T}_n$ is faster than for the individual estimators T_n^+ and T_n^-, respectively. Thus, combining the two estimators

leads to a method with much better asymptotic properties. Unfortunately, there is currently no limit theorem available in the literature for $\bar{T}_n$. As a second best choice one may therefore take another convex combination $\kappa T_n^- + (1-\kappa)T_n^+$ with $\kappa \in [0,1] \setminus \{\frac{1}{2}\}$. This is asymptotically equivalent to defining the M-estimator $T_{\kappa,n} = [\tau_{\kappa,n}]_1^2$ where $\tau_{\kappa,n}$ solves the equation

$$\lambda_{\kappa,F_n}(\tau) = \sum_{t=1}^{n} \psi_{\kappa;v,u}(X_t,\tau) = 0 \tag{5.154}$$

and

$$\psi_{\kappa;v,u}(x,\alpha) = \kappa\psi_{v,u}^-(x,\alpha) + (1-\kappa)\psi_{v,u}^+(x,\alpha). \tag{5.155}$$

The corresponding constants are

$$\mu(\alpha) = E\big[\dot{\psi}_{\kappa;v,u}(X_t,\alpha)\big] = \mu^+(\alpha) \tag{5.156}$$

and

$$h = -\int_{\mathbb{R}} \psi_{\kappa;v,u}(x,\alpha)\, p_X{}'(x)\, dx = \kappa h^- + (1-\kappa)h^+ = (1-2\kappa)h^+. \tag{5.157}$$

Since μ is the same as before, the constant in the asymptotic distribution of $T_{\kappa,n}$ is smaller (in absolute value) by the factor $|1-2\kappa|$:

$$n^{1-H}(T_{\kappa,n} - \alpha) \Rightarrow -\delta_\kappa \tilde{Z}_{H,\alpha}(1) \tag{5.158}$$

with

$$\delta_\kappa = (1-2\kappa)\frac{h^+ c_a}{\mu^+(\alpha_0)}.$$

Finally, note that estimating α for the left and right tail separately can also be used for testing the null hypothesis that the tail index is the same on both sides. For instance, suppose that $X_t = X_{1,t}^+ - X_{2,t}^-$ with

$$X_{1,t} = \sum_{j=0}^{\infty} a_j \varepsilon_{1,t-j},\ X_{2,t} = \sum_{j=0}^{\infty} a_j \varepsilon_{2,t-j} \tag{5.159}$$

where $\varepsilon_{j,i}$ are i.i.d. SαS with $\alpha = \alpha_j$ $(j = 1,2)$. A natural statistic for testing $H_0 : \alpha_1 = \alpha_2$ is given by $\Delta T = T_n^+ - T_n^-$. However, in order to obtain a unique distribution under H_0, one has to narrow down the null hypothesis to a more concrete situation. For example, under $H_0 : X_{1,t} = X_{2,t}$ $(t \in \mathbb{Z})$ (with $\alpha = \alpha_1$) we have

$$n^{1-H}\Delta T_n \Rightarrow \left(\frac{2h^+ c_a}{\mu^+(\alpha_1)}\right)\tilde{Z}_{H,\alpha_1}(1), \tag{5.160}$$

where now $H = d + \alpha_1^{-1}$. Another possibility is that the two processes $\{X_{1,t}; t \in \mathbb{Z}\}$, $\{X_{2,t}; t \in \mathbb{Z}\}$ are independent of each other. Then, we obtain, under the assumption that $\alpha_1 = \alpha_2$,

$$n^{1-H} \Delta T_n \Rightarrow \left(\frac{2^{1/\alpha_1} h^+ c_a}{\mu^+(\alpha_1)} \right) \tilde{Z}_{H,\alpha_1}(1). \tag{5.161}$$

For further details, see Beran et al. (2012).

5.16 Goodness-of-Fit Tests

Let $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ be a linear process with $a_j \sim c_a j^{d-1}$ $(j \to \infty)$, and such that the assumptions of Theorem 4.33 hold. We denote by $F(x) = P(X_t \le x)$ and $p_X = F'$ its marginal distribution and density function, respectively. Suppose we would like to test the null hypothesis H_0 that $F(x)$ belongs to a certain class of distributions $\mathscr{F}_0$ against the alternative that $F \notin \mathscr{F}_0$. Theorem 4.33 implies that under H_0 test statistics based on the empirical distribution function $F_n(x) = n^{-1} \sum 1\{X_t \le x\}$ converge to distributions that are much simpler than for short-memory series. The reason is that the standardized empirical process degenerates to a random variable multiplied by a constant that depends on x.

Let us start with the simplest situation of a simple hypothesis $\mathscr{F}_0 = \{F_0\}$, i.e. we test whether F is equal to *one* specific distribution F_0. Consider, for instance, the Kolmogorov–Smirnov statistic

$$T_{\mathrm{KS}} = \sup_{x \in \mathbb{R}} \big| F_n(x) - F_0(x) \big|.$$

Theorem 4.33 implies that

$$n^{\frac{1}{2}-d} \sqrt{d(2d+1)} c_\gamma^{-\frac{1}{2}} T_{\mathrm{KS}} \underset{d}{\to} \sup_{x \in \mathbb{R}} p_X(x) \cdot |Z| \tag{5.162}$$

where Z is a standard normal variable and c_γ is the constant in $\gamma_X(k) \sim c_\gamma k^{2d-1}$ $(k \to \infty)$. Thus, H_0 is rejected at the level of significance α if

$$T_{\mathrm{KS}} > n^{d-\frac{1}{2}} \sqrt{\frac{c_\gamma}{d(2d+1)}} \sup_{x \in \mathbb{R}} p_X(x) \cdot z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ standard normal quantile. In practice, testing a simple hypothesis rarely happens. Instead, one usually needs to test whether F belongs to a certain parametric family of distributions characterized by an unknown parameter, say τ. For instance, we may want to test whether F is a normal distribution with unknown mean and variance $\tau = (\mu_X, \sigma_X^2)$. The Kolmogorov–Smirnov statistic is then given by

$$T_{\mathrm{KS}} = \sup_{x \in \mathbb{R}} \big| F_n(x) - F_0(x; \hat{\tau}) \big|$$

where $\hat{\tau}$ is a consistent estimator of τ. The problem that complicates inference is now that estimating τ often changes the asymptotic distribution. This is also the case for i.i.d. and short-range dependent processes. However, under long memory even the rate of convergence may change. This has been discussed in Theorem 4.34 for the case where $\tau = \mu$ is estimated by the sample mean. As illustrated by this theorem, and first discussed in Beran and Ghosh (1990, 1991), estimating μ actually *improves* the rate of convergence. If $0 < d < \frac{1}{4}$, then the rate improves to $n^{-\frac{1}{2}}$ which is the same as under independence or short memory. The limiting distribution of $n^{-\frac{1}{2}} T_{\mathrm{KS}}$ is complicated. However, due to the $n^{-\frac{1}{2}}$-rate, statistical inference is possible because, for instance, blockwise bootstrap procedures are applicable (possibly under some additional regularity conditions). For $\frac{1}{4} < d < \frac{1}{2}$, the rate improves to n^{2d-1}, and we have

$$n^{1-2d}\sqrt{d(2d+1)}c_\gamma^{-\frac{1}{2}} T_{\mathrm{KS}} \underset{d}{\to} \sup_{x\in\mathbb{R}} \left|p'_X(x-\mu)\right| \cdot \left|Z_2 + \frac{1}{2}Z^2\right| \tag{5.163}$$

where Z and Z_2 are uncorrelated variables, Z is standard normal and Z_2 is the value of the Hermite–Rosenblatt process at time 1 (see Theorem 4.34). The rate improves further if, for instance, the variance is estimated by the sample variance (see Beran and Ghosh 1990, 1991; Ho 2002; Kulik 2009). Analogous results can also be obtained for subordinated processes. Inference based on (5.163) is, of course, more complicated than using (5.162). Under suitable regularity conditions, one may avoid using the asymptotic formulas directly by applying suitable bootstrap procedures (note, however, that traditional blockwise bootstrap methods do not work for $d > \frac{1}{4}$). For instance, if X_t is generated by Gaussian subordination, then a suitable sampling window bootstrap may be designed (see Sect. 10.5).

Analogous results can be obtained for other goodness-of-fit tests for the marginal distribution. Interesting is, for example, the empirical characteristic function

$$m_n(u) = n^{-1}\sum_{t=1}^{n} \exp\left(iu\frac{X_t-\mu}{\sigma}\right) = n^{-1}\left[T_{\mathrm{re}}(u) + iT_{\mathrm{im}}(u)\right]$$

with

$$T_{\mathrm{re}}(u) = n^{-1}\sum_{t=1}^{n} \cos\left(iu\frac{X_t-\mu}{\sigma}\right),$$

$$T_{\mathrm{im}}(u) = n^{-1}\sum_{t=1}^{n} \sin\left(iu\frac{X_t-\mu}{\sigma}\right).$$

Suppose, for example, that under the null hypothesis X_t is normally distributed, and we assume μ and σ to be known. Since $\cos(\cdot)$ is even and $\sin(\cdot)$ is odd, the real and imaginary parts have different rates of convergence. The Hermite rank of the sine is one so that $n^{\frac{1}{2}-d} T_{\mathrm{im}}$ converges in distribution to a normal random variable. For the cosine, the Hermite rank is two so that $n^{1-2d}[T_{\mathrm{re}} - E(T_{\mathrm{re}})]$ converges to a constant

times an Hermite–Rosenblatt variable (Rosenblatt process at time 1), provided that $\frac{1}{4} < d < \frac{1}{2}$. On the other hand, if $0 < d < \frac{1}{4}$, then $\sqrt{n}[T_{\mathrm{re}} - E(T_{\mathrm{re}})]$ converges to a centred normal variable. More generally, one can show functional convergence of the processes $\zeta_{\mathrm{im}}(u) = n^{\frac{1}{2}-d} T_{\mathrm{im}}(u)$, and $\zeta_{\mathrm{re}}(u) = n^{1-2d}[T_{\mathrm{re}}(u) - E(T_{\mathrm{re}}(u))]$ or $\zeta_{\mathrm{re}}(u) = \sqrt{n}[T_{\mathrm{re}}(u) - E(T_{\mathrm{re}}(u))]$ in $C[u_{\mathrm{low}}, u_{\mathrm{up}}]$ equipped with the supremum norm and u_{low}, u_{up} finite (Beran and Ghosh 1991). If μ and σ are estimated by $\bar{x}$ and s, respectively, all rates (except $\sqrt{n}$) improve and the improvement as well as the limiting processes depend on which subinterval d is in.

Another question in the context of goodness-of-fit is whether the assumed model for the spectral density may be correct. Thus we wish to test the null hypothesis $H_0 : f_X \equiv f_0$ where f_0 is a fixed spectral density against the alternative $H_1 : f_X \not\equiv f_0$. More generally, f_0 may depend on a finite parameter vector ϑ that has to be estimated. We will assume that X_t is a linear process. For short-memory time series, Milhoj (1981) suggested the statistic

$$T_{\mathrm{Milhoj}} = B_n^{-2} \int_{-\pi}^{\pi} \left(\frac{I_{X,n}(\lambda)}{f_0(\lambda)} \right)^2 d\lambda$$

with

$$B_n = \int_{-\pi}^{\pi} \frac{I_{X,n}(\lambda)}{f_0(\lambda)}\, d\lambda$$

and showed that under H_0, and some regularity conditions, the standardized statistic $\sqrt{n}(T - \pi^{-1})$ converges in distribution to an $N(0, 2\pi^{-2})$ variable. The application of this statistics in the case of long memory is discussed in Beran (1992). (Deo and Chen 2000) pointed out that the asymptotic distribution of T and of an approximation T^* where integrals are replaced by Riemann sums at Fourier frequencies is not the same. While T_{Milhoj} is related to the Box–Pierce portmanteau statistic (Box and Pierce 1970), an alternative test based on an information measure introduced in Mokkadem (1997) is studied in Faÿ and Philippe (2002). As a modification of the Kulback–Leibler divergence, Mokkadem (1997) defines an information theoretic quantity that measures in how far two spectral densities f and g differ by

$$M(f, g) = \log\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{f(\lambda)}{g(\lambda)}\, d\lambda \right) - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \frac{f(\lambda)}{g(\lambda)}\, d\lambda.$$

Note that $M \geq 0$ with equality if and only if $f \equiv g$. Also note that M is scale invariant since $\log((2\pi)^{-1} \int \sigma\, d\lambda) - (2\pi)^{-1} \int \log \sigma\, d\lambda = 0$. As in the context of maximum likelihood estimation (Sect. 5.5), we will use the notation $\vartheta = (\sigma_\varepsilon^2, \theta) = (\sigma_\varepsilon^2, d, \theta_2, \ldots, \theta_p)$. Under H_0, we assume the spectral density to be of the form

$$f(\lambda; \vartheta) = \frac{\sigma_\varepsilon^2}{2\pi} \left| 1 - e^{-i\lambda} \right|^{-2d} f_*(\lambda; \theta_2, \ldots) = \frac{\sigma_\varepsilon^2}{2\pi} g(\lambda; \theta)$$

and such that $\sigma_\varepsilon^2 > 0$, θ is in a compact subset of $\Theta^0 = [0, \frac{1}{2}) \times \mathbb{R}^{p-1}$. The deviation of a spectral density $f_X(\lambda)$ from $f(\lambda;\vartheta)$ is then measured by

$$S\big(f_X, f(\lambda;\vartheta)\big) = S\big(f_X, g(\lambda;\theta)\big) = \inf_{\vartheta\in\Theta} M\big(f_X, g(\lambda;\theta)\big). \tag{5.164}$$

The true density f_X is, however, not known. To define a test statistic based on observations $X_1, \ldots, X_n$, Fay and Philippe propose to first apply tapering and pooling. For instance, for a taper defined by

$$w_{n,t} = \frac{1}{\sqrt{2}}\big(1 - e^{i\frac{2\pi}{n}t}\big) \quad (t = 1, 2, \ldots, n)$$

we write

$$d_n^{(1)}(\lambda_j) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} w_{n,t} X_t e^{-it\lambda_j} \quad \left(\lambda_j = \frac{2\pi j}{n},\ j = 1, 2, \ldots, n\right)$$

and the tapered periodogram is $I_n^{(1)} = |d_n^{(1)}|^2$. Pooling is done by dividing the range $[0, \pi]$ into $K_n = [\frac{1}{2}(n-1)/(m+1)]$ disjoint intervals $[\lambda_{(k-1)(m+1)+1}, \lambda_{k(m+1)}]$ and taking the average of m tapered periodogram values in each interval. Thus, denoting by

$$\bar\lambda_k = (m+1)\frac{2\pi}{n}\left(k - \frac{1}{2}\right)$$

the middle of each interval, we define

$$\bar I_n(\bar\lambda_k) := m^{-1} \sum_{j=(k-1)(m+1)+1}^{k(m+1)-1} \big|d_n^{(1)}(\lambda_j)\big|^2 \quad (k = 1, 2, \ldots, K_n),$$

where $K_n = [\frac{1}{2}(n-1)/(m+1)]$. This means that the range $[0, \pi]$ is divided into K_n disjoint intervals $[\lambda_{(k-1)(m+1)+1}, \lambda_{k(m+1)-1}]$ containing m Fourier frequencies, and we take the average of the tapered periodogram values in each interval. Replacing f_X in (5.164) by $\bar I_n$ and approximating the integral by a Riemann sum leads to

$$\tilde S\big(\bar I_n, g(\lambda;\theta)\big) = \log\left(K_n^{-1}\sum_{k=1}^{K_n} \frac{\bar I_n(\bar\lambda_k)}{g(\bar\lambda_k;\theta)}\right) - K_n^{-1}\sum_{k=1}^{K_n} \log\left(\frac{\bar I_n(\bar\lambda_k)}{g(\bar\lambda_k;\theta)}\right).$$

We give a brief highly simplified account of results derived in Faÿ and Philippe (2002) for $m \geq 5$. If H_0 is true and θ is equal to the correct parameter θ^0, then under regularity conditions

$$E\big[\tilde S\big(\bar I_n, g\big(\lambda;\theta^0\big)\big)\big] = \mu_m + o\big(n^{-\frac{1}{2}}\big)$$

where

$$\mu_m = -E\big[\log\big(2\pi \bar I_n^Z(\bar\lambda_k)\big)\big]$$

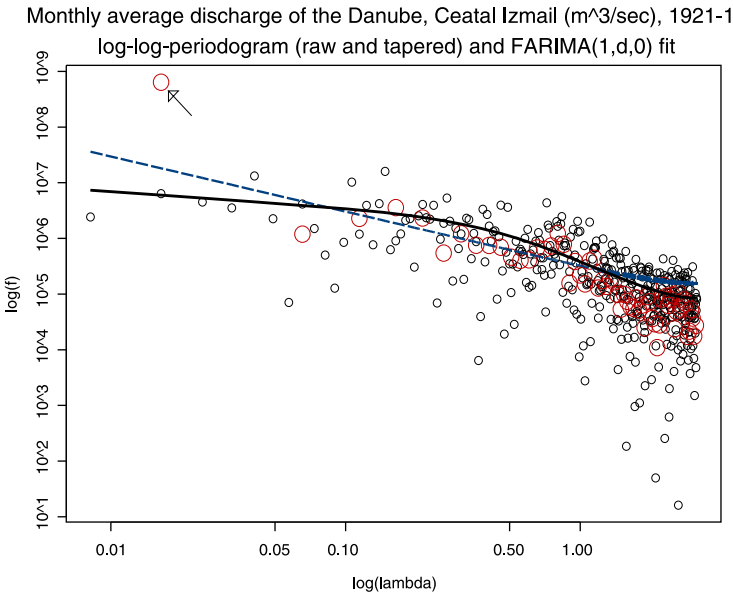

Fig. 5.16 For the deseasonalized monthly average discharge series of the Danube at Ceatal Izmail (m^3/s) (1921–1984), introduced in Sect. 1.2, the figure displays, in log–log-coordinates, the raw periodogram $I(\lambda_j)$ at Fourier frequencies, the tapered pooled periodogram $\bar{I}_n(\bar{\lambda}_k)$ (*large circles*) at the averaged frequencies $\bar{\lambda}_k$ and the spectral densities of a FARIMA$(1, d, 0)$ MLE fit (*full line*) and a FARIMA$(0, d, 0)$ MLE fit (*dashed line*)

with $\bar{I}_n^Z(\bar{\lambda}_k)$ calculated from n i.i.d. standard normal variables $Z_1, \ldots, Z_n$. This can be extended to the case where θ^0 is replaced by a $\sqrt{n}$-consistent estimator and, under suitable regularity conditions, one obtains a central limit theorem of the form

$$T_{\mathrm{FP}} = \sqrt{K_n}\big[\tilde{S}\big(\bar{I}_n, g(\lambda;\hat{\theta})\big) - \mu_m\big] \underset{d}{\to} N\big(0, \sigma_{\mathrm{FP}}^2\big)$$

with

$$\sigma_{\mathrm{FP}}^2 = \mathrm{var}\big\{2\pi\,\bar{I}_n^Z(\lambda_k) - \log\big(2\pi\,\bar{I}_n^Z(\lambda_k)\big)\big\}.$$

(Note that the right-hand side is the same for all k.) Applying the test, one has to bear in mind, however, that it is highly sensitive to "outliers" in the spectral domain. This may be a desirable feature when it comes to power considerations. However, the pooled tapered periodogram with the taper as defined above may sometimes yield an outlying value for the lowest frequency (or possibly even a few of the lowest frequencies). It may therefore advisable to remove the lowest or a few of the lowest frequencies first. This is illustrated by the following example.

Example 5.19 Suppose the parametric family consists of FARIMA(p, d, q) models with $d \in [0, \frac{1}{2})$, $p \le p_0$, $q \le q_0$ and p_0, q_0 fixed. In view of asymptotic theory, we may fit a FARIMA(p_0, d, q_0) process by one of the ML or Whittle methods

discussed in Sect. 5.5. For $m = 5$ one obtains $\mu_5 \approx 0.140$ and $\sigma^2_{\mathrm{FB}} \approx 0.378$ (see Dette and Sen 2010). Thus

$$T_{\mathrm{FP}} = \sqrt{K_n}\big[\tilde{S}\big(\bar{I}_n, g(\lambda; \hat{\theta})\big) - 0.14\big]$$

and the null hypothesis that f_X belongs to the FARIMA family with $p \le p_0$ and $q \le q_0$ is rejected if

$$|T_{\mathrm{FP}}| > \sqrt{0.378} z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$-quantile of the $N(0, 1)$ distribution. For $\alpha = 0.05$ the critical limit is about 1.2. To illustrate the method, we consider the deseasonalized monthly average discharge series of the Danube at Ceatal Izmail (m^3/s) (1921–1984) introduced in Sect. 1.2. Figure 5.16 displays, in log–log-coordinates, the raw periodogram $I_{X,n}(\lambda_j)$ at Fourier frequencies, the tapered pooled periodogram $\bar{I}_n(\bar{\lambda}_k)$ (large circles) at the averaged frequencies $\bar{\lambda}_k$, the spectral density of a FARIMA$(1, d, 0)$ MLE fit (full line) and also a FARIMA$(0, d, 0)$ fit (dashed line). The MLE fit ($\hat{d} = 0.146$, with a 95 %-confidence interval $[-0.044, 0.336]$, $\hat{\varphi}_1 = 0.614$, $[0.423, 0.806]$) is apparently good when compared to the raw periodogram. There is, however, a problem with the tapered periodogram at the lowest frequency $\bar{\lambda}_1$ since the value is far from all the other values. This has to do with the specific taper $(1 - \exp(i2\pi t n^{-1}))$. Calculating T_{FP} using all frequencies indeed yields a surprisingly large value of about 50. This contradicts the visually excellent fit, and is entirely due to the outlying value of $\bar{I}_n(\bar{\lambda}_1)$. If $\bar{\lambda}_1$ is omitted in the calculation, then we obtain $T_{\mathrm{FP}} \approx 1.145$ which corresponds to a p-value of about 0.25. Thus, there is no evidence for a departure from a FARIMA$(1, d, 0)$ model. In contrast, if a FARIMA$(0, d, 0)$ model is fitted instead, then the value of T_{FP} after omitting $\bar{\lambda}_1$ is equal to 7.219 which is significant at any reasonable level. This confirms the visual impression in Fig. 5.16 that a straight line (in log–log-coordinates) is not appropriate.

Finally, note that it can also be shown that T_{FP} is asymptotically normal under local (Faÿ and Philippe 2002) and fixed alternatives (Dette and Sen 2010); however, with a different expected value and variance. Another goodness-of-fit test for the autocovariance structure is proposed, for example, in Delgado et al. (2005).

Chapter 6
Statistical Inference for Nonlinear Processes

In this section, we consider nonlinear processes with long memory. We will mainly focus on volatility models: stochastic volatility (see Definitions 2.3–2.4 and Sect. 4.2.6 for limit theorems), ARCH(∞) processes (see Definition 2.1 and Sect. 4.2.7) and LARCH(∞) models (see (2.47) and (2.48), and Sect. 4.2.8). Statistical inference for traffic models is not well developed yet (see Faÿ et al. 2006, 2007; Hsieh et al. 2007 for some results in this direction).

Volatility models considered in this book have the general form $X_t = \xi_t \sigma_t$, where ξ_t ($t \in \mathbb{Z}$) is an i.i.d. sequence and σ_t depends on the past $(\xi_{t-1}, \xi_{t-2}, \ldots)$ and/or a latent process ζ_t. In particular, in the stochastic volatility model (SV),

$$\sigma_t = \sigma(\zeta_t), \qquad \zeta_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j},$$

where (ξ_t, ε_t) ($t \in \mathbb{Z}$) is a sequence of i.i.d. random vectors. If furthermore $\sigma(x) = \exp(x)$ and ζ_t is a long-memory Gaussian sequence independent of the i.i.d. centred sequence ξ_t, then the model is called LMSV.

If

$$\sigma_t = b_0 + \sum_{k=1}^{\infty} b_k X_{t-k}$$

and b_j decay slowly like a constant times j^{d-1} ($d \in (0, 1/2)$), then we obtain a LARCH(∞) model with long memory (recall that σ_t can be expressed explicitly in terms of $\xi_{t-1}, \xi_{t-2}, \ldots$). Finally, if

$$\sigma_t^2 = b_0 + \sum_{k=1}^{\infty} b_k X_{t-k}^2,$$

$\sum_{k=1}^{\infty} |b_k| < \infty$, we obtain a second-order stationary ARCH(∞) sequence. Other models, e.g. FIGARCH, are not discussed in this chapter.

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_6, © Springer-Verlag Berlin Heidelberg 2013

As in Chap. 5, we start our discussion with location estimation. In this case, the stochastic volatility (like LMSV) and LARCH(∞) models follow a similar pattern. The asymptotic distribution of the sample mean is not affected by long memory. The same applies to M-estimators, as long as the function ψ that defines the M-estimator is antisymmetric and the distribution of the noise variables ξ_t is symmetric. Otherwise, asymptotic properties of M-estimators are influenced by long memory. Such results were obtained in Beran (2006) and Beran and Schützner (2008), and are presented in Sects. 6.1.1 and 6.2.1, respectively, for SV and LARCH models. Finally, in Sect. 6.3.1, we discuss location estimation for ARCH(∞) processes. At the moment, a theory for M-estimators is not available.

As for estimation of memory parameters, one may note that long memory appears (if at all) in the squares. It is therefore tempting to apply methods described in Chap. 5 to the squared sequence X_t^2. However, it may be more natural to divide volatility processes into two groups: stochastic volatility-type models (with a possible leverage) and LARCH(∞)-type models.

In the first case, direct maximum likelihood estimation is not always feasible because of the presence of an unobserved latent process. Note, however, that, for instance, for a stochastic volatility model with an exponential volatility function $\sigma(x) = e^x$, one may consider a log-transformation. This approach is taken, among others, in Zaffaroni (2009) using parametric Whittle estimation and in Deo and Hurvich (2001), Hurvich and Soulier (2002), Hurvich et al. (2005b) or Dalla et al. (2006) who consider semiparametric estimation.

For the LARCH models, a maximum likelihood approach is feasible in principle because σ_t is an explicit function of past observations (see Beran and Schützner 2009). Up to date there are no theoretical results on semiparametric estimators in the Fourier or wavelet domain. Teyssière and Abry (2006) as well as Jach and Kokoszka (2008) study the numerical performance of wavelet estimators, in particular for LARCH models. For ARCH(∞) processes, σ_t^2 is again a direct function of past observations and MLE-type estimators are not difficult to calculate. In particular, one can show that the MLE is more efficient than Whittle estimation based on the squared observations (which is not really an approximate MLE), see Giraitis and Robinson (2001), Straumann (2004), Berkes and Horváth (2003).

Finally, we consider tail index estimation for heavy-tailed stochastic volatility models. Recall that for linear processes we considered in Sect. 5.15 the tail index M-estimation based on the assumption of stable innovations. Here we consider instead the Hill estimator which is consistent without specifying a particular model. Asymptotic normality of the Hill estimator for SV models was established in Kulik and Soulier (2011) and is presented in Sect. 6.1.3. For LARCH processes, a numerical, although wavelet-based, tail index estimation can be found in Jach and Kokoszka (2008).

6.1 Statistical Inference for Stochastic Volatility Models

In this section, we consider statistical inference for stochastic volatility models of the form

$$X_t = \sigma_t \xi_t \quad (t \in \mathbb{N}), \tag{6.1}$$

where $\sigma_t = \sigma(\zeta_t)$, $\zeta_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j}$ and (ξ_t, ε_t) $(t \in \mathbb{Z})$ is a sequence of i.i.d. random vectors. It is assumed that $E(\varepsilon_1) = 0$, however, there is no a priori assumption that the random variables ξ_t are centred.

In Sect. 6.1.1, we consider location estimation in a model $Y_t = \mu + X_t$, where X_t is an SV process. As mentioned in the introduction, the asymptotic distribution of the sample mean is not affected by long memory. The same applies to M-estimators, as long as the function ψ that defines the M-estimator is antisymmetric and the distribution of the noise variables ξ_t is symmetric (Beran and Schützner 2008).

We proceed with estimation of the memory parameter. Consider the volatility model (6.1). We recall that the memory parameter d appears in the asymptotics for the covariance function of the squares (see (2.61)). The graphical methods considered in Sect. 5.4 can be also applied in this case, by replacing X_t there by $Y_t = X_t^2$ here. For example, the R/S statistic can be defined as R_n/S_n, where

$$R_n = \max_{1 \le k \le n} \sum_{t=1}^{k} (Y_t - \bar{y}_n) - \min_{1 \le k \le n} \sum_{t=1}^{k} (Y_t - \bar{y}_n)$$

and $S_n^2 = (n-1)^{-1} \sum_{t=1}^{n} (Y_t - \bar{y}_n)^2$ is the sample variance of $Y_t = X_t^2$. The sample variance S_n^2 converges in probability to $\mathrm{var}(X_1^2)$ (provided it is finite). The same approach can be applied to all other methods considered in Sect. 5.4 (see, e.g. Giraitis et al. 2000b).

However, using the squares may not be appropriate for heavy-tailed data. For instance, the data may have a finite variance, but infinite fourth moments. Then the graphical methods can be quite misleading (see, e.g. Wright 2002).

In general, maximum likelihood estimation is not suitable for SV models because the likelihood function cannot be written in an explicit form (see, e.g. Robinson and Zaffaroni 1997, 1998). Asymptotic normality of the Whittle estimator applied to transformed data was considered explicitly in Breidt et al. (1998) and in case of leverage in Zaffaroni (2009). Some results can be deduced from earlier theory for models with signal and additive noise (Hosoya 1974; Hosoya and Taniguchi 1982). Note, however, that the Whittle approach does not have much to do with maximum likelihood estimation here because the (transformed) data the method is applied to are by definition far from normal.

As for semiparametric methods, asymptotic results in the SV case are a relatively simple generalization of the theory for linear processes considered in Chap. 5. Specifically, if $X_t = \xi_t \exp(\sum_{j=1}^{\infty} a_j \varepsilon_{t-j})$, then one can apply the log-transformation to X_t^2 and the resulting model has the form of a linear long-memory

process corrupted by i.i.d. noise. Asymptotic properties of semiparametric estimators in SV models were considered in Deo and Hurvich (2001), Hurvich and Soulier (2002), Hurvich et al. (2005b).

Finally, we discuss tail index estimation. In Sect. 5.2.3, we considered M-estimation for heavy-tailed long-memory processes. Such an approach requires strong assumptions on an innovation sequence of the linear process. Rates of convergence and the asymptotic distribution is affected by long memory and tail behaviour. In the present context, based on results on M-estimators in Sect. 6.1.1 below, one may be expected that the asymptotic behaviour of an M-estimator of the tail index is not affected by long memory. However, such results are not known at present. Instead, we consider the so-called Hill estimator (see, e.g. Embrechts et al. 1997). Its asymptotic properties are built upon results for the tail empirical process considered in Sect. 4.8.5. It is proven (see Kulik and Soulier 2011) that long memory does not affect the rate of convergence. This is confirmed in Jach et al. (2012) and Luo (2011), both in theory and numerical studies.

6.1.1 Location Estimation

Consider a time series $Y_t = \mu + X_t$ ($t \in \mathbb{N}$) such that the residuals X_t are generated by a stochastic volatility model (6.1). Furthermore, assume that the random variables ξ_t that appear in the model definition (6.1) are centred. Hence $E(X_t) = 0$. In Sect. 4.2.6, we found out that under appropriate moment assumptions,

$$n^{-1/2} \sum_{t=1}^{[nu]} X_t \Rightarrow vB(u),$$

where $v^2 = \mathrm{var}(X_1)$ and $B(u)$ ($u \in [0, 1]$) is a standard Brownian motion. In other words, long memory in volatility does not affect rates of convergence for the sample mean.

More generally, if ψ is a deterministic function such that $E[\psi(X_1)|\mathscr{G}_0] = 0$, where $\mathscr{G}_t$ is the sigma field generated by $(\xi_t, \varepsilon_t, \xi_{t-1}, \varepsilon_{t-1}, \ldots)$, then the central limit theorem above still holds with $v^2 = \mathrm{var}(\psi(X_1))$.

The condition $E[\psi(X_1)|\mathscr{G}_0] = 0$ is equivalent to

$$\int \psi\big(s\sigma(\zeta_1)\big)\, dF_\xi(s) = 0,$$

where F_ξ is the distribution function of ξ_1. If, for example, $\psi(x) = \mathrm{sign}(x)$, bearing in mind that $\sigma(\cdot) > 0$, this integral has the form

$$-\int_{-\infty}^{0} dF_\xi(s) + \int_{0}^{\infty} dF_\xi(s).$$

Thus, if the random variable ξ_1 is symmetric, then this expression vanishes. Recalling from Sect. 5.2.3 that the sign function yields the sample median (written down as an M-estimator), we can expect that in the particular case of symmetric random variables ξ_t and antisymmetric functions ψ, the asymptotic theory for M-estimators is the same as for i.i.d. data. To be more specific, if $\hat{\mu}$ is a solution of $\sum_{t=1}^{n} \psi(Y_t - \mu) = 0$, then

$$\sqrt{n}(\hat{\mu} - \mu) \to_d N\big(0, \sigma_\psi^2\big), \tag{6.2}$$

where $\sigma_\psi^2 = E[\psi^2(X_1)]/E^2[\psi'(X_1)]$. A general result was obtained in Beran and Schützner (2008) (cf. also Theorem 6.2 in Sect. 4.2.6). In particular, if

(A1) The random variables ξ_t are symmetric,
(A2) σ_t is a second-order stationary process with a finite variance such that ξ_t is independent of σ_s, $s \le t$ (but the sequences ξ_t and σ_t are not necessary independent),
(A3) The function $\psi(\cdot)$ is measurable and antisymmetric, that is, $\psi(x) = -\psi(-x)$, and $E[\psi^2(X_1)] < \infty$,

then (6.2) holds.

Theorem 6.1 *Consider the stochastic volatility model defined in* (6.1). *Assume that* (A1)–(A3) *above hold. Under additional regularity conditions*, (6.2) *holds*.

We note that "additional regularity conditions" refer to assumptions (A4)–(A8) in Beran and Schützner (2008).

Proof The proof differs from the proof of the central limit theorem for M-estimators based on linear processes with long-range dependence; see the proof of Theorem 5.1. The reason is that in the proof of that theorem we were looking for the asymptotic equivalence between an M-estimator and the sample mean.

To proceed, we expand

$$0 = \sum_{t=1}^{n} \psi(Y_t - \hat{\mu}) = \sum_{t=1}^{n} \psi(Y_t - \mu) + (\hat{\mu} - \mu) \sum_{t=1}^{n} \psi'\big(Y_t - \mu^*\big),$$

where $|\mu^* - \mu| \le |\hat{\mu} - \mu|$. Under appropriate differentiability properties of ψ, $|\hat{\mu} - \mu| < \delta$ implies $|\psi'(Y_t - \hat{\mu}) - \psi'(Y_t - \mu)| < k_1(\delta)$, where k_1 is a constant that depends on δ only. Hence, recalling that $Y_t - \mu = X_t$,

$$\sqrt{n}(\hat{\mu} - \mu) \approx \frac{n^{-1/2} \sum_{t=1}^{n} \psi(X_t)}{n^{-1} \sum_{t=1}^{n} \psi'(X_t)}.$$

One can argue that the denominator converges in probability to $E[\psi'(X_1)]$. Furthermore, a martingale central limit theorem yields asymptotic normality of the nominator. Hence, the result follows. For further details, we refer to Beran and Schützner (2008). □

The most general statement is given in Beran and Schützner (2008). The Gaussian assumption used in the statement of Theorem 4.10 is replaced by

(A1) The random variables ξ_t are symmetric;
(A2) σ_t is a second-order stationary process with a finite variance such that ξ_t is independent of σ_s, $s \le t$ (but the sequences ξ_t and σ_t are not necessary independent).

Furthermore, as in Theorem 4.10, it is assumed that

(A3) The function $\psi(\cdot)$ is measurable and antisymmetric, that is, $\psi(x) = -\psi(-x)$, and $E[\psi^2(X_1)] < \infty$.

Finally, there is an additional assumption on extremal behaviour of the sequence $\psi(X_t)$, as well as further technical conditions on function ψ, see (A4)–(A5) and (A6)–(A8) in Beran and Schützner (2008).

Theorem 6.2 *Consider the stochastic volatility model defined above. Assume that* (A1)–(A3) *above as well as* (A4)–(A5) *and* (A6)–(A8) *in Beran and Schützner* (2008). *Then* (4.68) *holds.*

6.1.2 Estimation of Dependence Parameters

As mentioned in the introduction to this section, maximum likelihood estimation does not seem to be feasible for models of the form (6.1). To be more specific, let us consider the LMSV model,

$$X_t = \xi_t \ \exp\left(\sum_{j=1}^{\infty} a_k \varepsilon_{t-j}\right), \tag{6.3}$$

where the sequences ξ_t $(t \in \mathbb{Z})$ and ε_t $(t \in \mathbb{Z})$ are mutually independent. Furthermore, we shall assume that all random variables are standard normal and $\sum_{j=1}^{\infty} a_j^2 = 1$. Then the density p_X of X_t is

$$p_X(x) = \int_0^{\infty} \phi\big(\log(x/y)\big)\phi(y)\, dy \quad (x > 0),$$

where ϕ is the standard normal density. An analogous formula is valid for $x < 0$. Furthermore, the joint density of $(X_1, \ldots, X_n)$ can be written as an n-fold integral with respect to $\phi(y_1) \cdots \phi(y_n) dy_1 \cdots dy_n$. Consequently, finding the maximum likelihood estimator is extremely difficult. Breidt et al. (1998) use the Whittle estimator (see Sect. 5.5.2) applied to the logarithm of the squares instead.

Much easier is the application of semiparametric methods to stochastic volatility models. We consider for simplicity the LMSV model (6.3). Applying the log-

transformation to X_t^2, we obtain a new model

$$Y_t = \mu + 2\sum_{k=1}^{\infty} a_k \varepsilon_{t-k} + Z_t,$$

where $Z_t = \log \xi_t^2 - E(\log \xi_t^2)$, $\mu = E(\log \xi_t^2)$. The semiparametric estimators (in the Fourier or wavelet domain) can be applied directly to the sequence Y_t. We note that Y_t has the form of a long-memory sequence plus i.i.d. noise Z_t. Hence, we are exactly in the situation of the additive noise model considered in Example 5.13. Specifically, if we assume that the spectral density $f_{\tilde{X}}$ of the linear process $\tilde{X}_t := \sum_{k=1}^{\infty} a_k \varepsilon_{t-k}$ has the form $f_{\tilde{X}}(\lambda) = \lambda^{-2d} f_*(\lambda)$, then Y_t has the spectral density

$$\begin{aligned} f_Y(\lambda) &= f_{\tilde{X}}(\lambda) + \sigma_Z^2/(2\pi) = \lambda^{-2d} f_*(\lambda) + \sigma_Z^2/(2\pi) \\ &\approx \lambda^{-2d} f_*(0) + \sigma_Z^2/(2\pi) = \lambda^{-2d} f_*(0)\big(1 + O\big(\lambda^{2d}\big)\big), \end{aligned}$$

where $\sigma_Z^2 = \mathrm{var}(Z_1)$. According to the results in Sect. 5.8, the optimal mean squared error of a semiparametric estimator is then of order

$$m = O\big(n^{-\frac{4d}{4d+1}}\big), \qquad MSE(\hat{d}) = O\big(n^{-\frac{4d}{4d+1}}\big),$$

cf. Deo and Hurvich (2001), Hurvich and Soulier (2002) for log-periodogram regression (GPH), and Arteche (2004) for the local estimator. Hurvich et al. (2005b) show that a modified version of these estimators can outperform the GPH approach.

Furthermore, the techniques considered in Sect. 5.6.4 can be applied to the situation of additive noise as well. Consequently, we obtain the following asymptotic normality of the local Whittle estimator (see Hurvich et al. 2005b; Dalla et al. 2006). The result mimics Theorem 5.5. We have to adapt the bandwidth condition (LW3) there to the present context.

Theorem 6.3 *Consider the LMSV model given in* (6.3). *If*

$$m^{-1} + m^{2d+1} n^{-2d} \to 0, \quad \text{(LW3–SV)}$$

then $m^{1/2}(\hat{d}_{\mathrm{LW}} - d) \to N(0, 1/4)$.

Proof The proof follows similar lines as in the case of a linear process without the additive noise (see Sect. 5.6.4). The main step is asymptotic normality of a weighted sum of periodogram ordinates. Let us recall some notation: $\lambda_j = 2\pi j/n$, $j = 1, \ldots, m$, are Fourier frequencies, $b_j = -2\log \lambda_j$, $I_{n,Y}(\cdot)$ is the periodogram associated with the sequence $Y_1, \ldots, Y_n$. We re-write the decomposition (5.67) in the present context to obtain

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - 1\right] + \sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,Y}(\lambda_j)}{g_Y(\lambda_j)} - \frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)}\right], \tag{6.4}$$

where $b_{j,m} = (b_j - \bar{b})/\sqrt{m}$ and $g_Y(\lambda) = |\lambda|^{-2d} f_*(\lambda)$. We deal with the first part only to illustrate the influence of the additive noise.

Let us decompose the difference between the normalized periodogram of Y_t and $\tilde{X}_t$:

$$\begin{aligned}\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)} &= \frac{I_{n,\tilde{X}}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)} + \frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)} \\ &= \frac{f_{\tilde{X}}(\lambda_j) - f_Y(\lambda_j)}{f_Y(\lambda_j)} \frac{I_{n,\tilde{X}}(\lambda)}{f_{\tilde{X}}(\lambda)} + \frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)} \\ &= \frac{\sigma_Z^2/2\pi}{f_Y(\lambda_j)} \frac{I_{n,\tilde{X}}(\lambda)}{f_{\tilde{X}}(\lambda)} + \frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)}.\end{aligned}$$

We start with the term $I_{n,Z}(\lambda_j)/f_Y(\lambda_j)$. Since the random variables Z_t are i.i.d., the expected value of the normalized periodogram is one (cf. (4.139)). Thus

$$E\left(\frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)}\right) = E\left(\frac{I_{n,Z}(\lambda_j)}{f_Z(\lambda_j)}\right)\frac{f_Z(\lambda_j)}{f_Y(\lambda_j)} \sim \frac{\sigma_Z^2}{2\pi}|\lambda_j|^{2d} f_*^{-1}(\lambda_j) \le C(j/n)^{2d}.$$

Furthermore, we recall that $E[I_{n,\tilde{X}}(\lambda_j)/f_{\tilde{X}}(\lambda_j)]$ is uniformly bounded (in j) and that $f_Y(\lambda_j) = O((j/n)^{-2d})$. Thus we conclude

$$E\left|\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_k)}\right| \le C(j/n)^{2d}.$$

Hence,

$$E\left|\sum_{j=1}^{m} b_{j,m}\left\{\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)}\right\}\right| \le \sum_{j=1}^{m} |b_{j,m}|(j/n)^{2d}.$$

The bound is $\max_{1\le j\le m}|b_{j,m}| \sum_{j=1}^{m}(j/n)^{2d} = o(1)n^{-2d}m^{2d+1}$ which converges to 0 if (LW3–SV) holds. Consequently, the asymptotic behaviour of

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - 1\right]$$

is the same as that of

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)} - 1\right].$$

The latter was studied in Sect. 5.6.4. □

The result of Theorem 6.3 can be extended to the case of stochastic volatility models with leverage, i.e. when $\rho_{Z,\varepsilon} = E[Z_t\varepsilon_t] \neq 0$. In this case, the spectral den-

sity of $Y_t = \log X_t^2$ behaves like

$$f_Y(\lambda) \sim f_{\tilde{X}}(\lambda) + \frac{\sigma_Z^2}{2\pi} + \mathrm{Re}\big((1-e^{i\lambda})^{-d}\big)\frac{2\rho_{Z,\varepsilon}\sigma_Z^2\sqrt{f_*(0)}}{\sqrt{2\pi}}.$$

6.1.3 Tail Index Estimation

Consider the stochastic volatility models $X_t = \xi_t\sigma_t$ given in (6.1), where ξ_t are i.i.d. random variables with

$$P(\xi_t > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\xi_t < -x) \sim A\frac{1-\beta}{2}x^{-\alpha},$$

as $x \to \infty$, and $\alpha > 0$ is the tail index. Furthermore, it is assumed that the sequence σ_t is independent of ξ_t.

One of the most important problems when dealing with heavy tails is to estimate the tail index α. A standard (though not quite unproblematic; see, e.g. Resnick 1997) method is Hill's estimator. Setting $\gamma = \alpha^{-1}$, the Hill estimator of γ is defined by

$$\hat{\gamma}_n = \frac{1}{k}\sum_{j=1}^{k}\log\left(\frac{X_{n-j+1:n}}{X_{n-k:n}}\right) = \int_0^\infty \frac{\hat{T}_n(s)}{1+s}\,ds,$$

where

$$\hat{T}_n(s) = \frac{1}{k}\sum_{j=1}^{n} 1\big\{X_j > X_{n-k:n}(1+s)\big\}, \qquad T(s) = (1+s)^{-\alpha},$$

and $X_{k:n}$ are the order statistics of the sample $X_1, \ldots, X_n$. Since $\gamma = \int_0^\infty (1+s)^{-1}T(s)\,ds$, we have

$$\hat{\gamma}_n - \gamma = \int_0^\infty \frac{\hat{e}_n^*(s)}{1+s}\,ds,$$

where $\hat{e}_n^*(s)$ is the tail empirical process defined in Sect. 4.8.5:

$$\hat{e}_n^*(s) = \hat{T}_n(s) - T(s) \quad \big(s \in [0,\infty)\big).$$

Thus we can apply Theorem 4.37 to obtain the asymptotic distribution of the Hill estimator. Heuristically,

$$\sqrt{k_n}(\hat{\gamma}_n - \gamma) \to_d \int_0^\infty \frac{\tilde{B}(T(s))}{1+s}\,ds$$

where $\tilde{B}(u) = B(u) - uB(1)$ $(u \in [0,1])$ is a Brownian bridge. This integral is a centred normal random variable with variance γ^2.

Table 6.1 Simulated average values of and standard deviations of the Hill estimator $\hat{\gamma}$ (where $\gamma = 1/\alpha$) for an LMSV model with standard deviation $\beta = 0.2$ and sample size $n = 1000$

$\gamma = 1/\alpha$	d =0	0.2	0.4	0.45
0.667	mean = 0.6631	0.6670	0.6717	0.6659
	Std. dev. = 0.0664	0.0682	0.0648	0.0648
0.5	0.5001	0.5010	0.4988	0.5010
	0.0506	0.0500	0.0515	0.0503
0.25	0.2513	0.2518	0.2511	0.2530
	0.0249	0.0251	0.0251	0.0246
0.167	0.1711	0.1718	0.1791	0.1833
	0.0174	0.0170	0.0174	0.0188
0.1	0.1208	0.1226	0.1379	0.1452
	0.0114	0.0111	0.0140	0.0165

Corollary 6.1 *Under the assumptions of Theorem 4.37, $\sqrt{k}(\hat{\gamma}_n - \gamma)$ converges weakly to the centred Gaussian distribution with variance γ^2.*

This result allows us to construct confidence intervals for γ, with a user-chosen number k of extreme observations. The result is, in fact, the best possible rate of convergence for the Hill estimator for i.i.d. data (see Drees 1998). The surprising result is that it is possible to achieve the i.i.d. rate in spite of long memory. A detailed proof is given in Kulik and Soulier (2011). Further results can be found in Jach et al. (2012). Also, Corollary 6.1 can be extended to stochastic volatility models with leverage, i.e. when the sequences σ_t and ξ_t are not mutually independent, see Luo (2011).

Example 6.1 We simulate an LMSV model $X_t = \xi_t \exp(\beta\zeta_t)$, $(t = 1, \ldots, n = 1000)$ with $\beta > 0$, ξ_t independent Pareto random variables with tail parameter α and ζ_t a long memory FARIMA$(0, d, 0)$ sequence with standard normal innovations and dependence parameter $d \in [0, 1/2)$. We assume that $\{\zeta_t, t = 1, \ldots, n\}$ and $\{\xi_t, t = 1, \ldots, n\}$ are mutually independent. Table 6.1 shows that dependence of ζ_t does not influence tail index estimation, unless α is very large. Note, however, that from a practical point of view, large values of α are not interesting (if $\alpha > 4$, then the squares X_t^2 have a finite variance). Note also that further simulations (not reported here) illustrate that, if the variability coefficient β is large, then dependence may start to play a role for finite samples, although this influence disappears asymptotically, as indicated in Corollary 6.1. We refer to Luo (2011) for further details.

6.2 Statistical Inference for LARCH Processes

In this section, we consider LARCH processes. As in Sect. 6.1, we start with location estimation showing again that the asymptotic distribution of the sample mean

as well as for M-estimators is not affected by long memory, as long as function ψ that defines the M-estimator is antisymmetric and the noise variables ξ_t are symmetrically distributed (Beran 2006).

As for parameter estimation, it is reported in Giraitis et al. (2000b) that the graphical methods (KPSS, V/S, R/S) perform well for LARCH processes. Giraitis et al. (2003) claim further that for LARCH(∞) models V/S is superior to R/S and *KPSS*. There is no existing theory for semiparametric estimators for LARCH processes. Teyssière and Abry (2006) and Jach and Kokoszka (2008) study the numerical performance of wavelet estimators. Giraitis and Robinson (2001) argue that for ARCH(∞)-type models (including LARCH), the Whittle approach is less motivated than the maximum likelihood procedure that yields explicit results. However, it turns out that the issue is actually more complex. This and other detailed theoretical results on MLE type estimation, including asymptotic normality, can be found in Beran and Schützner (2009), and will be discussed below.

6.2.1 Location Estimation

Consider a time series $Y_t = \mu + X_t$ with residuals X_t generated by a long-memory LARCH process

$$X_t = \sigma_t \varepsilon_t, \tag{6.5}$$

$$\sigma_t = a + \sum_{j=1}^{\infty} b_j X_{t-j}. \tag{6.6}$$

Here ε_t are i.i.d. random variables with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = 1$, and the coefficients are such that $a \neq 0$, $b_j \sim cj^{d-1}$ (as $j \to \infty$) for some $0 < d < \frac{1}{2}$ and $\sum b_j^2 < 1$. Since $cov(X_t, X_{t+k}) = 0$ $(k \neq 0)$, the variance of the sample mean is not affected by the dependence in volatility, i.e. $\text{var}(\bar{X}) = \sigma_X^2/n$ where $\sigma_X^2 = \text{var}(X_t)$ (note that $\sigma_X^2 = \sigma_Y^2 = \text{var}(Y_t)$). Beran (2006) defines sufficient moment conditions under which a functional limit theorem holds for partial sums, namely

$$n^{-\frac{1}{2}} \sigma_X^{-1} S_n(u) = n^{-\frac{1}{2}} \sigma_Z^{-1} \sum_{t=1}^{[nu]} X_t \underset{D[0,1]}{\to} B(u)$$

where convergence is in the space $D[0, 1]$ of càdlàg functions equipped with the Skorokhod metric and $B(u)$ denotes standard Brownian motion. More generally, for functions ψ satisfying some moment conditions, we can write

$$\begin{aligned} E\big[\psi(X_{t+k})\psi(X_t)\big] &= E\big\{E\big[\psi(X_{t+k})\psi(X_i) \mid \mathscr{F}_{t+k-1}\big]\big\} \\ &= E\big\{\psi(X_t)E\big[\psi(\varepsilon_{t+k}\sigma_{t+k}) \mid \mathscr{F}_{t+k-1}\big]\big\} \end{aligned}$$

where $\mathscr{F}_t$ denotes the σ-algebra generated by ε_j $(j \le t)$. In particular, if the distribution of ε_t is symmetric and ψ is antisymmetric, i.e. $\psi(-x) = -\psi(x)$, then $E[\psi(\varepsilon_{t+k}\sigma_{t+k}) \mid \mathscr{F}_{t+k-1}] = 0$ so that $\psi(X_t)$ $(t \in \mathbb{Z})$ is a martingale difference and $\mathrm{var}(\sum \psi(X_t)) = O(n)$. This has direct implications for M-estimators of the location parameter μ defined as solutions of $\sum_{i=1}^{n} \psi(Y_t - \hat{\mu}) = 0$. It can be shown that under regularity conditions, the asymptotic distribution of $\hat{\mu}$ is the same as for $S_{n;\psi} = E^{-1}[\psi'(X_1)]S_n(1)$. Thus, we have

$$\sqrt{n}(\hat{\mu} - \mu) \underset{d}{\to} N\big(0, \sigma_\psi^2\big)$$

where $\sigma_\psi^2 = E[\psi^2(X_1)]E^{-2}[\psi'(X_1)]$, cf. Sect. 5.2.3. In other words, the asymptotic distribution of M-estimators of location is undisturbed by LARCH type (long-range) dependence in volatility, and is, in fact, the same as if observations were i.i.d. For detailed conditions on ψ and ε_t, see Beran (2006). In conclusion, approximate $(1-\alpha)$-confidence intervals for μ may be given by

$$\hat{\mu} \pm z_{1-\alpha/2}\sigma_\psi n^{-\frac{1}{2}} \tag{6.7}$$

where $z_{1-\alpha/2}$ is the standard normal $(1-\alpha/2)$-quantile.

A completely different result is obtained, however, if ψ is not antisymmetric or if ε_t are not symmetrically distributed such that $E[X_1\psi(X_1)] \neq 0$. In this case, $\hat{\mu}$ has a slower rate of convergence and limit theorems derived in Berkes and Horváth (2003) apply; see also Sect. 4.2.8. From the applied point of view, this means that it is important to check symmetry of the innovation process.

6.2.2 Estimation of Dependence Parameters

6.2.2.1 Basic Definitions and Problems

Consider a parametric long-memory LARCH process $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ as in (6.5) and (6.6), where ε_t are i.i.d. continuous random variables with density function p_ε, $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = 1$, $a \neq 0$, $b_j \sim cj^{d-1}$ (as $j \to \infty$) for some $0 < d < \frac{1}{2}$, $\sum b_j^2 < 1$ and $b_j = b_j(\theta)$ with $\theta = (d, a, c, \ldots)$ denoting a finite dimensional parameter vector. The true parameter value will be denoted by θ^0. In the following, we summarize results from Beran and Schützner (2009). For simplicity of notation, we will consider the case with exact equality $b_j = cj^{d-1}$ $(j \ge 1)$ which implies that $\theta = (d, a, c)^T$.

Since σ_t is given explicitly as a function of past observations X_s $(s \le t-1)$, a plausible approach to estimating θ is to use the conditional likelihood function of $\varepsilon_t(\theta) = X_t/\sigma_t(\theta)$. If $\sigma_t(\theta)$ can be calculated exactly and θ is equal to the true parameter θ^0, then $\varepsilon_t(\theta)$ $(t \in \mathbb{Z})$ coincides with the innovations ε_t. Since ε_t $(t \in \mathbb{Z})$

are i.i.d. with density p_ε, the log-likelihood function can be written as

$$L_n(\theta) = \sum_{t=1}^{n} \log p_\varepsilon\big(\varepsilon_t(\theta)\big).$$

If differentiation with respect to θ is possible, then the maximum likelihood estimator of θ^0 can be defined as a solution of

$$\dot{L}_n(\hat{\theta}) := \dot{L}_n(\theta)|_{\theta=\hat{\theta}} = 0$$

(where "$\cdot$" denotes differentiation with respect to θ). In particular, if p_ε is a normal density function with mean zero, then

$$-\frac{2}{n} L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \frac{X_t^2}{\sigma_t^2(\theta)} + \log \sigma_t^2(\theta) + \log 2\pi \tag{6.8}$$

and

$$\begin{aligned} -\frac{2}{n} \dot{L}_n(\theta) &= \frac{\partial}{\partial \theta} \left[\sum_{t=1}^{n} \varepsilon_t^2(\theta) + \log \sigma_t^2(\theta) \right] \\ &= 2 \sum_{t=1}^{n} \dot{\varepsilon}_t(\theta) \varepsilon_t(\theta) + \frac{\dot{\sigma}_t^2(\theta)}{\sigma_t^2(\theta)}. \end{aligned}$$

If the innovations ε_t are not normally distributed, then this function can still be used to define an estimator $\hat{\theta}$, but the solution no longer coincides with the MLE and is therefore often called a pseudo- or quasi-maximum likelihood estimator (PMLE or QMLE).

If all quantities in the last equation are well defined, then the asymptotic distribution of $\hat{\theta}$ can be derived quite easily because $\dot{\varepsilon}_t(\theta^0)\varepsilon_t(\theta^0)$ is a martingale difference. However, in contrast to short-memory volatility models (Lee and Hansen 1994; Lumsdaine 1996; Berkes et al. 2003; Robinson and Zaffaroni 2006; Francq and Zakoian 2008; Truquet 2008), for LARCH processes with slowly decaying coefficients $b_j \sim cj^{d-1}$ $(0 < d < \frac{1}{2})$ several complications arise. First of all, it is not obvious whether $\sigma_t(\theta)$ is an ergodic process (see, e.g. Walters 2000; Krengel 1985; Petersen 1989). Moreover, for $\theta \neq \theta^0$, it is not even clear whether $\varepsilon_t(\theta) = \sum_{j=1}^{\infty} b_j(\theta) X_{t-j}$ is finite with probability one. (Note that for $\theta = \theta^0$ this problem disappears because $\varepsilon_t(\theta^0)$ is almost surely equal to the random variable ε_t.) The reason is that $\sum b_j(\theta) = \infty$ implies $\sum |b_j X_{t-j}| = \infty$ almost surely unless $P(\varepsilon_t = 0) = 1$. Similarly, it is not clear whether and in which sense the derivative of $\varepsilon_t(\theta)$ with respect to θ exists (this problem occurs even for $\theta = \theta^0$), and whether the derivative is equal to $\sum \dot{b}_j(\theta) X_{t-j}$. An additional technical property that has to be established when studying the asymptotic distribution of $\hat{\theta}$ is the measurability of infima involving $\sigma_t(\theta)$ on the (uncountable) set Θ.

Apart from these questions, there is also the problem that $\sigma_t^2(\theta)$ may become arbitrarily small. In particular, for $\theta \neq \theta^0$, $E[L_n(\theta)]$ may be infinite or not defined. In fact, Francq and Zakoian 2008 (also see Truquet 2008) showed that, because of this reason, even in the case of short memory with a finite number of nonzero coefficients b_j the estimator based on (6.8) is not consistent.

Finally, for long-memory LARCH models, the issue that has to be addressed is that $\sigma_t(\theta)$ depends on the entire past X_s ($s \leq t-1$), whereas the only available observations are $X_1, \ldots, X_n$. This means that σ_t cannot be calculated exactly. Because of the slow decay of b_j, finite approximations may not be very good.

6.2.2.2 Ergodicity

Let us start with the fundamental question of ergodicity. Ergodicity of the process $\sigma_t(\theta)$ ($t \in \mathbb{Z}$) follows once the existence of a measurable function $f : \mathbb{R}^\infty \to \mathbb{R}$ is established for which $\sigma_t(\theta) = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$ almost surely (Stout 1974, Theorem 3.5.8). In view of the definition

$$\sigma_t(\theta) = a + a \sum_{k=1}^{\infty} \sum_{j_1,\ldots,j_k=1}^{\infty} b_{j_1}(\theta) \cdots b_{j_k}(\theta) \varepsilon_{t-j_1} \cdots \varepsilon_{t-j_1-\cdots-j_k}, \tag{6.9}$$

the natural choice of f is

$$f = a + a \sum_{k=1}^{\infty} f_k$$

with

$$f_k(x_1, x_2, \ldots) = \sum_{1 \leq m \leq k} \sum_{\substack{j_1,\ldots,j_m=1 \\ j_1+\cdots+j_m=k}}^{\infty} b_{j_1} \cdots b_{j_m} x_{j_1} \cdots x_{j_1+\cdots+j_m}.$$

Almost sure convergence of $\sum f_k$ follows from the fact that, for each fixed t,

$$M_t(k) = f_k(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) \quad (k \in \mathbb{N})$$

is a martingale difference with respect to the sequence of σ-algebras $\mathcal{F}_k = \sigma(M_t(l), l \leq k)$. Measurability of f follows, for instance, from Corollary 2.1.3 in Straumann (2004).

6.2.2.3 Summability, Continuity and Differentiability

Next consider the existence of $\sigma_t(\theta)$ ($\theta \in \Theta$) and its derivatives. If the coefficients b_j were absolutely summable, then answering these questions would be straightforward because $\sum |b_j| < \infty$ implies absolute summability of the right-hand side of (6.9) which, in turn, implies that $\sigma_t(\theta)$ inherits the differentiability properties

of $b_j(\theta)$. For nonsummable coefficients, these arguments do not apply. The solution proposed in Beran and Schützner (2009) is to consider $\sigma_t(\theta)$ (for fixed t) as a stochastic process with index $\theta \in \Theta$. To carry over the properties of $b_j(\theta)$ to $\sigma_t(\theta)$, the process $\sigma_t(\theta)$ $(\theta \in \Theta)$ is assumed to be separable. More specifically, the technical condition can be written down as follows:

(S) For every $t \in \mathbb{Z}$, $(\sigma_t(\theta))_{\theta\in\Theta}$ is a separable stochastic process on Θ, i.e. for every open set $A \subset \Theta$ and closed interval B, the sets $\{\omega : \sigma_t(\theta) \in B, \forall\theta \in A\}$ and $\{\omega : \sigma_t(\theta) \in B, \forall\theta \in A \cap \mathbb{Q}^3\}$ differ only on a set $N \subset N_0$ where $P(N_0) = 0$.

Note that the original process $(\sigma_t(\theta))_{\theta\in\Theta}$ can always be replaced by a separable version (see Theorem 2.4 in Doob 1953). Before establishing differentiability of $\sigma_t(\theta)$, we recall two different definitions of derivatives that are particularly useful for stochastic processes.

Definition 6.1 A stochastic process $\xi(x)$ $(x \in [a, b])$ is uniformly mean squared differentiable (u.m.s.-differentiable), if there exists a process $\zeta(x) =: \xi'(x)$ $(x \in [a, b])$ such that

$$E\left[\left(\frac{\xi(x+h)-\xi(x)}{h} - \zeta(x)\right)^2\right] \underset{h\to 0}{\to} 0$$

uniformly in $x \in (a, b)$. The process $\xi'(x)$ is also called the L^2-derivative of $\xi(x)$.

Definition 6.2 Let $\Psi(a, b)$ be the set of (test) functions ψ that are infinitely continuously differentiable on (a, b) and such that the closure $\bar{K}_\psi$ of the support $K_\psi = \{x : \psi(x) \neq 0\}$ is a compact subset of (a, b). A function $g \in L^2(a, b)$ is called a generalized (or distributional) derivative of a function $f \in L^2(a, b)$, if

$$\int_a^b g(x)\psi(x)\,dx = -\int_a^b f(x)\psi'(x)\,dx$$

for all $\psi \in D(a, b)$.

Note that generalized derivatives extend differentiation to functions that are not differentiable in the usual sense (or more generally, to generalized functions). For an elementary introduction to generalized derivatives, see, e.g. Lighthill (1958). For a more detailed account of the theory and further references, see, e.g. Gelfand and Shilov (1966–1968), Kanwal (2004), Strichartz (1994), Vladimirov (2002), Zemanian (2010).

Example 6.2 Let $H(x) = 1\{x \geq 0\}$ be the Heaviside function defined on $(-\infty, \infty)$. For $\psi \in \Psi(-\infty, \infty)$, we then have

$$-\int_{-\infty}^{\infty} H(x)\psi'(x)\,dx = -\big[\psi(\infty) - \psi(0)\big] = \psi(0).$$

Thus, the generalized derivative H' is equal to the Dirac delta function δ defined by $\int \delta(x)\psi(x)\,dx = \psi(0)$ (for all $\psi \in \Psi(-\infty, \infty)$).

The following result is derived in Beran and Schützner (2009).

Theorem 6.4 *Suppose that there are constants $d_u < \frac{1}{2}$ and $0 < C < 1$ such that $b_j = cj^{d-1}$ with $d \in [0, d_u]$, $c \in [0, c_u(d)]$ and $c_u(d) = C/\sqrt{\sum_{j=1}^{\infty} j^{2d-2}}$. Assume furthermore that (S) holds. Then $\sigma_t(\theta)$ is almost surely infinitely many times differentiable in θ in the generalized sense, and the kth generalized partial derivative w.r.t. θ is given by*

$$\frac{\partial^k}{\partial\theta_{j_1}\cdots\partial\theta_{j_k}}\sigma_t(\theta) = \sum_{j=1}^{\infty} \frac{\partial^k}{\partial\theta_{j_1}\cdots\partial\theta_{j_k}} b_j(\theta) X_{t-j},$$

i.e. we can write $\dot{\sigma}_t := \partial/\partial\theta\sigma_t = \sum \dot{b}_j X_{t-j}$.

This theorem follows by applying the following results.

Lemma 6.1 *Let $\xi(x)$ $(x \in [a, b])$ be a separable and u.m.s.-differentiable process with the L^2-derivative $\xi'(x)$. Then $\xi'(x)$ is also a generalized derivative of $\xi(x)$.*

Lemma 6.2 (Kolmogorov) *Let $\xi(x)$ $(x \in [a, b])$ be such that $E[\xi(x)] = 0$, $E[\xi^2(x)] < \infty$ and*

$$E\left[\left|\xi(x_1) - \xi(x_2)\right|^{\alpha}\right] \le \text{const}|x_1 - x_2|^{1+\beta}$$

for some $\alpha, \beta > 0$. Then there exists a version of $\xi(x)$ with almost surely continuous paths.

Lemma 6.3 *Let $\xi(x)$ $(x \in [a, b])$ be a separable process, m times u.m.s.-differentiable with the L^2-derivatives $\xi^{(k)}$ $(k \le m)$ and such that the paths of $\xi^{(k)}$ $(k \le m)$ are almost surely continuous. Then $\xi(x)$ is also $(m-1)$-times continuously differentiable in the generalized sense.*

Note that the last lemma is essentially an application of Sobolev's famous embedding theorem (see, e.g. Adams and Fournier 2003). Using these lemmas, the theorem can be proved in three steps. First of all, it is obvious that the only problem with respect to differentiability occurs for d. The lemmas were therefore formulated for the case of a one-dimensional index x only. The other parameter components can be fixed, and we can write $\sigma_t = \sigma_t(d)$ and $\dot{b}_j = \frac{\partial}{\partial d} b_j$.

The first step of the proof is to show that $\sum \dot{b}_j X_{t-j}$ is indeed the L^2-derivative of σ_t in the u.m.s.-sense. This can be done directly by showing that

$$E\left[\left(\frac{\sigma_t(d+h) - \sigma_t(d)}{h} - \sum \dot{b}_j X_{t-j}\right)^2\right] \le \text{const} \cdot h^2$$

and similar inequalities for higher derivatives. In a second step, one shows in a similar way that the condition in Lemma 6.2 holds. Since $\sigma_t(d)$ $(d \in [d, d_u])$ is assumed to be separable, almost sure continuity of the paths of $\dot{\sigma}_t(d)$ $(d \in [d, d_u])$ and

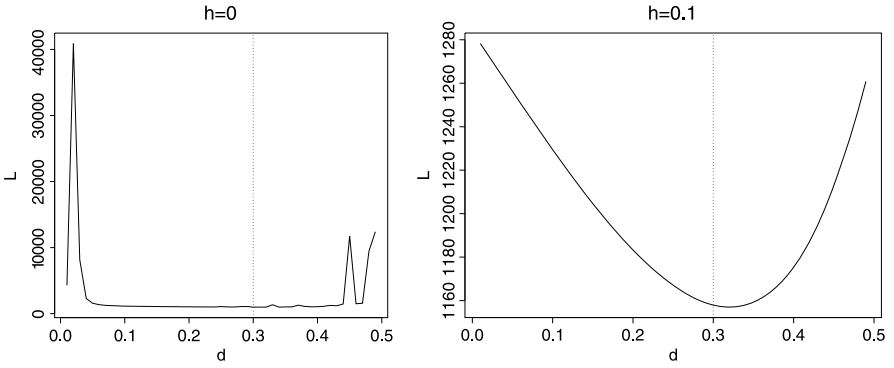

Fig. 6.1 Log-likelihood function $L_{n,h}(d)$ as a function of d for a simulated LARCH process with $b_j = cj^{d-1}$ and $d = 0.3$. The *left panel* shows $L_{n,h}$ for $h = 0$ whereas on the *right* $h = 0.1$ was used

higher order L^2-derivatives follows from Lemma 6.2. Finally, Lemma 6.3 implies that these are also derivatives in the generalized sense and the generalized derivatives $\sigma_t^{(k)}(d) = \frac{\partial^k}{\partial d^k}\sigma_t(d)$ $(k \le m-1)$ are almost surely continuous.

In a similar but slightly more involved manner, it can be shown that, under assumption (S), one can find bounds for $E(\sup_{\theta\in\Theta} |\sigma_t(\theta)|^m)$ $(m \ge 1)$ in terms of $\sup_{\theta\in\Theta} E(|\sigma_t(\theta)|^m)$ and $\sup_{\theta\in\Theta} E(|\dot{\sigma}_t(\theta)|^m)$. This is very useful for proving consistency (see below).

6.2.2.4 A Modified Log-likelihood Function

As mentioned above, a QMLE based on L_n in (6.8) is not consistent even in the case of short memory. The reason is that σ_t can be arbitrarily close to zero. Beran and Schützner (2009) therefore suggest a modified (quasi-) log-likelihood function. Multiplied by -1 it is given by

$$L_{n,h}(\theta) = n^{-1}\sum_{t=1}^{n}\left(\frac{X_t^2}{\sigma_t^2(\theta)+h} + \log\left[\sigma_t^2(\theta)+h\right]\right) \tag{6.10}$$

for some $h > 0$. Computationally, the effect of the correction is a regularization in the sense that the function $L_{n,h}$ becomes smoother, with clearly identifiable local minima. This is illustrated in Fig. 6.1 where $L_{n,h}$ is plotted against d (for fixed a and c) for $h = 0$ (left) and $h = 0.1$ (right), respectively. The correct value of $d = 0.3$ is indicated by a dotted vertical line. Obviously, for $h = 0$, the function is not suitable for minimization whereas the minimum for $h = 0.1$ is clearly visible and close to the true value.

The function $L_{n,h}$ can also be interpreted as a robust version of L_n in the following sense. Suppose that ε_t are Gaussian and instead of X_t we observe a perturbed process $Y_t = X_t + \zeta_t$ where ζ_t are i.i.d. $N(0,h)$-distributed, and independent of X_t.

Then $\operatorname{var}(Y_t \mid X_s, s \le t-1) = \sigma_t^2 + h$ so that the conditional log-likelihood function of $Y_1, \ldots, Y_n$ is given by

$$L_{n,Y}(\theta) = n^{-1} \sum_{t=1}^{n} \left[\frac{(X_t + \zeta_t)^2}{\sigma_t^2(\theta) + h} + \log\bigl(\sigma_t^2(\theta) + h\bigr) \right].$$

Integrating out ζ_t, we obtain $E_\zeta[L_{n,Y}(\theta)] = L_{n,h}(\theta)$.

6.2.2.5 Consistency

Let $\hat{\theta}_{n,h}$ be defined by minimizing $L_{n,h}$ with respect to θ and denote by θ^0 the true value of θ. Sufficient conditions for almost sure consistency of $\hat{\theta}_{n,h}$ are: (a) $\theta^0 \in \Theta^0$ (with θ^0 denoting the true parameter and Θ^0 the interior of Θ) and Θ is compact; (b) $L_{n,h}(\theta)$ is continuous and $\sup_\theta |L_{n,h}(\theta) - L_h(\theta)|$ converges a.s. to zero where

$$L_h(\theta) = E\bigl[L_{n,h}(\theta)\bigr]$$

and (c) $L_h(\theta)$ has a unique minimum at $\theta = \theta^0$.

Continuity of $L_{n,h}(\theta)$ follows from continuity of $\sigma_t^2(\theta)$ discussed in the previous section. Pointwise a.s. convergence of $|L_{n,h}(\theta) - L_h(\theta)|$ follows from ergodicity of $\sigma_t(\theta)$ (for each $\theta \in \Theta$) and

$$\sup_{\theta\in\Theta} E\left[\left| \frac{X_t^2 + h}{\sigma_t^2(\theta) + h} + \log\bigl(\sigma_t^2(\theta) + h\bigr) \right| \right] \le \text{const} \cdot \left\{ E\bigl[X_t^2\bigr] + h + \sup_{\theta\in\Theta} E\bigl[\sigma_t^2(\theta)\bigr] \right\}.$$

Since Θ is assumed to be compact, $\sup_{\theta\in\Theta} E[\sigma_t^2(\theta)] < \infty$ can be shown and thus Birkhoff's ergodic theorem implies $|L_{n,h}(\theta) - L_h(\theta)| \to 0$ almost surely. The convergence of $\sup_\theta |L_{n,h}(\theta) - L_h(\theta)|$ follows from equicontinuity of $L_{n,h}(\theta)$ which requires slightly more involved arguments (see Beran and Schützner 2009) involving certain moment conditions on ε_t.

The proof of (c) follows from

Lemma 6.4 *If ε_t are continuous random variables with density function p_ε, then*

$$P\bigl(\sigma_t(\theta) = 0\bigr) = 0 \quad (\textit{for all } t \textit{ and } \theta),$$
$$P\bigl(\sigma_t^2(\theta) = \sigma_t^2\bigl(\theta^0\bigr)\bigr) = 1 \implies \theta = \theta^0$$

and

$$\theta \neq \theta^0 \implies L_h(\theta) > L_h\bigl(\theta^0\bigr).$$

Proof Defining the set $N_t = \{\omega : \sigma_t(\theta) = 0\}$, the first equation means that $P(N_t) = 0$. To prove this, consider $\omega \in N_t \cap N_{t-1}^c$, i.e. we look at a realization

of the process such that $\sigma_t(\theta) = 0$ but $\sigma_{t-1}(\theta) \neq 0$. Then

$$0 = \sigma_t(\theta) = a + b_1(\theta)\overbrace{\varepsilon_{t-1}\sigma_{t-1}(\theta)}^{X_{t-1}} + \sum_{j=2}^{\infty} b_j(\theta) X_{t-j}$$

so that

$$\varepsilon_{t-1} = -\frac{1}{b_1(\theta)\sigma_{t-1}(\theta)}\left(a + \sum_{j=2}^{\infty} b_j(\theta) X_{t-j}\right).$$

However, the right-hand side involves only ε_s $(s \leq t-2)$ which is independent of the left-hand side ε_{t-1}. Therefore, since the ε_t's are assumed to be continuous variables, this equality can only occur with probability zero. In other words, $P(N_t \cap N_{t-1}^c) = 0$. The same arguments lead to $P(N_{t,k}) = 0$ where

$$N_{t,k} = \bigcap_{i=0}^{k-1} N_{t-i} \cap N_{t-k}^c \quad (k \geq 1).$$

Since $N_t = \bigcup_{k=1}^{\infty} N_{t,k}$, we obtain $P(N_t) = P(\sigma_t(\theta) = 0) = 0$.

Analogous arguments can be used to show that $P(\sigma_t^2(\theta) = \sigma_t^2(\theta^0)) = 1$ implies $\theta = \theta^0$. Finally, the last statement in the lemma follows from

$$L_h(\theta) - L_h\big(\theta^0\big) = E\left[\frac{\sigma_t^2(\theta^0) + h}{\sigma_t^2(\theta) + h} - \log\frac{\sigma_t^2(\theta^0) + h}{\sigma_t^2(\theta) + h} - 1\right]$$

and the inequality $u - \log u - 1 > 0$ $(u \neq 1)$. □

6.2.2.6 Asymptotic Normality

By similar arguments as for $L_{n,h}$, one can show that $\sup_\theta \|\dot{L}_{n,h}(\theta) - \dot{L}_h(\theta)\|$ and $\sup_\theta \|\ddot{L}_{n,h}(\theta) - \ddot{L}_h(\theta)\|$ (with the matrix norm $\|A\| = \sqrt{\mathrm{tr}(A^T A)}$) converge to zero almost surely. The asymptotic distribution of $\hat{\theta}_{n,h}$ can therefore be obtained by the Taylor approximation

$$\begin{aligned} 0 = L_{n,h}(\hat{\theta}_{n,h}) &\approx \dot{L}_{n,h}\big(\theta^0\big) + \ddot{L}_{n,h}\big(\theta^0\big)\big(\hat{\theta}_{n,h} - \theta^0\big) \\ &\approx \dot{L}_{n,h}\big(\theta^0\big) + \ddot{L}_h\big(\theta^0\big)\big(\hat{\theta}_{n,h} - \theta^0\big) \end{aligned} \tag{6.11}$$

implying

$$\hat{\theta}_{n,h} - \theta^0 \approx -\big[\ddot{L}_h\big(\theta^0\big)\big]^{-1}\dot{L}_{n,h}\big(\theta^0\big),$$

where $\ddot{L}_h(\theta) = E[\ddot{L}_{n,h}(\theta)]$. Thus, apart from the deterministic matrix $\ddot{L}_h(\theta^0)$, the asymptotic distribution of $\hat{\theta}_{n,h}$ is determined by the asymptotic distribution of

$\dot{L}_{n,h}(\theta^0)$ where

$$\dot{L}_{n,h}(\theta) = n^{-1}\frac{\partial}{\partial\theta}\left\{\sum_{t=1}^{n}\frac{X_t^2}{\sigma_t^2(\theta)+h} + \log\left[\sigma_t^2(\theta)+h\right]\right\}$$
$$= n^{-1}\sum_{t=1}^{n}\dot{\ell}_{t,h}(\theta)$$

with

$$\dot{\ell}_{t,h}(\theta) = 2\left(1 - \frac{X_t^2+h}{\sigma_t^2(\theta)+h}\right)\frac{\sigma_t(\theta)}{\sigma_t^2(\theta)+h}\dot{\sigma}_t(\theta).$$

For $\theta = \theta^0$, $E[\dot{\ell}_{t,h}(\theta^0) \mid \varepsilon_s, s \le t-1] = 0$ so that $\dot{\ell}_{t,h}(\theta^0)$ is a martingale difference. Therefore,

$$\sqrt{n}\dot{L}_{n,h}\left(\theta^0\right) \underset{d}{\to} Z_1$$

where Z_1 is a normal random vector with zero mean and covariance matrix

$$G_h = E\left[\dot{\ell}_{t,h}\left(\theta^0\right)\dot{\ell}_{t,h}^T\left(\theta^0\right)\right]$$
$$= 4E\left\{\frac{\sigma_t^6(\theta^0)[E(\varepsilon_t^4)-1]}{(\sigma_t^2(\theta^0)+h)^4}\dot{\sigma}_t\left(\theta^0\right)\dot{\sigma}_t^T\left(\theta^0\right)\right\}.$$

For the matrix $\ddot{L}_h(\theta^0)$, we have

$$\ddot{L}_h\left(\theta^0\right) = H_h = 4E\left[\frac{\sigma_t^2(\theta^0)}{(\sigma_t^2(\theta^0)+h)^2}\dot{\sigma}_t\left(\theta^0\right)\dot{\sigma}_t^T\left(\theta^0\right)\right].$$

Thus, we obtain (see Beran and Schützner 2009):

Theorem 6.5 *Suppose that H_h is nonsingular. Then, under suitable moment conditions,*

$$\sqrt{n}\left(\hat{\theta}_{n,h} - \theta^0\right) \to Z \underset{d}{\sim} N(0, V_h)$$

with covariance matrix

$$V_h = H_h^{-1}G_hH_h^{-1}.$$

It is interesting to see that in general H_h need not be of full rank. A sufficient condition for nonsingularity of H_h is that ε_t are continuous random variables. The proof essentially follows from $P(\sigma_t = 0) = 0$. To see this, we have to consider the quadratic form

$$u^T H_h u = 4E\left[\frac{\sigma_t^2}{(\sigma_t^2+h)^2}u^T\dot{\sigma}_t\dot{\sigma}_t^T u\right].$$

Since σ_t is not zero with probability one, the condition $u^T H_h u = 0$ can only be true if $P(\dot{\sigma}_t = 0) > 0$ or if $u = 0$. Considering, for instance, the specific case with $\theta = (a, c, d)^T$ and $b_j = cj^{d-1}$ $(j \geq 1)$, the equation $u^T \dot{\sigma}_t = 0$ can be written as

$$\begin{aligned} 0 &= u_1 \frac{\partial}{\partial a}\sigma_t + u_2 \frac{\partial}{\partial c}\sigma_t + u_3 \frac{\partial}{\partial d}\sigma_t \\ &= u_1 + \sum_{j=2}^{\infty} \bigl(u_2 j^{d-1} + u_3 c \log j \cdot j^{d-1}\bigr) X_{t-j} + u_2 \sigma_{t-1} \varepsilon_{t-1}. \end{aligned}$$

Since $P(\sigma_{t-1} = 0) = 0$, this can be rewritten as

$$-u_2 \varepsilon_{t-1} = \sigma_{t-1}^{-1} \left[u_1 + \sum_{j=2}^{\infty} \bigl(u_2 j^{d-1} + u_3 c \log j \cdot j^{d-1}\bigr) X_{t-j} \right].$$

However, the left-hand side is independent of the right-hand side. Since ε_t (and hence also X_t) has a continuous distribution, equality can only occur with positive probability if all components of u are zero. In other words, H_h is of full rank. Note that in a similar manner G_h can be shown to be positive definite.

It is interesting to look at the asymptotic covariance matrix of $\hat{\theta}_{n,h}$ for small values of h. Letting h tend to zero, we obtain in the limit

$$\lim_{h \to 0} V_h = \bigl[E\bigl(\varepsilon_t^4\bigr) - 1\bigr] H_0^{-1}$$

with

$$H_0 = 4E\left[\frac{\dot{\sigma}_t(\theta^0) \dot{\sigma}_t^T(\theta^0)}{\sigma_t^2(\theta^0)} \right].$$

In particular, if $E[\sigma_t^{-2}(\theta^0)] = \infty$, then the asymptotic variance of $\hat{\theta}_1 = \hat{a}$ is zero. (Note, however, that this does not necessarily follow for the other components $\hat{\theta}_2 = \hat{c}$ and $\hat{\theta}_3 = \hat{d}$.) It is also remarkable that $\hat{\theta}_{n,h}$ has the same rate of convergence, and formally also the same type of asymptotic covariance matrix, as estimators of comparable parameters for GARCH(p, q) and ARCH(∞) processes (cf. Berkes et al. 2003; Robinson and Zaffaroni 2006).

6.2.2.7 Estimation Given the Finite Past

Since σ_t depends on the complete past X_s $(s \leq t - 1)$, it cannot be calculated exactly. The simplest approximation is obtained by truncating the sum, i.e. setting all unobserved values X_s $(s \leq 0)$ equal to zero. This leads to the approximate estimator

$$\theta_{n,h}^* := \arg\min_{\theta \in \Theta} L_{n,h}^*(\theta),$$

where

$$L^*_{n,h}(\theta) := \frac{1}{n}\sum_{t=1}^{n} \frac{X_t^2 + h}{\bar\sigma_t^2(\theta) + h} + \ln\bigl(\bar\sigma_t^2(\theta) + h\bigr)$$

and

$$\bar\sigma_t(\theta) = a(\theta) + \sum_{j=1}^{t-1} b_j(\theta) X_{t-j}.$$

However, because of the slow decay $b_j \sim cj^{d-1}$, the error $\sigma_t(\theta) - \bar\sigma_t(\theta)$ may be quite large (note that the error is larger for smaller values of t). In fact, we have, as $t \to \infty$,

$$E\bigl[\bigl(\sigma_t(\theta) - \bar\sigma_t(\theta)\bigr)^2\bigr] = \sum_{j=t}^{\infty} b_j^2(c,d) \sim c_1 t^{2d-1}.$$

The question is therefore whether this approximation changes the asymptotic distribution of the estimator. As before, a Taylor expansion yields (cf. (6.11))

$$0 = \dot L^*_n\bigl(\theta^*_{n,h}\bigr) = \dot L^*_{n,h}(\theta_0) + \ddot L^*_{n,h}(\tilde\theta) \cdot \bigl(\theta^*_{n,h} - \theta^0\bigr)$$

so that the asymptotic distribution of $\theta^*_{n,h}$ follows from the asymptotic distribution of $\dot L^*_{n,h}(\theta^0)$. The latter is the same as for $\dot L_{n,h}(\theta^0)$ provided that

$$\Delta_n := \sqrt{n}\bigl(\dot L^*_{n,h}\bigl(\theta^0\bigr) - \dot L_{n,h}\bigl(\theta^0\bigr)\bigr) \xrightarrow{p} 0$$

as $n \to \infty$ which means that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \frac{\dot{\bar\sigma}_t(\theta)\bar\sigma_t(\theta)(X_t^2 + h)}{\bar\sigma_t^2(\theta) + h}\left(\frac{1}{\bar\sigma_t^2(\theta) + h} - \frac{1}{\sigma_t^2(\theta) + h}\right) \to_p 0.$$

Using the mean value theorem for $(x^2 + h)^{-1}$ and the asymptotic behaviour of $E[(\sigma_t(\theta) - \bar\sigma_t(\theta))^2]$, an upper bound for $E(|\Delta_n|)$ can be given by $E(|\Delta_n|) \le \text{const} \cdot n^d$. Unfortunately, for $d > 0$, this bound does not converge to zero. The errors $E[(\sigma_t(\theta) - \bar\sigma_t(\theta))^2]$ do not decay fast enough (in t) to be negligible when summing over all values of t. As a simple remedy, Beran and Schützner (2009) propose to use only those time points where a sufficient number of past observations is available. Specifically, let $m_n = [n^\beta] - 1$ for some $0 < \beta < 1$ where $[\cdot]$ is denotes the integer part,

$$L_{n,h;\beta}(\theta) := \frac{1}{m_n}\sum_{t=n-m_n}^{n} \frac{X_t^2 + \varepsilon}{\bar\sigma_t^2(\theta) + \varepsilon} + \ln\bigl(\bar\sigma_t^2(\theta) + \varepsilon\bigr)$$

and

$$\theta^{(\beta)}_{n,h} := \arg\min_{\theta\in\Theta} L_{n,h;\beta}(\theta).$$

Then, by similar arguments as before, and under suitable moment conditions,

$$n^{\frac{\beta}{2}}\big(\theta_{n,h}^{(\beta)} - \theta^0\big) \xrightarrow{d} N\big(0, H_h^{-1} G_h H_h^{-1}\big)$$

provided that $0 < \beta < 1 - 2d$. This means that the asymptotic normal distribution is the same as for $\hat{\theta}_{n,h}$; however, the rate of convergence is much slower than $n^{-\frac{1}{2}}$. For the "best" rate of $n^{d-\frac{1}{2}}$, one can at least show $E[|\theta_{n,h}^{(\beta)} - \theta^0|] \sim c_2 n^{-(\frac{1}{2}-d)}$, but it seems more difficult to derive the asymptotic distribution. The problem with a slower rate becomes worse if the long memory becomes stronger because β cannot exceed $1 - 2d$. For instance, for $d = 0.1$ we have $n^{\frac{1}{2}-d} = n^{-0.4}$ whereas for $d = 0.4$ the rate of convergence is $n^{-0.1}$ only. This makes a huge difference even for moderate sample sizes. For instance, for $n = 1000$, $n^{-0.1}/n^{-0.4} \approx 7.9$.

Although the explicit proofs in Beran and Schützner (2009) are written down for the specific case $b_j = cj^{d-1}$ ($\theta = (a, c, d)^T$) the generalization to general weights with $b_j \sim cj^{d-1}$ follows directly. A natural starting point is for instance given by coefficients defined by the fractional differencing operator, i.e. coefficients in the series (in $z \in \mathbb{C}$)

$$\sum_{j=1}^{\infty} b_j z^j = c(d)\big[(1-z)^{-d} - 1\big]$$

where

$$c^2(d) \le \left[\sum_{j=1}^{\infty} \binom{-d}{j}^2\right]^{-1}$$

(to ensure stationarity, see Sect. 2.1.3.6). This can easily be extended by multiplying the $\sum_{j=1}^{\infty} b_j z^j$ by a function $\psi(z)/\varphi(z)$ corresponding to an ARMA filter and adjusting the constant to satisfy the stationarity condition $\sum b_j^2 < 1$.

6.3 Statistical Inference for ARCH(∞) Processes

In this section, we briefly mention the existing theory for ARCH(∞) models. Location estimation mimics the results for SV and LARCH models; however, there are no available theorems for M-estimators. As for parametric estimation of dependence parameters, we note that the maximum likelihood estimation is much easier than in the LARCH(∞) case (Berkes and Horváth 2004). Furthermore, the MLE seems to be the most suitable approach. The Whittle estimator applied to squared sequences is no longer an approximation of the MLE and is indeed less efficient than the actual MLE (Giraitis and Robinson 2001).

6.3.1 Location Estimation

As in Sect. 6.2.1, we consider a time series $Y_t = \mu + X_t$; however, now the residuals X_t are generated by an ARCH(∞) process

$$X_t = \xi_t \sigma_t, \tag{6.12}$$

$$\sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2. \tag{6.13}$$

The random variables ξ_t are such that $E(\xi_t) = 0$ and $\sigma_\xi^2 = E(\xi_t^2) = 1$. Furthermore, $b_0 > 0$, $b_j \geq 0$ and $\sum b_j < 1$ (see Sect. 4.2.7). Then the central limit theorem holds for $S_n = \sum_{t=1}^n X_t$ (see Corollary 4.4) so that

$$\sqrt{n}(\bar{y} - \mu) \underset{d}{\to} N\left(0, \sigma_X^2\right)$$

with

$$\sigma_X^2 = \frac{b_0}{1 - \sum_{j=1}^{\infty} b_j}.$$

Thus, an approximate $(1-\alpha)$-confidence interval for μ can be given by

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}.$$

Since $\operatorname{var}(Y_1) = \operatorname{var}(X_1)$, the parameter σ_X can be estimated based on the observed data $Y_1, \ldots, Y_n$.

6.3.2 Estimation of Dependence Parameters

Consider a parametric ARCH(∞) process with $\mu = 0$ and coefficients $b_j = b_j(\theta^0)$ ($j \geq 0$) depending on a finite dimensional parameter vector $\theta^0 = (b_0^0, \vartheta^0)$. As in the LARCH case, quasi maximum likelihood estimation of θ^0 can be obtained by maximizing the Gaussian conditional log-likelihood function

$$-\frac{2}{n} L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \frac{X_t^2}{\sigma_t^2(\theta)} + \log \sigma_t^2(\theta) \tag{6.14}$$

where $\sigma_t^2(\theta) = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2$. In contrast to LARCH processes, no problems with respect to summability and differentiability of $\sigma_t^2(\theta)$ occur because the coefficients b_j are absolutely summable. For the same reason, the approximation of σ_t^2 by the truncated sum $b_0 + \sum_{j=1}^{t-1} b_j X_{t-j}^2$ is accurate enough to be negligible asymptotically. Moreover, by definition, σ_t^2 is bounded away from zero by b_0. Asymptotic

normality of $\hat{\theta}_{\mathrm{MLE}} = \arg\max L_n$ is shown in Weiss (1986) for ARCH(p) processes, Lee and Hansen (1994) and Lumsdaine (1996) for the GARCH(1, 1) model and Hall and Yao (2003) for GARCH(p, q) models. Similar results are also given in Berkes et al. (2003), Berkes and Horváth (2004). For more general ARCH(∞) processes, including the case of hyperbolically decaying coefficients b_j, Robinson and Zaffaroni (2006) derived the consistency of $\hat{\theta}_{\mathrm{MLE}}$.

Results on the asymptotic distribution for general ARCH(∞) processes are known for an alternative estimator (Giraitis and Robinson 2001), namely the Whittle estimator $\hat{\theta}_{\mathrm{Whittle}}$ based on the squared observations X_t^2 (see also Bollerslev 1986 and Robinson and Zaffaroni 1997, 1998 for earlier uses of Whittle estimation in volatility models). The idea is to write X_t^2 in the autoregressive form

$$\begin{aligned} X_t^2 &= E\big[X_t^2 \mid \mathscr{F}_{t-1}\big] + X_t^2 - E\big[X_t^2 \mid \mathscr{F}_{t-1}\big] \\ &= \sigma_t^2 + X_t^2 - \sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2 + \zeta_t \end{aligned}$$

with $\zeta_t = X_t^2 - \sigma_t^2$ and $\mathscr{F}_t$ the σ-algebra generated by X_s ($s \le t$). The residual process is a martingale difference with variance $\sigma_\zeta^2 = \mathrm{var}(\zeta_t)$. Since the equation can also be written as

$$\begin{aligned} \tilde{X}_t^2 = b_0^{-1} X_t^2 &= 1 + \sum_{j=1}^{\infty} b_0^{-1} b_j X_{t-j}^2 + b_0^{-1}\zeta_t \\ &= 1 + \sum_{j=1}^{\infty} \tilde{b}_j \tilde{X}_{t-j}^2 + \tilde{\zeta}_t, \end{aligned}$$

we may assume without loss of generality that $b_0 = 1$. Under moment assumptions (in particular, fourth-order stationarity of X_t), X_t^2 then has the spectral density

$$f_{X^2}(\lambda;\theta^0) = \frac{\sigma_\zeta^2}{2\pi} g_{X^2}(\lambda;\theta^0) = \frac{\sigma_\zeta^2}{2\pi}\left|1 - \sum_{j=1}^{\infty} b_j e^{-j\lambda}\right|^{-2}.$$

The Whittle estimator $\hat{\theta}_{\mathrm{Whittle}}$ of θ^0 based on this spectral density is obtained by minimizing

$$\mathscr{L}_{n,\mathrm{Whittle}}(\theta) = \frac{2}{n}\sum_{j=1}^{[(n-1)/2]} \frac{I_{n,X^2}(\lambda_j)}{g_{X^2}(\lambda_j;\theta)}$$

with respect to θ, where I_{n,X^2} is the periodogram of the sequence X_t^2 evaluated at the Fourier frequencies $\lambda_j = 2\pi j/n$ (cf. (5.42)). It should be noted, however, that, in contrast to L_n, the function $\mathscr{L}_{n,\mathrm{Whittle}}$ is not associated with a likelihood. In particular, for the case of Gaussian innovations ξ_t, L_n essentially corresponds to a (conditional) log-likelihood function whereas this is not the case for $\mathscr{L}_{n,\mathrm{Whittle}}$.

The reason is simply that the process X_t^2 is not Gaussian. This implies that, for Gaussian ξ_t, $\hat{\theta}_{\text{Whittle}}$ is asymptotically less efficient than $\hat{\theta}_{\text{MLE}}$. Specifically, Giraitis and Robinson (2001) derive for general ARCH(∞) processes (and suitable moment conditions) the limit

$$\sqrt{n}\big(\hat{\theta}_{\text{Whittle}} - \theta^0\big) \underset{d}{\to} N\big(0, 2W^{-1} + W^{-1} V W^{-1}\big)$$

where

$$W = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial\theta} \log g_{X^2}(\lambda) \left[\frac{\partial}{\partial\theta} \log g_{X^2}(\lambda)\right]^T d\lambda,$$

$$V = \frac{2\pi}{\sigma_\xi^2} \int_{-\pi}^{\pi} \frac{\partial}{\partial\theta} \frac{1}{g_{X^2}(\lambda_1)} \left[\frac{\partial}{\partial\theta} \frac{1}{g_{X^2}(\lambda_2)}\right]^T h(\lambda_1, -\lambda_2, \lambda_2)\, d\lambda_1\, d\lambda_2.$$

Here $h(\lambda_1, -\lambda_2, \lambda_2)$ denotes the fourth-order cumulant spectral density of X_t^2 defined by

$$h(\lambda_1, \lambda_2, \lambda_3) = \frac{1}{(2\pi)^3} \sum_{k_1,k_2,k_3=-\infty}^{\infty} \exp\big(-i(k_1\lambda_1 + k_2\lambda_2 + k_3\lambda_3)\big) c_{0,k_1,k_2,k_3}$$

where $c_{0,k_1,k_2,k_3} = cum(X_t^2, X_{t+k_1}^2, X_{t+k_2}^2, X_{t+k_3}^2)$ is the joint cumulant of the variables $Y_1 = X_t^2$, $Y_2 = X_{t+k_1}^2$, $Y_3 = X_{t+k_2}^2$, $Y_4 = X_{t+k_3}^2$. Recall that the cumulants $\kappa_{j_1,\ldots,j_m} = cum(Y_1^{j_1}, Y_2^{j_2}, \ldots)$ of a random vector $Y \in \mathbb{R}^m$ are the coefficients in the series expansion of the cumulant generating function

$$\kappa(u) = \log E\big[\exp\big(iu^T Y\big)\big] = \sum_{j_1,\ldots,j_m=0}^{\infty} \kappa_{j_1,\ldots,j_m} \frac{u_1^{j_1} \cdots u_m^{j_m}}{j_1! \cdots j_m!} i^{j_1+\cdots+j_m}.$$

For other estimators and a nice overview on estimation for ARCH(∞) processes, see, e.g. Giraitis et al. (2006).

Chapter 7
Statistical Inference for Nonstationary Processes

In this chapter, statistical inference for nonstationary processes is discussed. For long-memory, or, more generally, fractional stochastic processes this is of particular interest because long-range dependence often generates sample paths that mimic certain features of nonstationarity. It is therefore often not easy to distinguish between stationary long-memory behaviour and nonstationary structures. For statistical inference, including estimation, testing and forecasting, the distinction between stationary and nonstationary, as well as between stochastic and deterministic components, is essential.

The most obvious type of nonstationarity in time series is a deterministic trend. Related to that is the issue of parametric and nonparametric regression. Both topics will be addressed (Sects. 7.1, 7.2, 7.4, 7.5, 7.7). A common feature is that there is a distinct difference between fixed and random design regression. For most fixed designs, long memory influences the rate of convergence of parametric and nonparametric regression estimators. In contrast, random design often removes the effect of strong dependence. The issue is, however, more complex, and will be discussed in detail.

Standard techniques in nonparametric regression are kernel and local polynomial smoothing. The main question one has to address is the choice of a suitable bandwidth. In the context of fractional processes with an unknown long-memory parameter $d \in (-1/2, 1/2)$, this is a formidable task. The optimal bandwidth depends on the unknown long-memory parameter d. At the same time, using an inappropriate bandwidth leads to biased estimates of d. To complicate the matter, the possibility of nonstationarity due to integration (i.e. random walk type behaviour) cannot be excluded a priori, and may be masked by antipersistent dependence. Nevertheless, it is possible to design data driven algorithms for asymptotically optimal bandwidth selection and simultaneous estimation of dependence parameters as well as identification of random walk type structures (see Sect. 7.4.5.1). Extensions to nonlinear processes with trends are considered briefly in Sect. 7.4.10. As an alternative to kernel and local polynomial smoothing, trend estimation based on wavelets and the issue of optimal selection of the number of resolution levels is discussed in

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_7, © Springer-Verlag Berlin Heidelberg 2013

Sect. 7.5. Furthermore, a semiparametric regression model, also known as partial linear regression, is considered in Sect. 7.7.

Another important class of nonstationary models can be subsumed under the notion of local stationarity, in the sense that certain parameters change as a function of time. Quantile estimation along this line is discussed in Sect. 7.6. Local FARIMA type estimation is considered in Sect. 7.8.

The chapter concludes with a section on change point detection (Sect. 7.9). This is an important issue in the long-memory context because occasional structural changes often generate sample paths that resemble stationary processes with long-range dependence. A typical example is a model with occasional shifts in the mean. Various methods have been developed in the literature for distinguishing between structural changes and long-range dependence. We discuss a selection of typical methods.

7.1 Parametric Linear Fixed-Design Regression

In this section, we discuss estimation in fixed design linear regression with residuals exhibiting long memory. The least squares estimator (LSE) is compared with the BLUE. It turns out that under long memory (as well as under antipersistence) the LSE usually loses efficiency compared to the BLUE. This is in contrast to the case of weak dependence studied in Grenander (1954) and Grenander and Rosenblatt (1957). The concrete asymptotic results, however, depend on the combination of long-memory properties of the residuals and the type of regression functions (Yajima 1988, 1991). A practical problem with the BLUE is that the weights depend on the unknown autocovariance function of the residual process. For certain situations, Dahlhaus (1995) designed explicit weights that eliminate this problem. The asymptotic results for the LSE can be extended to robust estimation (see Giraitis et al. 1996a which is an extension of Beran 1991 to the regression context). Finally, we briefly discuss the question of optimal design in the linear (fixed-design) regression context.

7.1.1 Asymptotic Distribution of the LSE

We consider linear regression of the form

$$Y_t = \sum_{j=1}^{p} \beta_j x_{tj} + e_t \quad (t = 1, 2, \ldots, n) \tag{7.1}$$

where

$$e_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{7.2}$$

is a linear process with ε_t i.i.d., $E(\varepsilon_t) = 0$, $\operatorname{var}(\varepsilon_i) = \sigma_\varepsilon^2 < \infty$ and $a_j = c_a j^{d-1}$ $(0 < d < \frac{1}{2})$. The following notation will be used:

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad y(n) = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad e(n) = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix},$$

$$x_{t\cdot}(n) = \begin{pmatrix} x_{t1} \\ \vdots \\ x_t \end{pmatrix}, \qquad x_{\cdot j}(n) = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

and

$$\underset{n\times p}{X} = \big[x_{\cdot 1}(n), \ldots, x_{\cdot p}(n)\big] = \begin{bmatrix} x_{1\cdot}^T \\ \vdots \\ x_{n\cdot}^T \end{bmatrix}.$$

Then

$$y(n) = X\beta + e(n). \tag{7.3}$$

The least squares estimator of β is equal to

$$\hat{\beta}_{\mathrm{LSE}} = \big(X^T X\big)^{-1} X^T y(n) \tag{7.4}$$

so that

$$\hat{\beta}_{\mathrm{LSE}} - \beta = \big(X^T X\big)^{-1} X^T e(n) = \big(X^T X\big)^{-1} \begin{pmatrix} x_{\cdot 1}^T e(n) \\ \vdots \\ x_{\cdot p}^T e(n) \end{pmatrix}.$$

More generally, for a weighted least squares estimator with weights q_j $(j = 1, 2, \ldots, n)$ we have

$$\hat{\beta} = \big(X^T Q X\big)^{-1} X^T Q y(n) \tag{7.5}$$

and

$$\hat{\beta} - \beta = \big(X^T Q X\big)^{-1} X^T Q e(n) = \big(X^T Q X\big)^{-1} \begin{pmatrix} x_{\cdot 1}^T Q e(n) \\ \vdots \\ x_{\cdot p}^T Q e(n) \end{pmatrix} \tag{7.6}$$

where the $n \times n$ matrix Q is given by $Q = \operatorname{diag}(q_1, \ldots, q_n)$. The covariance matrix of $\hat{\beta}$ is equal to

$$\Sigma_{\hat{\beta}} = \operatorname{var}(\hat{\beta}) = \big(X^T Q X\big)^{-1} X^T Q \Sigma_e Q^T X \big(X^T Q X\big)^{-1}$$

where $\Sigma_e = [cov(e_i, e_j)]$ is the covariance matrix of $e(n)$. In particular, the best linear unbiased estimator (BLUE) is given by

$$\hat{\beta}_{\text{BLUE}} = \left(X^T \Sigma_e^{-1} X\right)^{-1} X^T \Sigma_e^{-1} y(n) \tag{7.7}$$

and its covariance matrix is equal to

$$\Sigma_{\hat{\beta}} = \text{var}(\hat{\beta}) = \left(X^T \Sigma_e^{-1} X\right)^{-1}.$$

To obtain a nondegenerate limit theorem for $\hat{\beta}$ defined in (7.5), we need to standardize the estimator by a matrix that takes into account that $\text{var}(\hat{\beta})$ depends on the design matrix X, the matrix Q and on the covariance matrix Σ_e of the residuals. The first issue is taken into account by the normalizing diagonal $p \times p$ matrix

$$D_n = \text{diag}\left(X'X\right) = \begin{pmatrix} \|x_{\cdot 1}\|^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \|x_{\cdot p}\|^2 \end{pmatrix}$$

where for $a \in \mathbb{R}^p$, $\|a\| = \sqrt{a_1^2 + \cdots + a_p^2}$ denotes the Euclidian norm. Then we can write

$$\begin{aligned} D_n^{\frac{1}{2}} \Sigma_{\hat{\beta}} D_n^{\frac{1}{2}} &= \left(D_n^{-\frac{1}{2}} X^T Q X D_n^{\frac{1}{2}}\right)^{-1} \left(D_n^{-\frac{1}{2}} X^T Q \Sigma_e Q^T X D_n^{-\frac{1}{2}}\right) \left(D_n^{-\frac{1}{2}} X^T Q X D_n^{-\frac{1}{2}}\right)^{-1} \\ &= C_n^{-1} \left(D_n^{-\frac{1}{2}} X^T Q \Sigma_e Q^T X D_n^{-\frac{1}{2}}\right) C_n^{-1}. \end{aligned}$$

For most deterministic design matrices X and weights q_j (i.e. Q), C_n converges to a nondegenerate $p \times p$ matrix C so that

$$D_n^{\frac{1}{2}} \Sigma_{\hat{\beta}} D_n^{\frac{1}{2}} \approx C^{-1} \left(D_n^{-\frac{1}{2}} X^T Q \Sigma_e Q^T X D_n^{-\frac{1}{2}}\right) C^{-1}$$

and

$$\begin{aligned} D_n^{\frac{1}{2}} (\hat{\beta} - \beta) &\approx C^{-1} \left(D_n^{-\frac{1}{2}} X^T Q\right) e(n) \\ &= C^{-1} W_n e(n) =: Z_n. \end{aligned}$$

Thus it is sufficient to study the asymptotic behaviour of $W_n e(n)$. If the elements of

$$W_n = D_n^{-\frac{1}{2}} X^T Q = [w_{j,n}]_{i,j=1,\ldots,p}$$

can be written as a function of i/n, then this amounts to studying the joint distribution of weighted sums

$$Z_{n,j} = \sum_{i=1}^{n} w_{j,n}\left(\frac{i}{n}\right) e_i \quad (j = 1, \ldots, p).$$

If, in addition,

$$w_{j,n}(u) \approx n^{-\kappa} w_j(u)$$

for a fixed weight functions w_j and a suitable power $n^{-\kappa}$, then results from Pipiras and Taqqu (2000c) can be used to obtain

$$n^{\kappa-H} D_n^{\frac{1}{2}} (\hat{\beta} - \beta) \underset{d}{\to} Z = C^{-1} \tilde{Z}$$

where $H = d + \frac{1}{2}$ and

$$\tilde{Z} = \int_0^1 w(u)\, dB_H(u) = \begin{pmatrix} \int_0^1 w_1(u)\, dB_H(u) \\ \vdots \\ \int_0^1 w_p(u)\, dB_H(u) \end{pmatrix}.$$

The vector Z is normally distributed with zero mean and covariance matrix $\mathrm{var}(Z) = C^{-1} V C^{-1}$ where the elements of $V = (v_{ij})_{i,j=1,\ldots,p}$ are given by

$$\begin{aligned} v_{ij} &= E\left[\left(\int_0^1 w_i(x)\, dB_H(x)\right)\left(\int_0^1 w_j(y)\, dB_H(y)\right)\right] \\ &= \int_0^1 \int_0^u w_i(x) w_j(y) (x-y)^{2d-1}\, dy\, dx. \end{aligned} \tag{7.8}$$

In terms of fractional integrals (see Sect. 7.3) this can also be written as

$$v_{ij} = \left(\frac{\Gamma(d+1)}{c_1}\right)^2 \int_{-\infty}^{\infty} \left(I_-^d w_i\right)(s) \left(I_-^d w_j\right)(s)\, ds \tag{7.9}$$

where

$$\left(I_-^d w_j\right)(s) = \frac{1}{\Gamma(d)} \int_0^1 u^{j-1} (u-s)_+^{d-1}\, du$$

for $0 \le s \le 1$ and zero otherwise, and c_1 is a constant that depends on d. To make sure that v_{ij} are all finite, certain conditions on w_j must be imposed. For instance, Deo (1997) defines the conditions $w_j \in C(0,1)$ and $x^\alpha (1-x)^\alpha w_j(x)$ bounded for $x \in [0,1]$ and a some $0 < \alpha < \min(\frac{1}{2}, 2d)$.

Example 7.1 Consider a polynomial regression model of degree p defined by $Y_i = \sum_{j=0}^p \beta_0 i^j + e_i$. Note that, for obvious reasons, we deviate slightly from the previous notation by including $j = 0$. Here, we have $X = [x_{\cdot 1}(n), \ldots, x_{\cdot p+1}(n)]$,

$x_{\cdot j}(n) = (1, 2^{j-1}, \ldots, n^{j-1})^T$, $x_{i\cdot}(n) = (1, i^1, \ldots, i^p)^T$,

$$\begin{aligned}\left\|x_{\cdot j}(n)\right\|^2 &= \sum_{i=1}^{n} i^{2j-2} = n^{2j-1} \sum_{i=1}^{n} \left(\frac{i}{n}\right)^{2j-2} n^{-1} \\ &\sim n^{2j-1} \int_0^1 s^{2j-2}\, ds = \frac{n^{2j-1}}{2j-1}\end{aligned}$$

and the $(p+1) \times (p+1)$ matrix

$$D_n \approx \begin{pmatrix} n & 0 & \cdots & 0 \\ 0 & \frac{n^3}{3} & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{n^{2p-1}}{2p-1} \end{pmatrix}.$$

Furthermore,

$$\begin{aligned}\left(X^T X\right)_{kl} &= x_{\cdot k}^T(n) \cdot x_{\cdot l}(n) = \sum_{i=1}^{n} i^{k+l-2} \\ &\sim n^{k+l-1} \int_0^1 s^{k+l-2}\, ds = \frac{n^{k+l-1}}{k+l-1}.\end{aligned}$$

For the LSE the elements of $C_n = (c_{ij})_{i,j=1,\ldots,p+1}$ are then given by

$$\begin{aligned}c_{kl} &= \left(D_n^{-\frac{1}{2}} X^T X D_n^{-\frac{1}{2}}\right)_{kl} = \frac{(X^T X)_{kl}}{\|x_{\cdot k}(n)\| \|x_{\cdot l}\|} \\ &\sim \frac{\sqrt{(2k-1)(2l-1)}}{k+l-1}\end{aligned}$$

and

$$\begin{aligned}[W_n]_{ji} &= \left(D_n^{-\frac{1}{2}} X^T\right)_{ji} = \frac{x_{ij}}{\|x_{\cdot j}\|} = \frac{i^{j-1}}{n^{j-\frac{1}{2}}} \sqrt{2j-1} \\ &= n^{-\frac{1}{2}} \left(\frac{i}{n}\right)^{j-1} \sqrt{2j-1}\end{aligned}$$

so that

$$\begin{aligned}w_{j,n}(u) &= n^{-\frac{1}{2}} w_j(u), \\ w(u) &= u^{j-1} \sqrt{2j-1}.\end{aligned}$$

Thus, we have $\kappa = \frac{1}{2}$. Putting these results together and noting that $\kappa - H = -d$, we obtain

$$n^{-d} D_n^{\frac{1}{2}} (\hat{\beta} - \beta) \underset{d}{\to} Z = C^{-1} \tilde{Z} \sim N\big(0, C^{-1} V C\big).$$

The explicit form of V is given by (Yajima 1988)

$$v_{ij} = \frac{\sqrt{(2i-1)(2j-1)}\Gamma(1-2d)}{\Gamma(d)\Gamma(1-2d)} \int_0^1 \int_0^1 x^{i-1} y^{j-1} |x-y|^{2d-1} \, dy \, dx. \quad (7.10)$$

7.1.2 The Regression Spectrum and Efficiency of the LSE

A natural question is whether the least squares estimator should be replaced by the best linear unbiased estimator (BLUE) that is optimally adapted to the covariance structure. This issue was first addressed in a systematic manner by Grenander (1954) and Grenander and Rosenblatt (1957) (also see, e.g. Priestley 1981 for a nice summary). To study the asymptotic covariance matrix of $\hat{\beta}_{\text{LSE}}$ and $\hat{\beta}_{\text{BLUE}}$ for a general class of deterministic regression functions the following conditions are imposed: Let

$$x_{\cdot j}(k) = \begin{pmatrix} x_{1+k,j} \\ \vdots \\ x_{n+k,j} \end{pmatrix}$$

with $x_{i,j} := 0$ if $i \notin \{1, 2, \ldots, n\}$ and

$$\big\langle x_{\cdot j}(0), x_{\cdot l}(k) \big\rangle = \sum_{i=1}^{n} x_{ij}(0) x_{il}(k).$$

Then we assume, as $n \to \infty$,

- (R1) $\|x_{\cdot j}\|^2 \to \infty$;
- (R2)

$$\frac{x_{nj}^2}{\|x_{\cdot j}\|^2} \to 0;$$

- (R3)

$$r_{jl}^{(n)}(k) = \frac{\langle x_{\cdot j}(0), x_{\cdot l}(k) \rangle}{\|x_{\cdot j}\| \|x_{\cdot k}\|} \to r_{jl}(k) \in \mathbb{R};$$

- (R4) Define the $p \times p$ matrix $R(k) = [r_{jl}(k)]_{j,l=1,\ldots,p}$. Then $R(0)$ is nonsingular.

The first condition makes sure that x_{ij} does not vanish too fast as time i tends to infinity. The second condition means that the last observed value x_{nj} does not dominate all the previous ones. Condition (R3) defines a kind of a cross-correlation.

The last condition excludes asymptotic collinearity of the explanatory variables. From the definition of $R(k)$ it follows that there is a (complex-valued) function $M : \lambda \to M(\lambda)$ assigning every frequency in $[-\pi, \pi]$ a $p \times p$ matrix $M(\lambda)$ such that

$$M(\lambda_2) - M(\lambda_1) \geq 0$$

for all $\lambda_2 \geq \lambda_1$, where "≥ 0" means positive semidefiniteness, and

$$R(k) = \int_{-\pi}^{\pi} e^{ik\lambda} \, dM(\lambda)$$

for all k. The so-called (regression) spectral distribution function $M(\cdot)$ plays a key role when comparing the relative asymptotic efficiency of the least squares estimator compared to the BLUE.

The matrix $R(k)$ may be interpreted as a (noncentred) asymptotic correlation matrix for the regression functions $x_{\cdot j}$. In particular, $R_{jj}(0) = \int dM_{jj}(\lambda) = 1$. This implies a property of M that turns out to be important in the context of long-range dependence. Suppose that

$$dM_{jj}(0) = M_{jj}(0+) - M_{jj}(0) = 1. \tag{7.11}$$

Since $dM_{jj}(\lambda) \geq 0$ and $|dM_{jl}(\lambda)| \leq dM_{jj}(\lambda)\, dM_{ll}(\lambda)$ this implies for all j, l,

$$dM_{jl}(\lambda) = 0 \quad (\lambda \neq 0). \tag{7.12}$$

As we will see below, (7.11) causes particular difficulties under long memory.

Example 7.2 Let $p = 1$ and $x_{t1} = x_t \equiv 1$. This means that Y_t is stationary and $\beta = \mu$ is the expected value of Y_t. Conditions (R1)–(R4) hold for obvious reasons, and $r(k) = r_{11}(k) \equiv 1$. Hence,

$$R(k) = \int_{-\pi}^{\pi} e^{ik\lambda} \, dM(\lambda) \equiv 1$$

so that M has a point mass at the origin such that (7.11) and (7.12) hold.

Example 7.3 For polynomial regression of order k we have $x_{tj} = t^{j-1}$ $(j = 1, \ldots, p;\ p = k + 1)$. Then, as $n \to \infty$,

$$\|x_{\cdot j}\|^2 = \sum_{t=1}^{n} t^{2j-2} \sim n^{2j-1} \int_0^1 u^{2j-2} \, du = \frac{n^{2j-1}}{2j-1}$$

and

$$r_{jl}^{(n)}(k) = \frac{\langle x_{\cdot j}(0), x_{\cdot l}(k) \rangle}{\|x_{\cdot j}\| \, \|x_{\cdot k}\|} \sim \sqrt{(2j-1)(2l-1)}\, n^{j+l-1} \sum_{t=1}^{n} t^{j-1} (t+k)^{l-1}$$

$$\sim \sqrt{(2j-1)(2l-1)} \int_0^1 u^{j+l-2} \, du = \frac{\sqrt{(2j-1)(2l-1)}}{j+l-1}.$$

Thus, the "lag" k does not matter, i.e. for all k we have

$$r_{jl}(k) = \int_{-\pi}^{\pi} e^{ik\lambda}\, dM_{jl}(\lambda) \equiv \frac{\sqrt{(2j-1)(2l-1)}}{j+l-1}$$

which implies $dM(\lambda) = 0$ $(\lambda \neq 0)$ and

$$dM_{jl}(0) = \frac{\sqrt{(2j-1)(2l-1)}}{j+l-1}.$$

In particular,

$$dM_{jj}(0) = \frac{2j-1}{2j-1} = 1$$

so that again (7.11) and (7.12) hold.

Example 7.4 Let $p = 1$ and $x_{t1} = \cos\lambda_0 t$ for some $\lambda_0 \in (0, \pi)$. Then

$$\|x_{\cdot 1}\|^2 \sim \frac{n}{2}$$

and

$$r_{11}^{(n)}(k) = \frac{\langle x_{\cdot 1}(0), x_{\cdot 1}(k)\rangle}{\|x_{\cdot 1}\|^2} = 2n^{-1}\sum_{t=1}^{n} \cos(\lambda_0 t)\cos\big(\lambda_0(t+k)\big)$$

$$= \cos\lambda_0 k + n^{-1}\sum_{t=1}^{n} \cos(2\lambda_0 t + \lambda_0 k) \sim \cos\lambda_0 k.$$

Thus, $dM(\pm\lambda_0) = \frac{1}{2}$ and $dM(\lambda) = 0$ otherwise.

Example 7.5 Let $p = 1$ and $x_t = x_{t1} = (-1)^t = \cos\pi t$. Then $x_t x_{t+k} = (-1)^k = \cos\pi k$, $\|x_{\cdot 1}\|^2 = n$ so that $r(k) = (-1)^k$. This implies $dM(\pm\pi) = \frac{1}{2}$ and $dM(\lambda) = 0$ otherwise.

Example 7.6 Let $p = 1$ and $x_t = x_{t1} = t(1 + e^{-i\lambda_0 t})$ for some $\lambda_0 \in (0, \pi)$. Note that the definitions above can be extended in a natural way to complex valued x-variables, with $\langle x_{\cdot j}(0), x_{\cdot l}(k)\rangle = \sum x_{tj}(0)\bar{x}_{tl}(k)$. Then

$$\|x_{\cdot 1}\|^2 = 2\sum t^2(1 + \cos\lambda_0 t) \sim \frac{2}{3}n^3$$

and

$$\begin{aligned}\langle x_{\cdot 1}(0), x_{\cdot 1}(k)\rangle &= \sum_{t=1}^{n} t(t+k)\bigl(1+e^{-i\lambda_0 t}\bigr)\bigl(1+e^{i\lambda_0(t+k)}\bigr)\\ &\sim \bigl(1+e^{i\lambda_0 k}\bigr)\sum_{t=1}^{n} t^2 \sim \bigl(1+e^{i\lambda_0 k}\bigr)\frac{1}{3}n^3.\end{aligned}$$

Hence

$$r(k) = r_{11}(k) = \frac{1}{2}\bigl(1+e^{i\lambda_0 k}\bigr) = \int_{-\pi}^{\pi} e^{ik\lambda}\, dM(\lambda)$$

so that

$$\begin{aligned} dM(0) &= M(0+) - M(0) = \frac{1}{2},\\ dM(\lambda_0) &= \frac{1}{2}\end{aligned}$$

and $dM(\lambda) = 0$ otherwise.

For residual processes with short-range dependence and spectral density f_e, the asymptotic covariance matrix of $\hat{\beta}_{\mathrm{LSE}}$ and $\hat{\beta}_{\mathrm{BLUE}}$ can be expressed in terms of M and f_e as follows (Grenander 1954; Grenander and Rosenblatt 1957):

Theorem 7.1 *Let* $f_e \in C[-\pi,\pi]$, $D_n = \operatorname{diag}(\|x_{\cdot 1}\|, \ldots, \|x_{\cdot p}\|)$ *and assume that* (*R*1)–(*R*4) *hold. Then, as* $n \to \infty$,

$$D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n \to 2\pi R^{-1}(0) \int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda) R^{-1}(0). \tag{7.13}$$

Theorem 7.2 *Under same assumptions as in Theorem* 7.1, *and* $f_e > 0$,

$$D_n \operatorname{var}(\hat{\beta}_{\mathrm{BLUE}}) D_n \to \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{f_e(\lambda)}\, dM(\lambda)\right]^{-1}. \tag{7.14}$$

Theorem 7.1 includes not only the case of short memory (with f continuous) but also antipersistence with $f_e(\lambda) = L(\lambda)|\lambda|^{-2d}$ $(-\frac{1}{2} < d < 0)$, provided that $L(\lambda)$ is continuous. However, if M is such that $dM(\lambda) = 0$ for all $\lambda \neq 0$, then $\int dM(\lambda) = 0$. In other words, for such explanatory variables the actual rate of convergence is faster than captured by (7.13). Theorem 7.2 does not include antipersistence because $f_e(\lambda) = 0$. The reason for the condition $f_e > 0$ is to avoid a pole in the integral $\int f_e^{-1}\, dM$. It should be noted, however, that the conditions as stated here are sufficient but not necessary. For instance, piecewise continuous spectral distributions f_e may be considered or even cases where $f_e(0) = 0$ provided that dM is zero in the neighbourhood of the origin. Long memory is, however, not included in any of the

two theorems (or possible simple modifications) because f_e has a pole. This causes difficulties with some of the integrals. A partial extension of the results was obtained by Yajima (1991). The main problem caused by the pole of f_e at the origin occurs when $dM(0) > 0$. The reason is that then $\int f_e(\lambda)\, dM(\lambda)$ is infinite. Moreover, if $dM(\lambda) = 0$ outside the origin, then $\int f_e^{-1}(\lambda)\, dM(\lambda) = 0$ so that we would divide by zero in (7.14).

Two cases have to be distinguished when considering long memory, namely

$$M_{jj}(0+) - M_{jj}(0) = 0 \quad \text{(case 1)} \tag{7.15}$$

and

$$M_{jj}(0+) - M_{jj}(0) > 0 \quad \text{(case 2)}. \tag{7.16}$$

For the second case, a more refined distinction will have to be made, namely

$$0 < M_{jj}(0+) - M_{jj}(0) < 1 \quad \text{(case 2a)} \tag{7.17}$$

and

$$M_{jj}(0+) - M_{jj}(0) = 1 \quad \text{(case 2b)}. \tag{7.18}$$

First, we state the result for case 1. Since M does not have any mass at zero, the pole of f_e does not disturb, i.e. there is no "interference" between long memory and the regression function.

Theorem 7.3 *Let* $f_e(\lambda) = L(\lambda)|1 - e^{-i\lambda}|^{-2d}$ $(0 < d < \frac{1}{2})$, $L \in C[-\pi, \pi]$, *and suppose that* (7.15) *holds for all* $j = 1, \ldots, p$. *Moreover, for* $j, l = 1, \ldots, p$ *define*

$$M_{jl}^{(n)}(\lambda) = \int_{-\pi}^{\lambda} m_{jl}^{(n)}(u)\, du,$$

$$m_{jl}^{(n)}(u) = \frac{\sum_{t=1}^{n} x_{tj} e^{-itu} \sum_{s=1}^{n} x_{sl} e^{isu}}{2\pi \|x_{\cdot j}\| \|x_{\cdot l}\|}.$$

Then, under (R1)–(R4),

$$D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n \to 2\pi R^{-1}(0) \int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda) R^{-1}(0) \tag{7.19}$$

if and only if for all $\delta > 0$ *there exists a finite constant* $c > 0$ *and* $n_0 \in \mathbb{N}$ *such that*

$$\int_{-c}^{c} f_e(\lambda)\, dM_{jj}^{(n)}(\lambda) < \delta \tag{7.20}$$

for all $j = 1, \ldots, p$ *and* $n \geq n_0$.

Proof Suppose first that (7.19) holds. For the left-hand side of (7.19), we have

$$D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n = \left(D_n^{-1} X^T X D_n^{-1}\right)^{-1} \left(D_n^{-1} X^T \Sigma X D_n^{-1}\right) \left(D_n^{-1} X^T X D_n^{-1}\right)^{-1}.$$

Due to (R3), $D_n^{-1} X^T X D_n^{-1}$ converges to $R(0)$. Hence (7.19) and the definition of $M^{(n)}$ imply

$$D_n^{-1} X^T \Sigma X D_n^{-1} = 2\pi \int_{-\pi}^{\pi} f_e(\lambda)\, dM^{(n)}(\lambda) \to 2\pi \int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda). \tag{7.21}$$

Since $M_{jj}(0+) - M_{jj}(0) = 0$, there exists a $c > 0$ such that $\int_{-c}^{c} f_e(\lambda)\, dM_{jj}(\lambda) < \delta$ for all j. Moreover, $M^{(n)}$ converges weakly to M and f_e is continuous on $\{|\lambda| \geq c\}$ so that

$$\int_{|\lambda|\geq c} f_e(\lambda)\, dM^{(n)}(\lambda) \to \int_{|\lambda|\geq c} f_e(\lambda)\, dM(\lambda). \tag{7.22}$$

Since also $\int_{-\pi}^{\pi} f_e(\lambda)\, dM^{(n)}(\lambda)$ converges to $\int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda)$ (7.21), (7.20) follows for n large enough.

Suppose now that (7.20) holds. Again, by the same argument, (7.22) holds. Therefore, (7.20) implies that $\int_{-\pi}^{\pi} f_e(\lambda)\, dM^{(n)}(\lambda)$ converges to $\int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda)$. □

Condition (7.20) holds, for instance, if $dM(\lambda) = 0$ in an open neighbourhood of the origin.

In case 2, components where (7.16) holds have to be standardized by a larger power of n as follows.

Theorem 7.4 *Let f_e be as in Theorem 7.3, $c_f = L(0) > 0$ and M such that (7.16) and (7.20) hold for $j = 1, \ldots, p$. Define the $p \times p$ matrix $V^* = [v_{jl}^*]_{j,l=1,\ldots,k}$ with the elements*

$$v_{jl}^* = c_f \lim_{n\to\infty} n^{-2d} \int_{-\pi}^{\pi} \left|1 - e^{-i\lambda}\right|^{-2d} dM_{jl}^{(n)}(\lambda)$$

and assume that all v_{jl}^* are finite. Then

$$n^{-2d} D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n \to V_{\mathrm{LSE}} = 2\pi R^{-1}(0) V^* R^{-1}(0). \tag{7.23}$$

Proof First, note that, by setting

$$\tilde{D}_n = \operatorname{diag}\left(\|x_{\cdot 1}\| n^d, \ldots, \|x_{\cdot p}\| n^d\right) = n^d D_n,$$

we have

$$\begin{aligned}\tilde{D}_n^{-1}\left(X^T X\right) \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) \left(X^T X\right) \tilde{D}_n^{-1} &= n^{-2d} D_n^{-1}\left(X^T X\right) \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) \left(X^T X\right) D_n^{-1} \\ &\sim n^{-2d} R(0) D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n R(0).\end{aligned}$$

Thus, we may consider

$$\tilde{D}_n^{-1}\left(X^T X\right) \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) \left(X^T X\right) \tilde{D}_n^{-1} = \tilde{D}_n^{-1} X^T \Sigma X \tilde{D}_n^{-1}.$$

Now

$$\begin{aligned}\tilde{D}_n^{-1} X^T \Sigma X \tilde{D}_n^{-1} &= \sum_{t,s=1}^{n} \frac{x_{tj}}{\|x_{\cdot j}\|} \frac{x_{sl}}{\|x_{\cdot l}\|} \gamma(t-s) \\ &= \int_{-\pi}^{\pi} \left(\sum_{t,s=1}^{n} \frac{x_{tj}}{\|x_{\cdot j}\|} \frac{x_{sl}}{\|x_{\cdot l}\|} e^{-i(t-s)\lambda} \right) f(\lambda)\, d\lambda \\ &= 2\pi \int_{-\pi}^{\pi} f_e(\lambda)\, dM^{(n)}(\lambda),\end{aligned}$$

by definition of $M_{jl}^{(n)}(\lambda)$ and $m_{jl}^{(n)}(\lambda)$. For $j \geq k+1$ the result follows as in the previous theorem. Moreover, since f_e is continuous for $|\lambda| \geq c$ and $M^{(n)} \to M$ weakly, we have

$$\int_{|\lambda| \geq c} f_e(\lambda)\, dM_{jl}^{(n)}(\lambda) \to \int_{|\lambda| \geq c} f_e(\lambda)\, dM_{jl}(\lambda) < \infty.$$

The only integral we need to take care of is $\int_{-c}^{c} f_e(\lambda)\, dM_{jl}^{(n)}(\lambda)$. Using the property $f_e(\lambda) \sim c_f |1 - e^{-i\lambda}|^{-2d}$ $(\lambda \to 0)$, one can show that

$$n^{-2d} \int_{-c}^{c} f_e(\lambda)\, dM_{jl}^{(n)}(\lambda) \sim n^{-2d} \int_{-\pi}^{\pi} \left|1 - e^{-i\lambda}\right|^{-2d} dM_{jl}^{(n)}(\lambda)$$

which converges to v_{jl}^* by assumption. □

The difference to case 1 characterized by (7.15) (and also to short memory) is that an additional normalization by n^{-2d} is required and a different limiting matrix V_{LSE} is obtained. The reason for the slower rate of convergence is that under (7.16) the regression functions have a strong low-frequency component in the sense that M includes a point mass at the origin. This interferes with the pole of f_e so that it becomes difficult to distinguish the low-frequency signal of the regression functions from low-frequency components in the residual process. Heuristically, the point mass of M at zero implies $\int f_e(\lambda)\, dM(\lambda) \geq f_e(0)\, dM(0) = \infty$ so that n^{-2d} has to be introduced to obtain a finite limit. A further interesting feature of (7.23) is that the asymptotic covariance matrix does not depend on the shape of f_e outside the origin. Only c_f and d are relevant. This is convenient for statistical inference since only these two parameters need to be estimated.

The evaluation of the matrix V^* is not always easy. An explicit formula is available for polynomial regression (Yajima 1988; also see Example 7.3):

Theorem 7.5 *Let f_e be as in Theorem 7.3, $c_f = L(0) > 0$ and $x_{tj} = t^{j-1}$. Then*

$$n^{-2d} D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n \to V_{\mathrm{LSE}} = 2\pi R^{-1}(0) V^* R^{-1}(0). \tag{7.24}$$

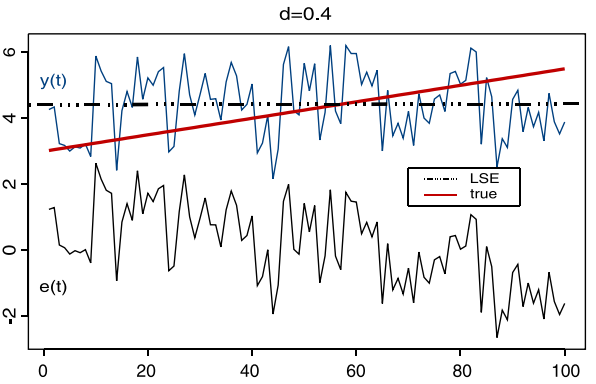

Fig. 7.1 $Y_t = 3 + 0.025t + e_t$ $(t = 1, 2, \ldots, 1000)$ where e_t is a FARIMA$(0, d, 0)$ process $e_t = (1 - B)^{-d}\varepsilon_t$ with $d = 0.4$ and $\mathrm{var}(\varepsilon_t) = 1$. The true trend function (*full line*) and the fitted least squares line (*dotted line*) are also plotted

where $[D_n]_{jj} \sim n^j/j$, *and* $R(0) = [r_{jl}]_{j,l=1,\ldots,p}$ *and* $V^* = [v^*_{jl}]_{j,l=1,\ldots,p}$ *have the elements*

$$r_{jl} \equiv \frac{\sqrt{(2j-1)(2l-1)}}{j+l-1}$$

and

$$v^*_{jl} = c_f \frac{\sqrt{(2j-1)(2l-1)}\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)} \int_0^1 \int_0^1 x^{j-1} y^{l-1} |x-y|^{2d-1}\, dy\, dx,$$

respectively.

Example 7.7 Figure 7.1 illustrates which problems long memory in the residual process may cause when the regression function has a zero-frequency component characterized by (7.16). Specifically, we observe $Y_t = 3 + 0.025t + e_t$ $(t = 1, 2, \ldots, 1000)$ where e_t is a FARIMA$(0, d, 0)$ process $e_t = (1 - B)^{-d}\varepsilon_t$ with $d = 0.4$ and $\mathrm{var}(\varepsilon_t) = 1$. The sample path of the residual process e_t (lower curve) has a spurious downward trend. The actual trend function with slope $\beta_1 = 0.025$ (full line) is therefore hardly visible in Y_t. The least squares estimate is indeed $\hat{\beta}_1 = 0.0002$ so that the fitted trend (dotted line) is practically horizontal. On the other hand, fitting a least squares line to the estimated residual process $\hat{e}_i$ yields $\hat{\beta}_1 = -0.025$. This is actually a spurious trend. If we use the usual t-test which assumes independence, then we come to the wrong conclusion that $\hat{\beta}_1$ is significantly different from zero with a p-value far below 1 %. Clearly, a correction of this test is needed to take into account the possibility of spurious trends in e_i. This is reflected in the additional norming constant n^{-2d} in Theorem 7.4. Theorem 7.5 leads to

$$V^* = \frac{2}{3} c_f \frac{\Gamma(1-2d)}{(2d+1)\Gamma(1-d)\Gamma(1+d)} = 1.29,$$

$D_n^2 \sim \frac{1}{3}n^3$ and $R(0) = 1$. Hence, an approximate corrected 95 %-confidence interval for β_1 is given by $-0.025 \pm 2\sqrt{3 \cdot 2\pi \cdot 1.29} n^{d-3/2} \approx [-0.09, 0.04]$ which includes zero.

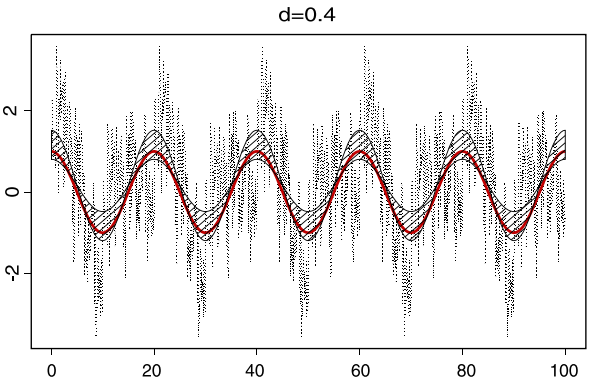

Fig. 7.2 $Y_t = \cos(2\pi t/100) + e_i$ $(t = 1, 2, \ldots, 1000)$ where e_t is a FARIMA$(0, d, 0)$ process $e_t = (1 - B)^{-d}\varepsilon_t$ with $d = 0.4$ and $\mathrm{var}(\varepsilon_t) = 1$. The true trend function (*full line*) is also plotted. The *shaded area* represents a 95 %-confidence region for the trend function, based on Theorem 7.3

Example 7.8 In Fig. 7.2, the same residuals as in the previous example are superimposed on a seasonal trend, namely $Y_t = \cos(2\pi t/100) + e_i$. In spite of the spurious trend in the residual sample path, it is not too difficult to distinguish the seasonal fluctuation from e_i. The reason is that the frequency $\lambda_0 = 2\pi/100 \approx 0.0628$ is isolated and relatively far from zero. Therefore, according to Theorem 7.3, $\hat{\beta}_{\mathrm{LSE}}$ has asymptotically the same rate of convergence as under independence. The only quantity that changes, depending on f_e, is the finite constant

$$D_n \,\mathrm{var}(\hat{\beta}_{\mathrm{LSE}}) D_n \to 2\pi R^{-1}(0) \int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda) R^{-1}(0),$$

$$2\pi \int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda) = 2\pi f_e(\lambda_0) = \left|1 - e^{-i\lambda_0}\right|^{-2d}.$$

The concrete estimate for the observed series in Fig. 7.2 is $\hat{\beta}_{\mathrm{LSE}} = 1.00$. Since

$$\sum_{t=1}^{n} \cos^2(\lambda_0 t) \approx \frac{1}{2} \sum_{t=1}^{n} \left|e^{i\lambda_0 t}\right|^2 = n/2,$$

we have $D_n^2 \sim \frac{1}{2}n$. An approximate 95 %-confidence interval for β_1 is therefore given by

$$\hat{\beta}_{\mathrm{LSE}} \pm 2\sqrt{2 \cdot 2\pi f_e(\lambda_0)} n^{-\frac{1}{2}} = 0.6 \pm 2\sqrt{31.9} n^{-\frac{1}{2}} = [0.64, 1.36].$$

This is shown in Fig. 7.2 as shaded area for the trend function.

A mixed result can also be obtained. If (7.15) holds for $j = 1, \ldots, k$ and (7.16) for $j = k + 1$, then, by setting

$$\tilde{D}_n = \mathrm{diag}\big(\|x_{\cdot 1}\|, \ldots, \|x_{\cdot k}\|, \|x_{\cdot k+1}\| n^d, \ldots, \|x_{\cdot p}\| n^d\big),$$

the asymptotic covariance matrix is of the form

$$V_{\mathrm{LSE}} = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}$$

where V_1 is as in Theorem 7.3 and V_2 as in 7.4.

The derivation of the asymptotic variance of $\hat{\beta}_{\mathrm{BLUE}}$ is a more challenging task. The first question is in how far formula (7.14) may be carried over to the long-memory case. The problem is that the integral $\int f_e^{-1}(\lambda)\, dM(\lambda)$ may be zero. More specifically, suppose that $M_{jj}(0+) - M_{jj}(0) = 1$. This implies $dM_{jl}(\lambda) = 0$ for all $\lambda \neq 0$ and $j, l = 1, \ldots, p$ (see (7.11) and (7.12)) so that $\int f_e^{-1}(\lambda)\, dM(\lambda) = 0$ and the inverse does not exist. Therefore, we have to distinguish between the cases 2a (7.17) and 2b (7.18), i.e. $0 < M_{jj}(0+) - M_{jj}(0) < 1$ and $M_{jj}(0+) - M_{jj}(0) = 1$, respectively. Under assumption (7.17), formula (7.14) indeed carries over to the long-memory case. The same is true for case 1 (7.15).

Theorem 7.6 *Let f_e be as in Theorem 7.3, $f_e > 0$ and M such that either (7.15) or (7.17) holds for $j = 1, \ldots, p$. Moreover, under (7.17) assume further that, for all $j = 1, \ldots, p$ and a suitable $\delta > 1 - 2d$,*

$$\max_{1 \le t \le n} \frac{x_{tj}^2}{\|x_{\cdot j}\|^2} = o\big(n^{-\delta}\big).$$

Then (7.14) holds, i.e.

$$D_n \operatorname{var}(\hat{\beta}_{\mathrm{BLUE}}) D_n \to V_{\mathrm{BLUE}} = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{f_e(\lambda)}\, dM(\lambda) \right]^{-1}. \tag{7.25}$$

Proof For case 1 with $M_{jj}(0+) - M_{jj}(0) = 0$, the result follows by analogous arguments as in the short-memory case because on $\{|\lambda| \ge c\}$ (with c arbitrary) f_e is continuous and such that $0 < f_e^{-1}(\lambda) < \infty$. For frequencies where $dM_{jj}(\lambda) > 0$, the function $f_e^{-1}(\lambda)$ is bounded away from zero.

Consider now case 2a, i.e. $0 < M_{jj}(0+) - M_{jj}(0) < 1$. Since

$$D_n \operatorname{var}(\hat{\beta}_{\mathrm{BLUE}}) D_n = \big(D_n^{-1} X^T \Sigma^{-1} X D_n^{-1}\big)^{-1},$$

we need to show that $D_n^{-1} X^T \Sigma^{-1} X D_n^{-1}$ converges to $(2\pi)^{-1} \int f_e^{-1}(\lambda)\, dM(\lambda)$. The essential problem is that we have to deal with the inverse of the covariance matrix. It can be shown by some extended algebra that indeed

$$D_n^{-1} X^T \big(\Sigma^{-1} - A_n\big) X D_n^{-1} \to 0 \tag{7.26}$$

where $A_n = [a_{jl}]_{j,l=1,\ldots,n}$ has the elements

$$a_{jl} = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} e^{i(j-l)\lambda} \frac{1}{f_e(\lambda)}\, d\lambda.$$

Showing (7.26) is the main difficulty of the proof (see Yajima 1991 for details). Using this approximation, we obtain for $C_n = [c_{jl}^{(n)}]_{j,l=1,\ldots,p} = D_n^{-1} X^T A_n X D_n^{-1}$,

$$c_{jl}^{(n)} = \sum_{t,s=1}^{n} \frac{x_{tj}}{\|x_{\cdot j}\|} \frac{x_{tl}}{\|x_{\cdot l}\|} \int_{-\pi}^{\pi} e^{i(j-l)\lambda} g(\lambda)\, d\lambda = \int_{-\pi}^{\pi} g(\lambda)\, dM_{jl}^{(n)}(\lambda)$$

where $2\pi g(\lambda) = 1/f_e(\lambda)$. Since $g(\lambda) \in C[-\pi, \pi]$ and $M^{(n)}$ converges weakly to M, this leads to

$$\lim_{n\to\infty} c_{jl}^{(n)} = \int_{-\pi}^{\pi} g(\lambda)\, dM(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{f_e(\lambda)}\, dM_{jl}(\lambda). \qquad \square$$

This result means that if the regression spectral distribution is not *completely* concentrated at the origin (cases 1 and 2a), then the pole of f_e at zero does not disturb the asymptotic covariance matrix of $\hat{\beta}_{\mathrm{BLUE}}$. In contrast, in order that the asymptotic covariance matrix of $\hat{\beta}_{\mathrm{LSE}}$ is unaffected by the pole of f_e, M must not have *any* mass at the origin. What happens otherwise is illustrated in Theorem 7.4.

A general result for $\hat{\beta}_{\mathrm{BLUE}}$ under condition (7.18) does not seem to be available currently. For polynomial regression, Yajima derived the following expression.

Theorem 7.7 *Let f_e be as in Theorem 7.3, $f_e > 0$ and $x_{tj} = t^{j-1}$ $(j = 1, \ldots, p)$. Then*

$$n^{-2d} D_n \operatorname{var}(\hat{\beta}_{\mathrm{BLUE}}) D_n \to V_{\mathrm{BLUE}} \tag{7.27}$$

where $V_{\mathrm{BLUE}} = 2\pi c_f W^{-1}$ and $W = [w_{jl}]_{j,l=1,\ldots,p}$ with

$$w_{jl} = \frac{\sqrt{(2j-1)(2l-1)}}{j+l-1-2d} \frac{\Gamma(j-d)\Gamma(l-d)}{\Gamma(j-2d)\Gamma(l-2d)}. \tag{7.28}$$

Note that, as for the LSE in case 2, the asymptotic covariance matrix V in (7.27) does not depend on the shape of f_e outside the origin.

Example 7.9 For $Y_t = \mu + e_t = \beta_0 + e_t$ with e_t generated by any stationary long-memory process with long-memory parameter d and a constant c_f, we have

$$W = w_{11} = \frac{1}{1-2d} \left[\frac{\Gamma(1-d)}{\Gamma(1-2d)} \right]^2 = \frac{\Gamma^2(1-d)}{\Gamma(1-2d)\Gamma(2-2d)}$$

so that

$$V_{\mathrm{BLUE}} = 2\pi c_f W^{-1} = 2\pi c_f \frac{\Gamma(1-2d)\Gamma(2-2d)}{\Gamma^2(1-d)}.$$

In comparison, for the LSE which is the sample mean $\bar{y}$, $R(0) = 1$ and

$$V_{\mathrm{LSE}} = 2\pi c_f \frac{\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)} \int_0^1 \int_0^1 |x-y|^{2d-1}\, dy\, dx$$

with

$$\int_0^1 \int_0^1 |x-y|^{2d-1}\,dy\,dx = \frac{2}{2d(2d+1)}.$$

Thus,

$$V_{\mathrm{LSE}} = 2\pi c_f \frac{\Gamma(1-2d)}{d(2d+1)\Gamma(d)\Gamma(1-d)}.$$

Note that in Sect. 1.3.1 we derived the asymptotic variance of the sample mean to be equal to

$$v(d)c_f = c_f \frac{2\Gamma(1-2d)\sin\pi d}{d(2d+1)}.$$

This is indeed the same as the previous formula because

$$\Gamma(d)\Gamma(1-d) = \frac{\pi}{\sin\pi d}.$$

The asymptotic relative efficiency of the LSE compared with the BLUE is equal to

$$e(d) = \frac{V_{\mathrm{BLUE}}}{V_{\mathrm{LSE}}} = \frac{(2d+1)\Gamma(2-2d)\Gamma(d+1)}{\Gamma(1-d)}. \tag{7.29}$$

This formula was first obtained by Adenstedt (1974) (also see Samarov and Taqqu 1988 and Beran and Künsch 1985), and holds for the whole range $-1/2 < d < 1/2$. We refer to the discussion in Sect. 5.2.2.

Example 7.10 Next, consider a linear trend model $Y_t = \beta_0 + \beta_1 t + e_t$ with e_t generated by any stationary long-memory process. Then

$$w_{11} = \frac{1}{1-2d}\left[\frac{\Gamma(1-d)}{\Gamma(1-2d)}\right]^2,$$

$$w_{22} = \frac{3}{3-2d}\left[\frac{\Gamma(2-d)}{\Gamma(2-2d)}\right]^2 = \frac{3(1-d)^2}{(3-2d)(1-2d)}w_{11}$$

and

$$w_{12} = w_{21} = \frac{\sqrt{3}}{2-2d}\frac{\Gamma(1-d)\Gamma(2-d)}{\Gamma(1-2d)\Gamma(2-2d)}$$

$$= \frac{\sqrt{3}(1-d)}{2-2d}w_{11}.$$

Thus

$$W = w_{11}\begin{pmatrix} 1 & \frac{\sqrt{3}(1-d)}{2-2d} \\ \frac{\sqrt{3}(1-d)}{2-2d} & \frac{3(1-d)^2}{(3-2d)(1-2d)} \end{pmatrix}.$$

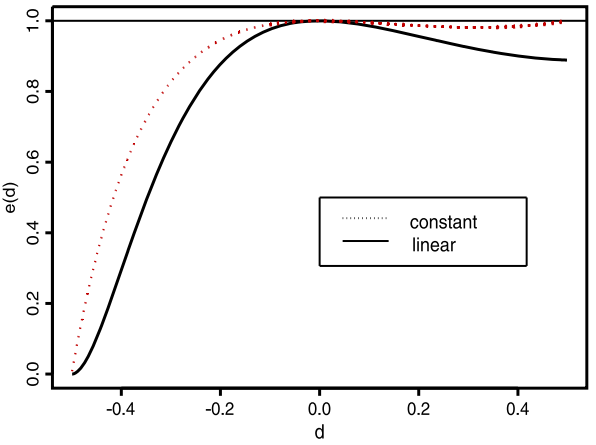

Fig. 7.3 Relative asymptotic efficiency $e(d) = \det(V_{\mathrm{BLUE}})/\det(V_{\mathrm{LSE}})$ of the least squares estimator in a linear regression model $Y_t = \beta_0 + \beta_1 t + e_t$ (*full linear*) and a regression model with $\beta_1 = 0$, i.e. $Y_t = \beta_0 + e_t$ (*dotted line*)

The inverse of W is equal to

$$W^{-1} = w_{11}^{-1} \begin{pmatrix} 4(1-d)^2 & -\frac{2}{\sqrt{3}}(3-2d)(1-2d) \\ -\frac{2}{\sqrt{3}}(3-2d)(1-2d) & \frac{4}{3}(1-2d)(3-2d) \end{pmatrix}.$$

The determinant of W^{-1} is equal to

$$\det\bigl(W^{-1}\bigr) = w_{11}^{-2}\left(4 - \frac{32}{3}\,d + \frac{16}{3}\,d^2\right)$$

so that

$$\det(V_{\mathrm{BLUE}}) = \left(\frac{2\pi c_f}{w_{11}}\right)^2 \left(4 - \frac{32}{3}\,d + \frac{16}{3}\,d^2\right).$$

By similar calculations, one can derive an explicit formula for V_{LSE} and the relative efficiency

$$e(d) = \frac{\det(V_{\mathrm{BLUE}})}{\det(V_{\mathrm{LSE}})} = \frac{(3+2d)(3-2d)}{36}\left[\frac{(1+2d)\Gamma(1+d)\Gamma(3-2d)}{\Gamma(2-d)}\right]^2.$$

(Note that there is a typo in Yajima 1988 in that $1/e(d)$ instead of $e(d)$ is given.) Figure 7.3 shows slightly larger efficiency losses than for the previous case where $\beta_0 = 0$. However, qualitatively the behaviour of $e(d)$ is quite similar.

Example 7.11 Let $Y_t = \beta_1(1 + \cos\lambda_0 t) + e_t$. Then this corresponds to case 2a with $0 < M(0+) - M(0) < 1$. Thus, Theorem 7.6 can be applied.

The next question is the comparison of the asymptotic covariance matrices for $\hat{\beta}_{\mathrm{LSE}}$ and $\hat{\beta}_{\mathrm{BLUE}}$. The previous examples illustrated that for polynomial regression $\hat{\beta}_{\mathrm{LSE}}$ is asymptotically efficient under short memory whereas this is not the case

when $d \neq 0$. In how far is this a general phenomenon? The short-memory case has been considered by Grenander (1954) (also see Grenander and Rosenblatt 1957). An essential notion in this context is the so-called regression spectrum:

Definition 7.1 Let M be a regression spectral distribution function. Then

$$S = \{\lambda \in [-\pi, \pi] : dM(\lambda) > 0\}$$

is called the regression spectrum.

Each (regression) spectral distribution function M can be decomposed in the following way.

Lemma 7.1 *There exist disjoint subsets $S_1, \ldots, S_m$ (for some $m \leq p$) such that*

$$S = \bigcup_{j=1}^{m} S_j$$

and

$$M(S_j)M^{-1}(\pi)M(S_j) = M(S_j),$$
$$M(S_j)M^{-1}(\pi)M(S_l) = 0 \quad (j \neq l)$$

where $M(S_j) = \int_{S_j} dM(\lambda)$ and $M(\pi) = \int_{-\pi}^{\pi} dM(\lambda)$.

Lemma 7.1 leads to the following definition.

Definition 7.2 The sets S_j are called the elements of the regression spectrum.

Using these definitions, Grenander derived the following necessary and sufficient conditions for the asymptotic efficiency of the LSE.

Theorem 7.8 *Let $f_e \in C[-\pi, \pi]$, $f_e > 0$, $D_n = \operatorname{diag}(\|x_{\cdot 1}\|, \ldots, \|x_{\cdot p}\|)$, assume that* (R1)–(R4) *hold and denote by $S_1, \ldots, S_m$ the elements of the regression spectrum. Then*

$$\lim_{n\to\infty} \operatorname{var}(\hat{\beta}_{\mathrm{BLUE}})\big[\operatorname{var}(\hat{\beta}_{\mathrm{LSE}})\big]^{-1} = I$$

if and only if there are constants c_j $(j = 1, \ldots, m)$ such that $f_e(\lambda) \equiv c_j$ for $\lambda \in S_j$ (i.e. f_e is constant on each S_j). Moreover, this is equivalent to

$$|S| \leq p, \qquad \sum_{\lambda \in S} \operatorname{rank}\{dM(\lambda)\} = p.$$

This is a classical result (see, e.g. Grenander and Rosenblatt 1957), and we therefore only outline the basic idea only. Suppose that f_e is indeed constant on each

element of the regression spectrum. Then Theorems 7.1 and 7.2 imply

$$\begin{aligned}&\operatorname{var}(\hat{\beta}_{\mathrm{BLUE}})\big[\operatorname{var}(\hat{\beta}_{\mathrm{LSE}})\big]^{-1}\\&\quad\sim 2\pi R^{-1}(0)\int f_e(\lambda)\,dM(\lambda)R^{-1}(0)\cdot\frac{1}{2\pi}\int\frac{1}{f_e(\lambda)}\,dM(\lambda).\end{aligned}$$

Using $R(0) = M(\pi)$ and Lemma 7.1, the right-hand side is equal to

$$\begin{aligned}&M^{-1}(\pi)\sum_{j,l=1}^{m} c_j M(S_j)M^{-1}(\pi)M(S_k)c_k^{-1}\\&\quad= M^{-1}(\pi)\sum_{j=1}^{m} M(S_j) = M^{-1}(\pi)M(\pi) = I.\end{aligned}$$

The question is under which circumstances Theorem 7.8 can be carried over to the case where $d \neq 0$. As we saw in the examples discussed previously, Theorem 7.8 no longer holds for polynomial regression, whereas $\hat{\beta}_{\mathrm{LSE}}$ turns out to be fully efficient for a periodic component. The essential argument in Theorem 7.8 is based on formulas (7.13) and (7.14) for the asymptotic covariance matrix of $\hat{\beta}_{\mathrm{LSE}}$ and $\hat{\beta}_{\mathrm{BLUE}}$, respectively. However, it is assumed implicitly that all quantities involved are finite. This is no longer the case, if f_e has a pole at the origin and $dM(0) > 0$. It can therefore be concluded that the LSE is asymptotically efficient, compared to the BLUE, if Theorems 7.3 and 7.6 are applicable and $dM(0) = 0$:

Theorem 7.9 *Let f_e and x_{tj} be as in Theorem 7.6 and $D_n = \operatorname{diag}(\|x_{\cdot 1}\|, \ldots, \|x_{\cdot p}\|)$. Assume that* (R1)–(R4) *hold and denote by $S_1, \ldots, S_m$ the elements of the regression spectrum $S = \bigcup S_j$ $(m \leq p)$. Then*

$$\lim_{n\to\infty} \operatorname{var}(\hat{\beta}_{\mathrm{BLUE}})\big[\operatorname{var}(\hat{\beta}_{\mathrm{LSE}})\big]^{-1} = I$$

if and only if $S_j = \{\lambda_j\}$ with $\lambda_j \in (0, \pi]$ and

$$\sum_{\lambda\in S}\operatorname{rank}\big\{dM(\lambda)\big\} = p.$$

Formally, the result is due to the fact that if $dM(0) < 1$, then there is at least one nonzero frequency where $dM(\lambda) > 0$. The integral $\int f_e^{-1}(\lambda)\,dM(\lambda)$ is therefore no longer zero and the usual formula for the asymptotic covariance matrix (which relies on the inverse of this integral) is applicable. Thus, essentially the LSE does not lose efficiency as long as the regression spectrum does not include the frequency zero. A loss of efficiency usually occurs, if $dM(0) > 0$. The intuitive reason is that in this case both the regression function and the residual process have a strong zero-frequency component. Incorporating the covariance structure in the estimator relieves this problem up to a certain extent. In fact, comparing Theorems 7.2 and 7.6,

in cases where $0 < dM(0) < 1$, this even leads to an improvement of the rate of convergence, matching the rate under short range dependence! This is illustrated by the following example.

Example 7.12 Let $Y_t = \beta_1(-1)^t + e_t$ with long-memory residuals e_t as above. Then $dM(\pm\pi) = \frac{1}{2}$ and zero otherwise, $D_n = \sqrt{n}$ and $R(0) = 1$. Thus, by Theorem 7.9, the LSE is asymptotically efficient. The asymptotic variance is given by

$$n \cdot \operatorname{var}(\hat{\beta}_1) \to V = 2\pi \int_{-\pi}^{\pi} f_e(\lambda)\, dM(\lambda) = 2\pi f_e(\pi).$$

For instance, if e_t is a FARIMA$(0, d, 0)$ process with variance one, then

$$V = \left|1 - e^{-i\pi}\right|^{-2d} \frac{\Gamma^2(1-d)}{\Gamma(1-2d)} = 2^{-2d} \frac{\Gamma^2(1-d)}{\Gamma(1-2d)}.$$

This is a monotonically decreasing function of d. In particular, for $d = 0$, we have $V = 1$ whereas, for instance, for $d = 0.4$ one obtains $V = 0.28$. The intuitive explanation for the better performance under long memory is that the sample paths of e_t tend to be "smoother" so that it is easier to distinguish them from the alternating function $x_t = (-1)^t$.

In summary, one can say that the efficiency of the LSE compared to the BLUE very much depends on the combination of the long-memory properties of e_i and the type of regression functions x_{tj}. A practical problem with the BLUE is, however, that the weights depend on the autocovariance function γ_e of the residual process. For observed data, γ_e is usually unknown and has to be estimated from the same data. Thus, in cases where only minor efficiency gains are to be expected, the LSE is preferred. In other cases, the BLUE is much more efficient so that one would like to use it. However, since γ_e has to be estimated, a balance between efficiency gain due to weighing by Σ^{-1} and additional inaccuracy induced by estimation of Σ has to be found. A further complication is that for large sample sizes and strong long memory inversion of Σ may be computationally difficult. As an alternative, Dahlhaus (1995) suggested using explicit weights without the need of inverting an $n \times n$ matrix. In particular, for polynomial regression with $x_{tj} = t^{j-1}$ $(j = 1, \ldots, p)$ he shows that the weighted estimator

$$\hat{\beta}_G = \left(X^T G X\right)^{-1} X^T G y(n)$$

with

$$\underset{p\times p}{G} = \operatorname{diag}\left(g(t_1), g(t_2), \ldots, g(t_n)\right),$$

$t_i = i/n$ and $g(u) = u^{-d}(1-u)^{-d}$ has the same asymptotic covariance matrix as the BLUE. In applications, one would use, for instance, $g_n(u) = u^{-d}(1 - u - \frac{1}{2}n)^{-d}$ to avoid $g(1) = \infty$. This result can be generated to regressors generated by Jacobi polynomials (see Dahlhaus 1995 for details; also see Sect. 3.1.4 for the definition of Jacobi polynomials).

7.1.3 Robust Linear Regression

Consider

$$Y_t = \sum_{j=1}^{p} \beta_j x_{tj} + e_t = x_{t.}'\beta + e_t \quad (t = 1, 2, \ldots, n) \tag{7.30}$$

as in (7.1) and a long-memory residual process as in (7.2). Denote by p_e the probability density function of the marginal distribution of e_t. A standard class of robust estimators of β (robust in the y-direction, see Hampel et al. 1986) can be defined as M-estimators, i.e. as solutions of p equations

$$\sum_{t=1}^{n} \psi\big(Y_t - x_{t.}'\hat{\beta}\big)x_{t.} = \underset{p\times 1}{0} \tag{7.31}$$

where ψ is such that $E[\psi(Y_t - x_{t.}'\beta)x_{t.}] = 0$. By similar arguments as for location estimation, it can be shown that the limit theorem (Theorem 4.33) for the empirical process implies asymptotic equivalence of any M-estimator and the LSE. If ψ is continuously differentiable, then this can be seen even more directly since (7.31) and consistency imply

$$\sum_{t=1}^{n} \psi\big(Y_t - x_{t.}'\beta\big)x_{t.} - \sum_{t=1}^{n} \dot{\psi}\big(Y_t - x_{t.}'\beta\big)x_{t.}x_{t.}'(\hat{\beta} - \beta) \approx 0$$

so that

$$\hat{\beta} - \beta \approx \big\{E\big[\dot{\psi}(e)\big]X'X\big\}^{-1}\frac{1}{n}\sum_{t=1}^{n}\psi(e_t)x_{t.}. \tag{7.32}$$

If we can use the approximation

$$\psi(e_t) = -\int \psi(u)p_e'(u)\,du \cdot e_t + r_t = -a_{\text{app},1}e_t + r_t$$

with $a_{\text{app},1} = E[\dot{\psi}(e_t)]$ and r_t in (7.32) is negligible (for instance, when a unique Appell expansion is valid), then

$$\hat{\beta} - \beta \approx \big(X'X\big)^{-1}\frac{1}{n}\sum_{t=1}^{n}x_{t.}e_t = \big(X'X\big)^{-1}X'e(n) = \hat{\beta}_{\text{LSE}} - \beta.$$

For more general, not necessarily differentiable, functions ψ, the limit theorem for the empirical process has to be applied more directly, along the lines of the proof of Theorem 5.1. A simplified version of the result in Giraitis et al. (1996a) can be stated as follows:

Theorem 7.10 *Let ψ be nondecreasing, right-continuous and bounded. Furthermore, suppose that $(X'X)^{-1}$ exists for n large enough,*

$$\sqrt{n}\max_{1\le t\le n}\left|x_{t\cdot}'\left(X'X\right)^{-\frac{1}{2}}\right| = O(1), \tag{7.33}$$

$e_t = \sum a_j \varepsilon_{t-j}$ is a linear process with $a_j \sim c_a j^{d-1}$ $(0 < d < \frac{1}{2})$, $E[|\varepsilon_t|^k] < \infty$ for all $k \in \mathbb{N}$ and denote by I the $p \times p$ identity matrix. Then

$$\operatorname{var}(\hat{\beta}_{\mathrm{LSE}})\left[\operatorname{var}(\hat{\beta})\right]^{-1} \to \underset{p\times p}{I}$$

and

$$\left[\operatorname{var}(\hat{\beta}_{\mathrm{LSE}})\right]^{-\frac{1}{2}}(\hat{\beta} - \hat{\beta}_{\mathrm{LSE}}) \to 0.$$

Example 7.13 For polynomial regression

$$c_{kl} = \left(D_n^{-\frac{1}{2}} X'X D_n^{-\frac{1}{2}}\right)_{kl} = \frac{(X'X)_{kl}}{\|x_{\cdot k}(n)\|\,\|x_{\cdot l}\|} \sim \frac{\sqrt{(2k-1)(2l-1)}}{k+l-1}$$

so that

$$\begin{aligned}\left|x_{t\cdot}'\left(X'X\right)^{-\frac{1}{2}}\right|^2 &= x_{t\cdot}'\left(X'X\right)^{-1}x_{t\cdot} \sim x_{t\cdot}' D_n^{-1} C^{-1} D_n^{-1} x_{t\cdot} \\ &= 1'C^{-1}1 \le p^2 \max_{1\le j,l\le p}|c_{jl}|.\end{aligned}$$

Thus (7.33) holds and the theorem can be applied, for instance, if e_t are generated by a FARIMA$(0, d, 0)$ process, then Theorem 7.10 holds.

7.1.4 Optimal Deterministic Designs

So far, it was assumed that the regression functions were evaluated at equidistant (time) points. For instance, for polynomial regression we considered $x_{ij} = i^{j-1}$ $(i = 1, \ldots, n)$. Replacing the diagonal matrix $D_n = \operatorname{diag}(n^{\frac{1}{2}}, n^{\frac{3}{2}}, \ldots, n^{\frac{2p-1}{2}})$ by $\tilde{D}_n = n \cdot \operatorname{diag}(1, 1, \ldots, 1)$ we may consider an analogous regression with $x_{ij} = t_i^{j-1} = g_j(t_i)$ where $t_i = i/n$. In some situations, it is possible to choose the points t_i where the regression functions are observed. This can be modelled as follows. For a given $T \in \mathbb{R}$, let

$$h : [0, 1] \to [-T, T] \tag{7.34}$$

be a function such that $h(t)$ can be written as a quantile $h(t) = F_h^{-1}(t)$ of a distribution function $F_h(x) = \int_{-\infty}^{x} \varphi(u)\,du$. Then it is assumed that the regression functions are generated at points

$$t_{i,n} = h\left(\frac{i-1}{n-1}\right).$$

The collection of all points,

$$\Xi_n = \{t_{1,n}, \ldots, t_{n,n}\} = \big\{h(0), \ldots, h(1)\big\},$$

is called the experimental design of the regression model. To obtain asymptotic results regarding the variance of $\hat{\beta}$, observations are assumed to be given by

$$Y_t = \beta_1 g_1(t) + \cdots + \beta_p g_p(t) + e_n(t) \quad (t = 1, \ldots, n) \tag{7.35}$$

where $e_n(t) = e_n^{(1)}(t) + e_n^{(2)}(t)$, $e_n^{(1)}$ and $e_n^{(1)}$ are zero mean processes, independent of each other, with variances σ_j^2 $(j = 1, 2)$, $e_n^{(1)}(t)$ being uncorrelated and $e_n^{(2)}(t)$ having autocorrelations

$$corr\big(e_n^{(2)}(t), e_n^{(2)}(t+k)\big) = \rho_n(k) = \rho(nk) \tag{7.36}$$

with $\rho(u) \sim c_\rho u^{2d-1}$ $(0 < d < \frac{1}{2})$ as $u \to \infty$. Moreover, g_j are "explanatory" linearly independent functions. We will use the notation

$$\kappa = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Note that (7.36) is equivalent to letting T in (7.34) tend to infinity while keeping ρ_n fixed. By similar arguments as in the previous sections, it can be shown that, under suitable regularity conditions, the asymptotic covariance matrix of the least squares estimator is given by Dette et al. (2009)

$$n^{1-2d} \cdot \mathrm{var}(\hat{\beta}_{\mathrm{LSE}}) = 2\sigma^2 c_\rho \kappa R_h^{-1}(0) V_h R_h^{-1}(0) \tag{7.37}$$

$$= 2\sigma^2 c_\rho \kappa \Psi(\varphi) \tag{7.38}$$

where

$$\big[R_h(0)\big]_{jl} = \int_0^1 g_j\big(h(u)\big) g_l\big(h(u)\big)\, du,$$

$$[V_h]_{j,l} = \int_0^1 g_j\big(h(u)\big) g_l\big(h(u)\big) Q\big(h'(u)\big)\, du$$

and

$$Q(v) = c_\rho^{-1} \lim_{n\to\infty} n^{-2d} \sum_{j=1}^{n} \rho(jv) = \frac{v^{2d-1}}{2d}.$$

Note in particular, that for an equidistant design with $h(u) = (2u-1)T$ (and hence $h'(u) \equiv 2T$), (7.37) gives back the asymptotic formulas in the previous section. An asymptotically optimal design is obtained by minimizing the function Ψ with respect to the design density φ.

Example 7.14 For $Y_t = \beta t + e_n(t)$, Dette et al. (2009) derived explicit expressions for the optimal design density φ_{opt}. Essentially, as d approaches 0, φ_{opt} tends to the uniform distribution on $[-T, T]$. This result is directly related to the fact that for short-memory processes the LSE is asymptotically efficient. Recall that for the same regression (however, with $t \in [0, 1]$), $w(u) = u^{-d}(1-u)^{-d}$ was the weight function yielding the same efficiency as the BLUE (Dahlhaus 1995). As $d \to 0$, w also converges to a constant function $w \equiv 1$. On the other hand, when d approaches $\frac{1}{2}$, then the optimal design density φ_{opt} puts more and more weight close to the left and right end of the interval. This is in correspondence with Dahlhaus' optimal weight function $w(u)$ in the equidistant case to having increasingly steeper poles at the ends of the interval. Intuitively, this means that one tries to estimate β from two parts of the series (the beginning and the end) that are as far apart in time as possible—thus avoiding too much correlation.

7.2 Parametric Linear Random-Design Regression

In this section, we address the problem of parameter estimation in a linear regression model

$$Y_t = \sum_{j=1}^{p} \beta_j X_{tj} + e_t \quad (t = 1, \ldots, n), \tag{7.39}$$

where the explanatory variables $X_{t,j}$ are random, and the processes $X_{t,j}$ $(t \in \mathbb{Z})$ and/or e_t $(t \in \mathbb{Z})$ may be strongly dependent or nonstationary. In Sect. 7.2.1, we start with two examples that illustrate possible effects of long memory in errors and predictors on parameter estimation in the random design case. These examples will provide some intuition for asymptotic results on contrast estimation. Estimation of contrasts is, historically, one of the first illustrations of the phenomenon that estimators in random design regression tend to perform better than in a typical fixed design case (Künsch et al. 1993, also see Beran 1994a, Chap. 9).

In Sect. 7.2.2, we focus on the heteroskedastic case

$$Y_t = \beta_0 + \beta_1 X_t + \sigma(X_t) e_t,$$

where $\sigma(\cdot)$ is a positive function. We assume that predictors and errors are stationary with possible long memory, independent from each other. The general theory for the LSE is based on randomly weighted partial sums (see Sect. 7.2.3) as presented in Kulik and Wichelhaus (2012), see also Guo and Koul (2008). Other approaches, tailored for the homoscedastic case $\sigma(\cdot) \equiv \sigma$ are presented, following Robinson and Hidalgo (1997) and Choy and Taniguchi (2001). Further results can be found in Koul (1992), Koul and Mukherjee (1993), Giraitis et al. (1996a), Koul and Surgailis (1997, 2000), Hallin et al. (1999), Chung (2002), Koul et al. (2004), Lazarova (2005).

Section 7.2.4 addresses the problem of spurious correlation between nonstationary series X_t, Y_t that are independent of each other. In the case of a random walk and

related integrated processes, it is well known that the sample correlation between two independent series does not converge to zero (see, e.g. Granger and Newbold 1974 and Phillips 1986). The same is true for fractionally integrated processes. We summarize detailed results including various combinations of nonstationarity, stationarity and long-range dependence as derived in Tsay and Chung (2000). Related results have been established in Phillips (1986, 1995), Phillips and Loretan (1991), Marmol (1995), Jeganathan (1999), Robinson and Marinucci (2003, 2003), Buchmann and Chan (2007).

Finally, Sect. 7.2.5 briefly addresses the problem of fractional cointegration. The idea of cointegration dates back to Granger (1981, 1983) and Engle and Granger (1987). In fractional cointegration, the reduction of the degree of integration is allowed to assume noninteger values. In some situations, this can lead to the lack of consistency of the LSE so that modifications are required (see, e.g. Robinson 1994a, 1994b and Marinucci 2000). Because the issue is of major interest in economics, there is meanwhile an extended literature. Important references are, for instance, Marinucci and Robinson (1999, 2001), Velasco (1999a, 1999b, 2003), Chen and Hurvich (2003a, 2003b, 2006) among others.

7.2.1 Some Examples, Estimation of Contrasts

As we saw in the previous section, the rate of convergence of (weighted) least squares estimators of β depends on the properties of the explanatory variables, i.e. on the regression design matrix X. If the explanatory themselves are random, then this means that the properties of $\hat{\beta}$ depend on the distribution of X_{tj} $(j = 1, \ldots, p)$. Relevant are mainly two questions:

1. Is $\mu_j = E(X_{tj})$ zero?
2. What is the temporal dependence structure of X_{tj}?

This is illustrated by the following examples.

Example 7.15 Let $Y_t = \beta X_t + e_t$ with X_t uncorrelated, $E(X_t) = 0$, $\mathrm{var}(X_t) = \sigma_X^2 < \infty$, e_t a zero mean stationary process with spectral density $f_e(\lambda) \sim c_f |\lambda|^{-2d}$ $(0 < d < \frac{1}{2})$ and independent of the process X_t. Then, by the law of large numbers, the asymptotic distribution of

$$\hat{\beta}_{\mathrm{LSE}} = \frac{\sum_{t=1}^{n} X_t Y_t}{\sum X_t^2} \sim \sigma_X^{-2} n^{-1} \sum_{t=1}^{n} X_t Y_t$$

is the same as that of

$$\sigma_X^{-2} n^{-1} \sum_{t=1}^{n} X_t Y_t.$$

Furthermore,

$$\operatorname{var}\left(\sigma_X^{-2} n^{-1} \sum_{t=1}^{n} X_t Y_t\right) = \operatorname{var}\left(\sigma_X^{-2} n^{-1} \sum_{t=1}^{n} X_t e_t\right) \sim \sigma_X^{-4} n^{-2} \cdot n \sigma_X^2 \sigma_e^2 = \frac{\sigma_e^2}{\sigma_X^2} n^{-1}.$$

Thus, X_t having zero mean and being uncorrelated removes a possible effect of (long-range) dependence in the residual process.

Example 7.16 Consider the same process as in the previous example; however, with $\mu = E(X_t) \neq 0$. Then the asymptotic distribution of $\hat{\beta}_{\mathrm{LSE}}$ is the same as that of

$$\left(\sigma_X + \mu_X^2\right)^{-2} n^{-1} \sum_{t=1}^{n} X_t Y_t.$$

Furthermore,

$$\begin{aligned}
\operatorname{var}\left(\sum_{t=1}^{n} X_t Y_t\right) &= \sum_{t,s=1}^{n} E[e_t e_s X_t X_s] \\
&= 2\mu_X^2 \sum_{k=1}^{n-1} \left(n - |k|\right) \gamma_e(k) + \left(\sigma_X + \mu_X^2\right) n \sigma_e^2 \\
&\sim \mu_X^2 \cdot \text{const} \cdot n^{2d+1} + o\left(n^{2d+1}\right).
\end{aligned}$$

Hence, even though X_t are uncorrelated, the possible long-range dependence stemming from the residuals is not removed.

Example 7.17 Let $X_t = (-1)^{Z_t}$ where Z_t are i.i.d. Bernoulli random variables with $P(Z_t = 1) = P(Z_t = 0) = \frac{1}{2}$ and independent of e_t. Then $\sigma_X^2 = 1$ and

$$\operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) \sim \sigma_e^2 n^{-1} = n^{-1} \int_{-\pi}^{\pi} f_e(\lambda)\, d\lambda.$$

It is in particular interesting to compare this with the asymptotic variance of $\hat{\beta}_{\mathrm{LSE}}$ for the fixed-design regression with $X_t = (-1)^t = \cos \pi t$ where, from Theorem 7.3, one obtains $n^{-1} 2\pi f_e(\pi)$. If f_e achieves its minimum at $\lambda = \pi$, then this means that alternating the sign systematically yields a better estimate of β than if assigning the sign purely randomly. For instance, for a fractional ARIMA$(0, d, 0)$ model with $d > 0$, $f_e(\pi)$ coincides with minimum of f_e whereas the contrary is true for $d < 0$. For $d = 0$, f_e is constant so that $2\pi f_e(\pi)$ and $\int_{-\pi}^{\pi} f_e(\lambda)\, d\lambda$ are the same.

From the applied point of view, a simple principle that may be deduced from these examples is that estimation of 'absolute' constants is more difficult than estimation of contrasts (for the definition of contrasts, see (7.43)). Or in other words, it is easier to compare constants than to estimate their individual values. This has been

known to applied statisticians for a long time. In the context of long-memory processes and simple experimental designs, this principle can be formulated explicitly as follows (see Künsch et al. 1993). Suppose p treatments are assigned randomly to n observational units that are observed in a certain temporal (or other) sequence. Assuming an additive effect of the treatments leads to the regression model

$$Y_t = \sum_{j=1}^{p} \beta_j x_{t,j} + e_t = x_{t\cdot}^T \beta + e_t \tag{7.40}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$, β_j is the jth treatment effect and e_t is a zero mean process with spectral density $f_e \sim c_e |\lambda|^{-2d}$ $(\lambda \to 0)$. The explanatory variables are defined by

$$x_{t,j} = 1\{a_t = j\}$$

with $a_t \in \{1, \ldots, p\}$ defining the treatment used. The question is now in how far long memory in the residuals affects the estimation of β and, in particular, whether the least squares estimator is asymptotically efficient. Furthermore, one may ask whether there are designs (random allocations of treatments) that improve the accuracy of estimates.

Künsch et al. (1993) considered the following standard designs:

(a) Complete randomization: a_t are i.i.d. with

$$P(a_t = j) = \pi_j.$$

(b) Restricted randomization: Given n, the number of assignments to treatment j $(j = 1, \ldots, p)$ is fixed, i.e. $n = n_1 + \cdots + n_p$ and

$$\sum_{t=1}^{n} x_{t,j} = n_j,$$

and all possible allocations of this type have the same probability

$$P(a_1, \ldots, a_n \mid n_1, \ldots, n_p) = p(a_1, \ldots, a_n) = \frac{n!}{n_1! \cdots n_p!}.$$

(c) Complete blockwise randomization: Restricted randomization within blocks, i.e. define $b = [n/l]$ blocks of length l,

$$B_k = \{(k-1)l + 1, \ldots, kl\}$$

and, within each block (and independently of other blocks), apply restricted randomization subject to

$$\sum_{t \in B_k} x_{t,j} = l_j \geq 1,$$

$$l = l_1 + \cdots + l_p.$$

The main difference between (a) and (b) is that in (a) n_j ($j = 1, \ldots, p$) are random whereas they are fixed in (b). However, in (a) n_j/n converges to π_j almost surely so that for n large enough, n_j is "in the neighbourhood" of the fixed number $n\pi_j$. The randomization in case (c) is even more restricted than in (b) because the number of assignments to treatment j is also fixed within each block. A typical choice of l and l_j in (c) is $l = p$ and $l_j = 1$.

In vector form, model (7.40) can be written as

$$Y(n) = X\beta + e(n) \tag{7.41}$$

with $Y(n) = (Y_1, \ldots, Y_n)^T$,

$$X = (x_{\cdot 1}, \ldots, x_{\cdot p}) = \begin{pmatrix} x_{1\cdot}^T \\ \vdots \\ x_{n\cdot}^T \end{pmatrix},$$

and column and row vectors $x_{\cdot j} = (x_{1j}, \ldots, x_{nj})^T$ and $x_{t\cdot} = (x_{t1}, \ldots, x_{tj})^T$, respectively such that

$$1^T x_{t\cdot} = \sum_{j=1}^{p} x_{tj} = 1, \qquad 1^T x_{\cdot j} = \sum_{t=1}^{n} x_{tj} = n_j.$$

By definition, column vectors are orthogonal, i.e.

$$\langle x_{\cdot j}, x_{\cdot l} \rangle = \sum_{t=1}^{n} x_{tj} x_{tl} = n_j \cdot \delta_{jl}$$

so that

$$X^T X = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n_p \end{pmatrix}.$$

Therefore, the least squares estimator of β can be written in a simple form

$$\hat{\beta}_{\mathrm{LSE}} = \left(X^T X\right)^{-1} X^T y(n) = \begin{pmatrix} n_1^{-1} \sum_{t=1}^{n} x_{t1} y_t \\ \vdots \\ n_p^{-1} \sum_{t=1}^{n} x_{tp} y_t \end{pmatrix}. \tag{7.42}$$

For the BLUE, we have the usual formula

$$\hat{\beta}_{\mathrm{BLUE}} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} y(n).$$

Now, instead of β itself, we are interested in estimation of contrasts. A contrast is defined by

$$c = \eta^T \beta = \sum_{j=1}^{p} \eta_j \beta_j, \tag{7.43}$$

where η is a deterministic vector such that

$$1^T \eta = \sum_{j=1}^{p} \eta_j = 0.$$

The variance of any estimated contrast can be written in terms of variances of estimates of the simple contrasts

$$c_{jk} = \beta_j - \beta_k.$$

It is therefore sufficient to study the variance of $\hat{c}_{jk} = \hat{\beta}_j - \hat{\beta}_k$. Since usually inference is carried out conditionally on the given (randomly generated) design, one has to consider the asymptotic behaviour of the conditional variance $V_n(\hat{c}_{jk} \mid X) = \mathrm{var}(\hat{c}_{jk} \mid X)$. Comparing the LSE and the BLUE of c_{jk}, the corresponding conditional variances $V_n(\hat{c}_{jk;\mathrm{LSE}} \mid X)$ and $V_n(\hat{c}_{jk;\mathrm{BLUE}} \mid X)$ will be denoted by $V_{n,\mathrm{LSE}}(X)$ and $V_{n,\mathrm{BLUE}}(X)$, respectively. The following result can be obtained by relatively simple approximations of the second moment.

Theorem 7.11 *Let f_e satisfy one of the following conditions*: (i) *f_e is piecewise continuous and $0 < c \le f_e \le C$ for suitable finite constants c and C, or* (ii) *$f_e(\lambda) = L(\lambda)|\lambda|^{-2d}$ with $0 < d < \frac{1}{2}$, $L(\cdot)$ continuous, of bounded variation and $0 < c \le L \le C$. Then, under complete randomization* (*design* (a)), *we have, as $n \to \infty$,*

$$\begin{aligned} nV_{n,\mathrm{LSE}}(X) &\underset{\text{a.s.}}{\to} \sigma_e^2 \left(\frac{1}{\pi_j} + \frac{1}{\pi_k} \right), \\ nV_{n,\mathrm{BLUE}}(X) &\underset{\text{a.s.}}{\to} \sigma_e^2 \left(\frac{1}{\pi_j} + \frac{1}{\pi_k} \right) \left[\frac{\sigma_e^2}{(2\pi)^2} \int_{-\pi}^{\pi} \frac{1}{f_e(\lambda)} \, d\lambda \right]^{-1}. \end{aligned} \tag{7.44}$$

The first remarkable result in this theorem is that contrasts can be estimated with the same rate of convergence as under independence, since $V_n = O(n^{-1})$. This is in sharp contrast to estimates of the slope parameters β_j themselves. Since the expected value of the explanatory variables is not zero, the rate of convergence of $\hat{\beta}_{j,\mathrm{LSE}}$ and $\hat{\beta}_{k,\mathrm{BLUE}}$ is slower, namely $\mathrm{var}(\hat{\beta}) \sim \mathrm{const} \cdot n^{2d-1}$. In contrast to the case of uncorrelated residuals, however, $\hat{\beta}_{j,\mathrm{LSE}}$ and $\hat{c}_{jk,\mathrm{LSE}}$ loses efficiency compared to $\hat{\beta}_{j,\mathrm{BLUE}}$ and $\hat{c}_{jk,\mathrm{BLUE}}$. This is even true for cases where $d = 0$ but f_e is not constant. Note that this is very much in contrast to fixed-design regression under Grenander's conditions. There, under short memory, $\hat{\beta}_{j,\mathrm{LSE}}$ (and hence also $\hat{c}_{jk,\mathrm{LSE}}$) does not lose efficiency. Here, under the given random design, conditionally on X (and hence

also unconditionally), the asymptotic efficiency of $\hat{c}_{jk,\mathrm{LSE}} = \hat{\beta}_{j,\mathrm{LSE}} - \hat{\beta}_{k,\mathrm{LSE}}$ compared to the best linear unbiased estimator $\hat{c}_{jk,\mathrm{BLUE}} = \hat{\beta}_{j,\mathrm{BLUE}} - \hat{\beta}_{k,\mathrm{BLUE}}$ can be written as

$$\mathit{eff}(\hat{c}_{jk,\mathrm{LSE}}) = \left[\frac{\sigma_e^2}{(2\pi)^2} \int_{-\pi}^{\pi} \frac{1}{f_e(\lambda)}\, d\lambda \right]^{-1}.$$

Note that although the result was derived originally for $d > 0$ only and $d = 0$ under the given assumptions, analogous arguments lead to (7.44) for $d < 0$.

Example 7.18 For e_t generated by a FARIMA$(0, d, 0)$ process with variance $\sigma_e^2 = 1$, we have

$$\begin{aligned} f_e(\lambda) &= \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d} \cdot \frac{\Gamma^2(1-d)}{\Gamma(1-2d)}, \\ \frac{1}{f_e(\lambda)} &= 2\pi \left|1 - e^{-i\lambda}\right|^{2d} \cdot \frac{\Gamma(1-2d)}{\Gamma^2(1-d)} \\ &= (2\pi)^2 \frac{\Gamma(1-2d)}{\Gamma^2(1-d)} \cdot \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{2d}. \end{aligned}$$

Using the equality $\int |1 - e^{-i\lambda}|^{2d}\, d\lambda = 2\pi\, \Gamma(1+2d)/\Gamma^2(1+d)$, we obtain

$$\begin{aligned} \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \frac{1}{f_e(\lambda)}\, d\lambda &= \frac{\Gamma(1-2d)}{\Gamma^2(1-d)} \int_{-\pi}^{\pi} \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{2d} d\lambda \\ &= \frac{\Gamma(1-2d)\Gamma(1+2d)}{[\Gamma(1-d)\Gamma(1+d)]^2}, \end{aligned}$$

and the relative asymptotic efficiency

$$\mathit{eff}(\hat{c}_{jk,\mathrm{LSE}}) = \frac{[\Gamma(1-d)\Gamma(1+d)]^2}{\Gamma(1-2d)\Gamma(1+2d)}.$$

Figure 7.4 shows $\mathit{eff}(\hat{c}_{jk,\mathrm{LSE}})$ for all values of d. Towards the two extremes $d \to \pm\frac{1}{2}$, the efficiency converges to zero. Thus, although the LSE keeps the same rate of convergence, it may be worthwhile using the BLUE, when d is far away from zero.

Similarly, for restricted and blockwise randomisation (designs (b) and (c)) it can be shown that the same asymptotic formulas for $V_{n,\mathrm{LSE}}$ hold as under independence (see Künsch et al. 1993). For $V_{n,\mathrm{BLUE}}$ this is conjectured to be true.

A possibility of improving the variance of the LSE is to apply blockwise randomization. The reason is that, under design (c), we have

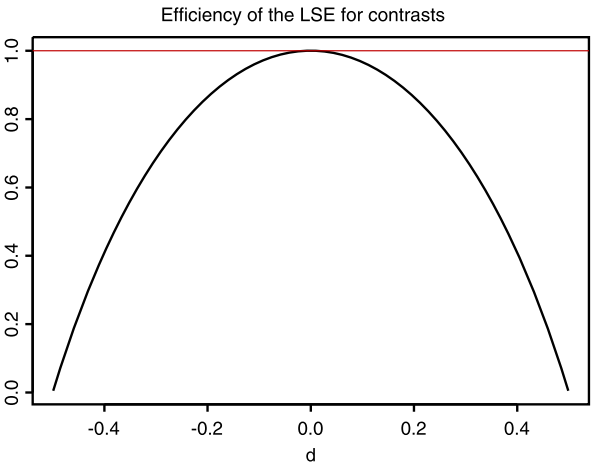

Fig. 7.4 Relative asymptotic efficiency of the LSE of a contrast $\beta_j - \beta_k$ compared to the BLUE, as a function of d. The model considered here is a FARIMA$(0, d, 0)$ process

$$E[V_{n,\mathrm{LSE}}] = \left(\frac{1}{n_j} + \frac{1}{n_k}\right)\left[\sigma_e^2 - \frac{2}{l-1}\sum_{k=1}^{l-1}\left(1 - \frac{k}{l}\right)\gamma_e(k)\right]$$
$$= \left(\frac{1}{n_j} + \frac{1}{n_k}\right)\sigma_l^2.$$

If the autocovariance function $\gamma_e(k)$ is strictly positive and (strictly) monotonically decreasing with limit zero, then σ_l^2 is strictly increasing in l and $\sigma_l^2 \to \sigma_e^2$ (see, e.g. Cochran 1946). Therefore, the smallest variance is expected under blockwise randomization with blocks of length $l = p$. Note, however, that this does not mean necessarily that, under this design, the efficiency of the LSE (compared to the BLUE) is better.

7.2.2 Some General Results and the Heteroskedastic Case

In this section, we consider a parametric random design regression model given by

$$Y_t = \beta_0 + \beta_1 X_t + \sigma(X_t)e_t \quad (t = 1, \ldots, n), \tag{7.45}$$

where $\sigma(\cdot)$ is a positive, deterministic function. As illustrated above, under random design, regression estimators may have a faster rate of convergence than in most fixed design cases. General results including the heteroskedastic case with $\sigma(\cdot)$ not constant can be derived, for instance, under the following conditions:

- (P1) The sequence X_t $(t \in \mathbb{Z})$ is i.i.d.;
- (P2) The sequence X_t $(t \in \mathbb{Z})$ is a linear process

$$X_t = \mu_X + \sum_{j=0}^{\infty} b_j \xi_{t-j},$$

where ξ_t $(t \in \mathbb{Z})$ are centred, i.i.d. random variables such that $\mathrm{var}(X_t) = \sigma_X^2 = 1$. Moreover, we assume $b_j = j^{d_X - 1} L_b(j)$, $d_X \in (0, 1/2)$. Unless stated otherwise, we assume $\mu_X = 0$;

- (E1) The sequence e_t $(t \in \mathbb{Z})$ is i.i.d.;
- (E2) The sequence e_t $(t \in \mathbb{Z})$ is a linear process

$$e_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j},$$

where ε_t $(t \in \mathbb{Z})$ are centred, i.i.d. random variables, $\mathrm{var}(\varepsilon_t) = \sigma_\varepsilon^2$ and $a_j = j^{d_e - 1} L_a(j)$, $d_e \in (0, 1/2)$.

Let f_X and f_e be the spectral densities of X_t and e_t, respectively. Under (P2) and (E2), we have $f_X(\lambda) = |\lambda|^{-2d_X} L_{f_X}(\lambda^{-1})$, $f_e(\lambda) = |\lambda|^{-2d_e} L_{\tilde{f}}(\lambda^{-1})$, where the functions L_{f_X} and L_{f_e} are slowly varying at infinity. Furthermore,

$$\mathrm{var}\left(n^{-1} \sum_{t=1}^{n} e_t \right) \sim n^{2d_e - 1} L_e(n), \qquad \mathrm{var}\left(n^{-1} \sum_{t=1}^{n} X_t \right) \sim n^{2d_X - 1} L_X(n),$$

where

$$L_e(n) = \frac{2L_a^2(n)}{2d_e(2d_e+1)} \sigma_\varepsilon^2 \int_0^\infty \left(u + u^2\right)^{d_e - 1} du = \frac{2\Gamma(1-2d_e)\sin \pi d_e}{d_e(2d_e+1)} L_{f_e}(n), \tag{7.46}$$

$$L_X(n) = \frac{2L_b^2(n)}{2d_X(2d_X+1)} \sigma_\xi^2 \int_0^\infty \left(u + u^2\right)^{d_X - 1} du = \frac{2\Gamma(1-2d_X)\sin \pi d_X}{d_X(2d_X+1)} L_{f_X}(n). \tag{7.47}$$

Recall also that (see Sect. 4.2.4)

$$n^{d_e - 1} L_e^{-1/2}(n) \sum_{t=1}^{n} e_t \xrightarrow{\mathrm{d}} Z_0, \qquad n^{d_X - 1} L_X^{-1/2}(n) \sum_{t=1}^{n} X_t \xrightarrow{\mathrm{d}} Z_1, \tag{7.48}$$

where Z_0 and Z_1 are standard normal random variables. Throughout this section, it is also assumed that the sequences X_t and e_t $(t \in \mathbb{Z})$ are mutually independent (the results are not applicable otherwise, see Sect. 7.2.5). Thus, Z_0 and Z_1 are independent. We recall also that

$$E[e_0 e_k] = \gamma_e(k) = L_a^2(k) \sigma_\varepsilon^2 \int_0^\infty \left(u + u^2\right)^{d_e - 1} du. \tag{7.49}$$

We start our discussion with the classical least squares estimator (LSE), which leads to

$$\hat{\beta}_1 - \beta_1 = \frac{1}{V_n^2} \frac{1}{n} \sum_{t=1}^{n} X_t \sigma(X_t) e_t, \tag{7.50}$$

$$\hat{\beta}_0 - \beta_0 = \frac{1}{n}\sum_{t=1}^{n} \sigma(X_t)e_t, \tag{7.51}$$

where

$$V_n^2 = \frac{1}{n}\sum_{t=1}^{n} X_t^2.$$

If $\sigma_X^2 = 1$, then the sample standard deviation V_n converges (in probability) to σ_X. For the purpose of limit theorems, we can replace V_n^2 by $\sigma_X^2 = 1$ in the expression for $\hat{\beta}_1$.

As we will see in Theorem 7.12, for stochastic regression, the rate of convergence of $\hat{\beta}_0$ is always influenced by a possible memory in the errors e_t. However, the rate of convergence of $\hat{\beta}_1$ depends properties of the regressors X_t ($t \in \mathbb{Z}$), the errors e_t ($t \in \mathbb{Z}$) and on the function $\sigma(\cdot)$. We start with a simple example.

Example 7.19 Consider the homoskedastic linear regression model without intercept,

$$Y_t = \beta_1 X_t + e_t \quad (t = 1, \ldots, n), \tag{7.52}$$

and assume that (P1) and (E2) hold. We note that

$$\operatorname{var}\left(n^{-1}\sum_{t=1}^{n} X_t e_t\right) = n^{-2}\sum_{t,s=1}^{n} E[X_t X_s]E[e_t e_s] = n^{-1}\sigma_e^2.$$

According to the law of large numbers, $n^{-1}\sum_{t=1}^{n} X_t^2 \xrightarrow{p} \sigma_X^2 = 1$. Therefore, the asymptotic behaviour of $\hat{\beta}_1 - \beta_1$ is the same as that of $n^{-1}\sum_{t=1}^{n} X_t e_t$. The formula for the variance suggests that $\hat{\beta}_1$ behaves as if the errors e_t were uncorrelated. We expect that $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ converges in distribution to a normal random variable; see (7.58) of Theorem 7.13.

Example 7.20 We consider the heteroskedastic linear regression model without intercept:

$$Y_t = \beta_1 X_t + \sigma(X_t)e_t \quad (t = 1, \ldots, n). \tag{7.53}$$

We assume again that (P1) and (E2) hold, and furthermore $0 \neq E[\sigma(X_1)X_1] < \infty$. Then

$$\operatorname{Var}\left(n^{-1}\sum_{t=1}^{n} X_t \sigma(X_t)e_t\right) \sim E^2\big[\sigma(X_1)X_1\big]n^{2d_e-1}L_e(n)$$

so that the rate of convergence of $\hat{\beta}_1$ is influenced by long memory in e_t.

Example 7.21 Consider the homoscedastic model without intercept (7.52) and assume that the errors and predictors fulfill (E2) and (P2), respectively. If $2(d_e +$

$d_X) > 1$

$$\begin{aligned}\operatorname{var}\left(n^{-1}\sum_{t=1}^{n} X_t e_t\right) &= n^{-2}\sum_{t,s=1}^{n} E[X_t X_s]E[e_t e_s] \\ &= n^{-2}\sum_{k=-(n-1)}^{n-1}\big(n-|k|\big)\gamma_e(k)\gamma_X(k) \\ &\sim n^{2(d_e+d_X)-2}L_e(n)L_X(n).\end{aligned}$$

Otherwise, if $2(d_e + d_X) < 1$, then the variance is of order n^{-1}. Thus, long memory in both errors and predictors may influence the limiting behaviour of $\hat{\beta}_1$; see Theorem 7.12.

The complete convergence of the least squares estimators (7.51) and (7.50) is characterized in the following two theorems. These theorems were proven in Guo and Koul (2008) and Kulik and Wichelhaus (2012). The proof is given in Sect. 7.2.3 in a general context of randomly weighted partial sums.

Theorem 7.12 *Consider the random design regression model* (7.45) *and let* $\hat{\beta}_1$, $\hat{\beta}_0$ *be least squares estimators defined in* (7.50) *and* (7.51).

- *Assume that* (P1) *or* (P2), *and* (E1) *hold. Then*

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \xrightarrow{\mathrm{d}} \sqrt{E\big[\sigma^2(X_1)\big]\sigma_e^2}\,Z_0 \tag{7.54}$$

and

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} \sqrt{E\big[\sigma^2(X_1)X_1^2\big]\sigma_e^2}\,Z_1, \tag{7.55}$$

where Z_0, Z_1 *are independent standard normal random variables.*
- *Assume that* (P1) *and* (E2) *hold. If* $E[\sigma(X_1)X_1] \neq 0$, *then*

$$n^{\frac{1}{2}-d_e}L_e^{-1/2}(n)(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} E\big[\sigma(X_1)X_1\big]Z_0 \tag{7.56}$$

and

$$n^{\frac{1}{2}-d_e}L_e^{-1/2}(n)(\hat{\beta}_0 - \beta_0) \xrightarrow{\mathrm{d}} E\big[\sigma(X_1)\big]Z_1, \tag{7.57}$$

where Z_0, Z_1 *are independent standard normal random variables.*
- *Assume that* (P2) *and* (E2) *hold and that* X_t, e_t *are Gaussian. If* $E[\sigma(X_1)X_1] \neq 0$, *then* (7.56) *and* (7.57) *hold.*

If $E[\sigma(X_1)X_1] = 0$, then the limiting behaviour of LS estimators changes.

Theorem 7.13 *Consider the random design regression model* (7.45) *and let* $\hat{\beta}_1$, $\hat{\beta}_0$ *be LS estimators defined in* (7.50) *and* (7.51). *Assume that* (P1) *or* (P2) *and* (E2) *hold with* $E[\sigma(X_1)X_1] = 0$ *and that* X_t, e_t *are Gaussian.*

- *If* $2(d_X + d_e) > 1$ *and* $E[\sigma(X_1)X_1^2] < \infty$, *then*

$$n^{1-(d_e+d_X)}\big(L_{f_X}(n)L_{f_e}(n)\big)^{-1/2}(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} E\big[\sigma(X_1)X_1^2\big]Z_{1,1} \tag{7.58}$$

where the random variable $Z_{1,1}$ *is defined in* (7.63).
- *If* $2(d_X + d_\varepsilon) < 1$ *and* $E[\sigma^2(X_1)X_1^2] < \infty$, *then*

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} N\big(0, C_0^2\big), \tag{7.59}$$

where $C_0^2 = \lim_{n\to\infty} \sum_{k=0}^{\infty} E[X_0\sigma(X_0)X_k\sigma(X_k)]E[\varepsilon_0\varepsilon_k]$.

Of course, the LSE is not the only possible method. In the homoscedastic model without intercept it is possible to remove the dependence in e_t first before estimating β_1. This way one can achieve $\sqrt{n}$-convergence. This is the case by definition for the BLUE. An alternative method that does not require inversion of the covariance matrix was suggested by Robinson and Hidalgo (1997). Thus, consider the homoscedastic regression model (7.52). Assume that (P2) and (E2) hold, possibly with $\mu_X \neq 0$. Define the following *weighted least squares* estimator of β_1:

$$\hat{\beta}_{\phi,\mathrm{LSE}} = \frac{\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(X_t - \bar{x})(Y_s - \bar{y})\phi_{t-s}}{\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(X_t - \bar{x})(X_s - \bar{x})\phi_{t-s}},$$

where

$$\phi_j = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi} \phi(\lambda)\cos(j\lambda)\,d\lambda,$$

and $\phi(\cdot)$ is some function such that $\phi_j = O(j^{-\gamma})$, $\gamma \geq 2d_e + 1$. This holds in particular if $\phi = f_e^{-1}$ is the reciprocal of the spectral density of e_t ($t \in \mathbb{Z}$). One can verify that

$$\mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(X_t - \bar{x})(Y_s - \bar{y})\phi_{t-s}\right) = O\big(n^{-1}\big).$$

Consequently, the asymptotic variance of $\hat{\beta}_{\phi,\mathrm{LSE}}$ is not influenced by LRD in X_t or e_t. This observation leads to the following result, proven in Robinson and Hidalgo (1997).

Theorem 7.14 *Consider the model* (7.52). *Assume that* (P2) *and* (E2) *hold. Under appropriate technical conditions*,

$$\sqrt{n}(\hat{\beta}_{\phi,\mathrm{LSE}} - \beta_1) \xrightarrow{\mathrm{d}} N\big(0, \Sigma_\phi^{-1}\Sigma_\psi\Sigma_\phi^{-1}\big),$$

where $\psi(\lambda) = \phi^2(\lambda)f_e(\lambda)$ *and we use the notation* $\Sigma_h = (2\pi)^{-1}\int_{-\pi}^{\pi} h(\lambda)\,d\lambda$ *for* $h = \psi, \phi$.

The "appropriate technical conditions" are in particular continuity of $\psi(\cdot)$ and independence between errors and predictors. Moreover, it has to be mentioned that $\sqrt{n}$-consistency does not hold, in general, in the heteroskedastic case. To see this, assume for simplicity that (P1) holds and $\mu_X = 0$. Then

$$\operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n} X_t \sigma(X_t) e_s \phi_{t-s}\right) \sim \phi_0^2 E^2\big[\sigma(X_1)X_1\big] \operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} e_t\right).$$

Finally, we consider again the model (7.52) and the following estimators:

$$\hat{\beta}_R := \sum_{t=1}^{n} Y_t \Big/ \sum_{t=1}^{n} X_t$$

and

$$\hat{\beta}_{\mathrm{BLUE}} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} Y,$$

with column vectors of $X = (X_1, \ldots, X_n)'$, $X = (Y_1, \ldots, Y_n)'$, respectively, and Σ being the covariance matrix of $e_1, \ldots, e_n$. The following result (under a slightly different set of assumptions) was proven in Choy and Taniguchi (2001).

Theorem 7.15 *Consider the model* (7.52). *Assume that* (P2) *and* (E2) *hold and that* $\mu_X = E[X_1] \neq 0$. *Then*

$$n^{1/2-d_e} L_e^{-1/2}(n)(\hat{\beta}_R - \beta_1) \xrightarrow{\mathrm{d}} \mu_X^{-1} Z_0$$

and

$$\sqrt{n}(\hat{\beta}_{\mathrm{BLUE}} - \beta_1) \xrightarrow{\mathrm{d}} C Z_0,$$

where $C^{-1} = (2\pi)^{-1} \int_{-\pi}^{\pi} f_e^{-1}(\lambda) f_X(\lambda)\, d\lambda$.

Proof We prove only the convergence of $\hat{\beta}_R$. We have

$$\hat{\beta}_R - \beta_1 = \frac{n^{-1}\sum_{t=1}^{n} e_t}{n^{-1}\sum_{t=1}^{n} X_t}.$$

By the law of large numbers, we may replace the denominator by μ_X. The convergence of the nominator, and hence of $\hat{\beta}_R$, follows from (7.48). □

By definition, $\hat{\beta}_{\mathrm{BLUE}}$ is better than $\hat{\beta}_R$ and $\hat{\beta}_{\mathrm{LSE}}$ (in the sense of a smaller variance of the asymptotic distribution). However, in the heteroskedastic case, Σ is the covariance matrix of $\sigma(X_1)e_1, \ldots, \sigma(X_n)e_n$. This involves knowledge of $\sigma(\cdot)$. In most situations with heteroskedastic errors, one may therefore prefer to use the LSE.

7.2.3 Randomly Weighted Partial Sums

Asymptotic results in the context of regression with stochastic explanatory variables are usually based on limit theorems for weighted sums, where weights are stochastic. It is therefore useful to consider such sums in general. Thus let

$$R_n := \frac{1}{n}\sum_{t=1}^{n} v(X_t)e_t \tag{7.60}$$

where $v(\cdot)$ is a deterministic function such that $E[v(X_t)] \neq 0$. Also, define the σ-algebras $\mathscr{X}_t = \sigma(X_1, \ldots, X_t)$, $\mathscr{H}_i = \sigma(\varepsilon_t, \varepsilon_{t-1}, \ldots)$. The following properties will be used under different combinations of (E1), (E2), (P1) and (P2)[1] (we used some of these properties also in Sect. 5.14 on density estimation):

- (M) If (E1) holds, then R_n ($n \geq 1$) is a martingale with respect to a sigma-field $\mathscr{X}_n \vee \mathscr{H}_n$.
- (M/L) If (P1) holds, we use the decomposition

$$\frac{1}{n}\sum_{t=1}^{n}\left\{v(X_t)e_t - E\left[v(X_t)e_t|\mathscr{X}_{t-1} \vee \mathscr{H}_{t-1}\right]\right\} + E\left[v(X_1)\right]\frac{1}{n}\sum_{t=1}^{n} E[e_t|\mathscr{H}_{t-1}]. \tag{7.61}$$

  The first part is a martingale, so that its convergence with scaling $\sqrt{n}$ can be described by an appropriate martingale central limit theorem. Furthermore, $E[e_t|\mathscr{H}_{t-1}] = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j}$ so that the second sum is just the sum of long-memory moving averages and the asymptotic behaviour of $\sum_{t=1}^{n} E[e_t|\mathscr{H}_{t-1}]$ is the same as that of $\sum_{i=1}^{n} e_t$ (cf. (7.48)):

$$n^{-d_e-\frac{1}{2}} L_e^{-1/2}(n) \sum_{t=1}^{n} E[e_t|\mathscr{H}_{t-1}] \xrightarrow{\mathrm{d}} Z_0.$$

  We will call the second term the *LRD part*. It contributes (and dominates) only if $E[v(X_1)] \neq 0$.
- (H) In general, under (E2) and (P2), we assume for simplicity that X_t are standard Gaussian. We decompose R_n as

$$R_n = E\left[v(X_1)\right]\frac{1}{n}\sum_{t=1}^{n} e_t + \sum_{m=1}^{\infty}\frac{J(m)}{m!}\frac{1}{n}\sum_{t=1}^{n} e_t H_m(X_t), \tag{7.62}$$

  where $J(m)$ is the mth Hermite coefficient of $z \to v(z)$. If $E[v(X_1)] \neq 0$, then the first term dominates, and convergence of R_n is equivalent to convergence of the sum $n^{-1}\sum_{i=1}^{n} e_t$. Indeed, let us note that from Lemma 3.5 the random

[1] (M), (M/L) and (H) stand for martingale property, martingale/long-memory decomposition and Hermite expansion, respectively.

variables $H_m(X_t)$, $(m \geq 1)$ are uncorrelated. Since the sequences X_t and e_t are independent, we have for each $m \neq k$ and all t, s,

$$cov\big(H_m(X_t)e_t, H_k(X_s)\big) = E\big(H_m(X_t)H_m(X_s)\big)E(e_t e_s) = 0.$$

Thus,

$$\operatorname{var}\left(\sum_{m=1}^{\infty} \frac{J(m)}{m!}\frac{1}{n}\sum_{t=1}^{n} e_t H_m(X_t)\right) = \sum_{m=1}^{\infty} \frac{J^2(m)}{(m!)^2}\operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} e_t H_m(X_t)\right).$$

Furthermore, for a given $m \in \mathbb{N}$ we have

$$\begin{aligned}\operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} e_t H_m(X_t)\right) &= n^{-2}\sum_{t,s=1}^{n} E\big[H_m(X_t)H_m(X_s)\big]E[e_t e_s] \\ &= m!n^{-2}\sum_{k=-(n-1)}^{n-1}\big(n-|k|\big)\gamma_X^m(k)\gamma_e(k) \\ &= O\big(\max\big\{n^{(2d_X-1)m+(2d_e-1)}L(n), n^{-1}\big\}\big),\end{aligned}$$

where L is a slowly varying function.

These decompositions provide a general framework that will be used several times. In particular, we will use it to prove Theorem 7.12. We note, however, that the situation with $E[\sigma(X_1)X_1] = 0$ and (E2) is not covered by any of these cases. To study this situation, we shall consider

$$T_n := n^{-1}\sum_{t=1}^{n} X_t e_t$$

directly, assuming (P2), (E2), and also that X_t, e_t $(t \in \mathbb{Z})$ are two independent centred Gaussian sequences. We recall some spectral theory from Sect. 4.1.3, see also proof of Theorem 4.2. The innovation processes ξ_t and ε_t have the spectral representation

$$\xi_t = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{it\lambda}\, dM_{0,\xi}(\lambda), \qquad \varepsilon_t = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{it\lambda}\, dM_{0,\varepsilon}(\lambda) \quad (t \in \mathbb{Z}),$$

where $M_{0,\xi}$ and $M_{0,\varepsilon}$ are two independent complex-valued Gaussian random measures with independent increments such that $E[|dM_\xi(\lambda)|^2] = \sigma_\xi^2\, d\lambda$, $E[|dM_\varepsilon(\lambda)|^2] = \sigma_\varepsilon^2\, d\lambda$. Furthermore,

$$X_t = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_X(\lambda), \qquad e_t = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_e(\lambda),$$

where

$$dM_X(\lambda) = \frac{1}{\sqrt{2\pi}}\left(\sum_{j=0}^{\infty} b_j e^{-ij\lambda}\right) dM_{0,\xi}(\lambda) = b(\lambda)\, dM_{0,\xi}(\lambda),$$

$$dM_e(\lambda) = \frac{1}{\sqrt{2\pi}}\left(\sum_{j=0}^{\infty} a_j e^{-ij\lambda}\right) dM_{0,\varepsilon}(\lambda) = a(\lambda)\, dM_{0,\varepsilon}(\lambda).$$

Repeating the same argument as in the proof of Theorem 4.2,

$$\begin{aligned} T_n &= \frac{1}{n}\sum_{t=1}^{n} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} b(\lambda)a(\omega)e^{it\lambda}e^{it\omega}\, dM_{0,\xi}(\lambda)\, dM_{0,\varepsilon}(\omega) \\ &= \frac{1}{n}\sum_{t=1}^{n} \int_{-n\pi}^{n\pi}\int_{-n\pi}^{n\pi} b\left(\frac{\lambda}{n}\right)a\left(\frac{\omega}{n}\right)D_n\left(\frac{\lambda+\omega}{n}\right) \\ &\quad \times n^{1/2}\, dM_{0,\xi}\big(n^{-1}\lambda\big)n^{1/2}\, dM_{0,\varepsilon}\big(n^{-1}\omega\big). \end{aligned}$$

If f_X and f_e are spectral densities of the two sequences, respectively, then by taking

$$b(\lambda) = L_{f_X}^{1/2}\big(\lambda^{-1}\big)|\lambda|^{-d_X}, \qquad a(\omega) = L_{f_e}^{1/2}\big(\omega^{-1}\big)|\omega|^{-d_e},$$

we may conclude for $d_X + d_e > 1/2$ that

$$\begin{aligned} &n^{1-(d_X+d_e)}\big(L_{f_X}(n)L_{f_e}(n)\big)^{-1/2}T_n \\ &\quad \xrightarrow{\mathrm{d}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{|\lambda|^{d_X}}\frac{1}{|\omega|^{d_e}}\frac{e^{i(\lambda+\omega)}}{i(\lambda+\omega)}\, dM_{0,\xi}(\lambda)\, dM_{0,\varepsilon}(\omega) =: Z_{1,1}. \end{aligned} \tag{7.63}$$

Having this general framework, we are ready to prove Theorems 7.12 and 7.13.

Proof of Theorem 7.12 Recall the formulas (7.50) and (7.51) for $\hat{\beta}_1$ and $\hat{\beta}_0$, and also that we may replace V_n^2 by $\sigma_X^2 = 1$.

1. If (E1) holds, i.e. the errors are i.i.d., we apply the (M)-decomposition to (7.60) with $\nu(X_t) = \sigma(X_t)X_t$ and $\nu(X_t) = \sigma(X_t)$, respectively. The martingale central limit theorem (Lemma 4.2) yields (7.54) and (7.55).

2. If (P1) and (E2) hold and $E[\sigma(X_1)X_1] \neq 0$, then we apply the (M/L)-decomposition to (7.60) with $\nu(X_t) = \sigma(X_t)X_t$. The limiting behaviour of $\hat{\beta}_1 - \beta_1$ is determined by

$$E\big[\sigma(X_t)X_t\big]\frac{1}{n}\sum_{t=1}^{n} E[e_t|\mathscr{H}_{t-1}]. \tag{7.64}$$

Similarly, the limiting behaviour of $\hat{\beta}_0 - \beta_0$ is determined by

$$E\big[\sigma(X_t)\big]\frac{1}{n}\sum_{t=1}^{n} E[e_t|\mathscr{H}_{t-1}]. \tag{7.65}$$

We conclude (7.56) and (7.57). Independence of the limiting random variables follows from

$$cov(\hat{\beta}_1, \hat{\beta}_0) \to 0.$$

3. Under the conditions (E2) and (P2), and $E[\sigma(X_1)X_1] \neq 0$, we apply (7.62) to $\nu(X_t) = \sigma(X_t)X_t$ and to $\nu(X_t) = \sigma(X_t)$. Convergence of the regression estimates can be concluded the same way as under (P1) and (E2). □

Proof of Theorem 7.13 Under the conditions (E2), (P2) and $E[\sigma(X_1)X_1] = 0$, we apply the (H)-decomposition (7.62) with $\nu(X_t) = \sigma(X_t)X_t$. Since $E[\nu(X_1)] = 0$, the limiting behaviour of $\hat{\beta}_1 - \beta_1$ is determined by

$$J(1)\frac{1}{n}\sum_{t=1}^{n} X_t e_t + \sum_{m=2}^{\infty}\frac{J(m)}{m!}\frac{1}{n}\sum_{t=1}^{n} e_t H_m(X_t),$$

where $J(1) = E[\sigma(X_1)X_1^2]$ is the first Hermite coefficient of $\nu(z) = \sigma(z)z$. Clearly, the first part dominates. Applying (7.63),

$$n^{1-(d_e+d_X)}\big(L_{f_X}(n)L_{f_e}(n)\big)^{-1/2}(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} J(1)Z_{1,1}. \tag{7.66}$$

□

Finally, it is worth mentioning another possibility. Consider assumptions (P2) and (E2), but with the modification $\mu_X \neq 0$ and instead of $E[\sigma(X_1)X_1] = 0$ (which was used in Theorem 7.13) the condition $E[\sigma(X_1)(X_1 - \mu_X)] = 0$. Then, the estimator of β_1 has to be replaced by

$$\hat{\beta}_1 - \beta_1 = \frac{1}{V_n^2}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\sigma(X_t)e_t - \frac{1}{n}\sum_{t=1}^{n} X_t\frac{1}{n}\sum_{t=1}^{n}\sigma(X_t)e_t\right), \tag{7.67}$$

with $V_n^2 = n^{-1}\sum_{t=1}^{n}(X_t - \bar{x})^2$. Again, we may replace V_n^2 by $\sigma_X^2 = 1$ asymptotically. Applying the (H)-decomposition to $n^{-1}\sum_{t=1}^{n}\sigma(X_t)e_t$ yields

$$\frac{1}{n}\sum_{t=1}^{n}\sigma(X_t)e_t = E\big[\sigma(X_t)\big]\frac{1}{n}\sum_{t=1}^{n} e_t + \sum_{m=1}^{\infty}\frac{J^*(m)}{m!}\frac{1}{n}\sum_{t=1}^{n} e_t H_m(X_t),$$

where now $J^*(m) = E[\sigma(X_1)H_m(X_1)]$. As in the proof of Theorem 7.13 (see also proof of Theorem 4.2),

$$n^{\frac{1}{2}-d_e} L_{f_e}^{-1/2}(n) \frac{1}{n} \sum_{t=1}^{n} e_t \xrightarrow{d} Z_0, \qquad n^{\frac{1}{2}-d_X} L_{f_X}^{-1/2}(n) \frac{1}{n} \sum_{t=1}^{n} X_t \xrightarrow{d} Z_1,$$

where Z_0 and Z_1 are independent and standard normal. Independence is clear since $E[X_t, \sigma(X_s)e_s] = 0$ for all s, t. Combining this with (7.66), we obtain

$$n^{1-(d_e+d_X)} \big(L_{f_X}(n) L_{f_e}(n)\big)^{-1/2} (\hat{\beta}_1 - \beta_1) \xrightarrow{d} \big(J(1)Z_{1,1} - E\big[\sigma(X_1)\big]Z_0 Z_1\big).$$

7.2.4 Spurious Correlations

So far it has been assumed that the explanatory variable(s) X_t and the residual process e_t are stationary. In practice, this is not always clear. In some applications, such as financial time series, it is, in fact, often more likely that none of the observed series is stationary. This is known to cause considerable problems for regression, even without introducing the complication of long memory or antipersistence. For instance, Granger and Newbold (1974) and Phillips (1986) considered two independent random walks

$$X_t = \sum_{j=1}^{t} \xi_j, \qquad Y_t = \sum_{j=1}^{t} \eta_j,$$

i.e. with ξ_j, η_j, i.i.d. and independent of each other. Suppose we set up an equation of the form

$$Y_t = \beta X_t + e_t$$

with e_t zero mean stationary. Since e_t is stationary but Y_t and X_t are not, we certainly cannot have $\beta = 0$. Of course, the model is misspecified. However, in practice we do not know that. The problem is then to see what happens if we actually fit a linear regression to the $x - y$-observations. For instance, if $\xi_t \sim N(0, \sigma_\xi^2)$ and $\eta_t \sim N(0, \sigma_\eta^2)$, then $\sum_{s=1}^{t} \xi_t =_d B_1(t)$, $\sum_{s=1}^{t} \eta_t =_d B_2(t)$ where B_1, B_2 are two Brownian motions that are independent from each other. Hence,

$$\sum_{t=1}^{n} X_t Y_t = \sum_{t=1}^{n} \left(\sum_{s=1}^{t} \xi_t \right) \left(\sum_{s=1}^{t} \eta_t \right) =_d \sum_{t=1}^{n} B_1(t) B_2(t)$$

$$=_d n^2 \sum_{i=1}^{n} B_1(u_i) B_2(u_i) \frac{1}{n}$$

where $u_i = in^{-1}$ so that

$$n^{-2} \sum X_t Y_t \underset{d}{\to} \int_0^1 B_1(u) B_2(u)\, du.$$

Similarly,

$$\sum_{t=1}^{n} X_t^2 \underset{d}{=} n \sum_{i=1}^{n} B_1^2(u_i) = n^2 \sum_{i=1}^{n} B_1^2(u_i) \frac{1}{n}$$

implies

$$n^{-2} \sum_{t=1}^{n} X_t^2 \underset{d}{\to} \int_0^1 B_1^2(u)\, du.$$

Thus,

$$\hat{\beta}_{\mathrm{LSE}} = \frac{\sum X_t Y_t}{\sum X_t^2} \underset{d}{\to} \frac{\int_0^1 B_1(u) B_2(u)\, du}{\int_0^1 B_1^2(u)\, du}.$$

In other words, instead of tending to zero, $\hat{\beta}_{\mathrm{LSE}}$ tends to a random variable that is not equal to zero with probability one. This means that, if a regression of Y on X is carried out, we will (for n large enough) always find a relationship even though it is not there. This is a famous phenomenon in econometrics, known as 'spurious correlation' or 'spurious regression'. Initiated by Granger and others, methods for determining the relationship between integrated time series has become an extended branch of the econometric literature, mostly subsumed under the label 'cointegration'.

Results on spurious correlations can be generalized to long-memory processes. For instance, Tsai (2006) and Tsay and Chung (2000) consider the following situation. Let η_t and ξ_t be i.i.d. and independent of each other, $E(\eta_t) = E(\xi_t) = 0$, $\mathrm{var}(\eta_t) = \sigma_\eta^2$ and $\mathrm{var}(\xi_t) = \sigma_\xi^2$. Furthermore, define the FARIMA processes

$$v_t = (1 - B)^{-d_1} \eta_t,$$
$$w_t = (1 - B)^{-d_2} \xi_t$$

with $0 < d_1, d_2 < \frac{1}{2}$, and the corresponding integrated processes, i.e. the FARIMA$(0, 1 + d_1, 0)$ and FARIMA$(0, 1 + d_2, 0)$ processes (starting at zero for $t = 0$),

$$v_t^* = v_{t-1}^* + v_t,$$
$$w_t^* = w_{t-1}^* + w_t.$$

Now we consider $\hat{\beta}_{\mathrm{LSE}}$ for the following regressions with intercept,

$$Y_t = \beta_0 + \beta_1 X_t + e_t,$$

Table 7.1 Models considered in the context of spurious correlation

	x_t stationary	x_t nonstationary
y_t stationary	M2	M4, M6
y_t nonstationary	M3	M1, M5

where X_t, Y_t are defined as follows:

- Model 1: $Y_t = v_t^*$, $X_t = w_t^*$;
- Model 2: $Y_t = v_t$, $X_t = w_t$ with $d_1 + d_2 > \frac{1}{2}$;
- Model 3: $Y_t = v_t^*$, $X_t = w_t$ with $d_2 > 0$;
- Model 4: $Y_t = v_t$, $X_t = w_t^*$ with $d_1 > 0$;
- Model 5: $Y_t = v_t^*$ on $X_t = t$;
- Model 6: $Y_t = v_t$ on $X_t = t$ with $d_1 > 0$.

Table 7.1 gives an overview. The following notation will be used:

$$\hat{\beta}_{\mathrm{LSE}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix},$$
$$\hat{\beta}_1 = \frac{\sum (X_t - \bar{x}) Y_t}{\sum (X_t - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t,$$
$$\sigma_y^2 = \mathrm{var}(Y_n), \qquad \sigma_x^2 = \mathrm{var}(X_n).$$

Moreover, $s^2 = (n-2)^{-2} \sum_{t=1}^n (y_t - \hat{y}_t)^2$ will denote the usual estimate of the variance of Y_t (note, however, that for a nonstationary Y_t, σ_y^2 grows with t, i.e. the estimate s^2 is actually meaningless) and similarly, $s_{\hat{\beta}_0}^2$ and $s_{\hat{\beta}_1}^2$ are the usual estimates of $\mathrm{var}(\beta_0)$ and $\mathrm{var}(\beta_1)$. Finally, $t_{\beta_0} = \hat{\beta}_0 / s_{\beta_0}$ and $t_{\beta_1} = \hat{\beta}_1 / s_{\beta_1}$ are the corresponding t-statistics for β_0 and β_1. For simplicity of presentation, we assume all moments of η_t and ξ_t to be finite.

For Model 1, the limit theorems in Sect. 4.2 can be applied to obtain

$$\sigma_y^2 \sim \sigma_\eta^2 c_1 n^{1+2d_1},$$
$$\sigma_x^2 \sim \sigma_\xi^2 c_2 n^{1+2d_2}$$

with

$$c_j = \frac{\Gamma(1-2d_j)}{(1+2d_j)\Gamma(1+d_j)\Gamma(1-d_j)} \quad (j = 1, 2).$$

Assume for a moment that our FARIMA sequences v_t and w_t are replaced by fGn, i.e. increments of two independent fractional Brownian motions B_{H_1}, B_{H_2} with

$H_j = d_j + \frac{1}{2}$. Then

$$\sum_{t=1}^{n} X_t =_d \sum_{t=1}^{n} B_{H_2}(t) =_d n^{1+H_2} \sum_{t=1}^{n} B_{H_2}\left(\frac{t}{n}\right)\frac{1}{n},$$

and an analogous embedding applies to $\sum_{t=1}^{n} Y_t$. Similarly, we can consider the other quantities in $\hat{\beta}_{\mathrm{LSE}}$, including $\sum_{t=1}^{n} X_t Y_t$ and $\sum_{t=1}^{n} X_t^2$:

$$\sum_{t=1}^{n} X_t Y_t =_d \sum_{t=1}^{n} B_{H_1}(t) B_{H_2}(t) =_d n^{1+H_1+H_2} \sum_{t=1}^{n} B_{H_1}\left(\frac{t}{n}\right) B_{H_2}\left(\frac{t}{n}\right)\frac{1}{n}.$$

Using the notation

$$\int_0^1 B_{H_i}(u) B_{H_j}(u)\, du = Z_{i,j}, \qquad \int_0^1 B_{H_j}(u)\, du = Z_i,$$

we have

$$n^{-(1+H_2)} \sum_{t=1}^{n} X_t = n^{-(\frac{3}{2}+d_2)} \sum_{t=1}^{n} X_t \to_d \int_0^1 B_{H_2}(u)\, du = Z_2,$$

$$n^{-(1+H_1)} \sum_{t=1}^{n} Y_t = n^{-(\frac{3}{2}+d_1)} \sum_{t=1}^{n} Y_t \to_d \int_0^1 B_{H_1}(u)\, du = Z_1,$$

$$n^{-(1+H_1+H_2)} \sum_{t=1}^{n} X_t Y_t = n^{-(2+d_1+d_2)} \sum_{t=1}^{n} X_t Y_t \to_d \int_0^1 B_{H_1}(u) B_{H_2}(u)\, du = Z_{1,2},$$

and similarly,

$$n^{-(1+2H_2)} \sum_{t=1}^{n} X_t^2 = n^{-(2+2d_2)} \sum_{t=1}^{n} X_t^2 \to_d \int_0^1 B_{H_2}^2(u)\, du = Z_{2,2}.$$

All asymptotic limits can be considered jointly. Since

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{n} X_t Y_t - \frac{1}{n}\sum_{t=1}^{n} X_t \sum_{t=1}^{n} Y_t}{\sum_{t=1}^{n} X_t^2 - \frac{1}{n}\sum_{t=1}^{n} X_t \sum_{t=1}^{n} X_t} \\ &= n^{d_1-d_2} \frac{n^{-(2+d_1+d_2)} \sum_{t=1}^{n} X_t Y_t - n^{-\frac{3}{2}+d_2} \sum_{t=1}^{n} X_t n^{-\frac{3}{2}+d_1} \sum_{t=1}^{n} Y_t}{n^{-(2+2d_2)} \sum_{t=1}^{n} X_t^2 - n^{-\frac{3}{2}+d_2} \sum_{t=1}^{n} X_t n^{-\frac{3}{2}+d_2} \sum_{t=1}^{n} X_t},\end{aligned}$$

we obtain

$$n^{d_2-d_1}\hat{\beta}_1 \to_d \frac{Z_{1,2} - Z_1 Z_2}{Z_{2,2} - Z_2^2} =: \beta_1^*.$$

Similar arguments apply to the other regression quantities of interest, and (due to convergence to fGn in $D[0,1]$) we may state the following result for general FARIMA models:

Theorem 7.16 *Assume that the FARIMA processes have all moments finite. Then, under Model* 1,

$$\frac{\sigma_{X_n}}{\sigma_{Y_n}}\hat{\beta}_1 \to_d \beta_1^*, \qquad \frac{1}{\sigma_{Y_n}}\hat{\beta}_0 \to_d Z_1 - \beta_1^* Z_2,$$

$$\frac{1}{\sigma_{Y_n}}s^2 \to_d Z_{1,1} - Z_1^2 - \left(\beta_1^*\right)^2\left(Z_{2,2} - Z_2^2\right) =: \sigma_*^2,$$

$$\frac{\sigma_{X_n}^2}{\sigma_{Y_n}^2}s_{\hat{\beta}_1}^2 \to_d \frac{\sigma_*^2}{Z_{2,2} - Z_2^2} =: \sigma_{*\beta_1}^2, \qquad \frac{n}{\sigma_{Y_n}^2}s_{\hat{\beta}_0}^2 \to_d \sigma_*^2\left\{1 + \frac{Z_2^2}{Z_{2,2} - Z_2^2}\right\} =: \sigma_{*\beta_0}^2,$$

$$\frac{1}{\sqrt{n}}t_{\beta_1} \to_d \frac{\beta_1^*}{\sigma_{*\beta_1}}, \qquad \frac{1}{\sqrt{n}}t_{\beta_0} \to_d \frac{\beta_0^*}{\sigma_{*\beta_0}},$$

$$R^2 \to_d \left(\beta_1^*\right)^2 \frac{Z_{2,2} - Z_2^2}{Z_{1,1} - Z_1^2}.$$

For related results, also see, e.g. Phillips (1995), Phillips and Loretan (1991), Marmol (1995), Jeganathan (1999), Robinson and Marinucci (2003, 2003), Buchmann and Chan (2007). Theorem 7.16 can be interpreted as follows. Model 1 deals with the case where Y_t and X_t are both integrated processes, independent of each other and such that the first difference exhibits (stationary) long memory. The estimated intercept $\hat{\beta}_0$ always diverges. For the slope, it is more complicated. If long memory in the dependent variable Y_t is at least as strong as in X_t (i.e. $d_1 \geq d_2$) then the estimated slope $\hat{\beta}_1$ does not converge to zero. In particular, if $d_1 = d_2$, we have spurious correlation in the standard sense, namely $\hat{\beta}_1$ converges to a non-constant random variable. If $d_1 > d_2$, then $\hat{\beta}_1$ assumes asymptotically the values $\pm\infty$ only. If X_t has stronger long memory than Y_t, then $\hat{\beta}_1$ does converge to zero; however, at a very slow rate. What is even worse is that the R^2-statistic does not converge to zero, irrespective of the concrete values of d_1 and d_2. Furthermore, we also have spurious correlation at a second-order level for all values of $d_1, d_2 > 0$, in the sense that the usual t-tests for β_0 and β_1 asymptotically reject the null hypothesis that these parameters are zero.

Example 7.22 Figures 7.5(a)–(f) display simulated distributions and boxplots of $\hat{\beta}_1$ for the cases $d_1 = d_2 = 0.4$ and $d_1 = 0.1$, $d_2 = 0.4$, respectively, and sample sizes $n = 20, 50, 100, 200, 400, 1000$ and 2000. As expected from Theorem 7.16, the results for the two cases are very different. In case 2, the distribution of $\hat{\beta}_1$ (Figs. 7.5(d)–(e)) is increasingly concentrated around the true value of β_1 as n grows. In case 1, however, the distribution remains essentially the same (Figs. 7.5

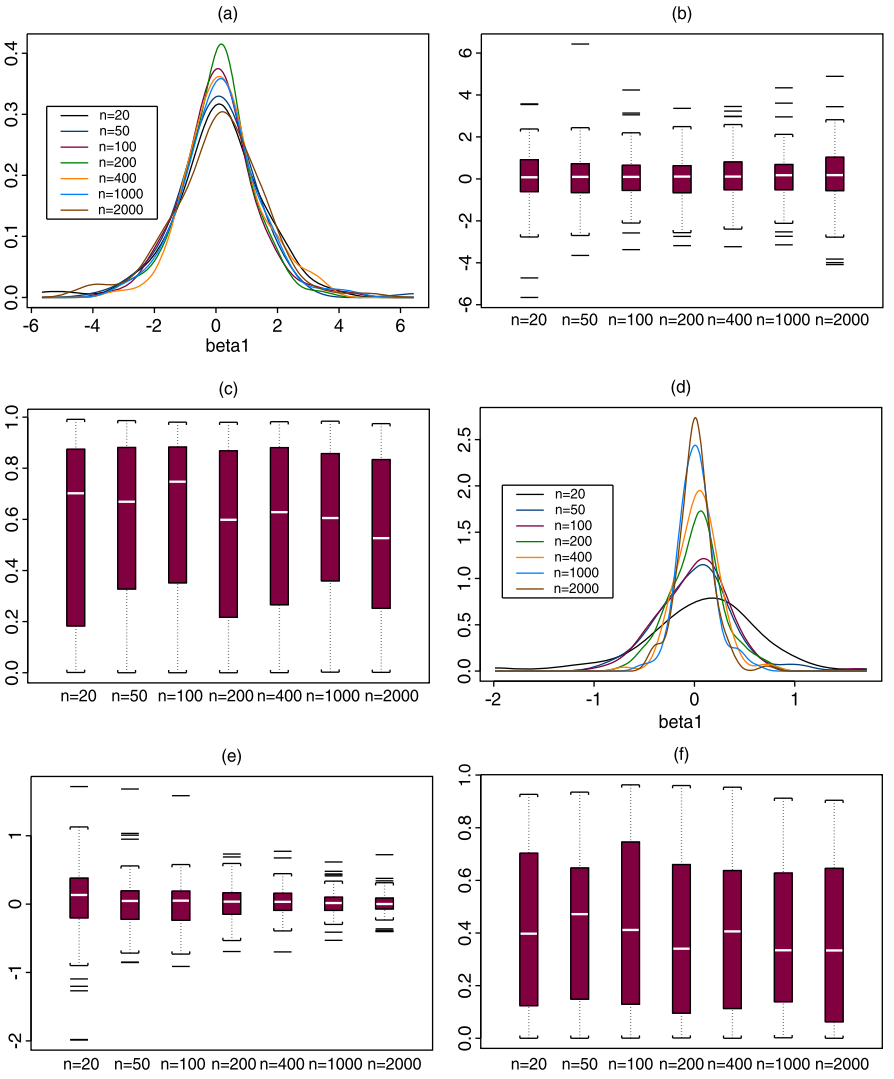

Fig. 7.5 Simulated distributions and boxplots of $\hat{\beta}_1$ in a regression of two independent integrated FARIMA$(0, d, 0)$ processes with $d_1 = d_2 = 0.4$ ((**a**) and (**b**)) and $d_1 = 0.1$, $d_2 = 0.4$ ((**d**) and (**e**)), respectively. The sample sizes are $n = 20, 50, 100, 200, 400, 1000$ and 2000. Also shown are boxplots of the R^2-statistic ((**c**) and (**f**), respectively)

(a)–(b)). For R^2, the behaviour is the same in both cases. As expected from the asymptotic result, the distribution of R^2 stabilizes at a nondegenerate level (Figs. 7.5(c) and (f)). In other words, one is led to believe that there is a linear relationship between the two series, although in reality they are independent of each other.

The results for the other models (Models 2 through 6) can be obtained by similar arguments. In the following, only the order of the variables is written down since this is the essential part of the statements. To simplify notation, we will write "$O_p^*(n^\alpha)$" for a random quantity that is equal to n^α times a random variable with positive variance. In contrast to Model 1, Model 2 involves the estimated relationship between two stationary long-memory processes. For obvious reasons, the least squares estimators of β_0 and β_1, as well as R^2, do converge to zero (see also (7.58) in Theorem 7.13). However, if $d_1 + d_2 > \frac{1}{2}$, then

$$t_{\beta_1} = O_p^*\big(n^{d_1+d_2-\frac{1}{2}}\big).$$

Thus, if the two variables have enough "joint" long memory, then second-order spurious correlations occur in the sense that the usual t-test rejects $H_0 : \beta_1 = 0$ asymptotically. Long memory has to be taken into account to obtain correct rejection regions. This is analogous to tests and confidence intervals for the location parameter, as considered in Sect. 5.2.1.

A different result is obtained in Model 3 where a nonstationary series Y_t is regressed on a stationary series X_t. Here, nonstationarity of the response series alone leads to spurious correlations, as described in the following theorem.

Theorem 7.17 *Under Model* 3,

$$\begin{aligned}
&\hat\beta_1 = O_p^*\big(n^{d_1+d_2}\big), \qquad \hat\beta_0 = O_p^*\big(n^{\frac{1}{2}+d_1}\big),\\
&s^2 = O_p^*\big(n^{1+2d_1}\big), \qquad s^2_{\beta_1} = O_p^*\big(n^{2d_1}\big), \qquad s^2_{\beta_0} = O_p^*\big(n^{2d_1}\big),\\
&t_{\beta_1} = O_p^*\big(n^{d_2}\big), \qquad t_{\beta_0} = O_p^*\big(n^{\frac{1}{2}}\big),\\
&R^2 = O_p^*\big(n^{2d_2-1}\big).
\end{aligned}$$

Thus, regressing a nonstationary long-memory process on an independent stationary long-memory series leads to spurious correlations in the sense that $|\hat\beta_1|$ diverges to infinity, and the t-test for β_1 needs adjustment. On the other hand, there is no spurious correlation as such because R^2 (which is in the case of simple linear regression equal to the square of the sample correlation) converges to zero. In contrast, regressing a stationary process on a nonstationary series leads to a spurious effect only when considering the (unadjusted) t-test.

Theorem 7.18 *Under Model* 4,

$$\begin{aligned}
&\hat\beta_1 = O_p^*\big(n^{d_1-d_2-1}\big), \qquad \hat\beta_0 = O_p^*\big(n^{d_1-\frac{1}{2}}\big),\\
&s^2 \underset{p}{\to} \sigma_v^2, \qquad s^2_{\beta_1} = O_p^*\big(n^{-2-2d_2}\big), \qquad s^2_{\beta_0} = O_p^*\big(n^{-1}\big),\\
&t_{\beta_1} = O_p^*\big(n^{d_1}\big), \qquad t_{\beta_0} = O_p^*\big(n^{d_1}\big),\\
&R^2 = O_p^*\big(n^{2d_1-1}\big).
\end{aligned}$$

Thus, apart from the need for an adjustment in the t-test, nothing too serious happens when regressing a stationary series on an unrelated nonstationary one.

The situation is different, when fitting a liner trend function to an integrated process:

Theorem 7.19 *Under Model* 5,

$$\hat{\beta}_1 = O_p^*\big(n^{d_1-\frac{1}{2}}\big), \qquad \hat{\beta}_0 = O_p^*\big(n^{\frac{1}{2}+d_1}\big),$$
$$t_{\beta_1} \sim O_p^*(\sqrt{n}), \qquad t_{\beta_0} = O_p^*(\sqrt{n}),$$
$$R^2 = O_p^*(1).$$

Thus, the t-test and the value of R^2 indicate asymptotically the presence of a linear trend. On the other hand, $\hat{\beta}_1$ itself is asymptotically zero with probability one, but the convergence to zero is very slow. Finally, if the differenced series (i.e. a stationary long-memory process) is regressed on a linear trend, then the only remaining problem is that the t-test would need adjustment. Specifically, one obtains for Model 6

$$t_{\beta_1} = O_p^*\big(n^{d_1}\big).$$

7.2.5 Fractional Cointegration

The problem of spurious correlations leads to the natural question how to recognize which (linear) relationships between observed nonstationary time series are real and which ones are spurious. The original definition of cointegration of random walk type processes (or integrated processes with an integer valued degree of integration) was introduced by Granger (1981, 1983) and further developed in Engle and Granger (1987) and many subsequent papers. Qualitative considerations suggesting that certain nonstationary time series should not drift arbitrarily far apart existed before, for instance, in Davidson et al. (1978). Much later, cointegration was extended to fractionally integrated processes. There is an extended literature on this topic, and fractional cointegration is still somewhat controversial among economists. Here, only a very brief introduction is given.

For simplicity, we consider the bivariate case, i.e. two series Y_t and X_t. The first step is to specify exactly what kind of nonstationarity is considered. This leads to the notion of integrated processes. There are at least two possible ways of defining such processes, and these definitions are, in fact, quite different (see, e.g. Chen and Hurvich 2009). The first definition was used, for instance, in Velasco (1999a, 1999b), Chen and Hurvich (2003a, 2003b, 2006) and Velasco (2003):

Definition 7.3 A univariate process X_t is called $I(d)$ of Type I or integrated of order $d > -\frac{1}{2}$ if either (a) $-\frac{1}{2} < d < \frac{1}{2}$, X_t is stationary and with spectral density

$f_X(\lambda) \sim c_f|\lambda|^{-2d}$ $(\lambda \to 0)$, or (b) $d > \frac{1}{2}$ and there is an integer m such that $-\frac{1}{2} < d^* = d - m < \frac{1}{2}$ and $(1-B)^m X_t$ is $I(d^*)$.

The second definition was used in Marinucci and Robinson (2000):

Definition 7.4 A univariate process X_t $(t \geq 1)$ is called $I(d)$ of Type II or integrated of order $d > -\frac{1}{2}$ if, for $t \geq 1$,

$$X_t = \sum_{j=0}^{t-1} a_j \xi_{t-j} = \sum_{j=0}^{\infty} a_j \xi^*_{t-j} = (1-B)^{-d} \xi^*_t$$

where ξ_t are zero mean i.i.d. with finite variance, $\xi^*_t = \xi_t \cdot 1\{t \geq 1\}$, and

$$a_j = \delta_{0j} \quad (d = 0),$$
$$a_j = \binom{-d}{j} = \frac{\Gamma(1-d)}{\Gamma(j+1)\Gamma(1-d-j)} \sim c \cdot j^{d-1}.$$

The second definition may be generalized by imposing the asymptotic condition on a_j only. It should be noted that the two definitions are quite different. For $d > \frac{1}{2}$, both imply a nonstationary process. For $-\frac{1}{2} < d < \frac{1}{2}$, X_t obtained from Definition 7.3 is stationary, whereas this is only the case asymptotically when Definition 7.4 is used. Moreover, different limits for partial sums are obtained. For example, if X_t is $I(d)$ according to Definition 7.4 with $\frac{1}{2} < d < \frac{3}{2}$, then

$$X_n = X^*_1 + X^*_2 + \cdots + X^*_n$$

where

$$X^*_t = (1-B)^{-(d-1)} \xi^*_t,$$

and the partial sums

$$S_n(u) = \sum_{i=1}^{[nu]} X^*_i \quad (0 \leq u \leq 1)$$

are such that $Z_n(u) = S_n(u)/\sqrt{\text{var}(S_n(1))}$ converges to a so-called Type II or Riemann–Liouville fractional Brownian motion (Marinucci and Robinson 2000; also see Akonom and Gourieroux 1987; Silveira 1991) which is defined for all $H = d + \frac{1}{2} > 0$. On the other hand, if X_t is obtained from Definition 7.3, then $Z_n(u)$ converges to the usual fractional Brownian motion as in Mandelbrot and van Ness (1968) (see Sect. 1.3.5) which is defined for $0 < H < 1$ only. For limit theorems for Fourier transforms under the two definitions, see, e.g. Velasco (2007).

More generally, $I(d)$ may be defined for bivariate (or multivariate) processes $X_t = (X_{t1}, X_{t2})$ as follows. Using the spectral representation

$$X_{t,j} = \int_{-\pi}^{\pi} e^{it\lambda}\, dM_j(\lambda) \quad (j = 1, 2),$$

the cross-covariance is

$$\gamma_{12}(k) = cov(X_{t+k,1}, X_{t,2}) = \int_{-\pi}^{\pi} f_{12}(\lambda) e^{ik\lambda}\, d\lambda$$
$$= \int e^{ik\lambda} E\big[dM_1(\lambda)\,\overline{dM_2(\lambda)}\big].$$

Thus, in this notation,

$$f_{12}(\lambda) = E\big[dM_1(\lambda)\,\overline{dM_2(\lambda)}\big].$$

If, for instance, $dM_2(\lambda) = e^{-i\phi_{12}(\lambda)}\, dM_1(\lambda)$ with $\phi_{12}(\lambda) = \phi\lambda$ and $\phi > 0$, then this means that $X_{t,2}$ is delayed with respect to $X_{t,1}$ by the time span ϕ. For the cross-spectral density, we have

$$f_{12}(\lambda) = e^{i\phi_{12}(\lambda)}\big|f_{12}(\lambda)\big| = e^{i\phi\lambda}\big|f_{12}(\lambda)\big|.$$

Thus, in the notation used here, the slope of the phase, $\phi'_{12}(\lambda)$, corresponds to the time delay of $dM_2(\lambda)$ with respect to $dM_1(\lambda)$ (see, e.g. Brockwell and Davis 1991). A possible definition of bivariate fractionally integrated processes is as follows:

Definition 7.5 A stationary process $X_t = (X_{t,1}, X_{t,2})^T \in \mathbb{R}^2$ is called $I(d_1, d_2)$ of Type I if there exist $-\frac{1}{2} < d_1, d_2 < \frac{1}{2}$ such that X_t has a 2×2 spectral density

$$f_X(\lambda) \sim \Lambda(\lambda) C_f \bar{\Lambda}(\lambda) \quad (\lambda \to 0)$$

with C_f a constant, real, positive semidefinite and symmetric $p \times p$ matrix such that $[C_f]_{ii} \neq 0$, and

$$\Lambda(\lambda) = \begin{pmatrix} |\lambda|^{-d_1} & 0 \\ 0 & e^{-i\phi_{12}(\lambda)}|\lambda|^{-d_2} \end{pmatrix}$$

for some differentiable function ϕ_{12} with derivative ϕ'_{12} such that $\lim_{\lambda\to 0}\phi'_{12}(\lambda) = \phi_0 \in (0, \pi]$. A nonstationary process X_t is called $I(d_1, d_2)$ of Type I if there is an integer m such that $-\frac{1}{2} < d_i^* = d_i - m < \frac{1}{2}$ and $(1-B)^m X_t = ((1-B)^m X_{t,1}, (1-B)^m X_{t,2})^T$ is $I(d_1^*, d_2^*)$.

The generalization to p-dimensional cointegrated vector series is obvious. More explicitly, a stationary $I(d_1, d_2)$ process has a spectral density that behaves at the origin like

$$f(\lambda) \sim \begin{pmatrix} |\lambda|^{-d_1} & 0 \\ 0 & e^{-i\phi_0\lambda}|\lambda|^{-d_2} \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{pmatrix} \begin{pmatrix} |\lambda|^{-d_1} & 0 \\ 0 & e^{i\phi_0\lambda}|\lambda|^{-d_2} \end{pmatrix}$$
$$= \begin{pmatrix} C_{11}|\lambda|^{-2d_1} & C_{12}|\lambda|^{-d_1-d_2} e^{i\phi_0\lambda} \\ C_{12}|\lambda|^{-d_1-d_2} e^{-i\phi_0\lambda} & C_{22}|\lambda|^{-2d_2} \end{pmatrix}.$$

In particular, this means that for low frequency components of X_t there is an approximately constant phase shift corresponding to $X_{t,2}$ being behind by $\Delta t = \phi_0$. In the simplest case with $\lim_{\lambda\to 0}\phi'_{12}(\lambda) = 0$ (see, e.g. Christensen and Nielsen 2006), there is no phase shift for very low frequencies (more precisely, for $\lambda \to 0$).

Example 7.23 Consider a multivariate FARIMA model defined as the stationary solution of

$$\begin{pmatrix} (1-B)^{d_1} & 0 \\ 0 & (1-B)^{d_2} \end{pmatrix} X_t = \varphi^{-1}(B)\psi(B)\xi_t = \eta_t = \begin{pmatrix} \eta_{t,1} \\ \eta_{t,2} \end{pmatrix} \tag{7.68}$$

(see, e.g. Lobato 1999; Robinson and Yajima 2002; Shimotsu 2006) with i.i.d. $\xi_t = (\xi_{t,1}, \xi_{t,2})^T$, zero mean random variables and $\xi_{t,1}$ independent of $\xi_{s,2}$ for all s, t. The spectral density of X_t is given by

$$f(\lambda) = \begin{pmatrix} (1-e^{-i\lambda})^{-d_1} & 0 \\ 0 & (1-e^{-i\lambda})^{-d_2} \end{pmatrix} f_\eta(\lambda) \begin{pmatrix} (1-e^{i\lambda})^{-d_1} & 0 \\ 0 & (1-e^{i\lambda})^{-d_2} \end{pmatrix}$$

where

$$\begin{aligned} \underset{2\times 2}{f_\eta}(\lambda) &= \frac{\sigma_\xi^2}{2\pi}\psi\left(e^{-i\lambda}\right)\varphi^{-1}\left(e^{-i\lambda}\right)\varphi^{-1}\left(e^{i\lambda}\right)\psi\left(e^{i\lambda}\right) \\ &=: \frac{\sigma_\xi^2}{2\pi}\left|\psi\left(e^{-i\lambda}\right)\varphi^{-1}\left(e^{-i\lambda}\right)\right|^2. \end{aligned}$$

For $\lambda \to 0$,

$$f_\eta(\lambda) \to C_f = \frac{\sigma_\xi^2}{2\pi}\left|\psi(1)\varphi^{-1}(1)\right|^2$$

and

$$\left(1-e^{i\lambda}\right)^d \sim (1-1-i\lambda)^d = \lambda^d e^{-i\frac{\pi}{2}d}.$$

Thus,

$$\begin{aligned} f(\lambda) &\sim \begin{pmatrix} \lambda^{-d_1}e^{i\frac{\pi}{2}d_1} & 0 \\ 0 & \lambda^{-d_2}e^{i\frac{\pi}{2}d_2} \end{pmatrix} C_f \begin{pmatrix} \lambda^{-d_1}e^{-i\frac{\pi}{2}d_1} & 0 \\ 0 & \lambda^{d_2}e^{-i\frac{\pi}{2}d_2} \end{pmatrix} \\ &= \begin{pmatrix} \lambda^{-d_1} & 0 \\ 0 & \lambda^{-d_2}e^{i\frac{\pi}{2}(d_2-d_1)} \end{pmatrix} C_f \begin{pmatrix} \lambda^{-d_1} & 0 \\ 0 & \lambda^{d_2}e^{-i\frac{\pi}{2}(d_2-d_1)} \end{pmatrix} \end{aligned}$$

so that Definition 7.5 applies with

$$\phi_{12}(\lambda) \equiv \frac{\pi}{2}(d_1 - d_2)$$

and

$$\phi_0 = \phi'_{12}(\lambda) \equiv 0.$$

This means that for FARIMA models as defined above there is no time shift, although the phase ϕ_{12} itself is not zero except for $d_1 = d_2$. (For less restrictive models, see, e.g. Robinson 2007). Note, however, that this only refers to $\lambda \to 0$. Outside any open neighbourhood of the origin, the AR- and MA-matrices φ and ψ can model any kind of phase shifts with $\phi'_{12} \neq 0$.

Similarly, a Type II $I(d_1, d_2)$-process can be defined (see, e.g. Robinson and Marinucci 2001, 2003, Marinucci and Robinson 2000; Marmol and Velasco 2004; Nielsen and Shimotsu 2007).

A simple, though not most general, definition of cointegration can be given as follows (Chen and Hurvich 2003a, 2003b, 2006).

Definition 7.6 Let $X_t \in \mathbb{R}^2$ be $I(d_1, d_2)$ with $d_1 = d_2 = d > -\frac{1}{2}$. Then X_t is cointegrated of order d, b (or $CI(d, b)$) if there exists a vector $\beta \in \mathbb{R}^2$ such that $\beta \neq 0$ and $Y_t(\beta) = \beta^T X_t \in \mathbb{R}$ is $I(d^*)$ with $d^* = d - b < d$. Any such vector β is called a cointegrating vector.

By definition, β is determined up to a scaling constant. Thus, for a bivariate series, there is at most one β with $\|\beta\| = \sqrt{\beta_1^2 + \beta_2^2} = 1$. More generally, for p-dimensional series, there are at most $p - 1$ such vectors. The number of linearly independent cointegrating vectors is then called the cointegrating rank. Note that originally, cointegration was defined for integer valued differencing parameters d_j only (Engle and Granger 1987): the components of $X_t \in \mathbb{R}^p$ are said to be cointegrated of order $d, b \in \mathbb{N}$ in the sense of Engle and Granger ($X_t \sim CI(d, b)$) if all components of X_t are $I(d)$ and there exists a vector $\beta \in \mathbb{R}^p$ such that $\beta^T X_t \sim I(d - b), b > 0$. Definition 7.6 is applicable to any d and $b = d - d^*$. The possibility of extending cointegration to fractional differences was suggested before by Granger (Granger 1981, 1986). Note also that d^* may be less or equal $-\frac{1}{2}$. This means that $Y_t(\beta)$ may turn out to be non-invertible. More general definitions that allow for $d_1 \neq d_2$ were also introduced in the literature, but are more complicated due to the variety of possible subsets with equal d_j's (see, e.g. Robinson and Yajima 2002; Robinson and Marinucci 2003, 2003).

Example 7.24 Suppose that X_{t1} and X_{t2} are both Type I $I(d)$ with $d \in (0, \frac{1}{2})$ and $e_t \in \mathbb{R}$ is Type I $I(d_e)$ with $0 < d_e < d < \frac{1}{2}$. If there is an $\alpha \neq 0$ such that

$$X_{t2} = \alpha X_{t1} + e_t, \tag{7.69}$$

then $X_t = (X_{t1}, X_{t2})^T$ is fractionally cointegrated with cointegrating vector $\beta = (1, -\alpha)^T$ and fractional integration parameters d and d_e (see, e.g. Robinson 1994b).

Example 7.25 Let X_t be defined as in the previous example and $\tilde{X}_t$ be such that $(1 - B)\tilde{X}_t = X_t$. Also denote by $\tilde{e}_t$ an $I(d_e + 1)$ process such that $(1 - B)\tilde{e}_t = e_t$.

Then

$$\tilde{X}_{t,2} = \mu + \alpha \tilde{X}_{t,1} + \tilde{e}_t \tag{7.70}$$

where μ is an arbitrary constant. The integrated process $\tilde{X}_t$ is cointegrated with cointegrating vector $\beta = (1, -\alpha)^T$ and fractional integration parameters $d + 1$ and $d_e + 1$ (see Chen and Hurvich 2003a for a generalization to $d + m$).

Example 7.26 A Type I p-dimensional fractional common component model proposed in Chen and Hurvich (2006) is defined as

$$X_t = A_0 \xi_t^{(0)} + A_1 \xi_t^{(1)} + \cdots + A_s \xi_t^{(s)}$$

with latent (unobserved) $I(d_j)$-processes $\xi_t^{(j)} \in \mathbb{R}^{p_j}$ such that

$$-m_0 + \frac{1}{2} < d_s < \cdots < d_0 < \frac{1}{2},$$

$A_0, \ldots, A_s$ are $p \times p_j$ full-rank matrices with all columns linearly independent, $p_0 + \cdots + p_s = r$, $1 \le r < p$ and $1 \le s \le r$. This means that X_t can be decomposed orthogonally into s cointegrating subspaces defined by $A_1, \ldots, A_s$ and the cointegration rank is r. Moreover, by definition, X_t is $I(d_0)$. If we choose β as a linear combination of the columns of matrix A_j $(j \neq 0)$, then—due to orthogonality—

$$Y_t(\beta) = \beta^T X_t = \beta^T A_j \xi_t^{(j)}$$

so that $Y_t(\beta)$ is $I(d_j)$.

Example 7.27 Sowell (1990) and Dueker and Startz (1998) consider a cointegrated FARIMA process of the form $X_t = (X_{t1}, X_{t2})^T$ with

$$\underset{2\times 2}{\varphi}(B) \begin{pmatrix} (1-B)^{d_1} & 0 \\ 0 & (1-B)^{d_2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\alpha & 1 \end{pmatrix} X_t = \underset{2\times 2}{\psi}(B)\xi_t \tag{7.71}$$

where $-\frac{1}{2} < d_2 < d_1 < \frac{1}{2}$, and φ and ψ are AR- and MA-operators of order p and q. This means that $X_t^* = (X_{t1}, X_{t2} - \alpha X_{t1})^T$ is the usual multivariate FARIMA process. The bivariate process X_t is cointegrated with cointegrating vector $\beta = (-\alpha, 1)^T$. If the i.i.d. innovation variables ξ_t are assumed to be Gaussian, then, in principle, the parameters in (7.71) can be estimated by a maximum likelihood type method. For non-Gaussian innovations, the same method may be used (under moment assumptions), though it may not be optimal (see, e.g. Dueker and Startz 1998; Jeganathan 1999).

For further results, discussions and literature, see, e.g. Chan and Terrin (1995), Breitung and Hassler (2002), Davidson (2002), Dolado et al. (2003), Robinson and Hualde (2003), Nielsen (2005a, 2005b), Johansen (2008, 2008), Lasak (2010).

In classical cointegration with integer valued d and b, the cointegrating vector $\beta = (1, -\alpha)^T$ can be estimated by minimizing $\sum(X_{1t} - \mu - \alpha X_{2t})^2$ with respect to μ and α. (The generalization to higher dimensions $p > 2$ is obvious.) In addition, because of the problem of spurious correlation, one has to test whether $\hat{\beta}$ is "real" or spurious. The classical method suggested by Engle and Granger is to test for unit roots in the residuals $\hat{e}_t = X_{1t} - \hat{\mu} - \hat{\alpha} X_{2t}$ (i.e. $H_0 : \varphi = 1$ vs. $H_1 : |\varphi| < 1$ where we assume $e_t = \varphi e_{t-1} + u_t$). This is typically done by a suitable version of the Dickey–Fuller test (Dickey and Fuller 1981). If H_0 is not rejected, then cointegration is assumed to be real. An alternative method is based on reduced rank regression of a multivariate ARMA process the cointegration model can be embedded in (see, e.g. Johansen 1996).

At first sight, the generalization of estimation and identification techniques to fractional cointegration is not obvious because unit root testing is not sufficient. The first question is estimation of β in the case where cointegration applies. The second question is how to guard against spurious correlations. In particular, the usual Dickey–Fuller test is not applicable. With respect to estimation no fundamentally new problem occurs if a parametric model, such as (7.71), is acceptable. In this case, maximum likelihood estimation of the cointegration vector β and other parameters of the model (including d_1, d_2) can be carried out in principle because everything is specified. However, in models where only the behaviour of the (cross-) spectrum near the origin is specified (see some of the examples above), the task is more difficult. Consider, for example, (7.69) with

$$X_{t2} = \alpha X_{t1} + e_t, \tag{7.72}$$

X_{t1} stationary with autocovariance function $\gamma_{11}(k)$, variance $\mathrm{var}(X_{t1}) = \gamma_{11}(0) = \sigma_1^2$ and $I(d)$ for some $0 < d < \frac{1}{2}$, and e_t stationary and $I(d_e)$ with $d_e < d$. For the least squares estimator of α, we then have

$$\hat{\alpha}_{\mathrm{LSE}} = \alpha + \frac{\sum_{t=1}^n X_{t1} e_t}{\sum_{t=1}^n X_{t1}^2} \underset{p}{\to} \alpha + \frac{cov(X_{t1}, e_t)}{\sigma_1^2}.$$

This is equal to zero only if X_{t1} and e_t are uncorrelated. The result is different from nonfractional cointegration where, for instance, $X_{t,1}$, $X_{t,2}$ are $CI(1, 1)$ which implies that $\sum_{t=1}^n X_{t1}^2$ is of a larger order than $\sum_{t=1}^n X_{t1} e_t$. A possible solution for the fractional cointegration model here is to apply least squares regression to low frequency components only. The reason is that

$$cov(X_{t1}, e_t) = \int_{-\pi}^{\pi} f_{1,e}(\lambda)\, d\lambda,$$

$$\mathrm{var}(X_{t1}) = \int_{-\pi}^{\pi} f_{11}(\lambda)\, d\lambda$$

where

$$f(\lambda) = \begin{pmatrix} f_{11}(\lambda) & f_{1,e}(\lambda) \\ f_{e,1}(\lambda) & f_{ee}(\lambda) \end{pmatrix}$$

is the (real-valued) bivariate spectral density of $(X_{t1}, e_t)'$. Since $0 \le |f_{1,e}| \le \sqrt{f_{11} f_{ee}}$ and $d_e < d$, we have for $\lambda \to 0$,

$$f_{1,e}(\lambda) = O\big(\lambda^{-d-d_e}\big) = o\big(\lambda^{-2d}\big).$$

Denote by

$$Z_j(\lambda_k) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_{tj} e^{i\lambda_k t} \quad (j = 1, 2)$$

the discrete Fourier transform of X_{tj} at Fourier frequencies $\lambda_k = 2\pi k/n$ and define

$$\hat{\alpha}_{\mathrm{LSE}}(m_n) = \frac{\sum_{k=1}^{m_n} Re(Z_1(\lambda_k)\overline{Z_2(\lambda_k)})}{\sum_{k=1}^{m_n} |Z_1(\lambda_k)|^2} \tag{7.73}$$

with $m_n \to \infty$ such that $m_n/n \to 0$. For Z_j we have

$$\begin{aligned} E\big[Z_1(\lambda_k)\overline{Z_2(\lambda_k)}\big] &= \frac{1}{2\pi n} \sum_{t,s=1}^{n} E\big[X_{t1}(\alpha X_{s1} + e_s)\big] e^{i\lambda_k(t-s)} \\ &= \alpha \frac{1}{2\pi n} \sum_{t,s=1}^{n} \gamma_{11}(t-s) e^{i\lambda_k(t-s)} \\ &\quad + \frac{1}{2\pi n} \sum_{t,s=1}^{n} cov(X_{t1}, e_s) e^{i\lambda_k(t-s)} \\ &\sim \alpha \cdot O\big(\lambda_k^{-2d}\big) + O\big(\lambda_k^{-d_e-d}\big) \end{aligned}$$

and

$$E\big[\big|Z_1(\lambda_k)\big|^2\big] = \frac{1}{2\pi n} \sum_{t,s=1}^{n} \gamma_{11}(t-s) e^{i\lambda_k(t-s)} = O\big(\lambda_k^{-2d}\big).$$

Similar arguments apply to the variance of the enumerator and denominator in (7.73) so that, under suitable detailed regularity conditions,

$$\hat{\alpha}_{\mathrm{LSE}}(m_n) = \alpha + O_p\big(\lambda^{d-d_e}\big) = \alpha + o_p(1)$$

(see Robinson 1994b). Robinson and Marinucci (2001) showed that $\hat{\alpha}_{\mathrm{LSE}}(m_n)$ is also consistent for a Type II nonstationary cointegration model. Similarly, Chen and Hurvich (2003a) showed consistency and derived the asymptotic distribution of $\hat{\alpha}_{\mathrm{LSE}}(m_n)$ refined by tapering, under a Type I cointegration model with arbitrary integer integration parameter (also see, e.g. Chen and Hurvich 2006; Robinson and Yajima 2002; Velasco 2003; Nielsen and Shimotsu 2007). Also note that an alternative estimator based on the Whittle approximation is proposed in Robinson (2008).

Moreover, Johansen and Nielsen (2010a, 2010b) show how to generalize reduced rank regression to fractional cointegration (also see Johansen 2010a, 2010b, 1996, 2008, Lütkepohl 2006).

The second question is how to design "unit roots" tests that detect fractional departures from stationarity. More generally, the question is how to identify the cointegration rank in the fractional cointegration context. Tests along this line are discussed, for instance, in Breitung and Hassler (2002, 2006), Davidson (2002, 2006), Robinson and Yajima (2002), Marmol and Velasco (2004), Nielsen (2004b, 2004c, 2004a, 2005a, 2005b), Chen and Hurvich (2006), Nielsen and Shimotsu (2007), Hualde and Velasco (2008), Avarucci and Velasco (2009), Lasak (2010), MacKinnon and Nielsen (2010). For additional references to fractional cointegration, see, e.g. Cheung and Lai (1993), Baillie and Bollerslev (1994), Ravishanker and Ray (1997, 2002), Kim and Phillips (2001), Gil-Alana (2004), Nielsen (2004b, 2004c), Robinson and Iacone (2005), Hualde and Robinson (2007, 2010), Robinson (2008), Berger et al. (2009), Davidson and Hashimzade (2009a, 2009b), Gil-Alana and Hualde (2009), Sela and Hurvich (2009), Franchi (2010), Nielsen (2010, 2011), Nielsen and Frederiksen (2011).

7.3 Piecewise Polynomial and Spline Regression

We consider a process of the form

$$X_t = m\left(\frac{t}{n}\right) + e_t \quad (t = 1, \ldots, n) \tag{7.74}$$

where e_t is a zero mean second-order stationary process. In some situations, a natural model for the expected value m is a piecewise polynomial. For instance, Fig. 1.18 in Sect. 1.2 shows typical olfactory response curves to an odorant stimulus administered at a known time point t_0. In this case, a continuous piecewise linear polynomial (or in other words, a linear spline function) with one known knot at time t_0 and one subsequent unknown knot characterizes the essential features of the expected value as a function of time. The residual processes e_t often exhibit long memory.

More generally, we may consider an arbitrary continuous piecewise polynomial function

$$m(s) = \sum_{k=0}^{l} \sum_{j=1}^{p_k} a_{k,j} (s - \eta_k)_+^{\beta_{j,k}}$$

with $\beta_{j,k} < \beta_{j+1,k}$, knots $0 = \eta_0 < \eta_1 < \cdots < \eta_l < 1$ of which some (but not necessarily all) are unknown. Note that m is continuous if $\beta_{j,k} \geq 1$ for $k \geq 1$. The definition includes splines, but is more general since apart from continuity no differentiability conditions are imposed. For simplicity of presentation, we will discuss the case with one unknown knot η only. As we will see, however, results can be formulated in a general form so that all cases with an arbitrary number of knots and

arbitrary polynomials are included. Thus, suppose that there is one unknown knot η. Then $m(s)$ has the representation

$$m(s) = \sum_{j=1}^{p} \alpha_j f_j(s) \quad \big(s \in [0,1]\big) \tag{7.75}$$

with $\alpha^T = (\alpha_1, \ldots, \alpha_p)$ denoting unknown regression coefficients and

$$\begin{aligned} &f_1(s) = 1, \qquad f_2(s) = s, \qquad \ldots, \qquad f_q(s) = s^{q-1}, \\ &f_{q+1}(s) = (s-\eta)_+, \qquad \ldots, \qquad f_p(s) = (s-\eta)_+^{p-q} \end{aligned} \tag{7.76}$$

(where $(s-\eta)_+^l := \max(0, (s-\eta)^l)$). The unknown parameter vector is $\theta = (\alpha^T, \eta)^T$. The true value of θ will be denoted by θ^o. Note that for identifiability of η^0, one needs the condition that $\alpha_j^0 \neq 0$ for at least one $j \geq q+1$. Beran and Weiershäuser (2011) and Beran et al. (2013) derived the asymptotic distribution of the least squares estimator of θ^0 under long memory, short memory and antipersistence of the residual process e_t. In particular, if e_t is linear, then unified formulas applicable to all three cases can be derived. The key to obtaining these results is a linearization of the nonlinear regression estimator of θ and convergence of weighted sums of e_t to integrals with respect to fractional Brownian motion. Combined with fractional calculus unified formulas follow.

We will use the notation $v(d)$ as in Corollary 1.2. Minimizing the sum of the squared residuals, $Q(\theta) = \sum_{t=1}^{n} [X_t - m(s_n; \theta)]^2$ (with $s_n = t/n$) with respect to θ can be done in two steps. First of all, for each value of η, the optimal value of α is obtained by standard linear least squares regression on the functions f_j defined by using knot η. Thus, for each $\eta \in (0,1)$ we define the $n \times p$ matrix

$$\mathbf{W}_n = \mathbf{W}_n(\eta) = (w_{ij})_{i=1,\ldots,n;\, j=1,\ldots,p} = (\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{p,n}) \tag{7.77}$$

with $w_{i,j} = f_j(\frac{i}{n})$ $(1 \leq i \leq n;\, 1 \leq j \leq p)$, and column vectors denoted by $\mathbf{w}_{j,n}$ $(j = 1, \ldots, p)$. For n large enough, $\mathbf{W}_n^T \mathbf{W}_n$ is invertible so that the projection matrix on the column space of $\mathbf{W}_n(\eta)$ may be written as

$$P_{\mathbf{W}_n} = P_{\mathbf{W}_n}(\eta) = \mathbf{W}_n \big(\mathbf{W}_n^T \mathbf{W}_n\big)^{-1} \mathbf{W}_n^T. \tag{7.78}$$

Thus, given observations $\mathbf{X} = (X_1, \ldots, X_n)^T$, $\hat{\eta}$ is obtained by minimizing $\|\mathbf{X} - P_{\mathbf{W}_n}(\eta)\mathbf{X}\|^2$ with respect to η. The slope estimates are given by

$$\hat{\alpha} = \big(\mathbf{W}_n^T \mathbf{W}_n\big)^{-1} \mathbf{W}_n^T \mathbf{X}$$

and $m(s_1), \ldots, m(s_n)$ are estimated by

$$\left[m\left(\frac{1}{n}; \hat{\theta}\right), m\left(\frac{2}{n}; \hat{\theta}\right), \ldots, m(1; \hat{\theta}) \right]^T = P_{\mathbf{W}_n(\hat{\eta})} \mathbf{X}. \tag{7.79}$$

Note that, in spite of the projection, neither $\hat{\alpha}$ nor $\hat{\eta}$ are linear in $\mathbf{X}$. For general piecewise polynomials, linearization of $\hat{\theta}$ has to take into account that derivatives of m with respect to η may not exist for $t = \eta$. Denoting by $m_{(j+)}$ the right-hand partial derivatives of m with respect to θ_j and defining the $n \times (p+1)$ matrix

$$\mathbf{M}_{n+} = \left[m_{(j+)}(t/n)\right]_{t=1,\ldots,n;\, j=1,\ldots,p+1} \in \mathbb{R}^{n\times(p+1)} \tag{7.80}$$

the limit

$$\lim_{n\to\infty} n^{-1}\left(\mathbf{M}_{n+}^T\mathbf{M}_{n+}\right)_{jk} = \int_0^1 m_{(j+)}(s,\theta)m_{(k+)}(s,\theta)\,ds \tag{7.81}$$

exists. Therefore, the matrix $\mathbf{M}_{n+}^T\mathbf{M}_{n+}$ is of full rank for n large enough, and we can also define the asymptotic matrix

$$\Lambda = \lim_n n\left(\mathbf{M}_{n+}^T\mathbf{M}_{n+}\right)^{-1}. \tag{7.82}$$

Suppose now that the spectral density of e_t is of the form $f_e(\lambda) \sim c_f|\lambda|^{-2d}$ for $\lambda \to 0$ where $d \in (-\frac{1}{2}, \frac{1}{2})$. Using the notation $e(n) = (e_1, \ldots, e_n)^T$ it can then be shown that $\|\hat{\theta} - \theta - (\mathbf{M}_{n+}^T\mathbf{M}_{n+})^{-1}\mathbf{M}_{n+}e(n)\| = o_p(n^{d-\frac{1}{2}})$ and

$$\lim_{n\to\infty} cov\left(n^{\frac{1}{2}-d}\nu^{-\frac{1}{2}}(d)\left(\mathbf{M}_{n+}^T\mathbf{M}_{n+}\right)^{-1}\mathbf{M}_{n+}^T e(n)\right) = \Lambda\Sigma_0\Lambda \tag{7.83}$$

where Σ_0 depends on d. At first sight, the formulas for Σ_0 seem to be quite different depending on whether we have long memory, short memory or antipersistence:

1. $d > 0$:

$$\Sigma_0 = d(1-2d)\left(\int_0^1\int_0^1 \frac{m_{(j)}(s)m_{(k)}(t)\,dt\,ds}{|s-t|^{1-2d}}\right)_{j,k=1,\ldots,p+1}. \tag{7.84}$$

2. $d = 0$:

$$\Sigma_0 = \left(\int_0^1 m_{(j)}(t)m_{(k)}(t)\,dt\right)_{j,k=1,\ldots,p+1}. \tag{7.85}$$

3. $d < 0$:

$$\begin{aligned}\Sigma_0 = c\Bigg(&\int_0^1 m_{(j)}(t)\int_{\mathbb{R}\setminus[0,1]} \frac{m_{(k)}(t)}{|s-t|^{1-2d}}\,ds \\ &- \int_0^1 \frac{m_{(k)}(s) - m_{(k)}(t)}{|s-t|^{1-2d}}\,ds\,dt\Bigg)_{j,k=1,\ldots,p+1}\end{aligned} \tag{7.86}$$

with $c = d(1-2d)$.

However, using fractional calculus (as discussed in Sect. 3.7.3), one formula for all three cases can be given. This approach also helps deriving the asymptotic distribution of $\hat{\theta}$ in an elegant way similar to Pipiras and Taqqu (2000a, 2000c, 2003).

Extending $m_{(j+)}$ to the real axis by setting $m_{(j+)}(t) = 0$ $(j = 1, \ldots, p+1)$ for $t \notin [0, 1)$, the unified formula for Σ_0 can be given as follows (Beran et al. 2013):

Theorem 7.20 *Define*

$$c_1^2(d) := \int_{\mathbb{R}} \left((1+s)^d - s^d\right)^2 ds + \frac{1}{2d+1}.$$

Then

$$\Sigma_0 = \left[\frac{\Gamma(d+1)^2}{c_1^2(d)} \int_{\mathbb{R}} \left(I_-^d m_{(j+)}\right)(s) \left(I_-^d m_{(k+)}\right)(s)\, ds \right]_{j,k=1,\ldots,p+1}.$$

Finally, recalling the linearization

$$n^{\frac{1}{2}-d} v^{-\frac{1}{2}}(d)(\hat{\theta} - \theta) \approx n^{\frac{1}{2}-d} v^{-\frac{1}{2}}(d) \left(\mathbf{M}_{n+}^T \mathbf{M}_{n+}\right)^{-1} \mathbf{M}_{n+} e(n),$$

convergence to a normal distribution can be derived by extending limit theorems for weighted sums given in Pipiras and Taqqu (2000a, 2000c). The limit is a linear transformation of the $(p+1)$-dimensional Gaussian variable

$$Z := \left(\int m_{(j+)}(s)\, dB_H(s) \right)_{j=1,\ldots,p+1}$$

where $B_H(s)$ denotes a fractional Brownian motion with Hurst parameter $H = d + 0.5$ and the integral $\int \cdot\, dB_H(s)$ is understood in the sense of Pipiras and Taqqu (2000a, 2000c). The asymptotic distribution can then be expressed as follows.

Theorem 7.21 *Under the assumptions summarized above* (*see Beran and Weiershäuser* 2011 *and Beran et al.* 2013 *for detailed assumptions*) *we have, as* $n \to \infty$,

$$n^{\frac{1}{2}-d} v^{-\frac{1}{2}}(d)(\hat{\theta} - \theta) \underset{d}{\to} \Lambda Z \sim N(0, \Lambda \Sigma_0 \Lambda). \tag{7.87}$$

Note that the formulation of the asymptotic distribution in terms of fractional integration is general so that it directly applies to any continuous piecewise polynomial function $m(s) = \sum_{k=0}^{l} \sum_{j=1}^{p_k} a_{k,j} (s - \eta_k)_+^{\beta_{j,k}}$ as specified above.

An application of these results to calcium imaging data in the context of olfactory research was introduced in Sect. 1.2. The data displayed in Fig. 1.18 are part of a data set consisting of estimated entropy series for 25 adult forager bees (*Apis mellifera carnica*). The original series were based on calcium imaging data reflecting the response in the antennal lobe of bees to an odorant stimulus (more specifically, hexanol). For the response series in Fig. 1.18, a linear spline function (i.e. a continuous piecewise linear function) with one known knot at the time of intervention and two subsequent unknown knots provides a rather accurate approximation of the

main characteristics. For each bee, two response series were measured under two different conditions, namely without and with the addition of the neurotransmitter octopamine. The research hypothesis was that under the influence of the neurotransmitter, the change in entropy should be faster. Using a linear splines fit with one known knot η_0 at the time of intervention and two subsequent unknown knots η_1,η_2, we have $m(s) = \alpha_0 + \alpha_1 s + \alpha_2(s-\eta_0)_+ + \alpha_3(s-\eta_1) + \alpha_4(s-\eta_2)_+$ with unknown parameter vector $\theta = (\alpha_0, \ldots, \alpha_4, \eta_1, \eta_2)$. Let θ_{without} and θ_{with} be the parameters without and with octopamine. Then checking the research hypothesis can be interpreted as testing the null hypothesis $H_0 : \alpha_{2,\text{without}} = \alpha_{2,\text{with}}$. Using least squares estimation for each of the response series, the distribution of $\hat{\alpha}_{2,\text{without}}$ and $\hat{\alpha}_{2,\text{with}}$, respectively, follows from the theorem above. Since the two series are always measured within one individual bee, the estimates are correlated so that a paired test has to be applied that takes into account the correlation ρ between the two estimates. The difference $\hat{\Delta} = \hat{\alpha}_{2,\text{with}} - \hat{\alpha}_{2,\text{without}}$ is then approximately normal with variance $\text{var}(\hat{\Delta}) = \text{var}(\hat{\alpha}_{2,\text{with}}) + \text{var}(\hat{\alpha}_{2,\text{without}}) - \rho\sqrt{\text{var}(\hat{\alpha}_{2,\text{with}})\,\text{var}(\hat{\alpha}_{2,\text{without}})}$. The variances are obtained from the asymptotic results above whereas ρ may be replaced by the sample correlation based on all bees in the data set. Beran et al. (2013) used these estimates to calculate an optimally weighted mean as an estimate of $\mu_\Delta = E(\hat{\Delta})$. Using asymptotic normality or bootstrap, it could indeed be shown that $\mu_\Delta > 0$ with a p-value below 1 %.

7.4 Nonparametric Regression with LRD Errors—Kernel and Local Polynomial Smoothing

In this section, we consider the nonparametric regression model

$$Y_i = m(X_i) + \sigma(X_i)e_i \quad (i = 1, \ldots, n), \tag{7.88}$$

where $m(\cdot)$, $\sigma(\cdot)$ are unknown functions, X_i are predictors (deterministic or random), and e_i is a second-order stationary process. First, in Sect. 7.4.1, we give a brief introduction to kernel (Priestley–Chao, Nadaraya–Watson) and local polynomial smoothing. We provide some preliminary calculations of the bias and variance and point out important differences between fixed and random design. It turns out that random design may improve rates of convergence. We have observed this already for parametric regression in Sects. 7.1 and 7.2 . Methods for estimating derivatives and boundary effects are also discussed.

In Sects. 7.4.2–7.4.3, we present general results for fixed design kernel and local polynomial estimation. In particular, it is shown that long memory or antipersistence influences rates of convergence. Hall and Hart (1990b) were the first to derive an asymptotic formula for the mean squared error of kernel estimators of the trend function in fixed-design regression with long-memory errors. This result was extended further in Beran and Feng (2001a, 2001b, 2002a, 2002b, 2002c), including kernel estimation with boundary corrections, local polynomial estimation of derivatives and integrated processes. Further results have been obtained in

Csörgő and Mielniczuk (1995b, 1995a), Robinson (1997), Beran and Feng (2001a, 2007), Pawlak and Stadtmüller (2007), Feng et al. (2007). Extensions to LARCH-type residuals are given in Beran and Feng (2007). Optimal convergence rates are derived in Feng and Beran (2012), but will not be discussed here. The nonexistence of optimal kernels in the long-memory setting is shown in Beran and Feng (2007). Sections 7.4.4 and 7.4.6 are devoted to bandwidth choice in nonparametric kernel and local polynomial regression. Bandwidth choice in the long-memory context by cross-validation originates from Hall et al. (1995a), whereas the plug-in approach is discussed in Ray and Tsay (1997), Beran and Feng (2002a, 2002b, 2002c). Sections 7.4.5 and 7.4.6 include a discussion of the so-called SEMIFAR models and iterative procedures to estimate the trend function and, in particular, the long-memory parameter simultaneously (Beran 1999; Beran and Feng 2001a, 2001b, 2002a, 2002b, 2007, Beran and Ocker 2001). Furthermore, robust versions of local polynomial estimators in the long-memory context are considered in Beran et al. (2002) and Beran et al. (2003). Extensions to nonequidistant time series and tests for rapid change points are discussed in Sect. 7.10 (Menéndez et al. 2010).

Section 7.4.8 is devoted to random design regression. It turns out that the choice of a bandwidth is even more fundamental than for fixed design regression. We show a dichotomy between small and large bandwidths. This is the same phenomenon as observed already for density estimation (see Sect. 5.14). For small bandwidths, long-range dependence in the residuals has no influence and one obtains exactly the same asymptotic distribution as for i.i.d. data. This is in contrast to fixed-design kernel (and local polynomial) regression. For large bandwidths, we have a long-memory behaviour. We also show an improvement in the rate of convergence for shape functions. Such observations have its origin in the work by Cheng and Robinson (1994). Further references include Csörgő and Mielniczuk (1999, 2000), Mielniczuk and Wu (2004), Zhao and Wu (2008), Kulik and Lorek (2011). In the latter article, the authors consider a very general class of errors that includes FARIMA–GARCH and antipersistent processes. In Bryk and Mielniczuk (2008), the authors consider a randomization scheme for fixed-design regression. As a consequence, the resulting kernel estimator has a rate of convergence as in the random-design case. Results for the Nadaraya–Watson estimator have further extensions to local linear regression estimators (see Masry and Mielniczuk 1999 and Masry 2001). Furthermore, Benhenni et al. (2008) considered consistency of a kernel estimator in functional regression with stochastic regressors and long-memory errors.

In Sect. 7.4.9, we deal with estimation of the conditional variance $\sigma^2(\cdot)$ in random-design regression. Rates of convergence are different than for estimation of the conditional mean $m(\cdot)$ in the model (7.88). Such results are obtained in Guo and Koul (2008), Zhao and Wu (2008), Kulik and Wichelhaus (2011, 2012), and also have some connections to residual empirical processes. The latter topic is not discussed here, we refer to Chan and Ling (2008) and Kulik and Lorek (2012).

7.4.1 Introduction

Here we briefly recall some basic results from kernel- and local polynomial smoothing. Also some first heuristic comments are made on the role of long-range dependence and antipersistence in the context of nonparametric regression.

7.4.1.1 The Priestley–Chao Regression Estimator—Deterministic Design

We consider the nonparametric regression model with a response variable Y being a function of a deterministic design variable X. In the simplest case, we have the regression model

$$Y_i = m(x_i) + e_i \quad (i = 1, 2, \ldots, n) \tag{7.89}$$

with fixed (i.e. deterministic) equally spaced design variables $x_1, x_2, \ldots, x_n$. Often one uses $x_i = t_i = in^{-1} \in [0, 1]$. To emphasize that the "explanatory" variables x_i are deterministic and equally spaced, we will use the notation t_i instead of x_i. Note that, strictly speaking, one actually has a sequence of models $Y_{i,n}$ because the grid of t-values (x-values) changes slightly with each n, i.e.

$$Y_i = Y_{i,n} = m(t_i) + e_i.$$

The residual process e_i is assumed to be second-order stationary with $E(e_i) = 0$, autocovariances $\gamma_e(k)$ and variance $\sigma_e^2 = \gamma_e(0)$. The regression function $m(t_i)$ is not specified except for suitable regularity conditions. In kernel and local polynomial smoothing, one usually assumes that m is at least continuous, or even a few times continuously differentiable (see, e.g. standard books such as Härdle 1990a, 1990b; Wand and Jones 1994; Fan and Gijbels 1996; Simonoff 1996; Eubank 1999; Tsybakov 2010).

Effective estimation of m can be quite difficult in the presence of long-range dependence. The reason is that long-memory processes tend to exhibit spurious trends which may be mistaken for deterministic ones. At the same time, smooth trends can lead to increased values of the periodogram near the origin and to sample autocovariances with a high positive bias. For example, considering a sample autocovariance at a fixed lag $k \geq 0$,

$$\hat{\gamma}(k) = n^{-1} \sum_{i=1}^{n-k} (y_i - \bar{y})(y_{i+k} - \bar{y}) \tag{7.90}$$

we have, as $n \to \infty$, $\mathrm{var}(\hat{\gamma}(k)) = o(1)$, but

$$\mathrm{Bias} = E\big[\hat{\gamma}(k)\big] - \gamma_e(k) \sim \int \left[m(t) - \int m(s)\, ds \right]^2 dt, \tag{7.91}$$

which is a positive constant, unless m is constant almost everywhere. Thus, not removing the trend function leads to the overestimation of d. Related to this is the

problem that the choice of a good estimate of m depends on approximate knowledge of d. A feasible solution that will be described below (Sects. 7.4.4 and 7.4.6) can be given in terms of an iterative procedure where trend estimation and estimation of the dependence parameters of e_i are applied repeatedly (Beran and Feng 2002a, 2002b; Ray and Tsay 1997).

Suppose now that m is smooth (in a sense to be specified). The problem is nonparametric estimation of this function. The Priestley–Chao estimator ($0 < x < 1$) is given by

$$\widehat{m}_{\mathrm{PC}}(t) = \frac{1}{nb} \sum_{i=1}^{n} y_i K\left(\frac{t_i - t}{b}\right) \tag{7.92}$$

(Priestley and Chao 1972) where $b > 0$ is a bandwidth, and $K \geq 0$ is a symmetric kernel function with support $[-1, 1]$ and $\int K(u)\,du = 1$. The idea is that, since m is continuous, the value of $m(t)$ may be estimated by taking a weighted average over a neighbourhood of x. For instance, if $K(u) = \frac{1}{2}1\{-1 \leq u \leq 1\}$, then $\widehat{m}_{\mathrm{PC}}(t)$ is the average over all y_i with $t - b \leq t_i \leq t + b$. Since $t_i = in^{-1}$, this condition means $n(t - b) \leq i \leq n(t + b)$ so that we are taking an average over $2[nb] + 1$ observations. Since the grid of t-values is increasingly dense and m is continuous, the bias of $\widehat{m}_{\mathrm{PC}}(t)$ converges to zero, provided that the neighbourhood we are taking observations from shrinks. At the same time, however, one needs to make sure that the variance of $\widehat{m}_{\mathrm{PC}}(t)$ tends to zero which means that the number of observations in the weighted mean must increase to infinity. This leads to the conditions $b \to 0$ and $nb \to \infty$.

The most important decision in kernel regression is the choice of the bandwidth b. If b is chosen too small, then the number of averaged observations is small so that the variance is large. On the other hand, if b is too large, then one averages the function m over a large neighbourhood of x. For highly nonlinear functions, this leads to a large bias. This dilemma leads to a trade-off between minimizing bias and variance. If the mean squared error is used as a criterion, then the separation of the two effects is additive,

$$\begin{aligned} MSE &= E\left[\left(\widehat{m}_{\mathrm{PC}}(t) - m_{\mathrm{PC}}(t)\right)^2\right] \\ &= \left[E\left(\widehat{m}_{\mathrm{PC}}(t)\right) - m_{\mathrm{PC}}(t)\right]^2 + E\left[\left(\widehat{m}_{\mathrm{PC}}(t) - E\left(\widehat{m}_{\mathrm{PC}}(t)\right)\right)^2\right] \\ &= \mathrm{Bias}^2 + \mathrm{Variance}. \end{aligned}$$

Asymptotic expressions for the bias do not depend on the autocovariance structure of e_i. Suppose that m is twice continuously differentiable. Using the notation $i_0 := [nt]$ and $u_i = (t_i - t)/b$, the standard argument is a Taylor expansion of the form

$$\mathrm{Bias}\left(\widehat{m}_{\mathrm{PC}}(t)\right) = E\left(\widehat{m}_{\mathrm{PC}}(t)\right) - m(t) = \frac{1}{nb} \sum_{i=1}^{n} K(u_i) m(t + bu_i) - m(t)$$

$$= \frac{1}{nb}\sum_{i=1}^{n} K(u_i)\left[m(t) + bu_i m'(t) + \frac{1}{2}b^2 u_i^2 m''(t) - m(t) + o\big(b^2\big)\right]$$

$$= b^2 \frac{1}{2} m''(t) \int_{-1}^{1} u^2 K(u)\,du + o\big(b^2\big) + O\left(\frac{1}{nb}\right).$$

(Note that the symmetry of K implies $\int K(u)u\,du = 0$.) Thus, the bias is proportional to the squared bandwidth and to the second derivative of $m(t)$. If we can assume a higher degree of smoothness of $m(t)$, then an even better order of the bias can be achieved by using a different type of kernel. Suppose that $m(t)$ is k times differentiable. Using a Lipschitz continuous kernel with

$$\int K(u)u^i\,du = \begin{cases} 1, & i = 0, \\ 0, & i = 1, \ldots, k-1, \\ \beta_k, & i = k, \end{cases} \tag{7.93}$$

we obtain

$$\begin{aligned} \operatorname{Bias}\big(\widehat{m}_{\mathrm{PC}}(t)\big) &\approx \frac{1}{nb}\sum_{i=1}^{n} K(u_i)\left[bu_i m'(t) + \frac{1}{2}b^2 u_i^2 m''(t) + \cdots\right] \\ &= \sum_{j=1}^{k} b^j \frac{m^{(j)}(t)}{j!} \int_{-1}^{1} u^j K(u)\,du + o\big(b^k\big) + O\left(\frac{1}{nb}\right) \\ &= b^k \frac{m^{(k)}(t)}{k!}\beta_k + o\big(b^k\big) + O\left(\frac{1}{nb}\right), \end{aligned}$$

provided that the error term in the Taylor expansion can be controlled well. Thus the bias is order $O(b^k)$. Kernels with property (7.93) are called *kernels of order k*, the kth moment of K, denoted by $\beta_k = \int K(u)u^k\,du \neq 0$, is the so-called *kernel constant* in the asymptotic bias. In most cases, one uses kernels of order 2 for estimating $m(t)$ because one would like to keep the assumptions on the unknown function as general as possible. More comments on the choice of a kernel are given in the next section.

In contrast to the bias, the variance of $\widehat{m}_{\mathrm{PC}}(t)$,

$$\operatorname{var}\big(\widehat{m}_{\mathrm{PC}}(t)\big) = (nb)^{-2}\sum_{i,j=1}^{n} K\left(\frac{t_i - t}{b}\right)K\left(\frac{t_i - t}{b}\right)\gamma_e(i-j),$$

depends on the autocovariance structure of e_i. In particular, the distinction between short memory, long memory or antipersistence is essential because the variance turns out to be proportional to $(nb)^{2d-1}$. This implies that a bandwidth chosen by minimizing the *MSE* will be of a different order for different values of d. It should be noted that the choice of b is not only important for estimating m but also for reliable estimation of the parameters d and c_f which, in turn, determine the optimal

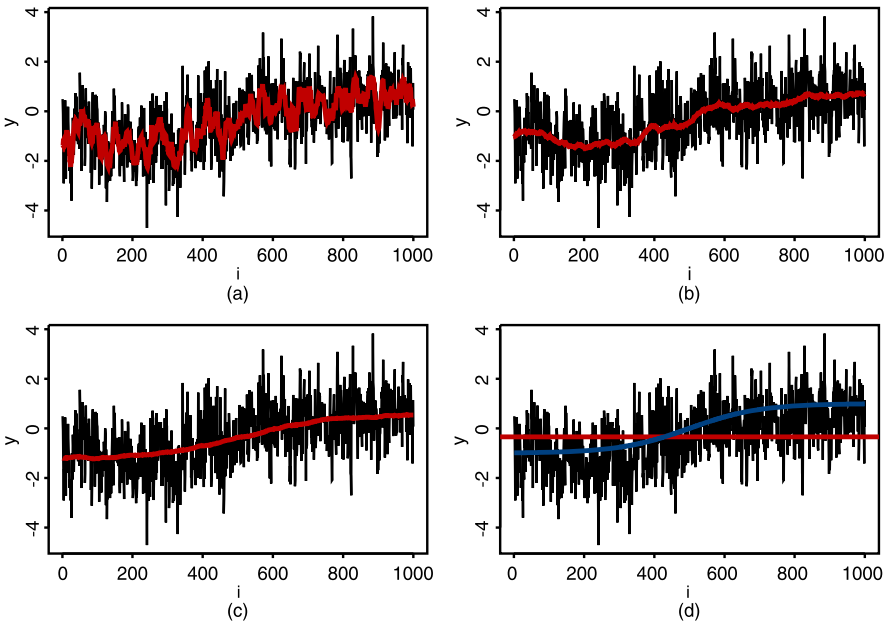

Fig. 7.6 The four pictures show the same series $Y_i = m(t_i) + e_i$ with $m(t) = \tanh(\frac{1}{2}(t - \frac{1}{2}))$ and e_i generated by a FARIMA$(0, 0.3, 0)$ process with innovation variance one. The four figures show nonparametric fits $\hat{m}(t)$ based on kernel regression with the rectangular kernel and different bandwidths: (**a**) very small bandwidth; (**b**) medium size bandwidths; (**c**) large bandwidth; (**d**) $b = \infty$. In (**d**), the true trend function is also shown

value of b. Moreover, knowledge of these two parameters is needed for tests and confidence intervals for m, as well as for forecasting.

If one lets d vary freely, then the choice of a good bandwidth is not only more difficult but also more important than in situations where one assumes short memory (i.e. $d = 0$) a priori. The reason is that, as mentioned above, the estimation of d from the residuals $\hat{e}_i = y_i - \hat{m}(t_i)$ very much depends on the choice of b. This is illustrated in Fig. 7.6 with $m(t) = \tanh(\frac{1}{2}(t - \frac{1}{2}))$ and e_i generated by a FARIMA$(0, 0.3, 0)$ process with innovation variance one. The four figures show nonparametric fits $\hat{m}(t)$ based on kernel regression with the rectangular kernel and different bandwidths: (a) very small bandwidth; (b) medium size bandwidths; (c) large bandwidth; (d) $b = \infty$ (so that $\hat{m}(t) \equiv \bar{y}$). The true trend function $m(t)$ is also displayed in Fig. 7.6(d). The bandwidth in (a) is clearly too small. The fitted line follows the data too closely. The corresponding residual series $\hat{e}_i$ (Fig. 7.7(a)) therefore resembles an antipersistent process. Fitting a FARIMA$(0, d, 0)$ process to $\hat{e}_i$ by maximum likelihood estimation (including model choice by the BIC) indeed yields a value of $\hat{d} = -0.34$. The moderate and large bandwidths used in (b) and (c) provide much better trend estimates. The corresponding values of $\hat{d}$ are equal 0.23 and 0.25, respectively, and thus much closer to the true value of $d = 0.3$. On the

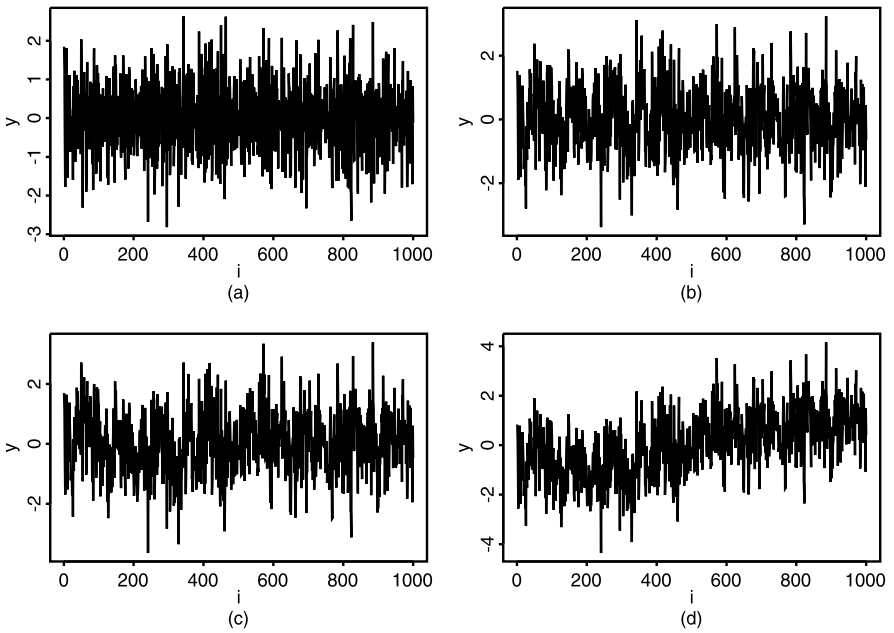

Fig. 7.7 Residuals $\hat{e}_i = Y_i - \hat{m}(t_i)$ based on the fits in Figs. 7.6(a)–(d)

other hand, choosing an infinite bandwidth, and thus not removing any trend estimate at all (Fig. 7.7(d)) leads to slight overestimation with $\hat{d} = 0.33$.

The easiest way to see the essential difference between long memory, short memory and antipersistence more formally is to look at the rectangular kernel $K(u) = \frac{1}{2}1\{-1 \le u \le 1\}$. For this second-order kernel, $\widehat{m}_{\mathrm{PC}}(t)$ is simply a sample mean of $2[nb] + 1$ consecutive observations. From Corollary 1.2, we know that the variance can be approximated by $c_f \nu(d) 2^{2d-1} (nb)^{2d-1}$ where the spectral density of e_i is assumed to be such that $f_e(\lambda) \sim c_f |\lambda|^{-2d}$, as $\lambda \to 0$, and

$$\nu(d) = \frac{\Gamma(1-2d)2\sin\pi d}{d(2d+1)} \quad (d \neq 0),\ \nu(0) = 2\pi.$$

Thus, for the mean squared error we have

$$MSE(t;b) \sim \tilde{C}_1(t) b^4 + \tilde{C}_2 (nb)^{2d-1} \tag{7.94}$$

with

$$\tilde{C}_1(t) = \left\{ \frac{1}{2} m''(t) \int_{-1}^{1} u^2 K(u)\, du \right\}^2 = \frac{1}{36} \{m''(t)\}^2$$

and $\tilde{C}_2 = v(d)2^{2d-1}c_f$. If the approximation is uniform in t (in a suitable sense), then we obtain an analogous formula for the integrated mean squared error

$$IMSE(b) = \int_0^1 MSE(t;b)\,dt \sim C_1 b^4 + C_2(nb)^{2d-1} \tag{7.95}$$

with

$$C_1 = \int_0^1 \tilde{C}_1(t)\,dt = \frac{1}{36}\int_0^1 \{m''(t)\}^2\,dt$$

and $C_2 = v(d)2^{2d-1}c_f$. Setting the derivative of the right-hand side of (7.95) equal to zero, we obtain the asymptotically optimal bandwidth

$$b_{\text{opt}} = C_{\text{opt}} n^{-\beta_{\text{opt}}} \tag{7.96}$$

with

$$\beta_{\text{opt}} = \frac{1-2d}{5-2d} = \frac{1}{5} - \frac{8d}{25-10d},$$

$$C_{\text{opt}} = \left[\frac{C_2(1-2d)}{4C_1}\right]^{\frac{1}{5-2d}} = \left[\frac{9(1-2d)v(d)2^{2d-1}c_f}{\int_0^1 \{m''(t)\}^2\,dt}\right]^{\frac{1}{5-2d}}.$$

The integrated squared curvature $\int_0^1 \{m''(t)\}^2\,dt$ is in the denominator. This means that a smaller bandwidth is required if m has various sharp turns. The reason is that the bias can become quite large when we average over a too large neighbourhood. In contrast, if m is close to a straight line, then the curvature is almost zero so that one may average with a large bandwidth without causing much damage. Note that b_{opt} is such that the bias and the variance terms in the *MSE* are of the same order. The optimal mean squared error is then of the order b^4 which means

$$MSE_{\text{opt}} \sim \text{const} \cdot n^{-4\beta_{\text{opt}}} = \text{const} \cdot n^{-\frac{4-8d}{5-2d}}. \tag{7.97}$$

Under short memory (including independence) with $d = 0$, one has the well known rates of $b_{\text{opt}} \sim \text{const} \cdot n^{-\frac{1}{5}}$ and $MSE_{\text{opt}} \sim \text{const} \cdot n^{-\frac{4}{5}}$. For long memory, β_{opt} is smaller than $\frac{1}{5}$ so that b_{opt} is larger and the MSE_{opt} converges to zero at a slower rate. The reason is that, due to long-term positive dependence, one needs more data to make the variance of the sample mean small. In contrast, under antipersistence ($d < 0$) β_{opt} is larger than $\frac{1}{5}$ so that the optimal bandwidth and mean squared error converge to zero faster than under short memory. These properties carry over to other kernels K. In summary, optimal bandwidth selection very much depends on the type of memory we have in the residual process. In the case of long memory, larger bandwidths are required. This is also related to the problem that it is often difficult to distinguish between long-range dependence and deterministic trend functions or change points in the mean (see also Sect. 7.9). The basic reason is that

trend functions tend to increase the values of the periodogram near the origin. This can be confounded with a pole due to long memory.

The practical application of (7.95) is not straightforward in practice because it involves the unknown quantities d, c_f and $m''(t)$. If we are willing to assume short memory, then the problem is less difficult because the long-memory parameter is fixed at $d = 0$. Various methods have been developed for obtaining a data driven approximation of the *IMSE* and thus an approximately optimal bandwidth. Well known methods are, for instance, cross-validation and iterative plug-in methods. If d is a free parameter in the interval $(-\frac{1}{2}, \frac{1}{2})$, then the problem is more involved. Data driven plug-in methods, however, have been developed, for instance, in Ray and Tsay (1997) and Beran and Feng (2002a, 2002b). The idea is to start with initial estimates of $m(\cdot)$ and $m''(t)$, estimate the parameters d and c_f from the residuals, obtain an estimate of b_{opt} and then iterate the procedure. This will be discussed below in the Sects. 7.4.4 and 7.4.6. In the short-memory context, similar methods are discussed in Gasser et al. (1991) and Ruppert et al. (1995).

7.4.1.2 Higher-Order Kernel Estimators and Estimation of Derivatives

So far we assumed that the kernel function K is given. More generally, not only the bandwidth but also the kernel K has to be chosen before carrying out a kernel regression. Although the choice of K is generally less important, it is still worth investigating the role of K in detail. In particular, one gains insight into the interplay between smoothness of the function and a suitable choice of the kernel, and it becomes more clear how to estimate derivatives.

Commonly used second-order kernels on $[-1, 1]$ are of the form

$$K_\mu(u) = C_\mu\big(1 - u^2\big)^\mu 1\{-1 \le u \le 1\} \tag{7.98}$$

for some nonnegative integer μ, where C_μ is such that $\int K(u)\,du = 1$. The parameter μ is called the *degree of smoothness* (or simply smoothness) of a kernel function of this type (see Müller 1984) which means that the $(\mu - 1)$th derivative of the kernel function is Lipschitz continuous. This also controls the degree of smoothness of the corresponding kernel estimator. For $\mu = 0, 1, 2, 3$, K_μ in (7.98) corresponds to the *Uniform kernel*, the *Epanechnikov kernel*, the *Bisquare kernel* and the *Triweight kernel,* respectively. Another commonly used kernel—which has, however, an unbounded support—is the Gaussian (or normal) kernel, i.e. the standard normal density function. It can also be considered as a rescaled limit of K_μ for $\mu \to \infty$. Explicit formulae of these kernel functions are given in Table 7.2.

The Uniform, the Epanechnikov and the Bisquare kernels are shown in Fig. 7.8. Corresponding higher-order kernels and kernels for estimating derivatives $m^{(j)}(t) = d^j/dt^j m(t)$ can be generated based on kernel functions defined in (7.98). This will be discussed below.

As already mentioned before, higher-order kernels as defined in (7.93) can be used to reduce the bias of $\hat{m}(t)$, if we are willing to assume stronger smoothness

Table 7.2 Some second-order kernels

Name	k	μ	Kernel (on $[-1, 1]$)
Uniform	2	0	$\frac{1}{2}$
Epanechnikov	2	1	$\frac{3}{4}(1-u^2)$
Bisquare	2	2	$\frac{15}{16}(1-2u^2+u^4)$
Triweight	2	3	$\frac{35}{32}(1-3u^2+3u^4-u^6)$
Gaussian	2	∞	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$ $(-\infty < u < \infty)$

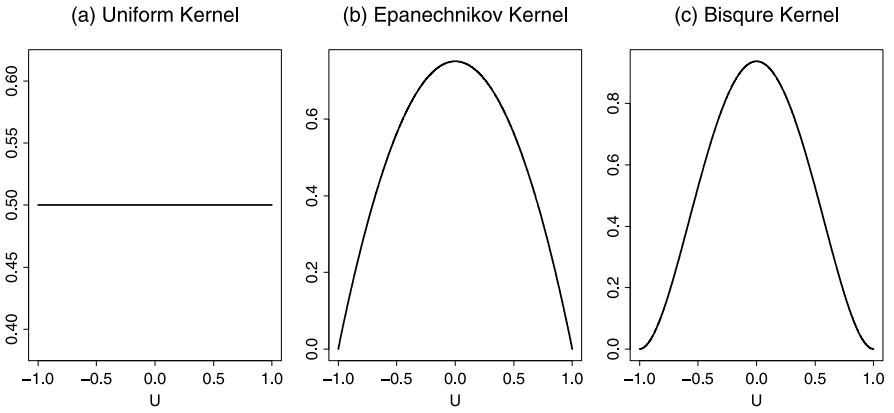

Fig. 7.8 Three commonly used second-order kernels with compact support

properties for m. Note that a high-order kernel with $k > 2$ (see (7.93)) is symmetric but not necessarily nonnegative. Thus, for

$$\hat{m}(t) = (nb)^{-1} \sum y_i K\big((t_i - t)/b\big) = \sum w_i y_i$$

the weights w_i are sometimes negative, although we still have $\sum w_i = 1$. Second-order kernels defined by (7.98) are special cases of (7.93) with $k = 2$. Most commonly used higher-order kernel functions are generated by the special kernels given in Table 7.2 (see Tables 5.7 of Müller 1988). Only kernels of polynomial form will be used for simplicity in the following. Most of the standard kernels proposed in the literature are of polynomial form.

Once the order of the kernel is fixed, its shape is less important and in particular does not influence the rate of convergence. If the residuals e_i are i.i.d., then the optimal second-order kernel is Epanechnikov's function $K(u) = \frac{3}{4}(1-u^2)$, in the sense that it minimizes the MSE when the optimal bandwidth is used (Epanechnikov 1969; Benedetti 1977). Similarly, higher-order kernels generated by the Epanechnikov kernel are also optimal for the corresponding order. These findings remain true under short memory. Despite its elegance this result is of little practical relevance because using suboptimal kernels does not lead to a substantial increase in the

asymptotic MSE (Rosenblatt 1971). Furthermore, it turns out that an optimal kernel function does not exist in the long-memory setting.

Slightly more important than the shape is the degree of smoothness of the kernel function because it carries over to $\hat{m}(t)$. If a kernel of smoothness μ is used, then $\hat{m}$ has the same degree of smoothness, i.e. the $(\mu - 1)$th derivative of $\hat{m}$ is Lipschitz continuous. Thus, the higher the μ the smoother the $\hat{m}$. For instance, $\hat{m}$ obtained with the uniform kernel is discontinuous because the kernel itself is discontinuous at both end points ($u = \pm 1$). Note in particular that this does not depend on the smoothness of the true function m, nor is it influenced by the dependence structure of e_i.

The most important feature of a kernel is its *order*. As demonstrated above, the optimal rate of convergence of $\hat{m}(t)$ is faster the higher the order k. One should bear in mind, however, that, in general, this is only true if $m(t)$ itself is smooth enough. Otherwise the asymptotic arguments leading to a bias of order $O(b^{2k})$ do not apply. Thus, using higher-order kernels and the corresponding asymptotic results involves rather strong assumptions on the unknown trend function m. Moreover, the finite sample variance of a higher order kernel estimator is usually larger than for a second-order kernel estimator. For small samples, the performance of a higher-order kernel estimator is therefore not necessarily better, even if m has the required smoothness properties. In practice, the order of the kernel is often chosen subjectively according to the data and further analysis. The safest choice that requires minimal assumptions is, however, a kernel of order 2.

Though the notion of higher-order kernels for estimating $m(t)$ may seem mainly of theoretical interest; the general approach of defining higher-order kernels via their moments becomes practically relevant when it comes to estimating derivatives. Estimation of derivatives is not only important in applications where the derivatives themselves are the object of interest. Even if the actual aim is to estimate $m(t)$, optimal data driven bandwidth selection based on the plug-in idea requires the estimation of higher-order derivatives (see, e.g. (7.96)). Kernel estimators of $m^{(j)}(t)$ in the i.i.d. case are investigated, for instance, in Gasser and Müller (1984), Rice (1986) and Ullah (1988, 1989). The simplest way of obtaining an estimate of the jth derivative is to start with $\hat{m}(t)$ based on a kernel of order $k > j$ (as in definition (7.93)) that is at least j times differentiable, and then take the derivative. Thus we define

$$\frac{d^j}{dt^j}\hat{m}_{\mathrm{PC}}(t) = \frac{1}{nb}\sum_{i=1}^{n}\frac{d^j}{dt^j}K\left(\frac{t_i - t}{b}\right)y_i \tag{7.99}$$

$$= \frac{1}{nb^{j+1}}\sum_{i=1}^{n}(-1)^j K^{(j)}\left(\frac{t_i - t}{b}\right)y_i. \tag{7.100}$$

A more systematic approach is to define a new class of kernels as follows. Let $j \geq 0$ be an integer and k such that $k - j \geq 2$ is an even number. A kernel function K of *order* (j, k) for estimating the jth derivative of $m(t)$ (Gasser et al. 1985; Müller

1984, 1988) is defined as a Lipschitz continuous function satisfying the moment conditions

$$\int K(u)u^i\,du = \begin{cases} 0, & 0 \le i \le k-1, i \ne j, \\ j!, & i = j, \\ \beta_k, & i = k, \end{cases} \tag{7.101}$$

where $\beta_k = \int K(u)u^k\,du \ne 0$ is again a *kernel constant* in the asymptotic bias. A kernel of order (j,k) with $k = j+2$ is called a standard kernel function. On the other hand, K is called a higher-order kernel, if $k > j+2$. The estimator of $m^{(j)}(t)$ is then given by

$$\hat{m}^{(j)}_{\mathrm{PC}}(t) = \frac{1}{nb^{j+1}} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) y_i = \sum_{i=1}^{n} w_i^j y_i \tag{7.102}$$

with $w_i^j = (nb^{j+1})^{-1} K((t_i - t)/b)$. As will be seen below, a necessary and sufficient condition for consistency of $\hat{m}^{(j)}_{\mathrm{PC}}(t)$, for $d \in (-0.5, 0.5)$, is that $b \to 0$ and $(nb)^{1-2d} b^{2j} \to \infty$. In particular, the second condition implies $nb^{1+j} \to \infty$ which is a necessary condition for w_i^j to tend to zero uniformly. More exactly, (7.102) is a good definition for interior points only. As discussed in the next section, the kernel has to be modified near the border to keep the bias small. This will be discussed below. A heuristic justification of definition (7.101) and (7.102) can be given as before, namely

$$\begin{aligned} E\big(\hat{m}^{(j)}_{\mathrm{PC}}(t)\big) &\approx \frac{1}{b^j} \sum_{i=0}^{k} b^i \frac{m^{(i)}(t)}{i!} \int_{-1}^{1} u^i K(u)\,du + o\big(b^{k-j}\big) + O\left(\frac{1}{nb}\right) \\ &= m^{(j)}(t) + b^{k-j} \frac{m^{(k)}(t)}{k!} \beta_k + o\big(b^{k-j}\big) + O\left(\frac{1}{nb}\right). \end{aligned}$$

Note that kernels of order $(0,k)$ coincide with kernels of order k according to the previous definition (7.93). Besides the moment conditions given in (7.101), some additional conditions are often required, such as the degree of smoothness and the minimal number of sign changes.

7.4.1.3 Boundary Effects and Boundary Kernels

Formula (7.102) does not yield good results for boundary points $t \in [0,b) \cup (1-b, 1]$ (see, e.g. Gasser and Müller 1979 and Müller 1984). The reason is that observations are not placed symmetrically on both sides of t. This increases the bias. While the bias of the estimator in (7.102) is of the order $O(b^2)$, it is the order $O(b)$ at boundary points. This problem can be solved by using the so-called boundary kernels. The solution is relatively complex in general though, in particular when higher order kernels are used or when estimation of the derivatives is

considered. A more elegant solution is provided by local polynomial regression discussed later, where adaptation at the boundary is automatic. Nevertheless, it is interesting to study the approach of boundary kernels because one gains a better understanding of boundary problems. Moreover, local polynomial fits can be represented asymptotically as kernel estimators with boundary kernels at boundary points (see Sect. 7.4.1.6).

Consider, for instance, a second-order kernel estimator $\hat{m}(t)$ of $m(t)$ and denote by $\Delta(t)$ its bias. The contribution of the bias to the IMSE is $B = \int_0^1 \Delta^2(t)\,dt$. Although the length of the boundary areas tends to zero, the contribution of $\Delta(t)$ in the boundary region is not negligible. The reason is that the contribution of interior points to the IMSE is

$$\int_b^{1-b} \Delta^2(t)\,dt = \int_b^{1-b} O\big(b^4\big)\,dt = O\big(b^4\big)$$

whereas for boundary points we have

$$\int_0^b \Delta^2(t)\,dt = \int_0^b O\big(b^2\big)\,dx = O\big(b^3\big)$$

and the same holds for $\int_{1-b}^1 \Delta^2(t)\,dt$. This means that the integrated squared bias is dominated by the bias in the boundary regions. In the extreme case with $t = 0$, the estimator in (7.102) even converges to $\frac{1}{2}m(0)$ because we have only half of the weights (Müller 1991). The boundary effect is even worse for higher-order kernel estimators and kernel estimators of derivatives.

The problem can be overcome by using boundary kernels that are designed to make the bias of the same order of magnitude for all $t \in [0, 1]$. To achieve that, the moment conditions given in (7.101) should be satisfied not only at interior but also at boundary points. Boundary kernels are solutions obtained from (7.101) and additional side conditions. Examples of boundary kernels may be found in Gasser and Müller (1979), Gasser et al. (1985), Müller (1991) and Müller and Wang (1994). In the following, the discussion will only be carried out for left boundary points $t \in [0, b)$. For the right boundary, arguments are analogous. Note that asymptotically any *fixed* point $t \in (0, 1)$ is an *interior* point because $b \to 0$. A left boundary point can be written as $t = cb$ with $0 \le c = c(t) < 1$. For interior points $t \in [b, 1-b]$, we define $c = 1$.

A left boundary kernel $K_c(u)$ of order (j, k) is defined as a Lipschitz continuous function with compact support $[-1, c]$ satisfying the moment conditions

$$\int_{-1}^{c} K_c(u)u^i\,du = \begin{cases} 0, & i = 0, \ldots, j-1,\ j+1, \ldots, k-1, \\ j!, & i = j, \\ \beta_{c,k} \neq 0, & i = k. \end{cases} \tag{7.103}$$

Boundary kernels for the right boundary $t \in (1-b, 1]$ are defined in an analogous manner.

Table 7.3 Three commonly used second-order μ-smooth boundary kernels

j	k	μ	Kernel function $K_c^{(\mu)}$ (on $[-1, c]$)
0	2	0	$\frac{1}{c+1}\{1+3(\frac{1-c}{1+c})^2+6\frac{1-c}{(1+c)^2}u\}$
0	2	1	$\frac{6}{(c+1)^3}\{1+5(\frac{1-c}{1+c})^2+10\frac{1-c}{(1+c)^2}u\}(1+u)(c-u)$
0	2	2	$\frac{30}{(c+1)^5}\{1+7(\frac{1-c}{1+c})^2+14\frac{1-c}{(1+c)^2}u\}(1+u)^2(c-u)^2$

Table 7.4 Three second-order boundary kernels proposed by Müller and Wang (1994)

j	k	μ	Kernel function $K_c^{(\mu,\mu-1)}$ (on $[-1, c]$)
0	2	0	$\frac{1}{c+1}\{1+3(\frac{1-c}{1+c})^2+6\frac{1-c}{(1+c)^2}u\}$
0	2	1	$\frac{12}{(c+1)^4}\{u(1-2c)+(3c^2-2c+1)/2\}(1+u)$
0	2	2	$\frac{15}{(c+1)^5}\{2u(5\frac{1-c}{1+c}-1)+(3c-1)+5\frac{(1-c)^2}{1+c}\}(1+u)^2(c-u)$

For the kernel function in the interior, some additional conditions are often required such as a certain degree of smoothness. Müller (1991) proposed a class of the so-called μ-smooth optimal boundary kernels which are obtained by solving (7.103) under the side condition that $\int_{-1}^{c}[K_c^{(\mu)}(u)]^2\,du$ is minimized. Such kernels have the same degree of smoothness in the boundary area as in the interior. Also, the degree of smoothness of such boundary kernels is always μ over the whole support $[-1, c]$. Second-order boundary kernels of this type (for estimating the regression function m itself) corresponding to the Uniform, the Epanechnikov and the Bisquare kernels in the interior (see Table 1 in Müller 1991) are listed in Table 7.3. For $c = 1$, these formulae reduce to the corresponding ones in the interior given in Table 7.2.

Another class of boundary kernels with a so-called $(\mu, \mu-1)$ degree of smoothness was proposed by Müller and Wang (1994). These are defined as solutions of (7.103) under certain smoothness conditions (see (K2) and (K3) in Müller and Wang 1994, with α and β there corresponding to μ and $\mu-1$, respectively). At a boundary point $t = cb$ with $0 \le c < 1$, the degree of smoothness of a boundary kernel in this class is μ at the left end point $u = -1$ and $\mu-1$ at the right end point $u = c$, provided that $\mu > 1$. In the interior, one obtains the same kernels as before. In particular, the kernels given in Table 7.3 may be called boundary kernels with a (μ, μ) degree of smoothness. The authors showed that these new boundary kernels have some advantages over those proposed in Müller (1991). Note that the boundary kernels given in Table 7.3 are polynomials of order $2\mu-2$ in the interior and of order $2\mu-1$ at the boundary. In contrast, for $\mu \ge 1$, the boundary kernels proposed by Müller and Wang (1994) are of the same order $2\mu-2$ in the interior and at the boundary. Boundary kernels in this class corresponding to the Uniform, the Epanechnikov and the Bisquare kernels in the interior are listed in Table 7.4. Note that here the boundary kernel corresponding to the Epanechnikov kernel with $c < 1$ is discontinuous at $u = c$. This means that the degree of smoothness at this end point is $\mu - 1 = 0$.

Further examples of boundary kernels can be found, for instance, in Gasser et al. (1985), Müller (1988, Sect. 5.8). Messer and Goldstein (1993) considered the continuation of equivalent spline kernels from the interior to the boundary. Gasser et al. (1985) also proposed some boundary kernels which, for any μ, are non-smooth at the end point $u = c$ ($c \neq 1$). Boundary kernels considered by Gasser et al. (1985) belong to another class generated by local polynomial regression with a truncated weight function at the boundary.

7.4.1.4 The Nadaraya–Watson Regression Estimator—Random Design

If we consider the same nonparametric regression model (7.89),

$$Y_i = m(x_i) + e_i \quad (i = 1, \ldots, n),$$

but with a design variable $X = x$ that is *random*, say with density function p_X, then the Priestley–Chao estimator has to be modified, in general. The reason is that by analogous arguments as above one obtains

$$E\big(\widehat{m}_{\mathrm{PC}}(x)\big) = p_X(x)m(x) + O\big(b^2\big) \quad \big(x \in (0,1)\big).$$

Thus, in general, one has a bias that does not disappear asymptotically, unless p_X is the uniform distribution on [0, 1]. (Note, in particular, that the equidistant fixed design considered previously can be seen as a special case, or rather an extended special case, in the sense of conditional inference given $x_1, \ldots, x_n$ and a uniform limiting design density p_X.) A simple solution is to divide $\widehat{m}_{\mathrm{PC}}(x)$ by a consistent estimate of $p_X(x)$. This is the idea of the Nadaraya–Watson estimator (Nadaraya 1964; Watson 1964)

$$\widehat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x_i - x}{b}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{b}\right)} = \frac{\widehat{m}_{\mathrm{PC}}(x)}{\hat{p}_X(x)} \tag{7.104}$$

where

$$\hat{p}_X(x) = \frac{1}{nb} \sum_{i=1}^n K\left(\frac{x_i - x}{b}\right)$$

is the so-called Parzen–Rosenblatt kernel estimator of $p_X(x)$ (Rosenblatt 1956; Parzen 1979) since, under standard conditions $\hat{p}_X(x) \to_p p_X(x)$ and $\widehat{m}_{\mathrm{PC}}(x) \to_p p_X(x)m(x)$, the Nadaraya–Watson estimator $\widehat{m}_{\mathrm{NW}}(x)$ converges in probability to $m(x)$. Expressions for the bias and variance are slightly more complicated than those for $\widehat{m}_{\mathrm{PC}}(x)$ in the deterministic equidistant case because the accuracy of $\hat{p}_X(x)$ also plays a role. However, the order of the bias is as before, namely $O(b^2)$ for second-order kernels. In how far the variance of $\widehat{m}_{\mathrm{NW}}(x)$ is influenced by the autocovariance structure depends on the random mechanism generating the values of X. This is similar to a parametric linear regression where, for instance, autocorrelations play no role when $Y_i = \beta x_i + e_i$ with $x_1, \ldots, x_n$ obtained by i.i.d. sampling of a zero-mean random variable X, whereas the opposite is true when $E(X) \neq 0$ (see Sect. 7.2).

7.4.1.5 Local Polynomial Smoothing

The main idea behind local polynomial smoothing (see, e.g. Ruppert and Wand 1994 and Fan and Gijbels 1995, 1996 and references therein) is based on a polynomial approximation of a $(p+1)$-times differentiable function $m(x)$ in a small neighbourhood of x. This is applicable to deterministic as well as to random designs. By a Taylor series expansion around x, a pth-degree polynomial approximation of $m(x_i)$ is given by

$$m(x_i) \approx m(x) + (x_i - x)m^{(1)}(x) + \frac{(x_i - x)^2}{2!}m^{(2)}(x) + \cdots + \frac{(x_i - x)^p}{p!}m^{(p)}(x).$$

As before, we use the notation $m^{(j)}$ for the jth derivative. Since the coefficients

$$\beta_j = \beta_j(x) = \frac{m^{(j)}(x)}{j!} \quad (j = 0, 1, 2, \ldots, p)$$

are fixed, we can rewrite $m(x_i)$ as

$$m(x_i) \approx \sum_{j=0}^{p} (x_i - x)^j \beta_j$$

where the coefficients $\beta_0, \ldots, \beta_p$ are the same for all x_i "close" to x. This enables us to estimate $m(x)$ and its derivatives $m^{(j)}(x)$ $(j = 1, 2, \ldots, p)$ by fitting a local polynomial of degree p to observations (x_i, y_i) with x_i (fixed or random) in the neighbourhood of x. Estimates of derivatives are then defined by

$$\hat{m}^{(j)}(x) = j!\hat{\beta}_j \quad (j = 0, 1, \ldots, p).$$

In other words, we apply a polynomial regression locally. The regression parameter $\beta = \beta(x) = (\beta_0, \ldots, \beta_p)^T$ is estimated by minimizing a weighted sum of squared residuals,

$$Q(x) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=0}^{p} (x_i - x)^j \beta_j \right\}^2 D\left(\frac{x_i - x}{b} \right),$$

with respect to β where the weights $D((x - x_i)/b)$ make sure that only values in the neighbourhood of x are included. In matrix form, Q can also be written as

$$Q(x) = (\mathbf{y} - \mathbf{X}\beta)' \mathbf{D}(x)(\mathbf{y} - \mathbf{X}\beta)$$

where

$$\mathbf{X} = (\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p+1}) = \begin{pmatrix} 1 & x_1 - x & \ldots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \ldots & (x_n - x)^p \end{pmatrix}$$

and

$$\mathbf{D} = \begin{pmatrix} D(\frac{x_1-x}{b}) & 0 & \dots & 0 \\ 0 & D(\frac{x_2-x}{b}) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & D(\frac{x_n-x}{b}) \end{pmatrix}. \tag{7.105}$$

The weighted least squares solution can be written as

$$\widehat{m^{(j)}}(x) = j!\hat{\beta}_j = j!\delta_{j+1}^T \left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}\mathbf{y} \tag{7.106}$$

where $\delta_j = (\delta_{1,j}, \dots, \delta_{p+1,j})^T$ $(j = 1, \dots, p+1)$ denote unit vectors with $\delta_{j,j} = 1$, $\delta_{i,j} = 0$ $(i \neq j)$.

To derive asymptotic properties of $\widehat{m^{(j)}}(x)$, it is often convenient to write (7.106) as a weighted sum. Defining the weighting system

$$\mathbf{w}_{j;b,n}^T = \left(w_{j;b,n}(x;1), \dots, w_{j;b,n}(x;n)\right) = j!\delta_{j+1}^T \left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}, \tag{7.107}$$

we have

$$\widehat{m^{(j)}}(x) = \mathbf{w}_{j;b,n}^T\mathbf{y} = \sum_{i=1}^n w_{j;b,n}(x;i)Y_i.$$

Note, that each weight $w_{j;b,n}(i)$ associated with Y_i changes with changing sample size n. Thus, investigating the asymptotic distribution of $\widehat{m^{(j)}}(x)$ amounts to studying the sequence of sums

$$S_n = \sum_{i=1}^n w_{j;b,n}(x;i)e_i = \sum_{i=1}^n \zeta_{i,n} \quad (n \in \mathbb{N}) \tag{7.108}$$

of a triangular array $\zeta_{i,n} = w_{v;b,n}(x;i)\varepsilon_i$ $(1 \leq i \leq n; n \in \mathbb{N})$. Since

$$\delta_{j+1}^T \left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}\mathbf{X} = \delta_{j+1}^T = (0, \dots, 0, 1, 0, \dots, 0)$$

(with 1 being the $(j+1)$st component), the weights have the property

$$\begin{aligned} \mathbf{w}_{j;b,n}^T\mathbf{x}_{\cdot j+1} &= j!\delta_{j+1}^T \left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}\mathbf{x}_{\cdot j+1} \\ &= \sum_{i=1}^n w_{j;b,n}(x;i)(x_i - x)^j = j! \end{aligned} \tag{7.109}$$

and

$$\mathbf{w}_{j;b,n}^T\mathbf{x}_{\cdot l+1} = \sum_{i=1}^n w_{j;b,n}(x;i)(x_i - x)^l = 0 \quad (l \neq j,\ 0 \leq l \leq p). \tag{7.110}$$

These equations hold under any design that makes $\hat{m}^{(j)}$ exactly unbiased in the case where m is a polynomial of degree $q \le p$.

The bias of local polynomial estimators is of the same order for interior and boundary points. For instance, if $j = 0$ and $p = 1$, then

$$\begin{aligned} E\big[\widehat{m}(x)\big] &= \sum_{i=1}^{n} w_{0;b,n}(x;i) m(x_i) \\ &= \sum_{i=1}^{n} w_{0;b,n}(x;i) \left[m(x) + (x_i - x) m^{(1)}(x) + \frac{1}{2} m^{(2)}(\tilde{x}_i)(x_i - x)^2 \right] \\ &= m(x) + 0 + \frac{1}{2} m^{(2)}(x) b^2 + o\big(b^2\big) = m(x) + O\big(b^2\big) \end{aligned}$$

where the latter equality follows from (7.110) and a detailed argument for the remainder term using the property $(x_i - x)^2 \le b^2$. More generally, local polynomial estimators of $m^{(j)}$ are automatically boundary corrected if $p - j$ is odd, in the sense that the bias at interior and boundary points is of the same order. In contrast, for kernel estimators (7.109) and (7.110) hold only approximately, and this leads to problems at the boundary. Furthermore, these properties show that local polynomial regression is design adaptive. In contrast to the Priestley–Chao kernel estimator, no adjustment by the design density is required.

More specifically, if $b \to 0$ and $nb^3 \to \infty$, then, under suitable conditions on D, expressions for the bias of $\widehat{m^{(j)}}(x)$ can be shown to be of the form

$$\operatorname{Bias}\big(\widehat{m^{(j)}}(x)\big) \sim c_1 \cdot \frac{m^{(p+1)}(x)}{(p+1)!} j! b^{p+1-j} \quad (\text{if } p - j \text{ odd}),$$

$$\operatorname{Bias}\big(\widehat{m^{(j)}}(x)\big) \sim c_2 \cdot \left\{ \frac{m^{(p+2)}(x)}{(p+2)!} + \frac{m^{(p+1)}(x)}{(p+1)!} \frac{p_X'(x)}{p_X(x)} \right\} j! b^{p+2-j} \quad (\text{if } p - j \text{ even})$$

with c_1 and c_2 not depending on m. In particular, this means that if $p - j$ is even, then the bias is affected by the design density. This can be problematic especially near the boundary of the x-space, and thus we have another reason for choosing $p - j$ odd. Moreover, one would like to choose p as small as possible in order to avoid unnecessary differentiability conditions on m. Therefore, the usual choice of p is $j + 1$ which leads to a bias of the order $O(b^2)$.

The variance of $\widehat{m^{(j)}}(x)$ depends on the autocovariance structure and the design. For asymptotic considerations, it is also useful to note that local polynomials can be approximated by kernel estimators. For instance, in the case of equidistant fixed design regression with $x_i = i/n =: t_i$, the asymptotically equivalent kernel estimator is (see Müller 1987 and Feng 1999)

$$\tilde{m}^{(j)}(t) = \frac{1}{nb} \sum K_{(j,p+1,c)}\left(\frac{t_i - t}{b} \right) Y_i$$

where the "equivalent kernel" $K_{(j,p+1,c)}$ has the following properties. As before, the notation is $t = cb$ and $1 - cb$ with $0 \le c < 1$ for boundary points $t = cb$ and $1 - cb$, and $c = 1$ for interior points $t \in [b, 1 - b]$. Then $K_{(j,p+1,c)}(u)$ is such that, for $0 \le j \le p$,

$$\int_{-c}^{1} K_{(j,p+1,c)}(u)u^l = 0 \quad (j \ne l),$$

$$\int_{-c}^{1} K_{(j,p+1,c)}(u)u^j = j!$$

and

$$\tau = \int_{-c}^{1} K_{(j,p+1,c)}(u)u^{p+1} \ne 0.$$

Note that the kernel is different for boundary points. This reflects the automatic boundary correction of local polynomials. Equivalence is expressed in terms of a uniform approximation of the weighting system $\mathbf{w}_{j;b,n}$ of $\hat{m}^{(j)}(t)$ by the weighting system $\tilde{\mathbf{w}}_{j;b,n}$ of $\tilde{m}^{(j)}(t)$, namely

$$\lim_{n\to\infty} \sup_{1\le i\le n} \left| \frac{w_{j;b,n}(t;i)}{\tilde{w}_{j;b,n}(t;i)} - 1 \right| = 0$$

where we define $0/0 := 1$ (Müller 1987; also see Lejeune 1985; Lejeune and Sarda 1992 and Ruppert and Wand 1994). Using the approximation by $\tilde{m}^{(j)}(t)$, one obtains the asymptotic variance of $\hat{m}^{(j)}(t)$ by similar arguments as for the Priestley–Chao kernel estimator,

$$\begin{aligned} \operatorname{var}\big(\tilde{m}^{(j)}(t)\big) &= (nb)^{-2} \sum_{i,j=1}^{n} K_{(j,p+1,c)}\left(\frac{t_i - t}{b}\right) K_{(j,p+1,c)}\left(\frac{t_i - t}{b}\right) \gamma_e(i-j) \\ &\sim \text{const} \cdot (nb)^{2d-1} b^{-2j} \end{aligned}$$

(Beran and Feng 2001a, 2001b, 2002c, 2007).

Example 7.28 Let $p = 0$. Then we obtain a local constant fit that minimizes

$$Q(x) = \sum_{i=1}^{n} \{y_i - \beta_0\}^2 D\left(\frac{t_i - t}{b}\right).$$

The solution is a weighted sample mean

$$\hat{\beta}_0(x) = \frac{1}{nb} \sum_{i=1}^{n} \tilde{D}\left(\frac{t_i - t}{b}\right) y_i$$

with

$$\tilde{D}(u) = \frac{D(u)}{(nb)^{-1}\sum_{i=1}^{n} D(u)}.$$

Thus, $\tilde{D}(u)$ is the equivalent kernel. Note that $\hat{\beta}_0(x)$ is the Nadaraya–Watson estimator discussed in the previous section. Explicit formulae of the weights for the local linear estimator of $m(t)$ are given by (2.3) and (2.4) in Fan (1992).

In summary, the main practical advantages of local polynomial estimation compared to direct kernel smoothing are the direct availability of estimated derivatives, the automatic bias correction at the border (for more discussion on this topic, see, e.g. Fan and Gijbels 1996) and design adaptivity. The calculation of $\hat{m}^{(j)}(x)$ is very simple because it essentially only requires a program for linear regression. The representation by an equivalent kernel estimator is useful for deriving asymptotic results.

7.4.1.6 Calculation of Equivalent Kernels

Here we provide some details on the calculation of the equivalent kernel introduced above. We consider the case of $j = 0$ only, i.e. estimation of $m(x)$ by

$$\widehat{m}(x) = \mathbf{w}^T \mathbf{y} = \sum_{i=1}^{n} w(i) Y_i$$

with

$$\mathbf{w} = \mathbf{w}_{0;b,n}^T = \delta_1^T \left(\mathbf{X}^T \mathbf{D} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D}.$$

Lejeune and Sarda (1992) showed that there is a kth order equivalent kernel function (for estimating m) where $k = p + 1$ if p is odd and $k = p + 2$ if p is even. It can be calculated as follows. Let

$$\mathbf{N}_p = \begin{pmatrix} 1 & \mu_1 & \dots & \mu_p \\ \mu_1 & \mu_2 & \dots & \mu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p+1} & \dots & \mu_{2p} \end{pmatrix}, \tag{7.111}$$

and

$$\mathbf{M}_p = \begin{pmatrix} 1 & \mu_1 & \dots & \mu_p \\ u & \mu_2 & \dots & \mu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ u^p & \mu_{p+1} & \dots & \mu_{2p} \end{pmatrix}, \tag{7.112}$$

where $\mu_j = \int_{-1}^{1} u^j D(u)\,du$ is the jth moment of $D(u)$. The equivalent kernel function is given by

$$K(u) = K_k(u) = \frac{\det(\mathbf{M}_p(u))}{\det(\mathbf{N}_p)} D(u). \tag{7.113}$$

Note that the kernel function is determined by the weight function $D(u)$ and the order of the polynomial p. It does not depend on the design and is therefore the same for fixed (equi- and nonequidistant) and random design. Another representation is

$$K(u) = \left(\sum_{j=1}^{p+1} a_{1j} u^{j-1} \right) W(u), \tag{7.114}$$

where $\mathbf{N}_p^{-1} = (a_{ij})_{i,j=1,\ldots,p+1}$. Note that for j even, $a_{1j} = 0$. Thus, all odd powers of u in (7.114) vanish. One can also see that $K(u)$ is a polynomial kernel whenever $D(u)$ is a polynomial. Moreover, if p is even, then $k = p + 2 = (p + 1) + 1$, and one can see that $K = K_k$ is the same for p and $p + 1$.

Let $w^{\mathrm{NW}}(x; i)$ denote the weights of the Nadaraya–Watson estimator of $m(\cdot)$ defined by $K_k(u)$. It can be shown that $w(x; i) = w^{\mathrm{NW}}(x; i)[1 + o_p(1)]$. Hence the kernel $K_k(u)$ is often called the (asymptotically) equivalent kernel function of the local polynomial regression. This interpretation is, however, somehow inaccurate because the detailed difference between the NW-estimator and the local polynomial estimator is only asymptotically negligible in the case of an equidistant design. This is not true for random or non-equidistant fixed design.

We conclude the discussion with two examples of equivalent kernels.

Example 7.29 Consider a local quadratic ($p = 2$) or local cubic ($p = 3$) estimator of $m(t)$ using the Epanechnikov kernel $D(u) = \frac{3}{4}(1 - u^2)$ ($|u| \le 1$) as weight function. We have $k = 4$, $a_{11} = \frac{15}{8}$ and $a_{13} = -\frac{35}{8}$. The resulting equivalent kernel is

$$K_4^{\mathrm{E}}(u) = \frac{15}{32}\left(3 - 10u^2 + 7u^4\right), \tag{7.115}$$

which is a well known fourth-order kernel used in the literature (Gasser et al. 1985).

Example 7.30 Consider a local quadratic ($p = 2$) or local cubic ($p = 3$) estimator of $m(t)$ using the Gaussian kernel $D(u) = \varphi(u) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}u^2)$ as weight function. We have $k = 4$, $a_{11} = \frac{3}{2}$ and $a_{13} = -\frac{1}{2}$. The resulting equivalent kernel is

$$K_4^{\mathrm{G}}(u) = \frac{1}{2}\left(3 - u^2\right)\varphi(u). \tag{7.116}$$

Further examples of equivalent kernel functions in the interior may be found in Gasser et al. (1985) and Müller (1988). Examples of equivalent kernels including boundary kernels and estimation of derivatives are given in Feng (1999, 2004a, 2004b).

7.4.2 Fixed-Design Regression with Homoscedastic LRD Errors

7.4.2.1 Bias and Variance of Kernel and Local Polynomial Estimators

We assume a nonparametric regression model (7.89) with a fixed equidistant design,

$$Y_i = Y_{i,n} = m(t_i) + e_i,$$

where $t_i = i/n$ and e_i is a second-order zero mean stationary process with spectral density $f_e(\lambda) \sim c_f|\lambda|^{-2d}$ for some $d \in (-\frac{1}{2}, \frac{1}{2})$. In view of the discussion above, essentially the same results are expected to hold for local polynomial estimators and kernel estimators with boundary kernels. The following results are therefore formulated under the assumption that $\hat{m}^{(j)}$ is either a local polynomial estimator (with polynomials of degree p) or a kernel estimator of the corresponding degree and boundary corrections.

For reasons discussed previously, we will assume $p - j$ to be odd. Moreover, we will use the notation $k = p + 1$. Thus $k \geq j + 2$ and $k - j$ is always even. If $\hat{m}^{(j)}$ is a local polynomial estimator with polynomials of order p, then it is asymptotically equivalent to a certain kth order kernel estimator with boundary corrections (see discussion above). The corresponding kernel is denoted by $K_{(j,p+1,c)}$. Otherwise, if we use a kernel estimator, then this denotes the kernel we use. To derive the asymptotic mean squared error, the following assumptions are sufficient (but not necessary).

A1. The errors e_i have the Wold decomposition

$$e_i = \sum_{s=0}^{\infty} a_s \varepsilon_{i-s}$$

where $E(\varepsilon_i) = 0$, $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_i) < \infty$,

$$f_e(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|A\left(e^{-i\lambda}\right)\right|^2 \sim c_f|\lambda|^{-2d} \quad (\lambda \to 0)$$

for some $d \in (-0.5, 0.5)$ and ε_i is a martingale difference.

A2. The trend function $m(t)$ is at least k $(= p + 1)$ times continuously differentiable on $[0, 1]$ with $k \geq j + 2$ and $k - j$ even, and $\hat{m}^{(j)}$ is either a pth order local polynomial or a kth order kernel estimator with a corresponding boundary correction.

A3. For the bandwidth we have, as n tends to infinity,

$$b \to 0, \qquad (nb)^{1-2d} b^{2j} \to \infty.$$

A4. For $y = x - (x - y)$ (with x and y in the support of $K_{(j,p+1,c)}$) the kernel $K_{(j,p+1,c)}$ can be written as

$$K_{(j,p+1,c)}(y) = K_{(j,p+1,c)}(x) + \tilde{K}_{(j,p+1,c)}(x - y), \tag{7.117}$$

where

$$\tilde{K}_{(j,p+1,c)}(x-y) = \sum_{j=1}^{r} \eta_j (x-y)^j,$$

with coefficients $\eta_j = \eta_j(x)$ determined by the value of x.

These conditions are sufficient for deriving the asymptotic results given below. Note, however, that for the derivation of the minimax lower bounds, for estimating the unknown dependence structure after subtracting a nonparametric trend estimate or for the development of data-driven algorithms, stronger conditions are required.

Assumption A1 defines the linear dependence structure, including short memory (with $d = 0$), long memory ($d > 0$) and antipersistence ($d < 0$). If ε_i are i.i.d., then e_i is a linear fractional process. However, linearity is not required. It is sufficient that the process e_i is a martingale difference. This is particularly useful when one would like to include short-range volatility dependence. For instance, Beran and Feng (2001a) consider the case where e_i is a FARIMA–GARCH with GARCH-innovations ε_i. In other words,

$$\begin{aligned} e_i &= (1-B)^{-d}\varphi^{-1}(B)\psi(B)\varepsilon_i, \\ \varepsilon_i &= \sqrt{v_i}\xi_i, \\ v_i &= \alpha_0 + \sum_{j=1}^{r} \alpha_j \varepsilon_{i-j}^2 + \sum_{j=1}^{s} \beta_j v_{i-j} \end{aligned}$$

where $A(B) = (1-B)^{-d}\varphi^{-1}(B)\psi(B)$ is the usual FARIMA(p,d,q) operator. If only the asymptotic variance of $\hat{m}^{(j)}$ is of interest, then weaker conditions than the martingale assumptions are sufficient. This assumption is useful when it comes to deriving the asymptotic distribution of $\hat{m}^{(j)}$. Assumption A2 is a regularity condition on the smoothness of m which, together with A3, is required for the derivation of the order of magnitude of the bias of $\hat{m}^{(j)}$. If only consistency is required, then it is sufficient that $m^{(j)}$ is continuous in a neighbourhood of x. As discussed previously, the first condition in A3 is needed so that the bias converges to zero. The second condition is needed for the variance to tend to zero. More specifically, $(nb)^{1-2d}b^{2j} \to \infty$ implies $nb^{j+1} \to \infty$ for all $d \in (-0.5, 0.5)$. This ensures that $w_{j;b,n}(t;i) \to 0$ (see (7.107)). Condition A4 is needed for the case of antipersistence (see the result below). For local polynomial estimation A4 can be achieved, for instance, by using a second-order weight function $K(u)$ in (7.105) that is μ-smooth and of the form

$$K(u) = C_\mu \left(1-u^2\right)^\mu 1\{-1 \le u \le 1\}$$

for some $\mu \in \mathbb{N}$. For kernel estimation a polynomial kernel can be chosen directly by taking into account (7.117).

For any point $t \in [0,1]$, the asymptotic mean squared error can be obtained by detailed arguments following along the line of the heuristic ideas outlined so far.

As before, for any interior point $t \in (0, 1)$ we write $c = 1$, and for boundary points $t = cb$ or $t = 1 - cb$ with $0 \le c < 1$. The corresponding support of $K_{(j,p+1,c)}$ is denoted by $\mathscr{S} = [-a_1, a_2]$ with $a_1 = c$ and $a_2 = 1$ for a left, and $a_1 = 1$ and $a_2 = c$ for a right boundary kernel. In the interior, we have $a_1 = a_2 = 1$.

Theorem 7.22 *Assume Conditions* A1–A4. *We define* $a_1 = b_1 = 1$ *for interior points* $t \in [b, 1 - b]$, $a_1 = c$, $a_2 = 1$ *for left boundary points* $t = cb \in [0, b)$ *and* $a_1 = 1$, $a_2 = c$ *for right boundary points* $t = 1 - cb \in (1 - b, 1]$. *Then for* $d \in (-0.5, 0.5)$ *and any* $t \in [0, 1]$ *we have*

(i) *Bias*:

$$E\big[\hat{m}^{(j)}(t) - m^{(j)}(t)\big] = b^{k-j} \frac{m^{(k)}(t)\beta_{(j,k,c)}}{k!}\big[1 + o(1)\big], \tag{7.118}$$

where $\beta_{(j,k,c)} = \int_{-a_1}^{a_2} u^k K_{(j,k,c)}(u)\, du$,

(ii) *Variance*:

$$\operatorname{var}\big(\hat{m}^{(j)}(t)\big) = (nb)^{2d-1} b^{-2j} V_{(j,k,c)}(d)\big[1 + o(1)\big], \tag{7.119}$$

where for $d = 0$ *we have*

$$V_{(j,k,c)}(0) = 2\pi c_f \int_{-a_1}^{a_2} K^2_{(j,k,c)}(x)\, dx, \tag{7.120}$$

for $d > 0$,

$$V_{(j,k,c)}(d) = 2c_f \Gamma(1 - 2d) \sin \pi d \times \int_{-a_1}^{a_2} \int_{-a_1}^{a_2} K_{(j,k,c)}(x) K_{(j,k,c)}(y) |x - y|^{(2d-1)}\, dx\, dy \tag{7.121}$$

and for $d < 0$,

$$V_{(j,k,c)}(d) = 2c_f \Gamma(1 - 2d) \sin(\pi d) I(j, k, c; d) \tag{7.122}$$

with

$$I(j, k, c; d) = \int_{-a_1}^{a_2} K_{(j,k,c)}(x) M(x)\, dx, \tag{7.123}$$

$$M(x) = \int_{-a_1}^{a_2} \tilde{K}_{(j,k,c)}(x - y) |x - y|^{2d-1}\, dy - K_{(j,k,c)}(x) \int_{y<-a_1 y>a_2} |x - y|^{2d-1}\, dy. \tag{7.124}$$

We note that for $j = 0$, $k = 2$ the results in Theorem 7.22 agree with the expressions for bias and variance given above. Note also that being in the boundary region not only affects the bias but also the variance. The reason is that having less data in

the boundary regions necessarily increases the variance, though the order does not change. A detailed proof of Theorem 7.22 can be found in Beran and Feng (2002a). For earlier partial results in the short- and long-memory context, respectively, see, e.g. Altman (1990), Hart (1991) and Hall and Hart (1990a). Note that, for $d < 0$, the integral on the right-hand side of (7.121) is not well defined. However, the two integrals on the right-hand side of (7.122) based on the decomposition of the kernel function given in (7.123) and (7.124) are both well defined, since $-0.5 < d < 0$ and the powers of $(y - x)$ in $\tilde{K}_{(j,k,c)}(x - y)$ are at least of order one. This is why the decomposition was needed.

Example 7.31 Let e_t be generated by a FARIMA$(0, d, 0)$ process. Consider the kernel estimation of m with the rectangular kernel for interior points and the corresponding boundary kernels for left and right boundary points. Thus, $j = 0$, and we choose $k = 2$. For interior points, we have

$$K_{(0,2,1)}(u) = \frac{1}{2} 1\{-1 \le u \le 1\}$$

and, for instance, for left boundary points we have the kernel

$$K_{(0,2,c)}(u) = \frac{1}{c+1}\left\{1 + 3\left(\frac{1-c}{1+c}\right)^2 + 6\frac{1-c}{(1+c)^2}u\right\}$$

with $0 \le c < 1$ (see Table 7.3). Note in particular that $K_{(j,k,c)}$ converges to the rectangular kernel as $c \to 1$. For $\beta_{(j,p+1,c)}$ we have

$$\beta_{(0,2,1)} = \int_{-1}^{1} u^2 K_{(0,2,1)}(u)\, du = \frac{1}{2}\int_{-1}^{1} u^2\, du = \frac{1}{3}$$

and, with $c < 1$,

$$\begin{aligned}
\beta_{(0,2,c)} &= \int_{-1}^{1} u^2 K_{(0,2,c)}(u)\, du \\
&= \frac{1}{c+1}\int_{-1}^{1} u^2 \left\{1 + 3\left(\frac{1-c}{1+c}\right)^2 + 6\frac{1-c}{(1+c)^2}u\right\} du \\
&= \frac{1}{c+1}\left\{\frac{1}{3} + \left(\frac{1-c}{1+c}\right)^2 + 3\frac{1-c}{(1+c)^2}\right\}.
\end{aligned}$$

Figure 7.9 shows how $\beta_{(0,2,c)}$ increases as c decreases to zero. The smallest value for $c = 0$ is equal to $\beta_{(0,2,0)} = \frac{13}{3}$. Thus, the bias of $\hat{m}(0)$ is more than four times larger than for interior points. More specifically, we have for $t \in [b, 1 - b]$,

$$\text{Bias} = E\big[\hat{m}(t)\big] - m(t) = b^2 \frac{1}{6} m^{(2)}(t) + o\big(b^2\big)$$

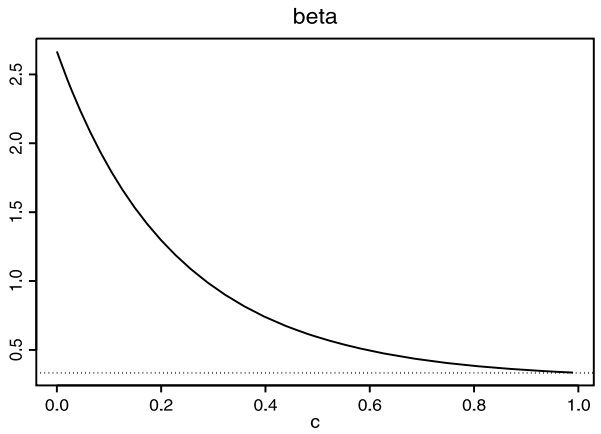

Fig. 7.9 Plot of $\beta_{(0,2,c)}$ for $0 \le c < 1$ and $K_{(0,2,c)}$ derived from the rectangular kernel

and for $t = 0$,

$$\text{Bias} = E\big[\hat{m}(0)\big] - m(0) = b^2 \frac{13}{8} m^{(2)}(0) + o\big(b^2\big).$$

The variance can be evaluated from (7.119) by inserting $K_{(0,2,c)}$ in the corresponding integral. Figure 7.10 shows $V_{(j,k,c)}(d)$ as a function of $c \in [0, 1]$ for different values of d. As for the bias, the variance increases the closer we are to the boundary. However, in contrast to the bias, the effect is stronger for higher values of d. This means that the increase in the variance near the border is much more dramatic in the presence of strong long memory so that, for instance, confidence intervals for $m(t)$ near the border can differ considerably from those at interior points. Note also that for $d < 0$, the function $\tilde{K}_{(j,p+1,c)} = \tilde{K}_{(0,2,1)}$ is given as follows. Let $y = (y - x) + x$. Then for interior points ($c = 1$) we have

$$K_{(0,2,1)}(y) = \frac{1}{2} 1\{-1 \le y \le 1\} = K_{(0,2,1)}(x) + \tilde{K}_{(0,2,1)}(x - y)$$

with $\tilde{K}_{(0,2,1)}$ being an indicator function determined by the value of x by

$$\tilde{K}_{(0,2,1)}(u) = -\frac{1}{2}\big(1\{u < x - 1\} + 1\{u > 1\}\big).$$

For $0 \le c < 1$ and left boundary points, we have

$$\tilde{K}_{(0,2,c)}(u) = 1\{-1 \le x \le c\} 1\{x - c \le x - y \le x + 1\},$$

and for right boundary points,

$$\tilde{K}_{(0,2,c)}(u) = 1\{-c \le x \le 1\} 1\{x - 1 \le x - y \le x + c\}.$$

Again, the variance increases with decreasing c.

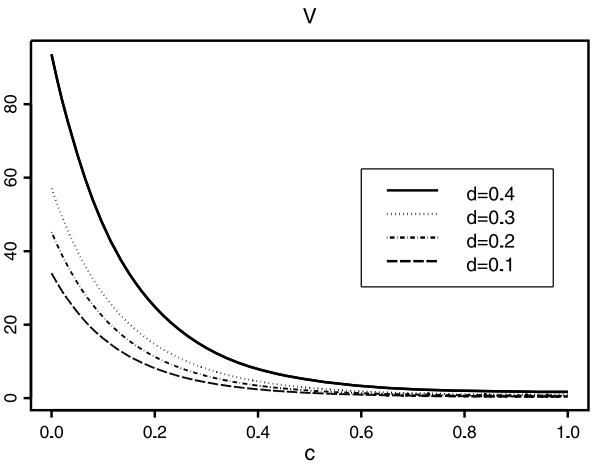

Fig. 7.10 $V_{(0,2,c)}(d)$ plotted as a function of $c \in [0, 1)$ for different values of $d \in (0, \frac{1}{2})$

Theorem 7.22 implies an asymptotic formula for the MSE at t of the form

$$MSE(t) = E\left[\left(\hat{m}^{(j)}(t) - m^{(j)}(t)\right)^2\right] \tag{7.125}$$

$$\sim b^{2(k-j)}\left(\frac{m^{(k)}(t)\beta_{(j,k,c)}}{k!}\right)^2 + (nb)^{2d-1}b^{-2j}V_{(j,k,c)}(d). \tag{7.126}$$

By minimizing this expression, we obtain the asymptotically optimal local bandwidth

$$b_{\text{opt}} = b_{\text{opt}}(t) = C_{\text{opt}}(t)n^{-\alpha_{\text{opt}}} \tag{7.127}$$

where

$$\alpha_{\text{opt}} = \frac{1-2d}{2k+1-2d}$$

and

$$C_{\text{opt}}(t) = \left\{\frac{2j+1-2d}{2(k-j)}\left(\frac{k!}{m^{(k)}(t)\beta_{(j,k,c)}}\right)^2 V_{(j,k,c)}(d)\right\}^{\frac{1}{2k+1-2d}}. \tag{7.128}$$

Here it was assumed tacitly that $m^{(k)}(x) \neq 0$. Note that a bandwidth of the optimal order $n^{-\alpha_{\text{opt}}}$ is such that the squared asymptotic bias and the asymptotic variance are of the same order of magnitude. Inserting $b_{\text{opt}}(x)$ in (7.125), we obtain an optimal MSE of the order

$$MSE_{\text{opt}} = O\left(n^{-r}\right), \tag{7.129}$$

with

$$r = 2(k-j)\alpha_{\text{opt}} = 2(k-j)\cdot\frac{1-2d}{2k+1-2d}. \tag{7.130}$$

Under the assumptions of Theorem 7.22, this rate turns out to be optimal among all possible nonparametric regression estimators (Feng and Beran 2012). Moreover, Beran and Feng (2007) show that there is no kernel (or weighting system) that would be optimal for all values of $d \in (0, \frac{1}{2})$. Thus, in contrast to the case where we restrict models to short-range autocorrelations, optimization with respect to the kernel is not meaningful because the value of d is not known a priori.

The standard choice of k is $k = j + 2$ which leads to

$$\begin{aligned}\alpha_{\text{opt}} = \alpha_{\text{opt}}(j,d) &= \frac{1-2d}{5+2j-2d} \\ &= \frac{1}{5+2j} - \frac{2d(4+2j)}{(5+2j-2d)(5+2j)} \\ &= \alpha_{\text{opt}}(j,0) - \frac{2d(4+2j)}{(5+2j-2d)(5+2j)}\end{aligned}$$

and

$$\begin{aligned}r_{\text{opt}} = r_{\text{opt}}(j,d) = 4\alpha_{\text{opt}}(j,d) &= \frac{4-8d}{5+2j-2d} \\ &= \frac{4}{5+2j} - \frac{8d(4+2j)}{(5+2j-2d)(5+2j)} \\ &= r_{\text{opt}}(j,0) - \Delta r_{\text{opt}}(j,d).\end{aligned}$$

Thus, compared to the case of short memory with $d = 0$, the optimal order of the MSE is increased for $d > 0$ and decreased for $d < 0$ by the factor $n^{\Delta r_{\text{opt}}(j,d)}$. In Fig. 7.11, $\Delta r_{\text{opt}}(j,d)$ is plotted against $j = 0, 1, 2, 3$ and 4 for $n = 1000$, and d ranging between -0.4 and 0.4. The effect is quite dramatic for low values of j and strong long memory. The largest increase within the range considered here is obtained for $j = 0$ and $d = 0.4$ with $\Delta r_{\text{opt}}(0, 0.4) \approx 0.61$. Note that, for instance, for $n = 1000$ this amounts to an increase by the factor $n^{\Delta r_{\text{opt}}(j,d)} \approx 67$.

If one prefers to use a global bandwidth instead of a local one, then one can minimize an integrated MSE (IMSE). If we use local polynomial estimation or a kernel estimator with boundary kernels, then the bias for boundary points is of the same order as in the interior. The contribution of boundary points to the IMSE is therefore asymptotically negligible because the boundary intervals shrink to length zero. (It should be emphasized, however, that this conclusion is wrong when one does *not* use boundary kernels—see the previous discussion.) The asymptotic expression therefore simplifies to

$$IMSE = \int_0^1 MSE(t)\,dt \tag{7.131}$$

$$\sim b^{2(k-j)} \left(\frac{\beta_{(j,k,1)}}{k!} \right)^2 I_k + (nb)^{2d-1} b^{-2j} V_{(j,k,1)}(d) \tag{7.132}$$

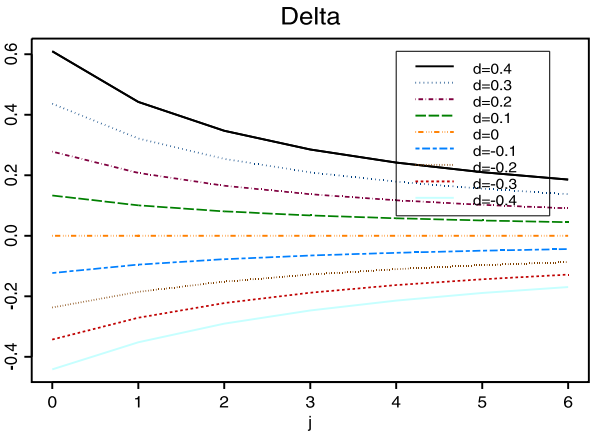

Fig. 7.11 Change Δr of the optimal exponent r_{opt} in $MSE_{\text{opt}}(\hat{m}^{(j)}) = O(n^{-r_{\text{opt}}})$ compared to the case of short memory, as a function of j, plotted for different values of d

where

$$I_k = \int_0^1 \left(m^{(k)}(t)\right)^2 dt. \tag{7.133}$$

The asymptotically optimal global bandwidth is then given by

$$b_{\text{opt}} = C_{\text{opt}} n^{-\alpha_{\text{opt}}} \tag{7.134}$$

where α_{opt} is as before and

$$C_{\text{opt}} = \left\{ \frac{2j+1-2d}{2(k-j)} \left(\frac{k!}{\beta_{(j,k,1)}} \right)^2 \frac{V_{(j,k,1)}(d)}{I_k} \right\}^{\frac{1}{2k+1-2d}}. \tag{7.135}$$

Example 7.32 Let e_t be generated by a FARIMA$(0, d, 0)$ process with $0 < d < \frac{1}{2}$. Consider kernel estimation of m with the rectangular kernel for interior points and the corresponding boundary kernels for left and right boundary points. Then $j = 0$, $k = 2$,

$$\begin{aligned} K_{(0,2,1)}(u) &= \frac{1}{2} 1\{-1 \le u \le 1\}, \\ V_{(0,k,1)}(d) &= \frac{\Gamma(1-2d)\sin \pi d}{4\pi} \int_{-1}^{1} \int_{-1}^{1} |x-y|^{(2d-1)}\, dx\, dy \\ &= \frac{\Gamma(1-2d)\sin \pi d}{4\pi} \frac{2^{2d+1}}{d(2d+1)}, \\ \beta_{(0,2,1)} &= \frac{1}{2} \int_{-1}^{1} u^2\, du = \frac{1}{3} \end{aligned}$$

and (with the notation from (7.133))

$$IMSE \sim b^{2(k-j)}\left(\frac{\beta_{(j,k,1)}}{k!}\right)^2 I_k + (nb)^{2d-1} b^{-2j} V_{(j,k,1)}(d) \tag{7.136}$$

$$= b^4\left(\frac{1}{6}\right)^2 I_2 + (nb)^{2d-1}\frac{\Gamma(1-2d)\sin\pi d}{\pi}\frac{2^{2d-1}}{d(2d+1)}. \tag{7.137}$$

This is the same expression we obtained in (7.95).

7.4.2.2 Asymptotic Distribution

As mentioned previously in (7.108), local polynomial and kernel estimators can be written as sums of triangular arrays. Investigating the asymptotic behaviour of $\hat{m}^{(j)}(t)$ amounts to studying a sequence of sums

$$S_n = \sum_{i=1}^{n} \zeta_{i,n} \quad (n \in \mathbb{N}) \tag{7.138}$$

with

$$\zeta_{i,n} = w_{j;b,n}(i)e_i$$

$(1 \le i \le n;\ n \in \mathbb{N})$. The asymptotic distribution of $\hat{m}^{(j)}(t)$ therefore follows as a corollary of a suitable limit theorem for triangular arrays. For instance, Beran and Feng (2002a) consider the case of a second order stationary residual process

$$e_i = \sum_{s=0}^{\infty} a_s \varepsilon_{i-s}$$

with square integrable martingale differences ε_i and

$$f_e(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|A\left(e^{-i\lambda}\right)\right|^2 \sim c_f|\lambda|^{-2d} \quad (\lambda \to 0)$$

for some $d \in (-0.5, 0.5)$. This includes not only second-order stationary linear processes but also nonlinear fractional processes such as FARIMA–GARCH models. Under relatively mild conditions on the marginal distribution of e_i, one has a limit theorem

$$\sigma_n^{-1}\sum_{i=1}^{n} e_i \underset{d}{\to} Z \sim N(0,1),$$

where

$$\sigma_n^2 = \mathrm{var}\left(\sum_{i=1}^{n} e_i\right).$$

This can be extended to sums of arrays $\zeta_{i,n} = w_{i,n}e_i$ as follows.

Theorem 7.23 *Under the conditions stated above* (*see Beran and Feng* 2002a), *the following holds. Let* $(w_{i,n})$ *be a triangular array of weights such that* $\sigma_{n,w}^2 := \operatorname{var}(\sum_{i=1}^n w_{i,n}e_i) > 0$ *for all* n. *If*

$$\max_{1\le i\le n} |w_{i,n}|/\sigma_{n,w} \to 0 \tag{7.139}$$

and

$$\sup_i \left| \sum_{j=1}^n w_{j,n}a_{i-j} \right| \Big/ \sigma_{n,w} \to 0, \tag{7.140}$$

then

$$\left[\sum_{i=1}^n w_{i,n}e_i \right] \Big/ \sigma_{n,w} \underset{d}{\to} Z \sim N(0,1). \tag{7.141}$$

The detailed proof of Theorem 7.23 can be found in Beran and Feng (2002a). Condition (7.139) means that the weights $w_{i,n}$ are uniformly negligible. Note that, if $\max|w_{i,n}| = O(1)$, then $\sigma_{n,w}^2 \to \infty$ as $n \to \infty$. Condition (7.140) on the weighted sum $\sum w_j a_{i-j}$ is often related to (7.139). Theorem 4.2 in Müller (1988) on the asymptotic normality of a weighted sum of i.i.d. random variables is a special case of Theorem 7.23. Related results on the asymptotic normality of weighted sums can be found, for instance, in Fuller (1996, Theorem 6.3.4).

Asymptotic normality for local polynomial and kernel estimators is now a corollary of (7.141). As before, we distinguish between interior points $t \in (0,1)$ with $c = 1$, and boundary points $t = ch$ or $t = 1 - ch$ with $c \in [0,1)$.

Corollary 7.1 *Let* $\hat{m}^{(j)}(t)$ $(t \in [0,1])$ *be a local polynomial estimator or a kernel estimator with boundary kernels. Suppose that the conditions of Theorem* 7.22 *and the conditions on* e_i *in Theorem* 7.23 *hold. Assume furthermore that the bandwidth is of the optimal order, i.e.*

$$b = c_b \cdot n^{-\alpha_{\text{opt}}}$$

(*for some* $0 < c_b < \infty$), *and let*

$$\mu_{(j,k,c)} = c_b^{\frac{1}{2}-d+k} \frac{m^{(k)}(t)\beta_{(j,k,c)}}{k!}. \tag{7.142}$$

Then, for any $d \in (-\frac{1}{2}, \frac{1}{2})$, *we have*

$$(nb)^{\frac{1}{2}-d} b^j \left[\hat{m}^{(j)}(t) - \hat{m}^{(j)}(t) \right] \underset{d}{\to} Z_{(j,k,c)} \sim N\left(\mu_{(j,k,c)}, V_{(j,k,c)}(d)\right), \tag{7.143}$$

where $V_{(j,k,c)}(d)$ *and* $\beta_{(j,k,c)}$ *are the constants defined in Theorem* 7.22.

Note that, as usual in nonparametric regression, using a bandwidth with the optimal rate leads to a non-negligible asymptotic bias after standardization. For statistical inference about $m^{(j)}(t)$, this means that one needs to include an estimate of this bias. The other option is to use a slightly faster rate for the bandwidth so that the bias disappears asymptotically because it is dominated by the variance.

A further result that is useful for simultaneous confidence bands for the function $m(t)$ has been shown in Csörgő and Mielniczuk (1995a) for the case of long memory. Assuming a spectral density $f_e(\lambda) \sim c_f|\lambda|^{-2d}$ or autocovariances $\gamma_e(k) \sim c_\gamma|k|^{2d-1}$ with $0 < d < \frac{1}{2}$, and a second-order kernel estimator $\hat{m}$, one can show that for interior points $0 < t_1 < \cdots < t_l < 1$ one has asymptotic independence. In other words,

$$(nb)^{1/2-d} V_{(0,2,1)}^{-\frac{1}{2}} \big(\hat{m}(t_1) - m(t_1), \ldots, \hat{m}(t_l) - m(t_l)\big) \xrightarrow[d]{} (Z_1, \ldots, Z_l) \qquad (7.144)$$

where Z_i are independent standard normal random variables and $V_{(0,2,1)}$ is defined in (7.121). The result is, of course, only valid, if the standardized sums of e_i are also asymptotically normal. Specifically, Csörgő and Mielniczuk (1995a) consider Gaussian residuals as well as Gaussian subordination. In the latter case, the Hermite rank of the transformation has to be one (see Sect. 4.2.3). The reason why we have asymptotic independence can be seen quite easily. For $t \neq s$, we have

$$\begin{aligned} cov\big(\hat{m}(t), \hat{m}(s)\big) &\sim c_\gamma n^{2d-1} b^{-2} \int_0^1 \int_0^1 K\left(\frac{x-t}{b}\right) K\left(\frac{y-s}{b}\right) |x-y|^{2d-1}\, dx\, dy \\ &\sim c_\gamma n^{2d-1} \int_{-1}^1 \int_{-1}^1 K(u) K(v) \big|t - s - b(u-v)\big|^{2d_\varepsilon - 1}\, du\, dv. \end{aligned}$$

Up to this point, the evaluation is almost the same as for the variance of $\hat{m}(t)$. However, the crucial difference is that with $b \to 0$ the function $g(u, v) = |t - s - b(u - v)|$ converges to $|x - y|$ uniformly in $(u, v) \in [-1, 1]^2$. Therefore,

$$cov\big(\hat{m}(t), \hat{m}(s)\big) \sim c_\gamma n^{2d-1} |t-s|^{2d-1}.$$

However, our standardization in (7.144) is $(nb)^{1/2-d}$ so that

$$(nb)^{1-2d} cov\big(\hat{m}(t), \hat{m}(s)\big) \sim c_\gamma b^{1-2d} |t-s|^{2d-1} \to 0.$$

Note finally that all asymptotic considerations above were made under the assumption that $f_e(\lambda) \sim c_f|\lambda|^{-2d}$ and $\gamma_e(k) \sim c_\gamma|k|^{2d-1}$. More generally, the same results follow when the constants c_f and c_γ are replaced by slowly varying functions. Also extensions to Gaussian subordination with non-Gaussian limits can be considered (see Csörgő and Mielniczuk 1995a). Further results can be found, for instance, in Robinson (1997).

7.4.3 Fixed-Design Regression with Heteroskedastic LRD Errors

Suppose we have a slightly more general model with a deterministic equidistant design, namely with a residual process that has a time-varying variance. More specifically, we assume

$$Y_i = m(t_i) + \sigma(t_i)e_i \tag{7.145}$$

with $\sigma(\cdot)$ continuous and e_i as before. Suppose moreover that, apart from possible heteroskedasticity modelled by σ, the assumptions of Theorem 7.22 hold. Since the bias is not influenced by the autocovariance structure, the asymptotic expression for the bias remains the same. For the variance, the assumption that σ is continuous implies that at point t only $\sigma^2(t)$ comes in asymptotically. Thus, in the formulas for the asymptotic variance given in Theorem 7.22, we just have to multiply $V_{(j,k,c)}$ by $\sigma^2(t)$. Formula (7.125) changes to

$$MSE(t) \sim b^{2(k-j)} \left(\frac{m^{(k)}(t)\beta_{(j,k,c)}}{k!} \right)^2 + (nb)^{2d-1} b^{-2j} \sigma^2(t) V_{(j,k,c)}(d). \tag{7.146}$$

All other formulas for b_{opt} and MSE_{opt}, Theorem 7.22, Corollary 7.1, and (7.144) have to be modified accordingly.

7.4.4 Bandwidth Choice for Fixed Design Nonparametric Regression—Part I

Nonparametric regression works well only if an appropriate bandwidth is chosen. Unfortunately, asymptotic expressions for the MSE and IMSE all involve unknown parameters. If we allow d to vary, instead of being fixed at zero, the situation is even worse because a good estimate of d is essential, in particular if $d > 0$ (see, e.g. Figs. 7.6 and 7.7). It is therefore very important to design a reliable data-adaptive method for the case of fractional residuals with unknown correlation structure.

Bandwidth selection in nonparametric regression with uncorrelated errors is well studied. Numerous results on this topic may be found in the literature. Standard bandwidth selection rules include cross-validation (CV; Clark 1975; Bowman 1984), generalized cross-validation (GCV; Craven and Wahba 1979) and the so-called R-Criterion (Rice 1984). Also see Härdle et al. (1988), Marron (1989) and Jones et al. (1996) for related surveys on bandwidth selection rules in the closely related context of nonparametric density estimation. The main drawback of those bandwidth selection rules is that their rate of convergence is just $O(n^{-1/10})$. Other, more recent, bandwidth selection rules in nonparametric regression have higher rates of convergence. These include, for instance, the iterative plug-in (IPI, Gasser et al. 1991), the direct plug-in (DPI, Ruppert et al. 1995) and the double smoothing approach (DS, Müller 1985; Härdle et al. 1992; Heiler and Feng 1998). Bandwidth selection in nonparametric regression with *dependent* errors is more difficult

because the bandwidth selection and the estimation of the dependence structure depend on each other. This problem is discussed, for instance, in Altman (1990), Hart (1991), Herrmann et al. (1992), Hall et al. (1995a), Ray and Tsay (1997), Opsomer et al. (2001) and Beran and Feng (2002a, 2002b, 2002c). The two main approaches discussed in the long-memory context are bootstrap based cross-validation (Hall et al. 1995b), and the iterative plug-in method (Ray and Tsay 1997; Beran and Feng 2002a, 2002b, 2002c).

Although the case of a fractional residual process is very general, it does have a clear structure due to the asymptotic dominance of the parameters d and c_f. An iterative plug-in (IPI) algorithm is therefore a natural approach. The first IPI algorithm in the long-memory context was proposed by Ray and Tsay (1997).

Specifically, consider a second-order kernel estimator of m. Ray and Tsay (1997) propose the following iteration.

1. Estimate an "optimal" bandwidth $\hat{b}_{\text{opt}}$, assuming only short-range dependent errors, using a standard method such as the procedure in Herrmann et al. (1992).
2. Set $b_0 = \hat{b}_{\text{opt}}$.
3. For $j \geq 1$ estimate $m(t)$ using b_{j-1} and let $\hat{e}_i = y_i - \hat{m}(t_i)$. Estimate d and c_f using the log-periodogram regression by Geweke and Porter-Hudak (or any other semiparametric method) applied to $\hat{e}_i$.
4. Let $b_{2,j} = b_{j-1} n^{(1-2\hat{d})/(2(5-2\hat{d}))}$, and estimate m'' and $I(m'') = \int (m''(t))^2 \, dt$ using a fourth-order kernel estimator for estimating the second derivative with the bandwidth $b_{2,j}$.
5. Improve b_{j-1} by setting

$$b_j = \hat{C}_{\text{opt}} n^{(2\hat{d}-1)/(5-2\hat{d})} \tag{7.147}$$

   where $\hat{C}_{\text{opt}}$ is obtained from the current estimates of d, c_f, and $I(m'')$.
6. Increase j by 1 and repeat Steps 3 to 5 until convergence is reached. Finally, at the end of the iteration set $\hat{b}_{\text{opt}} = b_j$.

This algorithm is based on the proposal of Herrmann et al. (1992). The formula $b_{2,j} = b_{j-1} n^{(1-2\hat{d})/(2(5-2\hat{d}))}$ in Step 4 is called an inflation method. An improved algorithm was proposed in Beran and Feng (2002a, 2002b, 2002c). This is discussed in more detail in Sect. 7.4.6.

7.4.5 The SEMIFAR Model

7.4.5.1 Introduction

As we have seen in this chapter, distinguishing between deterministic trend functions and random stationary fluctuations with long memory can be quite difficult. A further complication is that sometimes it may not even be clear whether the stochastic component of the observed series is stationary. For practical applications,

one would therefore like to have a data-driven methodology that is able to identify at least certain standard types of stochastic nonstationarities and distinguish them from stationary dependence (including short and long memory, and antipersistence) or deterministic trend functions. A semiparametric approach along this line, the so-called SEMIFAR (semiparametric autoregressive) models, has been developed in Beran (1999) and Beran and Feng (2001b, 2002a, 2002b). For applications, see, e.g. Beran and Ocker (2001), Beran et al. (2003), Beran (2007b) and Feng et al. (2007). An implementation is available in the S-Plus module $S+FinMetrics$ (see Zivot and Wang 2003).

The idea is to define a semiparametric model that incorporates a nonparametric trend function, parameters that determine whether the detrended series is integrated or stationary, and parameters determining the detailed dependence structure of the underlying stationary process. All parameters are estimated from the data, including an integer valued and a fractional differencing parameter. The SEMIFAR model, originally introduced in Beran (1999), extends the model in Beran (1995) by including a trend function.

7.4.5.2 Definition of the SEMIFAR Model

Assume that $m(t)$ ($t \in [0,1]$) is a trend function satisfying suitable smoothness conditions, let ε_i ($i \in \mathbb{N}$) be a sequence of i.i.d. zero mean random variables with finite variance $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_i)$, define $B^j m(t_i) = m(t_{i-j})$, where $t_i = i/n$ is rescaled time, and denote by $\varphi(z) = 1 - \sum_{j=1}^{p} \varphi_j z^j$ a polynomial with all roots outside the unit circle. A SEMIFAR model is defined as follows.

Definition 7.7 A process X_i is called a semiparametric fractional autoregressive (or SEMIFAR) model if there exist an integer $r \in \{0,1\}$ and a $d \in (-0.5, 0.5)$ such that

$$\varphi(B)(1-B)^d\big\{(1-B)^r X_i - m(t_i)\big\} = \varepsilon_i. \tag{7.148}$$

For $Y_i = (1-B)^r X_i$ we are back to the model with a nonparametric trend function and stationary errors generated by a FARIMA$(p,d,0)$ process, namely

$$Y_i = m(t_i) + e_i \quad (i = 1, 2, \ldots, n), \tag{7.149}$$

where $e_i = \varphi^{-1}(B)(1-B)^{-d}\varepsilon_i$. We will also use the notation

$$E_i = (1-B)^d e_i = \sum_{j=0}^{\infty} b_j e_{i-j} = \varphi^{-1}(B)\varepsilon_i \tag{7.150}$$

for the autoregressive residuals obtained after filtering out the fractional differencing component. Note, however, that we are assuming r to be unknown, so that taking the appropriate rth difference cannot be applied directly.

7.4.5.3 Fitting the SEMIFAR Model

Fitting a SEMIFAR models consists of two main parts: (a) nonparametric estimation of the trend function $m(t)$ and (b) estimation of the parameters σ_ε^2, r, d, p and $\varphi_1, \ldots, \varphi_p$. Since r is an integer and $d \in (-\frac{1}{2}, \frac{1}{2})$, r and d can be summarized by one parameter $d_{\text{total}} = d + r$ only. The two differencing parameters can be obtained from d_{total} by $r = [d_{\text{total}} + 0.5]$ and $d = d_{\text{total}} - r$, where $[\cdot]$ denotes the integer part. Parts (a) and (b) of SEMIFAR fitting depend on each other because for (b) we need to have subtracted a good estimate of the trend function, whereas for (a) one would need to know r in the first place, and also have some knowledge of d, σ_ε^2 and $\varphi_1, \ldots, \varphi_p$ (and the second derivative of m) to calculate the optimal bandwidth. The method considered in Beran (1999) and Beran and Feng (2002a, 2002b) is an iterative plug-in algorithm. This is related (but not identical) to similar methods in the short-memory context (Gasser et al. 1991; Ruppert et al. 1995) and to the method by Ray and Tsay (1997) introduced in Sect. 7.4.4. Note that, as discussed in Sect. 7.4.4, other methods like cross-validation seem less appropriate. Even in the i.i.d. context, it is well known that cross-validation and related methods (Clark 1975; Bowman 1984; Craven and Wahba 1979) lead to highly volatile bandwidths that converge to the optimal one at the slow rate of $O(n^{-\frac{1}{10}})$. Methods based on the plug-in principle are known to provide more reliable bandwidth estimates with a smaller variability and much faster convergence to the optimal bandwidth (Gasser et al. 1991; Ruppert et al. 1995; Müller 1985; Härdle et al. 1992; Heiler and Feng 1998). In the context of long memory, the situation is even worse since the estimate of the IMSE obtained by cross-validation converges to the actual IMSE only under very restrictive conditions. In contrast, the plug-in method (for fixed design) considered here can be shown to provide reasonable reliable estimates of the optimal bandwidths (see results below).

The key ingredient of the plug-in method is the possibility of estimating the unknown parameter vector consistently even though the trend estimate $\hat{m}(t)$ may not be optimal. More specifically, let $\vartheta^0 = (\sigma_{\varepsilon,0}^2, \theta^0) = (\sigma_{\varepsilon,0}^2, d_{\text{total}}^0, \varphi_1^0, \ldots, \varphi_{p^0}^0)$ be the true parameter vector defining the (possibly integrated) fractional ARIMA component. Suppose that $\hat{m}(x)$ is a kth order kernel regression estimator with a bandwidth $b = O(n^{-\alpha})$ such that $0 < \alpha < 1/2$. Then it can be shown that, under some regularity conditions and the assumption $k\alpha + d^0 > 0$ (which always holds for $d^0 > 0$), the parameter θ^0 (including the integer differencing parameter r^0) can be estimated consistently. The same is true when the autoregressive order p^0 is chosen by the BIC (Beran et al. 1998) as discussed in Sect. 5.5.6 (provided that p^0 does not exceed the maximal autoregressive order $p_{\max}$ used in the selection). Moreover, if $k\alpha + d^0 > \frac{1}{4}$, then the approximate MLE defined in Beran (1995) yields a $\sqrt{n}$-consistent estimator of θ^0 (for more details, see Beran and Feng 2002a and Feng 2004a, 2004b). Note that this is a specific condition for avoiding too large bandwidths.

7.4.6 Bandwidth Choice for Fixed Design Nonparametric Regression—Part II: Data-Driven SEMIFAR Algorithms

In the following, we present two data-driven algorithms within the SEMIFAR framework. The first algorithm (Algorithm A, AlgA) relies on a full search with respect to d, and was originally proposed in Beran (1999) (also see Beran and Ocker 2001). The second algorithm (Algorithm B, AlgB) was proposed in Beran and Feng (2002b) and runs much faster than Algorithm A because a full search is avoided. As explained below, both methods are superior to the plug-in procedure proposed by Ray and Tsay (1997) in different ways. To simplify the presentation, only local linear estimates of the trend function m will be considered here, and m'' (needed in the constant of the bias) will be calculated using a local cubic or a fourth-order kernel estimator.

Algorithm A

Step 1: Let $p_{\max}$ be the maximal order of $\varphi(B)$ that will be tried, and define a sufficiently fine grid $G \in (-0.5, 1.5) \setminus \{0.5\}$. First, carry out Steps 2 through 4 for $p = p_{\max}$ in order to select the integer differencing order r.

Step 2: For each $d_{\text{total}} \in G$, set $r = [d_{\text{total}} + 0.5]$, $d = d_{\text{total}} - r$, and $Y_i(r) = (1-B)^r X_i$, and carry out Step 3.

Step 3: Carry out the following iteration:

Step 3a: Let $b_0 = \Delta_0 \min(n^{(2d-1)/(5-2d)}, 0.5)$ (for some fixed $\Delta_0 > 0$) and set $j = 1$.

Step 3b: Calculate $\hat{m}(t_i; r)$ using the bandwidth b_{j-1}. Set $\hat{e}_i(r) = Y_i(r) - \hat{m}(t_i; r)$.

Step 3c: Set $\hat{E}_{i,d_{\text{total}}} = \sum_{j=0}^{i-1} b_j(d)\hat{e}_{i-j}$ $(\approx (1-B)^d \hat{e}_i)$, where $b_j = (-1)^j \binom{d}{j}$.

Step 3d: Estimate the autoregressive parameters $\varphi_1, \ldots, \varphi_p$, from $\hat{E}_{i,d_{\text{total}}}$ and obtain the estimates $\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_\varepsilon^2(d_{\text{total}}; j)$ and $\hat{c}_f = \hat{c}_f(j)$. Estimation of the parameters can be done, for instance, by using the S-PLUS function *ar.burg* or *arima.mle* or an analogous R-function for autoregressive MLE. If $p = 0$, set $\hat{\sigma}_\varepsilon^2$ equal to $n^{-1} \sum \hat{E}_{i,d_{\text{total}}}^2$ and $\hat{c}_f$ equal to $\hat{\sigma}_\varepsilon^2/(2\pi)$.

Step 3e: Set $b_{2,j} = (b_{j-1})^\alpha$ with $\alpha = \alpha_0 = (5-2d)/(9-2d)$, and improve b_{j-1} by defining

$$b_j = \left(\frac{1-2d}{I^2(K)} \frac{(1-2d)\hat{V}}{\hat{I}(m''(t; b_{2,j}))} \right)^{1/(5-2d)} \cdot n^{(2d-1)/(5-2d)} \quad (7.151)$$

where $I(K) = \int u^2 K(u)\, du$, $I(\hat{m}''(t; b_{2,j}))$ is an estimate of $I(m'') = \int [m''(t)]^2\, dt$ using bandwidth $b_{2,j}$ and $\hat{V}$ is an estimate of the constant in the asymptotic variance (see Theorem 7.22).

Step 3f: Increase j by one and repeat Steps 3b to 3e until convergence is reached or until a given number of iterations has been carried out. This yields, for each $d_{\text{total}} \in G$ separately, the ultimate value of $\hat{\sigma}_\varepsilon^2(d_{\text{total}})$, as a function of d_{total}.

Step 4: Define $\hat{d}_{\text{total}}$ to be the value of d_{total} for which $\hat{\sigma}_\varepsilon^2(d_{\text{total}})$ is minimal, and let $\hat{r} = [\hat{d}_{\text{total}} + 0.5]$.

Step 5: For each $p = 0, 1, \ldots, p_{\max}$, carry out Steps 2 through 4 for $l = \hat{r}$. Define $\hat{d}_{\text{total}}$ to be the value of d_{total} for which $\hat{\sigma}_\varepsilon^2(d_{\text{total}})$ is minimal. This, together with the corresponding estimates of the AR parameters, yields a value of an information criterion for the given order p, e.g. $\text{BIC}(p) = n \log \hat{\sigma}_\varepsilon^2(p) + p \log n$, as a function of p and the corresponding values of $\hat{\theta}$ and $\hat{m}$.

Step 6: Select the order p that minimizes the BIC(p). This yields the final estimates of θ^0 and m.

This algorithm differs from Ray and Tsay (1997) mainly in the inflation method and in the estimation of the integer differencing parameter r. The inflation method used here in Step 3e is $b_{2,j} = (b_{j-1})^\alpha$ with $\alpha = \alpha_0 = (5 - 2\hat{d})/(9 - 2\hat{d})$. This is also called an exponential inflation method (EIM). Ray and Tsay (1997) use instead a multiplicative inflation method (MIM) of the form $b_{2,j} = b_{j-1} n^\beta$ with $\beta = \beta_{\text{v}} = \frac{1}{2}(1 - 2\hat{d})/(5 - 2\hat{d})$. The constants α or β in the two inflation methods are called inflation factors. The asymptotic rate of convergence of $\hat{b}$ depends on the choice of the inflation factor only, not on the choice of the inflation method. However, an algorithm based on the EIM requires a smaller number of iterations to reach a consistent bandwidth estimate. Commonly used choices of the inflation factors are: (i) α_{v} or β_{v} such that the variance of $\hat{b}$ is minimized; (ii) α_{opt} or β_{opt} such that the MSE of $\hat{I}$ is minimized and the rate of convergence of $\hat{b}$ is optimized; or (iii) α_0 or β_0 such that the MSE of $\hat{m}''$ is minimized. Explicit formulae for these inflation factors may be found in Beran and Feng (2002b). The rate of convergence of $\hat{b}$ based on α_{v} or β_{v} is the worst of all three choices, namely $O(n^{(2d^0-1)/(5-2d^0)})$. The rate of convergence of AlgA—which is based on α_0—is of the order $O(n^{2(2d^0-1)/(9-2d^0)})$ which is slightly faster than for the algorithm in Ray and Tsay (1997). Another advantage of AlgA compared to Ray and Tsay (1997) is the choice of the initial bandwidth. Although it does not affect the rate of convergence of $\hat{b}$, the initial bandwidth in AlgA is already of the correct optimal order. This reduces the number of required iterations.

Algorithm B AlgA is straightforward and intuitive. However, the iterative procedure has to be carried out for each trial value $d \in G$. This makes the algorithm computationally slow. Beran and Feng (2002b) therefore proposed a much faster algorithm where all parameters, except for p and r, are estimated directly from the residuals by maximizing the likelihood function. In the practical implementation, the S-PLUS function *arima.fracdiff* or an analogous R-function can be used. The algorithm can essentially be described as follows:

Step 1: First, we obtain a bandwidth for estimating r^0:
- Step 1a: Set $r = 1$. Calculate $Y_i(r) = (1-B)^r X_i$, and estimate m from $Y_i(r)$ using the initial bandwidth $b_0 = n^{-1/3}$. Calculate the residuals.
- Step 1b: Set $p = p_{\max}$ and assume that the residual process follows a FARIMA$(p, d, 0)$ model. Calculate a second initial bandwidth b_1 following, e.g. AlgA or another simple bandwidth selection procedure, but with $\alpha = \hat{\alpha}_{\text{opt}} = (5 - 2\hat{d})/(7 - 2\hat{d})$.

Step 2: Estimate r^0:
- Step 2a: Carry out Steps 1a and 1b with the selected b_1 as new initial bandwidth for $r = 0$ and $r = 1$ separately.
- Step 2b: Select r following the BIC. Now we obtain an estimate $\hat{r}$ of r^0.
- Step 2c: Set $r = \hat{r}$.

Step 3: Further iterations: Carry out further iterations for each $p = 0, 1, \ldots, p_{\max}$ with $r = \hat{r}$ and a new starting bandwidth $b_2 := \frac{1}{3} n^{-1/3}$ (or $b_2 := n^{-5/7}$) until convergence is reached or a given number of iterations has been reached.

Step 4: Select the best AR order p following the BIC and take the parameter estimate corresponding to $\hat{p}$ as the final estimate.

In this algorithm, $r = 1$ is used at the first iteration as a starting value of r. The initial input of the S-PLUS function *arima.fracdiff* is therefore always stationary, no matter what the value of r^0 is. The purpose of this step is to obtain a starting bandwidth for estimating r. The estimated value of r^0 is then selected in the second iteration and is asymptotically consistent. The use of $p = p_{\max}$ avoids the selection of p in the first two steps. Afterwards, $\hat{r}^0$ is used as a known parameter. At the beginning, the starting bandwidth $b_0 = n^{-1/3}$ is used. Since $(2 \cdot (-0.5) - 1)/(5 - 2 \cdot (-0.5)) = -1/3$, this is the smallest possible order of optimal bandwidths for d in the range $(-0.5, 0.5)$. The order of magnitude of b_0 also ensures that, for any $r^0 \in \{0, 1\}$, the bandwidth selected at the end of Step 1 fulfills the basic assumptions on the bandwidth.

AlgB runs much quicker than AlgA. Furthermore, the rate of convergence of $\hat{b}$ is improved by choosing the inflation factor $\alpha_{\text{opt}} = (5 - 2\hat{d})/(7 - 2\hat{d})$. The resulting rate of convergence of $\hat{b}$ is now of the order $O_p(n^{2(2d^0-1)/(7-2d^0)})$, which is the highest known rate for an iterative plug-in bandwidth selector in the current context. More specifically, the following results can be shown (Beran and Feng 2002b).

Proposition 7.1 *Let X_i be a SEMIFAR process defined by (7.148). Suppose that $m(t) \in C^4[0, 1]$ and, as $n \to \infty$, $nb \to \infty$ and $b \to 0$. Denote by b_A the optimal asymptotic bandwidth obtained by minimizing the asymptotic formula for the IMSE and let b_M be the actually optimal bandwidth that minimizes the exact finite sample IMSE. Then*

$$\frac{b_A - b_M}{b_M} = O\left(b_M^2\right).$$

For the data driven bandwidths obtained by AlgA and AlgB, respectively, the following asymptotic formulas hold (Beran and Feng 2002b):

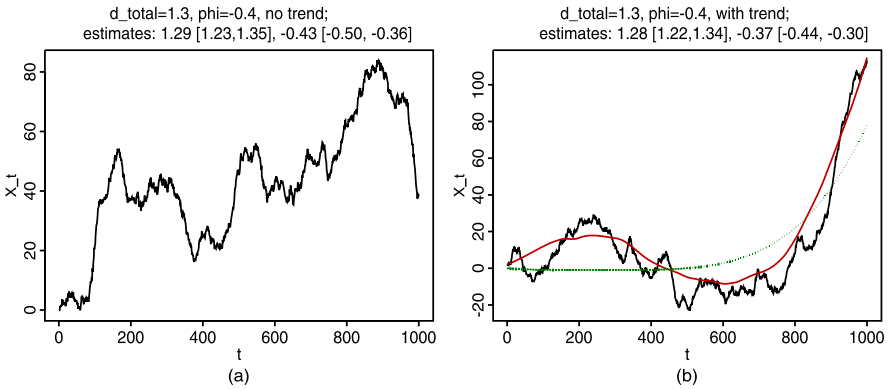

Fig. 7.12 (**a**) Simulated FARIMA$(p^0, d^0, 0)$ series with $p^0 = 1$, $d^0_{\text{total}} = 1.3$ ($r^0 = 1$, $d = 0.3$) and $\varphi^0_1 = -0.4$. This is the same as a SEMIFAR model with the same parameters and $m(t) \equiv 0$. (**b**) SEMIFAR process with the same parameters as in (**a**), but including a non-constant trend function $m(t)$. The estimated trend (*full line*) is also plotted together with the true (integrated) trend function (*dotted line*)

Theorem 7.24 *Let X_i be a SEMIFAR process with autoregressive order p_0, fractional differencing parameter d^0, and integer differencing parameter $r^0 \in \{0, 1\}$. Suppose that $m(t) \in C^4[0, 1]$, and denote by $\hat{b}_{\text{AlgA}}$ and $\hat{b}_{\text{AlgB}}$ the data driven bandwidths obtained by Algorithms A and B, respectively, with maximal AR-order $p_{\max} \geq p_0$. Then*

$$\hat{b}_{\text{AlgA}} = b_M \left\{1 + O_p\left(n^{(4d^0-2)/(9-2d^0)}\right)\right\},$$
$$\hat{b}_{\text{AlgB}} = b_M \left\{1 + O_p\left(n^{(4d^0-2)/(7-2d^0)}\right)\right\}.$$

For details, see Beran and Feng (2002a, 2002b). The iterative plug-in algorithms can easily be adapted to select bandwidths for estimating derivatives $\hat{m}^{(j)}$ ($j > 0$). Similar asymptotic results can be obtained for $\hat{b}$ as in Theorem 7.24.

Example 7.33 Figure 7.12 shows two simulated SEMIFAR series. In Fig. 7.12(a), the sample path was simulated by an integrated FARIMA process without trend. More specifically, we have $n = 1000$ observations of a FARIMA$(p^0, d^0, 0)$ series with $p^0 = 1$, $d^0_{\text{total}} = 1.3$ ($r^0 = 1$, $d = 0.3$) and $\varphi^0_1 = -0.4$. This is the same as a SEMIFAR model with the same parameters and $m(t) \equiv 0$. The SEMIFAR fit using Algorithm B is $\hat{p} = 1$, $\hat{d}_{\text{total}} = 1.29$ (hence $\hat{r} = 1$, $\hat{d} = 0.29$) and $\hat{\varphi} = -0.43$ with 95 %-confidence intervals $[1.23, 1.35]$ and $[-0.50, -0.36]$, respectively. Also no significant trend was found. The series in (b) is a SEMIFAR process with the same parameters for the stochastic part, but including a trend function $m(t)$. The estimated parameters obtained by AlgB are $\hat{p} = 1$, $\hat{d}_{\text{total}} = 1.28$ and $\hat{\varphi} = -0.37$, with 95 %-confidence intervals $[1.22, 1.34]$ and $[-0.44, -0.30]$, respectively. The estimated trend function is significant (at the 5 %-level) and also plotted, together with true trend function. Note that $m(t)$ is the trend function of the differenced process.

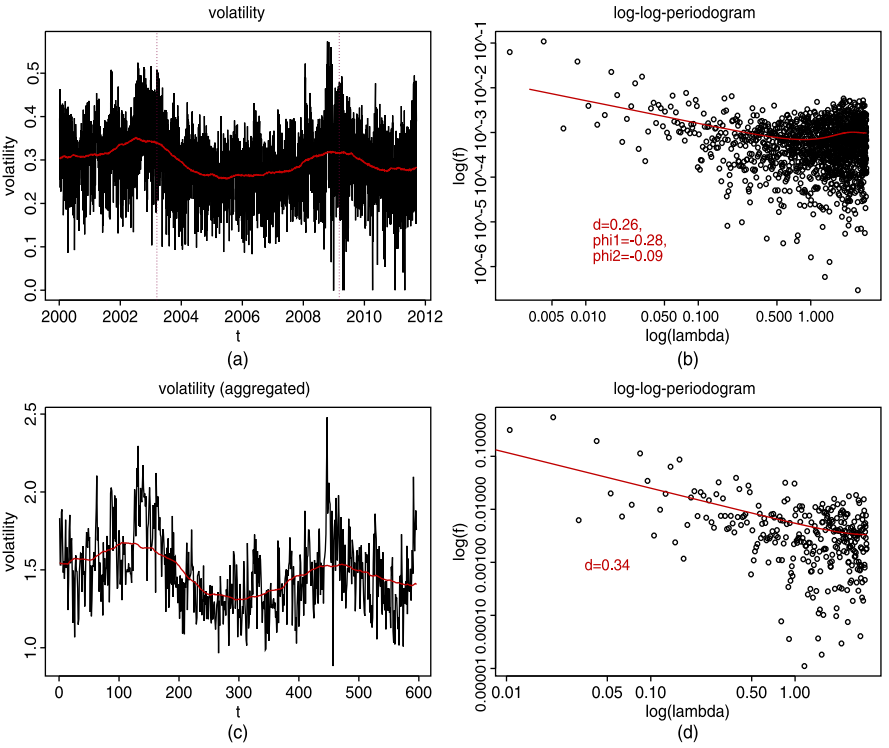

Fig. 7.13 Volatility series for the DAX between January 3, 2000 and September 12, 2011. (**a**) Shows daily data together with a nonparametric trend function fitted by Algorithm B. The corresponding log–log-plot of the periodogram together with the fitted spectral density is displayed in (**b**). (**c**) and (**d**) show analogous results, however, for weekly aggregates of the original data

Figure 7.12(b) shows, however, the integrated process. In contrast to m, the integrated trend function is not bounded. This explains why the estimated trend in the picture is relatively far from the true trend: errors $\hat{m}(t_i) - m(t_i)$ in the differenced domain have a long lasting effect in the integrated domain. This reflects the general uncertainty about trends when considering integrated processes.

Example 7.34 Figure 7.13(a) shows a volatility series of the DAX between January 3, 2000 and September 12, 2011 as defined in Sect. 1.2. A nonparametric trend function fitted by Algorithm B is also shown. The trend is significant at the 5 %-level. The parameter estimates are $\hat{p} = 2$, $\hat{d}_{\text{total}} = 0.26$ (i.e. $\hat{r} = 0$, $\hat{d} = 0.26$), $\hat{\varphi}_1 = -0.28$, $\hat{\varphi}_2 = -0.09$ with 95 %-confidence intervals $[0.21, 0.30]$, $[-0.33, -0.22]$ and $[-0.14, -0.04]$, respectively. The corresponding log–log-plot of the periodogram (of the detrended process) together with the fitted spectral density is displayed in (b). The results are confirmed when one looks at weekly aggregates. Figure 7.13(c) shows weakly averages of the original series displayed in (a). The SEMIFAR-fit again yields a significant trend which looks very much like

the function fitted in (a). As expected (see Sect. 2.2.1), due to temporal aggregation, the log–log-plot of the periodogram (of the detrended series) displayed in (d) is closer to a straight line. Applying Algorithm B indeed yields $\hat{p} = 0$ so that a pure FARIMA$(0, d, 0)$ model seems appropriate. (Note that the spectral density of a FARIMA$(0, d, 0)$ model is very close to the one of fractional Gaussian noise). The estimated value of d is 0.34 with a 95 %-confidence interval of $[0.27, 0.40]$.

Example 7.35 Figure 7.14(a) shows monthly precipitation anomalies for the Sahel region between January 1900 to December 2011 (data courtesy of Todd Mitchell, The Joint Institute for the Study of the Atmosphere and Ocean at the University of Washington, JISAO; the data source is the National Oceanic and Atmospheric Administration Global Historical Climatology Network (version 2), at the National Climatic Data Center of NOAA; http://www.ncdc.noaa.gov/temp-and-precip/ghcn-gridded-products.php). First, we try to fit a stationary FARIMA$(p, d, 0)$ process by selecting the order p using the BIC with $p \le p_{\max} = 16$. Figure 7.14(b) displays the periodogram of the data in log–log-coordinates, together with the fitted spectral density. The fit appears to be quite good, and mimics in particular the seasonal peaks. The estimated AR-order is $\hat{p} = 13$. The estimated long-memory parameter is equal to $\hat{d} = 0.35$ with a 95 %-confidence interval of $[0.14, 0.55]$. Note, however, that we used the restriction $d < 0.5$. Now the question is whether the apparent long memory may not rather be caused by a deterministic trend function or an integrated process (i.e. $d_{\text{total}} > 0.5$). We therefore fit a SEMIFAR process using AlgB and the BIC with $p \le p_{\max} = 16$. The fitted trend function indeed turns out to be significantly different from a constant (see (c), with horizontal lines demarking the critical limits). As suspected, the trend indicates a decline in precipitation starting around 1960. Subtracting the trend function seems to have removed long memory, since for the residuals we obtain a 95 %-confidence interval for d of $[-0.28, 0.18]$ (and $\hat{p} = 12$). The corresponding log–log-periodogram and fitted spectral density of the detrended data are shown in (d). Note also that the possibility of an integrated process ($d_{\text{total}} > 0.5$, $r = [d_{\text{total}} + 0.5]$) was excluded by the estimation procedure. A more detailed analysis can be obtained by separating the rainy season (June to October) from the rest of the year. Figure 7.14(e) shows the Sahel rainfall index with each year being represented by measurements form the rainy season only (i.e. we have June to October only for each year). The fitted trend function is very similar to the one in Fig. 7.14(c), and significant. Also as before, the estimated value of d is not longer significant, with a 95 %-confidence interval of $[-0.20, 0.13]$ (see (f) for the log–log-periodogram and spectral density). Note also that the selected autoregressive order of $\hat{p} = 3$ is much smaller than before because of the different (stochastic) periodicity. Finally, Fig. 7.14(g)–(h) show the results for the other months. This time the trend function is not quite significant at the 5 %-level. However, it is close to the critical limits and clearly monotonously decreasing. In contrast to the rainy season, $\hat{d} = 0.09$ with a 95 %-confidence interval of $[0.03, 0.15]$ indicates the possibility of slight long-range dependence in the residuals. Moreover, there does not appear to be any periodicity left (see Fig. 7.14(h)), and accordingly we have $\hat{p} = 0$. In summary, we may say that there is relatively clear evidence for a decline in precipitation in the

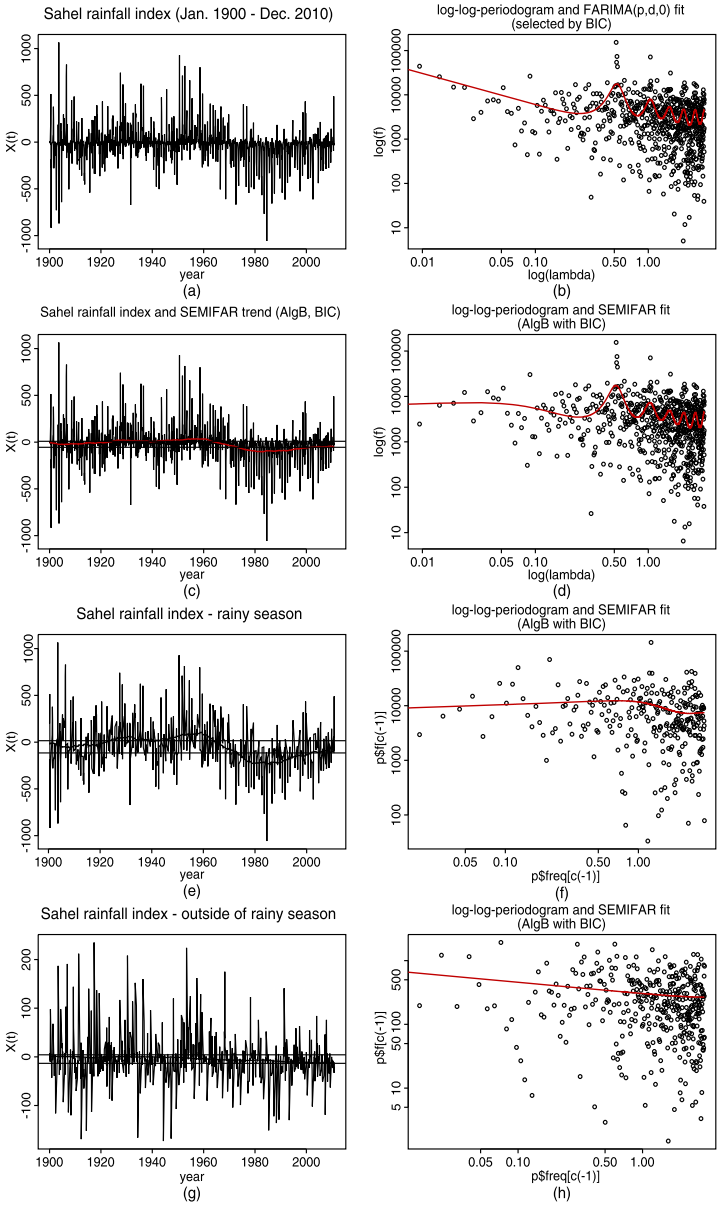

Fig. 7.14 Monthly precipitation anomalies for the Sahel region between January 1900 to December 2011 (data courtesy of Todd Mitchell, JISAO, University of Washington; http://www.ncdc.noaa.gov/temp-and-precip/ghcn-gridded-products.php): (**a**) original series; (**b**) log–log-periodogram and spectral density obtained by stationary fit; (**c**) data with fitted SEMIFAR trend (and critical limits); (**d**) log–log-periodogram and spectral density after SEMIFAR fit; (**e**) series with rainy seasons only; (**f**) log–log-periodogram and spectral density after SEMIFAR fit for data in (**e**); (**g**) series excluding rainy seasons; (**h**) log–log-periodogram and spectral density after SEMIFAR fit for data in (**g**)

Sahel zone starting around 1960. The alternative models of an integrated process or of stationarity with long memory can probably be excluded.

7.4.7 Trend Estimation from Replicates

Suppose that we have N time series $Y_j(i)$ where $j = 1, 2, \ldots, N$ denotes a replicate, $i = 1, 2, \ldots, n$ denotes time and the problem is estimation of the common trend $m(\cdot)$ in the nonparametric regression model

$$y_j(i) = m(t_i) + e_j(i) \quad \left(t_i = \frac{i}{n}\right)$$

by smoothing the average series $\bar{y}(i) = N^{-1}\sum_{j=1}^{N} y_j(i)$. The function $m(t)$ $(t \in (0, 1))$ is assumed to be smooth whereas $e_j(i)$ are random error terms that are stationary zero mean processes within each replicate but independent between replicates. In other words, $cov(e_j(i), e_l(i + k))$ is zero if $j \neq l$ and equals $\gamma_j(k)$ otherwise, where γ_j is a covariance function.

Specifically, we make the following assumptions on the jth error series $e_j(i)$:

- (A1) Mean: $E[e_j(i)] = 0$;
- (A2) Spectral density: $\lim_{\lambda\to 0}[f_j(\lambda)/\{D_j|\lambda|^{-2d_j}\}] = 1$ where $D_j > 0$, $0 < d_j < 1/2$ and the convergence is uniform;
- (A3) Covariances: $cov(e_j(i), e_j(i + k)) = \gamma_j(k) \sim C_j|k|^{2d_j-1}$ as $|k| \to \infty$, $d_j \neq 0$, $C_j > 0$ where, $C_j = \sin(\pi d_j)\Gamma(1 - 2d_j)D_j/(1 + 2d_j)$.

Consider the Priestley–Chao estimate of $m(t)$,

$$\hat{m}(t) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right)\bar{y}(i),$$

where the kernel K is a symmetric probability density function on $(-1, 1)$ and b is a bandwidth such that

$$b \to 0 \quad \text{and} \quad nb^3 \to \infty \quad \text{as } n \to \infty.$$

The uniform kernel $K(u) = \frac{1}{2}1\{|u| \leq 1\}$ is an example of such a kernel which we use in this section, but the arguments also hold for other kernels.

Clearly, the precision of such an estimator will depend on n as well as on N. Two different cases are of interest: (i) *N is fixed and finite* and (ii) $N \to \infty$.

Case (i) N is fixed and finite. As we shall see, in this case the mean squared error of the estimated trend function will be dominated by the largest fractional differencing parameter.

Theorem 7.25 *Let N be fixed and finite. Then, as $n \to \infty$, the asymptotic expression of the bias of $\widehat{m}(t)$ for $t \in (0,1)$ is*

$$E\big[\widehat{m}(t)\big] - m(t) = \frac{b^2}{2} m''(t) \int_{-1}^{1} u^2 K(u)\, du + o\big(b^2\big).$$

Proof Since $E[\bar{y}(i)] = m(t_i)$, the proof follows, as we have seen before in previous sections, from a two-term Taylor series expansion of $m(t_i)$ around t and in particular by noting that as $n \to \infty$,

$$\left| \frac{1}{nb} \sum_{j=1}^{n} \left(\frac{t_j - t}{b} \right)^p K\left(\frac{t_j - t}{b} \right) - \int_{-1}^{1} u^p K(u)\, du \right| = O\left(\frac{1}{nb} \right)$$

where p is a positive integer. To simplify further, the term $O((nb)^{-1})$ can be absorbed into $o(b^2)$ since $nb^3 \to \infty$. □

As an example, when K is the uniform kernel on $(-1, 1)$, since $\int_{-1}^{1} u^2 K(u)\, du = 1/3$ the asymptotic expression of the bias of $\widehat{m}(t)$ is

$$E\big[\widehat{m}(t)\big] - m(t) = \frac{b^2}{6} m^{(2)}(t) + o\big(b^2\big)$$

and for $\eta \in (0, 1/2)$, as $n \to \infty$, the integrated squared bias of $\widehat{g}$ is:

$$\int_{\eta}^{1-\eta} \big\{ E\big[\widehat{m}(t)\big] - m(t) \big\}^2 dt = \frac{b^4}{36} \int_{\eta}^{1-\eta} \big\{ m^{(2)}(t) \big\}^2 dt + o\big(b^4\big).$$

As for the covariances, note that when $d = \max\{d_1, \ldots, d_k\}$, N is fixed and finite and $\bar{e}(i) = N^{-1} \sum_{j=1}^{N} e_j(i)$, by (A2) and (A3),

$$cov\big(\bar{e}(i), \bar{e}(i+k)\big) = \gamma_{\bar{e}}(k) = \frac{1}{N^2} \sum_{j=1}^{N} \gamma_j(k) \sim \frac{1}{N^2} C_{d,N} |k|^{2d-1} \quad (\text{as } |k| \to \infty)$$

where

$$C_{d,N} = \sum_{j:d_j = d} C_j.$$

Similarly, the spectral density is

$$f_{\bar{e}}(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_{\bar{e}}(k) e^{-ik\lambda} = \frac{1}{N^2} \sum_{j=1}^{N} f_j(\lambda) \sim \frac{1}{N^2} D_{d,N} |\lambda|^{-2d} \quad (\text{as } \lambda \to 0)$$

where

$$D_{d,N} = \sum_{j:d_j = d} D_j.$$

These facts can be summarized as follows:

Lemma 7.2 *Let $d = \max\{d_1, \ldots, d_N\}$, and let N be fixed and finite. Then the largest fractional differencing parameter d is also the fractional differencing parameter for the sample mean process $\bar{e}(i)$ $(i = 1, 2, \ldots)$.*

Theorem 7.26 *Let N be fixed and finite. Let $K(u) = \frac{1}{2}1\{-1 \le u \le 1\}$, $d = \max\{d_1, \ldots, d_N\}$ and*

$$\beta(d, N) = \frac{2^{2d-1}}{d(2d+1)} C_{d,N}.$$

Then for $\eta \in (0, 1/2)$ and as $n \to \infty$, the integrated variance of $\widehat{m}$ is

$$\int_{\eta}^{1-\eta} \mathrm{Var}\big[\widehat{m}(t)\big]\, dt = \frac{1}{N^2}(1-2\eta)(nb)^{2d-1}\beta(d, N) + o\big((nb)^{2d-1}\big).$$

Proof For every fixed $t \in (0, 1)$,

$$\begin{aligned}\mathrm{Var}\big(\widehat{m}(t)\big) &= \frac{1}{(2nbN)^2} \sum_{j=1}^{N} \sum_{r,s=n(t-b)}^{n(t+b)} \gamma_j(r-s) \\ &= \frac{1}{(2nbN)^2} \sum_{j=1}^{N} \sum_{r_1,s_1=1}^{2nb+1} \gamma_j(r_1 - s_1)\end{aligned}$$

where the last expression is obtained by substituting $r_1 = r - n(t-b) + 1$ and $s_1 = s - n(t-b) + 1$. Thus, we get

$$\begin{aligned}\mathrm{Var}\big(\widehat{m}(t)\big) &= \frac{1}{(2nbN)^2} \sum_{j=1}^{N} \sum_{k=-2nb}^{2nb} \big(2nb + 1 - |k|\big)\gamma_j(k) \\ &= \frac{1}{N^2} \sum_{j=1}^{N} \big[V_{n,j}^{(1)} + V_{n,j}^{(2)} - V_{n,j}^{(3)}\big]\end{aligned}$$

where

$$V_{n,j}^{(1)} = \frac{1}{2nb} \sum_{k=-2nb}^{2nb} \gamma_j(k),$$

$$V_{n,j}^{(2)} = \frac{1}{(2nb)^2} \sum_{k=-2nb}^{2nb} \gamma_j(k),$$

$$V_{n,r}^{(3)} = \frac{1}{(2nb)^2} \sum_{k=-2nb}^{2nb} |k|\gamma_j(k).$$

We have $d_j \in (0, 1/2)$ so that $2d_j - 1 \in (-1, 0)$ and

$$\lim_{nb\to\infty} \sum_{k=-2nb}^{2nb} \gamma_j(k) = \gamma_j(0) + 2C_j \lim_{nb\to\infty} \sum_{k=1}^{2nb} |k|^{2d_j-1} = \infty.$$

Also as $nb \to \infty$,

$$\left| \sum_{u=1}^{2nb} |u|^{2d_j-1} - (2nb)^{2d_j} \int_0^1 x^{2d_j-1}\, dx \right| = O\big((nb)^{2d_j-1}\big).$$

Simplifying, and since $(nb)^{2d_j-2} = o((nb)^{2d_j-1})$,

$$V_{n,j}^{(1)} = \frac{C_j}{d_j}(2nb)^{2d_j-1} + o\big((nb)^{2d_j-1}\big)$$

and clearly $V_{n,j}^{(2)} = o(V_{n,j}^{(1)})$. As for $V_{n,j}^{(3)}$, $|k|\gamma_j(k) \sim C_j|k|^{2d_j}$ as $|k| \to \infty$, so that

$$V_{n,r}^{(3)} = \frac{2C_j}{2d_j+1}(2nb)^{2d_j-1} + o\big((nb)^{2d_j-1}\big).$$

The theorem follows by noting that $V_{n,j}^{(1)} - V_{n,j}^{(3)} = (2nb)^{2d_j-1}C_j/(d_j(2d_j+1)) + o((nb)^{2d_j-1})$ and, as $n \to \infty$, the sum $\sum_{j=1}^{N}\{V_{n,j}^{(1)} - V_{n,j}^{(3)}\}$ will be dominated by a multiple of $(nb)^{2d-1}$ where d is the largest fractional differencing parameter. □

Corollary 7.2 *Let $K(u) = \frac{1}{2}1\{-1 < u < 1\}$ and, as $n \to \infty$, $b \to 0$ and $nb^3 \to \infty$. If N is fixed and finite and d_j $(j = 1, 2, \ldots, N)$ are fractional differencing parameters with $d = \max\{d_1, \ldots, d_N\}$, $0 < d_j < \frac{1}{2}$, then for $\eta \in (0, 1/2)$, the asymptotic expression for the integrated mean squared error for $\widehat{m}$ is (as $n \to \infty$)*

$$IMSE(\widehat{m}) = \left[\frac{b^4}{36}\int_\eta^{1-\eta} \big\{m^{(2)}(t)\big\}^2\, dt + \frac{1}{N^2}(1-2\eta)(nb)^{2d-1}\beta(d, N)\right] + o\big(\max\big(b^4, (nb)^{2\delta-1}\big)\big)$$

and the global optimum bandwidth minimising $IMSE(\widehat{m})$ is

$$b_{\text{opt}} = \left[\frac{9(1-2\eta)(1-2d)\beta(d,N)}{\int_\eta^{1-\eta}\{m^{(2)}(t)\}^2\, dt}\right]^{1/(5-2d)} \times\ n^{(2d-1)/(5-2d)}N^{-2/(5-2\delta)}$$

where $\beta(d, N)$ is defined in Theorem 7.26.

Substituting b_{opt} in the leading term of $IMSE(\widehat{m})$ the optimum rate of convergence can be obtained as $O(n^{(8d-4)/(5-2d)}N^{-8/(5-2d)})$. Note that when $d \to 0$ (i.e. the process approaches short-memory or independence) and $N = 1$, the familiar

rate $n^{-4/5}$ for the integrated mean squared error for estimation of the trend function can be confirmed. As usual, the rate of convergence under long memory ($d > 0$) is slower than under independence ($d = 0$). Compare also with (7.97) which corresponds to the case $N = 1$.

Case (ii) In this case, infinitely many replicates are available asymptotically.

Theorem 7.27 *We assume that* $\lim_{N\to\infty} N^{-1} \sum_{j=1}^{N} f_j(\lambda) = f(\lambda)$ *uniformly in* $\lambda \in (0, \pi)$ *with* $f(\lambda) \sim L(\lambda)|\lambda|^{-2d}$, $0 < d < 1/2$ *where* L *is slowly-varying at zero in the sense of Zygmund. Let* $\gamma(k) = (2\pi)^{-1} \int_{-\pi}^{\pi} f(\lambda) e^{ik\lambda}\, d\lambda \sim L(1/|k|)|k|^{2d-1}$ ($|k| \to \infty$). *Then for* $\eta \in (0, 1/2)$, *the asymptotic expression for the integrated mean squared error of* $\widehat{m}$ (*as* $N \to \infty$, $n \to \infty$) *is*

$$\begin{aligned} IMSE(\widehat{m}) = {} & \frac{b^4}{36} \int_{\eta}^{1-\eta} \left\{m^{(2)}(t)\right\}^2 dt \\ & + \frac{1}{N} \frac{1}{d(2d+1)} (1-2\eta)(2nb)^{2d-1} L\left(\frac{1}{nb}\right) \\ & + o\left(\max\left(b^4, (nb)^{2d-1}\right)\right). \end{aligned} \tag{7.152}$$

Proof The expression for the bias term follows as in Theorem 7.25. As for the variance, first of all, j disappears due to the convergence of the mean $N^{-1} \sum_{j=1}^{N} \gamma_j(k)$ appearing in $\mathrm{var}(\widehat{m}(t))$ to the limit $\gamma(k)$ that follows a slow hyperbolic decay given by (A3). The proof follows from similar arguments as for Theorem 7.26. □

Corollary 7.3 *Under the conditions of Theorem* 7.27, *the global optimum bandwidth minimizing* $IMSE(\widehat{m})$ *is*

$$\begin{aligned} b_{\mathrm{opt}} = {} & \left[\frac{9(1-2\eta)(1-2d)2^{(2d-1)/(5-2d)} L(1/(nb))}{d(2d+1) \int_{\eta}^{1-\eta} \left\{g^{(2)}(t)\right\}^2 dt} \right]^{1/(5-2d)} \\ & \times n^{(2d-1)/(5-2d)} N^{-1/(5-2d)} \end{aligned}$$

where the slowly-varying function L *is defined in Theorem* 7.27.

Remark By assumption, the spectral density $f_j(\lambda)$ of the jth error process e_j behaves at zero like a constant D_j times $|\lambda|^{-2d_j}$. In the theorem above, however, we assume the average spectral density to be a product of a slowly varying function L and $|\lambda|^{-2d}$ where $0 < d < 1/2$. In particular, L need not be a constant. An insight into this may be gained, for instance, by considering the case of i.i.d. random fractional differencing parameters having a moment generating function M where $M(-2\log|u|) = L(u)|u|^{-2d}$; an example is the uniform distribution; see Ghosh (2001). In this case, the expected value of the spectral density function is directly proportional to $L(\lambda) \times |\lambda|^{-2\theta}$ where $1/2 > \theta > 0$, and $L(u) \propto 1/\log(|u|)$.

7.4.8 Random-Design Regression Under LRD

In this section, our goal is to estimate the conditional mean function $m(Y_t|X_t)$ in a random-design model with residuals exhibiting long-range dependence and a variance that may depend on X_t. Thus, we have

$$Y_i = m(X_i) + \sigma(X_i)e_i \tag{7.153}$$

where now X_i is a stationary process with marginal density p_X, e_i is a stationary zero mean process with long memory and σ is a continuous function of X_i. Since the design is random, we consider the Nadaraya–Watson estimator (7.104), i.e.

$$\widehat{m}_{\mathrm{NW}}(x) = \frac{\widehat{m}_{\mathrm{PC}}(x)}{\hat{p}_X(x)} = \frac{(nb)^{-1}\sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right) Y_i}{\hat{p}_X(x)} \tag{7.154}$$

where

$$\hat{p}_X(x) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right) \tag{7.155}$$

is a kernel density estimator of p_X.

We can summarize the limiting behaviour of $\widehat{m}_{\mathrm{NW}}$ in the following theorem. This theorem summarizes results obtained under different sets of assumptions and using different techniques in papers like Cheng and Robinson (1994), Csörgő and Mielniczuk (1999, 2000), Mielniczuk and Wu (2004), Zhao and Wu (2008) and Kulik and Lorek (2011).

Theorem 7.28 *Suppose that m and σ are twice continuously differentiable in a neighbourhood of x_0. Then the following holds*:

- *Suppose that X_i are i.i.d. and $e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}$ is a linear process with i.i.d. zero mean innovations ε_i, $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_i) < \infty$ and $a_j \sim c_a j^{d_e - 1}$ for some $0 < d < \frac{1}{2}$. Then, for a sequence of bandwidths*

$$b = o\left(n^{-2d_e}\right)$$

we have

$$\sqrt{nb}\sqrt{\hat{p}_X(x_0)}\left\{\hat{m}(x_0) - E\left[\hat{m}(x_0)\right]\right\} \xrightarrow{\mathrm{d}} Z\sqrt{\sigma^2(x_0)p(x_0)\int K^2(u)\,du} \tag{7.156}$$

where Z is a standard normal random variable.

- *Under the same assumptions, but with*

$$b \gg n^{-2d_e},$$

we have

$$n^{\frac{1}{2}-d_e} c_e^{-\frac{1}{2}} \big\{\hat{m}(x_0) - E\big[\hat{m}(x_0)\big]\big\} \xrightarrow{\mathrm{d}} \sigma(x_0) Z \tag{7.157}$$

where $c_e = c_{f_e} \nu(d_e)$ *is the constant in* $\operatorname{var}(\sum_{i=1}^n e_i) \sim c_e n^{2d_e+1}$.

- *Suppose that* $X_i = \sum_{j=0}^{\infty} a_{j,X} \xi_{i-j}$ *is a zero mean Gaussian process with long-range dependence such that* $\gamma_X(k) \sim c_\gamma |k|^{2d-1}$ $(0 < d < \frac{1}{2})$. *Then, keeping the other conditions as above, the same results follow for* $b = o(n^{-2d_e})$ *and* $b \gg n^{-2d_e}$, *respectively.*

Proof We write

$$\begin{aligned}
\hat{p}_X(x_0)\big\{\hat{m}(x_0) - E\big[\hat{m}(x_0)\big]\big\} &= \frac{1}{nb} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right) Y_i - E\big[\hat{m}(x_0)\big] \hat{p}_X(x_0) \\
&= \frac{1}{nb} \sum_{i=1}^n K\left(\frac{X_i - x_0}{b}\right) \big\{m(X_i) - E\big[\hat{m}(x_0)\big]\big\} \\
&\quad + \frac{1}{nb} \sum_{i=1}^n K\left(\frac{X_i - x_0}{b}\right) \sigma(X_i) e_i .
\end{aligned}$$

It can be shown that the first term is $o_p((nb)^{-1/2})$ and is hence asymptotically negligible. The second term has the structure $R_n := n^{-1} \sum_{i=1}^n \nu_n(X_i) e_i$ (cf. (7.60)), where

$$\nu_n(X_i) = b^{-1} K\left(\frac{x_0 - X_i}{b}\right) \sigma(X_i) = b^{-1} K\left(\frac{X_i - x_0}{b}\right) \sigma(X_i).$$

Note that

$$\begin{aligned}
E\big[\nu_n(X_1)\big] &= b^{-1} \int K\left(\frac{x_0 - u}{b}\right) \sigma(u) p_X(u)\, du \\
&= \int K(u) \sigma(x_0 - ub) p_X(x_0 - ub)\, du \neq 0.
\end{aligned} \tag{7.158}$$

Since σ and p_X are assumed to be twice continuously differentiable in a neighbourhood of x_0, with bounded second derivatives, we have

$$E\big[\nu_n(X_1)\big] \sim \sigma(x_0) p_X(x_0), \qquad \operatorname{var}\big(\nu_n(X_1)\big) \sim b^{-1} \sigma^2(x_0) p_X(x_0) \int K^2(u)\, du. \tag{7.159}$$

Thus, we can apply techniques from Sect. 7.2.3:

- If e_i are i.i.d., then R_n is a martingale. An application of a martingale central limit theorem (Lemma 4.2) yields

$$\sqrt{nb}\frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{b}\right)\sigma(X_i)e_i \xrightarrow{\mathrm{d}} \sigma(x_0)Z\sqrt{p_X(x_0)\int K^2(u)\,du}.$$

- If e_i is a linear long-memory process and X_i are i.i.d., then we apply the (M/L)-decomposition

$$R_n = n^{-1}E\big[\nu_n(X_1)\big]\sum_{i=1}^{n} E[e_i|\varepsilon_s, s \le i-1]$$
$$+ n^{-1}\sum_{t=1}^{n}\big\{\nu_n(X_i)e_i - E\big[\nu_n(X_i)e_i|X_s, \varepsilon_s, s \le i-1\big]\big\} =: R_{n,1} + R_{n,2}.$$

The second part is a martingale and again an application of the martingale CLT yields

$$\sqrt{nb}R_{n,2} \xrightarrow{\mathrm{d}} Z\sqrt{\sigma^2(x_0)p_X(x_0)\int K^2(u)\,du}. \tag{7.160}$$

For the first part, we have, recalling (7.48) and (7.159),

$$n^{\frac{1}{2}-d_e}c_e^{-\frac{1}{2}}R_{n,1} \xrightarrow{\mathrm{d}} \sigma(x_0)p_X(x_0)Z. \tag{7.161}$$

- If both, X_i and e_i are linear processes with long memory, then we proceed exactly the same way as in the case of parametric linear regression. The direct application of the Hermite polynomial decomposition does not lead to weakly dependent behaviour (7.156). However, conditioning on $\xi_i, \xi_{i-1}, \ldots$, we start with an (M/L)-decomposition

$$\frac{1}{n}\sum_{i=1}^{n}\big(\nu_n(X_i)e_i - E\big[\nu_n(X_i)e_i|\xi_s, \varepsilon_s, s \le i-1\big]\big)$$
$$+ \frac{1}{n}\sum_{i=1}^{n} E[e_i|\varepsilon_s, s \le i-1]\int K\left(\frac{x_0 - (u + \hat{X}_i)}{b}\right)\sigma(u + \hat{X}_i)p_\xi(u)\,du$$
$$=: \tilde{R}_{n,2} + \tilde{R}_{n,1}, \tag{7.162}$$

where $p_\xi(\cdot)$ is the density of ξ_i and $\hat{X}_i = X_i - \xi_i$ is the one-step forecast of X_i given ξ_s $(s \le i-1)$. Now, $\tilde{R}_{n,2}$ is a martingale and its limiting properties are described by (7.160). For $\tilde{R}_{n,1}$ we apply the Hermite polynomial decomposition (7.62) with

$$\tilde{\nu}_n(z) = \int K\left(\frac{x_0 - (u+z)}{b}\right)\sigma(u+z)p_\xi(u)\,du.$$

Let $p_{\hat{X}}$ be the density of $\hat{X}_i$. Note that p_X is the convolution of $p_{\hat{X}}$ and p_ξ, i.e. $p_X = p_{\hat{X}} * p_\xi$. Then

$$\begin{aligned} E\big[\tilde{\nu}_n(\hat{X}_t)\big] &= \int\int K\left(\frac{x_0-(u+z)}{b}\right)\sigma(u+z)p_\xi(u)p_{\hat{X}}(z)\,du\,dz \\ &= \int\int K(u)\sigma(x_0-bu)p_\xi(x_0-z-bu)p_{\hat{X}}(z)\,du\,dz \\ &\sim \sigma(x_0)\int K(u)\,du\int p_\xi(x_0-z)p_{\hat{X}}(z)\,dz = \sigma(x_0)p_X(x_0). \end{aligned}$$

Thus, using the same argument as for parametric regression, we are able to conclude that (7.161) holds for $\tilde{R}_{n,2}$. The result then follows by comparing the term $R_{n,1}$ with $R_{n,2}$, and $\tilde{R}_{n,1}$ with $\tilde{R}_{n,2}$, respectively, and noting that $\hat{p}_X$ is the consistent estimator of p_X (see Sect. 5.14). □

The theorem is remarkable in several ways. First of all, it reveals a dichotomy between *small* and *large* bandwidths. This is the same phenomenon as observed already for density estimation (see Sect. 5.14). For small bandwidths $b = cn^{-\alpha} = o(n^{-2d_e})$, the long-range dependence in the residuals has no influence, and one obtains exactly the same asymptotic distribution as for i.i.d. data. The optimal bandwidth is then of the form $b = cn^{-\frac{1}{5}}$, and optimal MSE has the order $O(n^{-\frac{4}{5}})$. This is in contrast to fixed-design kernel estimation. On the other hand, this behaviour is not unexpected in view of similar results for random design linear regression (Sect. 7.2) and kernel density estimation (Sect. 5.14). For large bandwidths $b \gg n^{-2d_e}$, the contribution of the bias is proportional to $n^{-4\alpha} \gg n^{-8d_e}$ whereas the variance is proportional to $n^{-(1-2d_e)}$. Since $1-2d_e < 8d_e$ is equivalent to $d_e > 0.1$, the first conclusion is that the optimal MSE is of the order $n^{-\frac{4}{5}}$ (with $b_{\text{opt}} = cn^{-\frac{1}{5}}$) only if $d_e < 0.1$. For $d_e > 0.1$, the optimal order is $n^{-(1-2d_e)}$ which is achieved as long as the variance dominates the bias. This is the case for a whole range of bandwidths $b = cn^{-\alpha}$ with $1-2d_e < 4\alpha < 8d_e$. These general results are the same as for density estimation. We therefore do not repeat the same comments and refer the reader to Sect. 5.14. The second remarkable aspect of Theorem 7.28 is that long memory in the explanatory process X_i does not influence the asymptotic behaviour.

The results can be generalized to multivariate time series. In the context of (7.160), the limit is multivariate normal with independent components; in the context of (7.161), the limit is multivariate normal with perfectly correlated components. Furthermore, one can also obtain analogous results for multivariate predictors.

The main conclusion is that for $d_e > 0.1$, the MSE is dominated by the variance as long as the bandwidth is not too large but of a larger order than n^{-2d_e}. An exact choice of b is not needed to achieve the optimal rate of $n^{-(1-2d_e)}$. However, as for density estimation, a higher-order expansion of the MSE can be used to derive a criterion for an optimal bandwidth—even though it may not have an influence

asymptotically. Considering a weighted integrated mean squared error

$$IMSE(\hat{m}, m; w) = \int E\big[\big(\hat{m}(x) - m(x)\big)^2\big] w(x)\, dx,$$

Kulik and Lorek (2011) obtained the following formula.

Proposition 7.2 *Under the assumptions of the third part of Theorem 7.28 (i.e. when both e_i and X_i have long memory), we have*

$$\begin{aligned} IMSE(\hat{m}, m; w) \sim{}& \frac{1}{nb}\kappa_1 \int \frac{\sigma^2(x)}{p_X(x)} w(x)\, dx \\ &+ b^4 \frac{\kappa_2^2}{4} \int \left(\frac{m''(x) p_X(x) + 2m'(x) p_X'(x)}{p_X(x)} \right)^2 w(x)\, dx \\ &+ n^{2d_\varepsilon - 1} c_\varepsilon \int \sigma^2(x) w(x)\, dx + b^2 n^{2d_\varepsilon - 1} c_e \kappa_2 \int \psi_e(x) w(x)\, dx, \end{aligned} \tag{7.163}$$

where $\kappa_1 = \int K^2(u)\, du$, $\kappa_2 = \int u^2 K(u)\, du$, and

$$\psi_e(x) = \sigma(x) \frac{(\sigma(x) p_X(x))''}{p_X(x)}.$$

Of course, the weight function w must be chosen in such the way that the integrals are finite. For example, if $\sigma(x) \equiv 1$ and p_X is the standard normal density, then

$$\int \frac{\sigma^2(x)}{p_X(x)} w(x)\, dx = \int \frac{w(x)}{p_X(x)}\, dx$$

would be infinite if we chose $w(x) \equiv 1$, whereas this is not the case, for instance, for $w(x) = p_X^2(x)$.

The first term in (7.163) is due to the bias, the second one describes i.i.d.-type behaviour. The term involving d_e describes a possible contribution of long memory. Note that we have to include the term $b^2 n^{2d_e - 1} c_e$ to obtain a criterion for bandwidth selection that can also be used for $d > 0.1$. For $d > 0.1$ this terms does not have an influence on the optimal behaviour of the *MISE*, but it improves the higher-order term in the expansion. Optimizing the higher order expansion with respect to b yields

$$b_{\text{opt}} \sim \begin{cases} C n^{-\frac{1}{5}} & \text{if } d_e < 0.3, \\ C n^{-\frac{2}{3} d_e} & \text{if } d_e > 0.3. \end{cases}$$

The optimal $IMSE(\hat{m}, m; w)$ with b_{opt} is then proportional to $n^{-4/5}$ if $d_e < 1/10$, and to $n^{2d_e - 1} c_e(n)$ if $d_e > 1/10$. However, as discussed above (also see Sect. 5.14),

for $d > 1/10$ the optimal order can be achieved even if b is not exactly of the order $O(n^{-\frac{2}{3}d_e})$.

The optimal bandwidth depends on unknown parameters. Moreover, for $d_e > 0.1$ data driven bandwidth choice is not quite trivial because b_{opt} is based on a higher order expansion of the IMSE. Given an observed series where we may not know much about the underlying process, it seems quite difficult to estimate the IMSE with sufficient accuracy to assess the contribution of higher-order terms. For instance, cross-validation turns out to be applicable for $d_e < 0.1$ only (for a precise statement, see Kulik and Lorek 2011).

An improved result can be obtained if one is interested in the shape of the function $m(x)$ only. This means that the aim is to estimate

$$m^*(x) = E[Y|X = x] - E[Y] = m(x) - \int m(x)\, p_X(x)\, dx.$$

The natural estimator is given by

$$\hat{m}^*(x) = \widehat{m}_{\text{NW}}(x) - \bar{y} \tag{7.164}$$

where $\bar{y} = n^{-1} \sum Y_i$. In contrast to Proposition 7.2, the mean squared error is now influenced by the dependence structure of X_i (Kulik and Lorek 2011) whereas the long-memory property of e_i disappears:

Theorem 7.29 *Suppose that m is twice continuously differentiable in a neighbourhood of x_0 and $\sigma(x) \equiv 1$. Then the following holds*:

- *Suppose that X_i are i.i.d. and $e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}$ is a linear process with i.i.d. zero mean innovations ε_i, $\sigma_\varepsilon^2 = \text{var}(\varepsilon_i) < \infty$ and $a_j \sim c_a j^{d_e - 1}$ for some $0 < d_e < \frac{1}{2}$. Then*

$$\begin{aligned} IMSE(\hat{m}, m; w) \sim b^4 \frac{\kappa_2^2}{4} \int \left(\frac{m''(x) p_X(x) + 2m'(x) p_X'(x)}{p_X(x)} \right)^2 w(x)\, dx \\ + \frac{1}{nb} \kappa_1 \int \frac{w(x)}{p_X(x)}\, dx, \end{aligned} \tag{7.165}$$

  where $\kappa_1 = \int K^2(u)\, du$, $\kappa_2 = \int u^2 K(u)\, du$.
- *Suppose that X_i is a zero mean Gaussian process with long-range dependence such that $\gamma_X(k) \sim c_\gamma |k|^{2d_X - 1}$ $(0 < d_X < \frac{1}{2})$ and $\text{var}(\sum_{i=1}^n X_i) \sim c_X n^{2d_X - 1}$. Then*

$$\begin{aligned} IMSE(\hat{m}, m; w) \sim b^4 \frac{\kappa_2^2}{4} \int \left(\frac{m''(x) p_X(x) + 2m'(x) p_X'(x)}{p_X(x)} \right)^2 w(x)\, dx \\ + \frac{1}{nb} \kappa_1 \int \frac{w(x)}{p_X(x)}\, dx + n^{2d_X - 1} c_X E^2\big[m(X)X\big]. \end{aligned} \tag{7.166}$$

The first part of Theorem 7.29 means that for i.i.d. explanatory variables the asymptotic mean squared error is exactly the same as for i.i.d. residuals. Thus, if we are interested in the shape of m only, then the optimal bandwidth is the same as under i.i.d. assumptions, namely $b_{\mathrm{opt}} = C_{\mathrm{opt}} n^{-\frac{1}{5}}$, and the optimal IMSE is of the order $O(n^{-\frac{4}{5}})$. This is similar to results on linear regression through the origin with explanatory variables having expected value zero. Note in particular that even if $\int m(x) p_X(x)\,dx = 0$, the rate can be improved by subtracting $\bar{y}$. This is similar to the improved rate of the empirical process when subtracting the sample mean (see Sect. 4.8.3) and results discussed in the context of goodness-of-fit testing where estimation of nuisance parameters improves the rate of convergence (Sect. 5.16). On the other hand, if X_i exhibits long memory, then the rate deteriorates for functions m whose Hermite rank is one. In terms of orders, we have $IMSE = O(b^4) + O((nb)^{-1}) + O(n^{2d_X - 1})$. Minimization with respect to $b = cn^{-\alpha}$ therefore yields exactly the same optimal value $b_{\mathrm{opt}} = C_{\mathrm{opt}} n^{-\frac{1}{5}}$ as for i.i.d. residuals. However, the optimal mean squared error is of the order $O(n^{-\frac{4}{5}})$ only if $\frac{4}{5} \le 1 - 2d_X$ which means $d_X \le 0.1$. For $d_X > 0.1$ the variance dominates the optimal IMSE which is asymptotically proportional to $n^{2d_X - 1}$. On the other hand, for very large bandwidths $b = cn^{-\alpha}$ with $\alpha < \frac{1}{4}(1 - 2d_X)$, the bias dominates the IMSE which is then, however, far from the optimal one. In summary, if X_i exhibits long memory, then the results are analogous to estimation of m; however, with d_e replaced by d_X.

7.4.9 Conditional Variance Estimation

We go back to the parametric regression model (7.45)

$$Y_i = \beta_0 + \beta_1 X_i + \sigma(X_i) e_i.$$

Our goal now is to estimate the conditional variance function $\sigma^2(\cdot)$ in a nonparametric way. To do so, we first estimate β_0 and β_1 by the least squares method studied in Sect. 7.2. Then, in analogy to conditional mean estimation, we estimate $\sigma^2(\cdot)$ by smoothing residuals with a kernel K and a bandwidth b,

$$\hat{\sigma}^2(x_0) = \frac{(nb)^{-1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 K(\frac{X_i - x_0}{b})}{\hat{p}_X(x_0)}, \tag{7.167}$$

where $\hat{p}_X(x_0)$ is the kernel density estimator defined in (7.155). It is known that in the case of weakly dependent errors and/or predictors, estimation of β_0 and β_1 does not influence the performance of $\hat{\sigma}^2(\cdot)$ (see Fan and Yao 1998; Zhao and Wu 2008).

To see what happens in the case of long memory, we will work under the condition that X_i are i.i.d. and $e_i = \sum a_j \varepsilon_{i-j}$ is a linear long-memory process with $a_j \sim c_a j^{d-1}$ $(0 < d < \frac{1}{2})$. Defining

$$\Delta_t = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) X_t =: \Delta_0 + \Delta_{1,t},$$

we can write down the decomposition

$$\begin{aligned}\hat{p}_X(x_0)\big(\hat{\sigma}^2(x_0)-\sigma^2(x_0)\big) &= \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{X_i-x_0}{b}\right)\big(\sigma^2(X_i)-\sigma^2(x_0)\big)\\ &\quad+\frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{X_i-x_0}{b}\right)\sigma^2(X_i)\big(e_i^2-1\big)\\ &\quad-\frac{2}{nb}\sum_{i=1}^{n}\Delta_i\sigma(X_i)K\left(\frac{X_i-x_0}{b}\right)e_i\\ &\quad+\frac{1}{nb}\sum_{i=1}^{n}\Delta_i^2 K\left(\frac{X_i-x_0}{b}\right)\\ &=: J_1+J_2-J_3+J_4.\end{aligned}$$

If β_0 and β_1 were known, then we would have $\Delta_i=0$ and thus $J_3=J_4\equiv 0$. Let us recall the proof of Theorem 7.28. The first two terms J_1 and J_2 are very similar to the terms appearing in the decomposition of $\hat{p}_X(x_0)(\hat{m}(x_0)-m(x_0))$. If we assume $nb^5\to 0$, then $\sqrt{nb}J_1=o_p(1)$ so that the term J_1 is negligible. The second term can be decomposed into two terms J_{21} and J_{22} with

$$\sqrt{nb}J_{21}\xrightarrow{\mathrm{d}} Z_1\sigma^2(x_0)\sqrt{p_X(x_0)\int K^2(u)\,du} \tag{7.168}$$

and, if $d\in(1/4,1/2)$,

$$n^{1-2d_\varepsilon}c_{e,2}^{-\frac{1}{2}}J_{22}\xrightarrow{\mathrm{d}}\sigma^2(x_0)p_X(x_0)Z_{2,H_0}(1) \tag{7.169}$$

where $Z_{2,H_0}(1)$ is the Hermite–Rosenblatt process at time 1 and $c_{e,2}$ is the constant in $\mathrm{var}(\sum(e_i^2-1))\sim c_{e,2}n^{4d+2}$. If $d\in(0,1/4)$, then $\sqrt{n}J_{22}=o_P(1)$. The reason for the difference between (7.161) and (7.169) is that the latter involves limiting behaviour of $\sum_{t=1}^{n}(e_t^2-1)$.

To deal with J_3, write

$$\begin{aligned}J_3 &= (\hat{\beta}_0-\beta_0)\frac{2}{nb}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{b}\right)\sigma(X_i)e_i\\ &\quad+(\hat{\beta}_1-\beta_1)\frac{2}{nb}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{b}\right)X_i\sigma(X_i)e_i\\ &=: (\hat{\beta}_0-\beta_0)\tilde{L}_3+(\hat{\beta}_1-\beta_1)\tilde{R}_3.\end{aligned}$$

Defining the quantity

$$\tilde{J}_3 := \frac{2}{n^2 b} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left(\frac{X_i - x_0}{b}\right) \sigma(X_i)\sigma(X_j) X_i X_j e_i e_j,$$

we may decompose J_3 into two parts,

$$J_3 = \tilde{L}_3 \frac{1}{n} \sum_{i=1}^{n} \sigma(X_i)\varepsilon_i + \frac{1}{V_n} \tilde{J}_3, \tag{7.170}$$

with $V_n^2 = n^{-1} \sum_{i=1}^{n} X_i^2$. Furthermore, in $\tilde{J}_3$ we may ignore summation over $i = j$. Since X_i are i.i.d., the (M/L)-decomposition suggests that J_3 behaves like

$$E\left[b^{-1} K\left(\frac{X_i - x_0}{b}\right) \sigma(X_i)\sigma(X_j) X_i X_j\right] n^{-2} \sum_{t=1}^{n} \sum_{s=1, s\neq t}^{n} e_t e_s.$$

Since the expected value above behaves like $E[\sigma(X_1)X_1]\sigma(x_0)x_0$, we conclude from (7.48) that

$$n^{(1-2d_e)} c_e^{-\frac{1}{2}} \tilde{J}_3 \xrightarrow{\mathrm{d}} 2E\big[\sigma(X_1)X_1\big]\sigma(x_0)x_0 p_X(x_0) \cdot Z_0^2. \tag{7.171}$$

Similar arguments yield

$$n^{(1-2d_e)} c_e^{-\frac{1}{2}} \tilde{L}_3 n^{-1} \sum_{i=1}^{n} \sigma(X_i) e_i \xrightarrow{\mathrm{d}} 2E\big[\sigma(X_1)\big]\sigma(x_0) p_X(x_0) \cdot Z_0^2. \tag{7.172}$$

Since V_n converges in probability to 1, the last two equations mean that $n^{1-2d_e} c_e^{-\frac{1}{2}} J_3$ converges in distribution to

$$2\big\{E\big[\sigma(X_1)X_1\big]x_0 + E\big[\sigma(X_1)\big]\big\}\sigma(x_0) p_X(x_0) \cdot Z_0^2.$$

We note that this conclusion is obtained by justifying that the convergence in (7.171) and (7.172) is joint. Similar considerations can be applied to J_4. Details can be found in Kulik and Wichelhaus (2011). There, the results are obtained under more general assumption on predictors; see also Guo and Koul (2008). Extension to conditional variance estimation in the model (7.153) are given in Kulik and Wichelhaus (2012) and Zhao and Wu (2008). In summary, the following dichotomy is obtained:

Theorem 7.30 *Consider the random design regression model* (7.45). *Assume that* $nb^5 \to 0$ *and* σ *is twice continuously differentiable in a neighbourhood of* x_0. *Furthermore, suppose that* X_i *are i.i.d. and* $e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}$ *is a second-order stationary linear process with* $a_j \sim c_a j^{d_e - 1}$ $(0 < d_e < \frac{1}{2})$, *and denote by* Z *and* Z_0 *standard normal variables and by* $Z_{2,H_0}(1)$ *an Hermite–Rosenblatt variable. Then the following holds*:

- *If* $b = o(n^{1-4d_e})$, *then*

$$\sqrt{nb}\sqrt{\hat{p}_X(x_0)}\big(\hat{\sigma}^2(x_0) - \sigma^2(x_0)\big) \xrightarrow{\mathrm{d}} Z\sigma^2(x_0)\sqrt{p_X(x_0)\int K^2(u)\,du};$$

- *If* $b \gg n^{1-4d_e}$, *then*

$$\begin{aligned} & n^{1-2d_e} c_e^{-\frac{1}{2}} \big(\hat{\sigma}^2(x_0) - \sigma^2(x_0)\big) \\ & \quad \underset{d}{\to} \sigma^2(x_0) Z_{2,H_0}(1) \\ & \qquad + \big\{E^2\big[\sigma(X_1)X_1\big]x_0^2 - 2\sigma(x_0)x_0 E\big[\sigma(X_1)X_1\big]\big\} Z_0^2 \\ & \qquad + \big\{E^2\big[\sigma(X_1)\big] - 2\sigma(x_0)E\big[\sigma(X_1)\big]\big\} Z_0^2. \end{aligned} \tag{7.173}$$

The last two terms quantify the price we have to pay due to estimation of β_0 and β_1 and due to the fact that the error process has long-range dependence. Note that the first of the two terms disappears, if $E^2[\sigma(X_1)X_1] = 0$. Finally, note that the assumption $nb^5 \to 0$ was used for convenience in order that the bias of $\hat{\sigma}^2(x_0)$ be asymptotically negligible. This assumption can be dropped, but then $\hat{\sigma}^2(x_0) - \sigma^2(x_0)$ has to be replaced by $\hat{\sigma}^2(x_0) - E[\hat{\sigma}^2(x_0)]$, and the bias of $\hat{\sigma}^2(x_0)$ has to be treated separately (as it was done previously when estimating the conditional mean function $m(x_0)$ nonparametrically).

7.4.10 Estimation of Trend Functions for LARCH Processes

Consider a time series model $Y_i = m(t_i) + e_i$ with a nonparametric trend function $m(t_i)$ ($t_i \in [0, 1]$) and residuals e_i that exhibit long-range dependence in volatility, and a linear dependence structure corresponding either to short memory, long memory or antipersistence. The main question addressed here is the asymptotic behaviour of nonparametric estimators of m. In particular, one is interested in characterizing the influence of linear and nonlinear dependence of $\hat{m}$.

More specifically, Beran and Feng (2007) consider residuals e_i having a Wold decomposition

$$e_i = \sum_{j=0}^{\infty} a_j X_{i-j} = A(B)Z_i$$

with $|A(e^{-i\lambda})|^2 \sim L_{f_e}(\lambda)|\lambda|^{-2d_1}$ ($-\frac{1}{2} < d_1 < \frac{1}{2}$) as $\lambda \to 0$, $L_{f_e}(\lambda) \in C[-\pi, \pi]$ slowly varying, and Z_i is a long-memory LARCH process with $b_j \sim cj^{d_2-1}$ (as $j \to \infty$) for some $0 < d_2 < \frac{1}{2}$ and $\sum b_j^2 < 1$. For the autocovariances of e_i, we have $\gamma_e(k) \sim L_{\gamma_e}(k)|k|^{2d_1-1}$ with L_{γ_e} slowly varying, whereas Z_i are uncorrelated

but the squares Z_i^2 have autocovariances of the form $\gamma_{Z^2}(k) \sim L_{\gamma_{Z^2}}(k)|k|^{2d_2-1}$ (as $j \to \infty$) where $L_{\gamma_{Z^2}}$ is another slowly varying function.

We recall that, given a polynomial degree $p \in \mathbb{N}$ and a bandwidth $b > 0$, a local polynomial estimator of the jth derivative $m^{(j)}(t_0)$ (for a fixed $t_0 \in [0, 1]$) can be written as

$$\widehat{m^{(j)}}(x) = j!\hat{\beta}_j = j!\delta_{j+1}^T \left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}\mathbf{y} \tag{7.174}$$

$$= \mathbf{w}_{j,b;n}^T\mathbf{y} = \sum_{i=1}^{n} w_{j,b;n}(i)Y_i \tag{7.175}$$

where $\delta_j = (\delta_{1,j}, \ldots, \delta_{p+1,j})^T$ $(j = 1, \ldots, p+1)$ denote unit vectors with $\delta_{j,j} = 1$, $\delta_{i,j} = 0$ $(i \neq j)$ (see (7.106)). Thus, investigating the asymptotic behaviour of $\hat{\mu}^{(j)}(t_0)$ amounts to studying the sequence of sums

$$S_n = \sum_{i=1}^{n} w_{j,b;n}(i)Y_i = \sum_{i=1}^{n} \zeta_{i,n} \quad (n \in \mathbb{N})$$

of a triangular array $\zeta_{i,n} = w_{j,b;n}(i)Y_i$ $(1 \leq i \leq n;\, n \in \mathbb{N})$. For the specific weights given by local polynomial estimation, Beran and Feng (2007) derive asymptotic normality of S_n under suitable conditions on the tail behaviour of e_i and on the weights $w_{j,b;n}$. In particular, one must make sure that the weights are balanced in the sense that $\max_{1\leq i\leq n} w_{j,b;n}^2(i)$ is asymptotically of a smaller order than $\mathrm{var}(S_n)$ (for the detailed assumptions, see Beran and Feng 2007). Also note that the results for the mean squared error are the same as in Theorem 7.22 because these depend on the linear dependence structure only.

7.4.11 Further Bibliographic Comments

Hall and Hart (1990b) were the first to derive an asymptotic formula for the mean squared error of kernel estimators of the trend function $m(t)$ in fixed-design regression with long-memory errors. This result was extended further in Beran and Feng (2001a, 2001b, 2002a, 2002b, 2002c), including kernel estimation with boundary corrections, local polynomial estimation of derivatives and integrated processes. Results along the line of (7.144) were proven in Csörgő and Mielniczuk (1995a) under the condition of a homoscedastic Gaussian residual process (the modification to the heteroskedastic case is obvious). See also Csörgő and Mielniczuk (1995b) and Robinson (1997). Nonparametric trend estimation in replicated long-memory time series is considered in Ghosh (2001). The general results applicable to local polynomial estimators of $m^{(j)}$ and kernel estimators with boundary correction was given in Beran and Feng (2001a, 2001b, 2002a) (also see Feng et al. 2007). Properties of cross-validation and plug-in bandwidth were studied in Hall et al. (1995a) and Beran and Feng (2002a, 2002b, 2002c), respectively. Data driven bandwidth

selection including asymptotic results on the convergence of the estimated bandwidth can also be found in Beran and Feng (2002a, 2002b, 2002c). Extensions to LARCH-type residuals are given in Beran and Feng (2007). Opsomer et al. (2001) give an overview of up-to-date existing results in nonparametric estimation with short- and long-memory errors. Robust versions of local polynomial estimators in the long-memory context are considered in Beran et al. (2002) and Beran et al. (2003). Optimal convergence rates in the long-memory setting are derived in Feng and Beran (2012). The nonexistence of optimal kernels in the long-memory setting is shown in Beran and Feng (2007). Extensions to nonequidistant time series and tests for rapid change points are derived in Menéndez et al. (2010).

Theorem 7.28 has its origin in work by Cheng and Robinson (1994). Further references include Csörgő and Mielniczuk (1999, 2000), Mielniczuk and Wu (2004), Zhao and Wu (2008), Kulik and Lorek (2011). In the latter article, the authors consider very general class of errors, which include FARIMA–GARCH or antipersistent processes. In Bryk and Mielniczuk (2008), the authors consider a randomization scheme for fixed-design regression. As a consequence, the resulting kernel estimator has a rate of convergence as in the random-design case. Results for the kernel Nadaraya–Watson estimator have further extensions to local linear regression estimators; see Masry and Mielniczuk (1999) and Masry (2001).

7.5 Trend Estimation Based on Wavelets

7.5.1 Introduction

In this section, we consider adaptive estimation of $m(t) = E(X)$ using wavelets. The advantage of the wavelet approach is evident for functions m that are inhomogeneous in time or not smooth. We start with the fixed-design case. As was shown for kernel and local polynomial estimation, the rates of convergence are affected by the presence of long memory. The same happens for wavelet methods (see, e.g. Wang 1996; Wang 1997; Johnstone and Silverman 1997; Johnstone 1999; Li and Xiao 2007; Kulik and Raimondo 2009a; Beran and Shumeyko 2012a). Again, in the random design case, it is possible to achieve the same rates as for weakly dependent data (Kulik and Raimondo 2009b).

7.5.2 Fixed Design

7.5.2.1 Data Adaptive Trend Estimation

As before, we consider a model with trend,

$$Y_i = m(t_i) + e_i, \tag{7.176}$$

with $t_i = i/n$, $m \in L^2[0,1]$ and e_i a zero mean stationary process with long-range dependence. Wavelet based trend estimation in the context of i.i.d. or short-range dependent residuals has been considered by many authors (see, e.g. a series of pioneering papers by Donoho and Johnstone). Most results deal with optimality in the sense of a minimax risk, and are partially also applicable in the long-memory setting. For an observed data set, however, the minimax principle often leads to estimates of m that may be far from optimal in the specific situation. A useful alternative is therefore to take a data adaptive approach where one tries to extract information about the dependence structure of e_i and preliminary information about m in order to come up with a (close to) optimal solution for $\hat{m}$. Results along this line are available in Li and Xiao (2007) and Beran and Shumeyko (2012a). For simplicity, suppose that e_i is a Gaussian process with autocovariance function $\gamma(k) = E(e_i e_{i+k}) \sim C_\gamma |k|^{2d-1}$ $(k \to \infty)$ and spectral density $f(\lambda) = (2\pi)^{-1} \sum \gamma(k) \exp(-ik\lambda) \sim C_f |\lambda|^{-2d}$ $(\lambda \to 0)$. To include a larger variety of wavelets, Beran and Shumeyko (2012a) assume that the support of the father and mother wavelets $\phi(t)$ and $\psi(t)$ is $[0, N]$ with N an arbitrary integer. Moreover, $\psi(0) = \psi(N) = 0$ and

$$\int_0^N \phi(t)\,dt = \int_0^N \phi^2(t)\,dt = \int_0^N \psi^2(t)\,dt = 1. \tag{7.177}$$

Then, for any $J \geq 0$, the system $\{\phi_{Jk}, \psi_{jk}, k \in \mathbb{Z}, j \geq 0\}$ with

$$\psi_{jk}(t) = N^{1/2} 2^{(J+j)/2} \psi\big(N2^{J+j}t - k\big), \qquad \phi_{Jk}(t) = N^{1/2} 2^{J/2} \phi\big(N2^J t - k\big),$$

is an orthonormal basis in $L^2(\mathbb{R})$ (see Sects. 3.5 and 3.5). An important role is played by the number $M_\psi \in \mathbb{N}$ of vanishing moments, defined by the properties

$$\int_0^N t^k \psi(t)\,dt = 0 \quad (k = 0, 1, \ldots, M_\psi - 1) \tag{7.178}$$

and

$$\int_0^N t^{M_\psi} \psi(t)\,dt = \nu_{M_\psi} \neq 0. \tag{7.179}$$

Recall that for every fixed, $J \geq 0$, every function $m \in L^2([0,1])$ has a unique orthogonal wavelet representation

$$m(t) = \sum_{k=-N+1}^{N2^J-1} s_{Jk} \phi_{Jk}(t) + \sum_{j=0}^{\infty} \sum_{k=-N+1}^{N2^{J+j}-1} d_{jk} \psi_{jk}(t), \tag{7.180}$$

with

$$s_{Jk} = \int_0^1 m(t) \phi_{Jk}(t)\,dt, \qquad d_{jk} = \int_0^1 m(t) \psi_{jk}(t)\,dt.$$

Setting

$$\hat{s}_{Jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_{Jk}(t_i), \qquad \hat{d}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_{jk}(t_i),$$

a (hard) thresholding wavelet estimator of m is defined by

$$\hat{g}(t) = \sum_{k=-N+1}^{N2^J-1} \hat{s}_{Jk}\phi_{Jk}(t) + \sum_{j=0}^{q}\sum_{k=-N+1}^{N2^{J+j}-1} \hat{d}_{jk}\, I\big(|\hat{d}_{jk}| > \delta_j\big)\psi_{jk}(t). \tag{7.181}$$

The constants J, q and δ_j are called the decomposition level, smoothing parameter and threshold, respectively, and can be chosen quite freely except for some minimal asymptotic requirements such as $\delta_j \to 0$ (with rates in a certain range), $q \to \infty$, etc. The decomposition level J may also tend to infinity, but a reasonable assumption is that $2^J = o(n)$. The reason is that the lowest resolution level which is of the order $O(2^{-J})$ should tend to zero at a slower rate than the distance n^{-1} between successive observational time points. This requirement corresponds to letting the length of the window of a kernel estimator tend to zero at a slower rate than n^{-1}. More specifically, $N2^J t \in [0, N]$ if and only if $0 \le t \le 2^{-J}$, so that we need $n^{-1} = o(2^{-J})$.

The question of interest is now how to choose the constants J, q and δ_j optimally for a given data set. An asymptotic answer is given, at least partially, in Beran and Shumeyko (2012a) (also see Li and Xiao 2007). The solution consists of an asymptotic expression for the integrated mean squared error $MISE = \int E[(\hat{m}(t) - m(t))^2]\,dt$ that can be minimized. The result depends on the differentiability of m, the number m_ψ of vanishing moments and further regularity properties of the mother wavelet ψ, and on the long-memory parameter d. A specific assumption used in Beran and Shumeyko (2012a) is a uniform Hölder condition with exponent $1/2$, i.e.

$$\big|\psi(x) - \psi(y)\big| \le C|x - y|^{1/2}, \quad \forall x, y \in [0, N]. \tag{7.182}$$

This is, however, not necessary since analogous results can be derived, for instance, for Haar wavelets.

In a first step, it can be shown that minimization with respect to J, q and $\{\delta_j\}$ yields the following optimal order of the *MISE*:

Theorem 7.31 *Suppose that $m \in C^r[0,1]$, $m^{(r)}(t) \neq 0$ for a non-zero set (w.r.t. Lebesgue measure), the process ε_i is Gaussian with covariance structure $\gamma(k) = E(e_i e_{i+k}) \sim C_\gamma |k|^{2d-1}$, and ψ is such that $M_\psi = r$. Then, minimizing the MISE with respect to J, q and $\{\delta_j\}$ yields the optimal order*

$$IMSE_{\text{opt}} = O\big(n^{-\frac{2r\alpha}{2r+\alpha}}\big) \tag{7.183}$$

where $\alpha = 1 - 2d$.

Since only the rate is given, Theorem 7.31 is not directly applicable in practice. Instead, an expression for the *IMSE* including all relevant constants is required. Moreover, the trend function (or its derivatives) should be allowed to have at least a finite number of jumps.

It turns out that the optimal order can be achieved without thresholding, i.e. setting $\delta_j = 0$ for all j. Using no thresholding simplifies asymptotic calculations. A detailed analysis of the *IMSE* yields the following optimal values of J and q.

Theorem 7.32 *Under the assumptions of the previous theorem and thresholds*

$$\delta_j = 0 \quad (0 \le j \le q),$$

the following holds: *Let*

$$C_\phi^2 = C_\gamma \int_0^N \int_0^N |x-y|^{-\alpha}\, \phi(x)\, \phi(y)\, dx\, dy, \tag{7.184}$$

$$C_\psi^2 = C_\gamma \int_0^N \int_0^N |x-y|^{-\alpha}\, \psi(x)\, \psi(y)\, dx\, dy. \tag{7.185}$$

(i) *If* $(2^\alpha - 1)C_\phi^2 > C_\psi^2$, *then the asymptotic IMSE is minimized by decomposition levels* J^* *satisfying* $2^{J^*} = o(n^{\frac{\alpha}{2r+\alpha}})$ *and smoothing parameters*

$$q^* = \left\lfloor \frac{\alpha}{2r+\alpha} \log_2 n + C_\psi^* \right\rfloor - J^* \tag{7.186}$$

where $\log_2$ *denotes logarithm to the base* 2. *The optimal IMSE is of the form*

$$MISE = A_1 A_2 \cdot n^{-\frac{2r\alpha}{2r+\alpha}} + o\big(n^{-\frac{2r\alpha}{2r+\alpha}}\big) \tag{7.187}$$

with constants A_1, A_2 *defined explicitly as functions of* d, *and the wavelet functions* (*see Beran and Shumeyko* 2012a).

(ii) *If* $(2^\alpha - 1)C_\phi^2 < C_\psi^2$, *then minimizing the asymptotic IMSE with respect to* J *and* q *yields*

$$\hat{g}(t) = \sum_{k=-N+1}^{N2^{J^*}-1} \hat{s}_{J^*k}\, \phi_{J^*k}(t), \tag{7.188}$$

with

$$J^* = \left\lfloor \frac{\alpha}{2r+\alpha} \log_2 n + C_\phi^* \right\rfloor + 1 \tag{7.189}$$

and C_ϕ^* *defined explicitly as a function of* d, *and the wavelet functions* (*see Beran and Shumeyko* 2012a). *The optimal IMSE is of the form*

$$IMSE = A_3 A_2 \cdot n^{-\frac{2r\alpha}{2r+\alpha}} + o\big(n^{-\frac{2r\alpha}{2r+\alpha}}\big), \tag{7.190}$$

where again A_1, A_2 *can be given explicitly.*

This result establishes an explicit asymptotic expression (and not just the order) for optimal choices of J^* and q^*, for the case where g is sufficiently smooth and when a wavelet basis is used that matches at least this degree of smoothness. Most interesting is part (i) where the optimal estimator does not contain any mother wavelets. Thus, smoothing is done solely by refining the resolution level J^* in the father wavelet decomposition. The optimal choice is a logarithmic increase of J^* with constants as given in (7.189).

If jumps in the function g are expected, then the same asymptotic formula for the *MISE* holds, when essentially using the same rules in this theorem; however, adding thresholded mother wavelet components to capture local disturbances. Thus, consider

$$\hat{g}(t) = \sum_{k=-N+1}^{N2^J-1} \hat{s}_{Jk}\phi_{Jk}(t) + \sum_{j=0}^{q}\sum_{k=-N+1}^{N2^{J+j}-1} \hat{d}_{jk}\, I\big(|\hat{d}_{jk}| > \delta_j\big)\psi_{jk}(t). \tag{7.191}$$

Then the following holds.

Theorem 7.33 *Suppose that $g^{(r)}$ exists on $[0,1]$ except for at most a finite number of points, and, where it exists, it is piecewise continuous and bounded. Furthermore, assume that $supp(g^{(r)})$ has positive Lebesgue measure, $M_\psi = r$ and the process e_i is Gaussian with long memory as specified above. Then the following holds*:

(i) *If $(2^\alpha - 1)C_\phi^2 > C_\psi^2$, J is such that $2^J = o(n^{\frac{\alpha}{2r+\alpha}})$, $q = \lfloor \log_2 n \rfloor - J$, q^* is defined by (7.186), and δ_j is such that for $0 \le j \le q^*$*

$$\delta_j = 0 \tag{7.192}$$

and for $q^* < j \le q$

$$2^{J+j}\delta_j^2 \to 0, 2^{(J+j)(2r+1)}\delta_j^2 \to \infty, \qquad \delta_j^2 \ge \frac{4eC_\psi^2 N^{-1+\alpha}(\ln n)^2}{n^\alpha 2^{(J+j)(1-\alpha)}}, \tag{7.193}$$

then (7.187) holds.

(ii) *If $(2^\alpha - 1)C_\phi^2 < C_\psi^2$, $J = J^*$ with J^* defined by (7.189), $q = \lfloor \log_2 n \rfloor - J$ and δ_j such that*

$$\begin{aligned} &2^{J+j}\delta_j^2 \to 0, 2^{(J+j)(2r+1)}\delta_j^2 \to \infty, \\ &\delta_j^2 \ge \frac{4eC_\psi^2 N^{-1+\alpha}(\ln n)^2}{n^\alpha 2^{(J+j)(1-\alpha)}} \quad (0 \le j \le q), \end{aligned} \tag{7.194}$$

then (7.190) holds.

7.5.2.2 Convergence in Besov Classes

An alternative approach to convergence rates of wavelet estimators in the long-memory context was initiated by Wang (1996). Assume that the error sequence e_i

is Gaussian with covariance function $\gamma(k) \sim c_\gamma k^{2d-1}$, $d \in (0, 1/2)$. As before, set $\alpha = 1 - 2d$. Then, in continuous time, a model that is analogous to $Y_i = m(t_i) + e_i$ discussed above is given by

$$dY(t) = m(t)\,dt + \varepsilon^\alpha\,dB_H(t), \tag{7.195}$$

where $B_H(t)$ ($t \in [0, 1]$) is a standard fractional Brownian motion (fBm) with Hurst index $H = d + 1/2$, and $\varepsilon = n^{-1/2}$ is the "noise level".

Recall that the function $m(t)$ can be expanded as

$$m(t) = \sum_{k=-\infty}^{\infty} \alpha_{jk}\phi_{Jk}(t) + \sum_{j\geq J}\sum_{k=0}^{\infty} \beta_{jk}\psi_{jk}(t).$$

Equivalently, we may write

$$m(t) = \alpha_{00}\phi_{00}(t) + \sum_{j\geq 0}\sum_{k=0}^{\infty} \beta_{jk}\psi_{jk}(t)$$

where $\phi_{00}(t)$ is a suitable father wavelet. To characterize properties of m, one considers the so-called Besov spaces, characterised by the behaviour of the wavelet coefficients as follows:

Definition 7.8 Assume that $m \in L^\lambda([0, 1])$. We say that m belongs to the Besov space $\mathscr{B}^r_{\lambda,s}([0, 1])$ if

$$\sum_{j\geq 0} 2^{j(r+1/2-1/\lambda)s}\left[\sum_{0\leq k\leq 2^j} |\beta_{jk}|^\lambda\right]^{s/\lambda} < \infty. \tag{7.196}$$

The parameter r can be thought of as related to the number of derivatives of m. With different values of λ and s, Besov spaces capture a variety of smoothness features in a function, including spatially inhomogeneous behaviour.

The wavelet estimator is constructed similarly to (7.181):

$$\hat{m}(t) = \hat{\alpha}_{00}\phi_{00}(t) + \sum_{j=0}^{J}\sum_{k=0}^{2^j-1} \hat{\beta}_{jk}1\big(|\hat{\beta}_{jk}| > \delta_j\big)\psi_{jk}(t),$$

where in the continuous time model (7.195) we set

$$\hat{\beta}_{jk} := \hat{\beta}^C_{jk} := \int \psi_{jk}(t)\,dY_t. \tag{7.197}$$

Of course, in the original model we have to take instead

$$\hat{\beta}_{jk} := \hat{\beta}^D_{jk} := \frac{1}{n}\sum_{i=1}^{n} \psi_{jk}(t_i)Y_i. \tag{7.198}$$

The tuning parameters J and δ_j are chosen as follows:

- *Fine resolution level* J:

$$2^J = \left(\frac{n}{\log n}\right)^{\alpha} = \left(\frac{n}{\log n}\right)^{1-2d}. \tag{7.199}$$

- *Threshold*: The threshold value $\delta = \delta_j$ has three input parameters and is written as

$$\delta_j = \eta \sigma_j c_n \tag{7.200}$$

  - η: $\eta > \sqrt{8\alpha}\sqrt{2 \vee p}$;
  - σ_j: a level-dependent scaling factor

$$\sigma_j = \tau 2^{-j(1-\alpha)/2}, \tag{7.201}$$

$$\tau^2 = (1-\alpha/2)(1-\alpha)\int_0^1 \int_0^1 \psi(u)\psi(v)|u-v|^{-\alpha}\,du\,dv; \tag{7.202}$$

  - c_n: a sample size-dependent scaling factor

$$c_n = (\log n)^{\frac{1}{2}} n^{-\frac{\alpha}{2}}. \tag{7.203}$$

The following comments have to be made here. First, in the definition of η, we have a new parameter p that is connected to the loss function we would like to use. Specifically, let

$$\|f-g\|_\nu^\nu = \int \left|f(t)-g(t)\right|^\nu dt$$

be the νth norm. Then we will measure accuracy of the estimator $\hat{m}$ by computing

$$E\left(\|\hat{m}-m\|_\nu^\nu\right).$$

Clearly, if $\nu = 2$, this definition agrees with the IMSE, as considered in Theorem 7.31. The value of σ_j comes from

$$\sigma_j^2 = \operatorname{var}\left(\int \psi_{jk}(t)\,dB_H(t)\right).$$

Furthermore, the parameter τ in (7.202) is chosen for the continuous model (7.195). For the original discrete time model, the parameter should be changed to

$$\tau^2 = c_f \int_0^1 \int_0^1 \psi(u)\psi(v)|u-v|^{-\alpha}\,du\,dv.$$

We note that the estimator is adaptive with respect to the smoothness class as our tuning paradigm does not depend on r.

The following result was proven in Kulik and Raimondo (2009a), see also Wang (1996), Wang (1997), Johnstone and Silverman (1997), Johnstone (1999) and Li and Xiao (2007).

Theorem 7.34 *Consider the continuous time model (7.195) with $\varepsilon = n^{-1/2}$, and the wavelet estimator with (7.199), (7.200), (7.201), (7.202) and (7.203). Assume $p > 1$ and $m \in \mathscr{B}^r_{\lambda,s}$ with $r \geq \frac{1}{\lambda}$. There exists a constant $C > 0$ such that for all $n \geq 0$,*

$$E\left(\|\hat{m} - m\|_v^v\right) \leq C\left(\frac{(\log n)^{\frac{1}{\alpha}}}{n}\right)^{\gamma},$$

with

$$\gamma = \frac{vr\alpha}{2r+\alpha} \quad \text{if } r \geq \frac{\alpha}{2}\left(\frac{v}{\lambda} - 1\right), \tag{7.204}$$

$$r - \left(\frac{1}{\lambda} - \frac{1}{v}\right)_+ > \frac{r}{2r+\alpha}, \tag{7.205}$$

$$\gamma = \frac{\alpha v(r - \frac{1}{\lambda} + \frac{1}{v})}{2(r - \frac{1}{\lambda} + \frac{\alpha}{2})} \quad \text{if } \frac{1}{\lambda} < r < \frac{\alpha}{2}\left(\frac{v}{\lambda} - 1\right). \tag{7.206}$$

The proof of this result is based on the so-called maxiset theorem, see Kerkyacharian and Picard (2000). In particular, the following estimates are crucial. First, $E(\hat{\beta}_{jk}) = \beta_{jk}$ and

$$\operatorname{var}(\hat{\beta}_{jk}) = \operatorname{var}\left(\varepsilon^{\alpha} \int \psi_{\kappa}(t)\, dB_H(t)\right) = n^{-\alpha} 2^{-j(1-\alpha)} \tau^2 \leq C\sigma_j^2 c_n^2.$$

Since the random variables $\hat{\beta}_{jk} - \beta_{jk}$ are Gaussian, we have the following large deviations inequality

$$P\left(|\hat{\beta}_{jk} - \beta_{jk}| > \eta\sigma_j c_n/2\right) \leq \exp\left(-\log n \frac{\eta^2}{8}\right) \leq C\left(c_n^{2p} \wedge c_n^4\right) \tag{7.207}$$

provided $\eta > \sqrt{8\alpha}\sqrt{p \vee 2}$.

The two rate regimes (7.204) and (7.206) are referred as the 'dense' and 'sparse' phases (see, e.g. Kerkyacharian and Picard 2000 in the i.i.d. case). The result above shows that the boundary region $r = \frac{\alpha}{2}(\frac{p}{\lambda} - 1)$ depends on the LRD index α, and the sparse region is smaller for dependent data. In other words, some inhomogeneous properties of the trend function are "hidden" in the LRD noise. We note further that the condition $p > \frac{2}{\alpha} + \lambda$ is required for the sparse regime to be visible. In particular, if $p = 2$ then there is no sparse region and the rate results agree (up to a logarithmic term) with the result in Theorem 7.31.

7.5.3 Random Design

In this part, we are interested in estimating the conditional mean function $m(\cdot)$ in the heteroskedastic model

$$Y_i = m(X_i) + \sigma(X_i)e_i \quad (i = 1, \ldots, n). \tag{7.208}$$

Again, the rates of convergence will be analysed using Besov classes, although in the random-design context we cannot change this model to a continuous set-up as we did before. Furthermore, the fact that we consider random design has to be addressed appropriately. This can be done using the so-called *warped wavelets*. The wavelet expansion of $m(t)$ is replaced by

$$m(x) = \alpha_{0,0}\phi_{00}\big(F(x)\big) + \sum_{j\geq 0}\sum_{k=0}^{\infty}\beta_{jk}\psi_{jk}\big(F(x)\big), \tag{7.209}$$

with

$$\beta_{jk} = \int_0^1 m(x)p(x)\psi_{jk}\big(F(x)\big)\,dx, \tag{7.210}$$

and $F(\cdot)$, $p = F'$ being a cumulative distribution and density function of X_1, respectively.

The partially adaptive wavelet estimator we are going to consider is

$$\hat{m}(t) = \hat{\alpha}_{00}\phi_{00}\big(F(t)\big) + \sum_{j=0}^{J}\sum_{k=0}^{2^j-1}\hat{\beta}_{jk}1\big(|\hat{\beta}_{jk}| \geq \delta_j\big)\psi_{jk}\big(F(t)\big), \tag{7.211}$$

where

$$\hat{\alpha}_{00} := \frac{1}{n}\sum_{i=1}^{n}\phi_{00}\big(F(X_i)\big)Y_i, \qquad \hat{\beta}_{jk} := \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}\big(F(X_i)\big)Y_i. \tag{7.212}$$

The highest resolution level is chosen as

$$2^J \sim \frac{n}{\log n}.$$

The theoretical level-dependent threshold parameter is set to be

$$\delta_j = \tau_0\left(\frac{\log n}{\sqrt{n}} \vee 1\big\{E\big(\psi_{jk}\big(F(X_1)\big)\sigma(X_1)\big) \neq 0\big\}\frac{(\log n)^{1/2}}{n^{\alpha/2}}\right)$$

where τ_0 is *large enough* and $\alpha = 1 - 2d$. We note the significant difference between fixed and random design. The choice of the highest resolution level J in the case

of a random design does not involve LRD. Furthermore, in most regular cases the threshold δ_j does not depend on α. Indeed, we have

$$E\big[\psi_{jk}\big(F(X_1)\big)\sigma(X_1)\big] = \int \psi_{jk}(u)\sigma\big(F^{-1}(u)\big)\,du.$$

Note first that if $\sigma(\cdot) \equiv \sigma$, then the above integral vanishes. Furthermore, this is also the case if $\sigma(\cdot)$ has polynomial-like behaviour and appropriately regular wavelets are used. Consequently, in most practical cases the parameters of the wavelet estimator can be tuned without knowledge of α.

Since we deal with warped wavelets, we have to consider the following weighted norm

$$\|f-g\|^{\nu}_{L^{\nu}(p)} = \left(\int \big|f(x)-g(x)\big|^{\nu} p(x)\,dx\right).$$

Using the notation

$$\alpha_D := \frac{2r}{2r+1}, \qquad \alpha_S := \frac{2(r-(\frac{1}{\lambda}-\frac{1}{\nu}))}{2(r-\frac{1}{\lambda})+1}, \tag{7.213}$$

the following rates of convergence can be derived (Kulik and Raimondo 2009b):

Theorem 7.35 *Consider the random-design regression model* (7.208) *such that* X_i *are i.i.d. and* e_i *is a long-range dependent Gaussian sequence such that* $\gamma_e(k) \sim c_\gamma k^{2d-1}$. *Both sequences are assumed to be independent from each other. Assume furthermore that* $m \circ F^{-1} \in \mathscr{B}^r_{\lambda,s}([0,1]), \lambda \geq 1$, *where* $r > \max\{\frac{1}{\lambda}, \frac{1}{2}\}$. *Then*

$$E\big(\|\hat{m}-m\|^{\nu}_{L^{\nu}(p)}\big) \leq Cn^{-\frac{\nu}{2}\gamma}(\log n)^{\kappa},$$

where

$$\gamma = \begin{cases} \alpha_D & \text{if } \alpha > \alpha_D \text{ and } r > \frac{\nu-\pi}{2\pi}, \text{ dense phase;} \\ \alpha_S & \text{if } \alpha > \alpha_S \text{ and } \frac{1}{\pi} < r < \frac{p-\pi}{2\pi}, \text{ sparse phase;} \\ \alpha & \text{if } \alpha \leq \min(\alpha_S, \alpha_D), \text{ LRD phase,} \end{cases}$$

α_S, α_D *are given in* (7.213), *and* $\kappa > 0$. *If* $\alpha = 1$, *then the LRD phase is not relevant.*

The proof is based on the M/L technique, as discussed before in the context of random-design regression. The main tool is a large deviation inequality for LRD processes. Informally speaking, LRD appears at low resolution levels only and is suppressed by the additional threshold term.

Furthermore, as in the case of kernel estimators, the rates of convergence improve when once considers estimation of the shape function $m^*(t) = m(t) - E(m(X_1))$.

To get full adaptiveness $F(\cdot)$ has to be replaced by its empirical counterpart $F_n(\cdot)$. The results of Theorem 7.35 continue to hold. However, the highest resolution level must be chosen according to $2^J \sim \sqrt{n/\log n}$.

The results in Theorem 7.35 are optimal. It other words, it is not possible to find estimators that achieve better rates of convergence.

7.6 Estimation of Time Dependent Distribution Functions and Quantiles

Limit theorems for empirical quantiles of stationary long-memory processes, and their direct application to quantile estimation have been discussed in Sect. 4.8.2.1. Here we consider the more complicated situation where quantiles may change with time. The approach introduced in the following is nonparametric.

Consider time series observations $Y_1, Y_2, \ldots, Y_n$ such that $Y_i = G(Z_i, t_i)$ where $t_i = i/n$ are rescaled times and $\{Z_i,\ i = 1, 2, \ldots\}$ is a zero mean stationary Gaussian process with long-memory. The function $G(x, \cdot)$ is assumed to be an unknown square integrable function (with respect to the $N(0, 1)$ density). As for the Gaussian process Z_i, we assume that

$$cov(Z_i, Z_{i+k}) = \gamma(k) \sim C|k|^{2H-2}, \quad \text{as } |k| \to \infty,$$

H being the long-memory parameter with $1/2 < H < 1$ and C is a positive constant. For $y \in \mathbb{R}$, $t_i = i/n$, define the cumulative distribution function of Y at rescaled time t_i to be

$$F_{t_i}(y) = P(Y_i \le y).$$

For simplicity of arguments, let F_t, $t \in (0, 1)$ be continuous with a probability density function f_t defined by

$$f_t(y) = \frac{\partial}{\partial y} F_t(y).$$

The problem is the nonparametric estimation of $F_t(\cdot)$, $t \in (0, 1)$ and consequently the estimation of the α-quantile $(0 < \alpha < 1)$

$$\theta_t(\alpha) = \inf_y \{y | F_t(y) \ge \alpha\},$$

and deriving asymptotic confidence bands for these functions. The results summarized in this section can be found in Ghosh et al. (1997). As for applicability of these ideas, estimation and prediction of the time dependent probability function $F_t(y)$ can be of practical relevance in various situations. For instance, if Y_i is precipitation at time i (rescaled time t_i), then $1 - F_t(y)$ is the probability that the amount of rain at time t will exceed a previously specified level y, having implications for regions where heavy rainfall is the primary factor leading to floods. Equivalently, quantile functions may be considered. Very low values of $\theta_t(\alpha)$ for low α may be indicative of a drought, also having serious implications for agriculture.

The time dependent Gaussian subordination model considered here is a model for processes that are nonstationary in the sense that the marginal distribution function may change with time. Moreover, the distribution may be Gaussian or non-Gaussian. Some simple examples are:

(i) $Y_i = \mu(t_i) + \sigma(t_i) Z_i$, where μ and σ are real-valued functions;
(ii) $Y_i = \mu_1(t_i) Z_i^2 + \mu_2(t_i) Z_i^3$ where μ_1 and μ_2 are real-valued functions;

(iii) $Y_i = 1\{Z_i < z\} - P(Z_i < z)$, $z \in \mathbb{R}$, etc.

Let $K(u)$, $u \in (-1, 1)$ be a symmetric probability density function on $(-1, 1)$. Also let $b_n = b$ be a sequence of bandwidths such that $b \to 0$ and $nb^3 \to \infty$ as $n \to \infty$. Define the Priestley–Chao estimator

$$\widehat{F}_t(y) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) I_i(y)$$

where

$$I_i(y) = 1 \quad \text{if } Y_i \le y \quad \text{and} \quad I_i(y) = 0 \quad \text{otherwise.}$$

Since the indicator function $I_i(y)$ is a function of Y_i, it is also Gaussian subordinated. We assume that the following Hermite polynomial expansion holds

$$I_i(y) - P(Y_i \le y) = \sum_{l=m}^{\infty} \frac{c_l(t_i, y)}{l!} H_l(Z_i).$$

In the above expansion, m is the Hermite rank of G, the functions c_l are the Hermite coefficients, and H_l denotes the Hermite polynomial of degree l. Note that when $H > 1 - 1/(2m)$, $I_i(y) - P(Y_i \le y)$, $i = 1, 2, \ldots$ will have long-memory.

Theorem 7.36 *Under the conditions stated above for $H > 1 - 1/(2m)$ and under further regularity conditions on the Hermite coefficients and assuming that the distribution function $F_t(y)$ is twice differentiable with respect to t, for fixed t and y and as $n \to \infty$, $\widehat{F}_t(y)$ will have the following asymptotic properties*:

$$\begin{aligned}
\operatorname{Bias}\big(\widehat{F}_t(y)\big) &= \frac{b^2}{2} A(t, y) + o\big(b^2\big), \\
\operatorname{Var}\big(\widehat{F}_t(y)\big) &= (nb)^{m(2H-2)} B(t, y) \\
&\quad + o\big((nb)^{m(2H-2)}\big), \\
MSE\big(\widehat{F}_t(y)\big) &= A^2(t, y) b^4 + B(t, y)(nb)^{m(2H-2)} \\
&\quad + o\big(\max\big(b^4, (nb)^{m(2H-2)}\big)\big)
\end{aligned}$$

where

$$A(t, y) = \frac{1}{2} \frac{\partial^2}{\partial t^2} F_t(y) \int_{-1}^{1} u^2 K(u)\, du,$$

$$B(t, y) = C^m \frac{c_m^2(t, y)}{m!} \int_{-1}^{1} \int_{-1}^{1} K(u) K(v) |u - v|^{m(2H-2)}\, du\, dv.$$

Proof We have,

$$E\big[\widehat{F}_t(y)\big] = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) E\big[I_i(y)\big] = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) F_{t_i}(y).$$

The proof for bias of $\widehat{F}_t(y)$ then follows by a Taylor series expansion of $F_{t_i}(y)$ around t and by noting that as $n \to \infty$,

$$\left| \frac{1}{nb}\sum_{i=1}^{n} \left(\frac{t_i - t}{b}\right)^p K\left(\frac{t_i - t}{b}\right) - \int_{-1}^{1} u^p K(u)\, du \right| = O\left(\frac{1}{nb}\right)$$

where p is a positive integer, and also $O(\frac{1}{nb}) = o(b^2)$ since $nb^3 \to \infty$. Moreover, since K is a symmetric probability density function, $\int_{-1}^{1} u^p K(u)\, du$ equals 1 when $p = 0$ and equals 0 when p is odd.

As for the variance, since $cov[H_{l_1}(Z_i), H_{l_2}(Z_j)] = 0$ if $l_1 \neq l_2$ and equals $l![\gamma(i-j)]^l$ if $l_1 = l_2 = l$,

$$\begin{aligned}
\operatorname{var}\big(\widehat{F}_t(y)\big) &= \frac{1}{(nb)^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{t_i - t}{b}\right) K\left(\frac{t_j - t}{b}\right) cov\big[G(Z_i, t_j), G(Z_j, t_j)\big] \\
&= \frac{1}{(nb)^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{t_i - t}{b}\right) K\left(\frac{t_j - t}{b}\right) \sum_{l=m}^{\infty} \frac{c_l(t_i)c_l(t_j)}{l!}\big[\gamma(i-j)\big]^l \\
&\sim \frac{1}{(nb)^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} K\left(\frac{t_i - t}{b}\right) K\left(\frac{t_j - t}{b}\right) \sum_{l=m}^{\infty} \frac{c_l(t_i)c_l(t_j)}{l!} C^l |i-j|^{l(2H-2)}.
\end{aligned}$$

The last step follows since $\sum_{i,j} |i-j|^{l(2H-2)}$ diverges as $n \to \infty$. Now using a one-term Taylor series expansion of $c_l(t_i)$ and $c_l(t_j)$ around t and due to the convergence of the Riemann sums involving the kernel K, the expression for the variance follows. The formula for the mean squared error (MSE) follows from definition. □

By differentiating the asymptotic expression for the MSE with respect to b, a formula for an optimal bandwidth for estimating $F_t(y)$ can be derived as

$$b_t^{(\mathrm{opt})}(y) = Q_t(y) \times n^{m(2H-2)/(4+m(2-2H))}$$

where

$$Q_t(y) = \left[\frac{m(2-2H)B(t,y)}{4A^2(t,y)}\right]^{1/[4+m(2-2H)]}.$$

Thus, for instance, when $m = 1$ and $H \approx 1/2$, $b_t^{(\mathrm{opt})}(y) \propto n^{-1/5}$. As H moves away from 0.5 and approaches 1, $b_t^{(\mathrm{opt})}(y)$ becomes large as well. This has to do with

the fact that long memory creates an apparent smoothness in the data as a result of which larger bandwidths suffice for optimum smoothing.

The quantile function $\theta_t(\alpha)$ for a given α can be estimated by inverting the estimated distribution function $\widehat{F}_t(y),\ y \in \mathbb{R}$ as follows:

$$\hat{\theta}_t(\alpha) = \inf_y \{y | \widehat{F}_t(y) \geq \alpha\}.$$

It turns out that the estimator $\hat{\theta}_t$ inherits the asymptotic properties of $\widehat{F}_t$. Specifically, we have the following result:

Theorem 7.37 *Let $\theta_t(\alpha)$ be unique and $f_t(\theta_t(\alpha)) > 0$. Then,*

$$\begin{aligned}
\text{Bias}\big(\hat{\theta}_t(\alpha)\big) &= \frac{b^2}{f_t(\theta_t(\alpha))} A\big(t, \theta_t(\alpha)\big) + o\big(b^2\big), \\
\text{Var}\big(\hat{\theta}_t(\alpha)\big) &= (nb)^{m(2H-2)} \frac{B(t, \theta_t(\alpha))}{f_t^2(\theta_t(\alpha))} + o\big((nb)^{m(2H-2)}\big), \\
MSE\big(\hat{\theta}_t(\alpha)\big) &= \left[\frac{A^2(t, \theta_t(\alpha))}{f_t^2(\theta_t(\alpha))} b^4 + \frac{B(t, \theta_t(\alpha))}{f_t^2(\theta_t(\alpha))} (nb)^{m(2H-2)}\right] \\
&\quad + o\big(\max\big(b^4, (nb)^{m(2H-2)}\big)\big).
\end{aligned}$$

Proof For additional information, refer to Rao (1973, Chap. 6f.2) and Serfling (1980, Chap. 2.3). First of all, as $n \to \infty$, $\hat{\theta}_t(\alpha) \to \theta_t(\alpha)$ in probability. Secondly, as in Pollard (1984, p. 98),

$$(nb)^{m(2-2H)}\big[\hat{\theta}_t(\alpha) - \theta_t(\alpha)\big] = \frac{-(nb)^{m(2-2H)}[\widehat{F}_t(\hat{\theta}_t(\alpha)) - F_t(\hat{\theta}_t(\alpha))] - o_p(1)}{f_t(\theta_t(\alpha)) + o_p(1)}.$$

The result follows from the continuous mapping theorem. □

Remark It is easy to see that the asymptotically optimal local bandwidth that minimizes the leading term in the MSE of $\hat{\theta}_t(\alpha)$ (term inside the square brackets) is the same as the bandwidth needed for the estimation of $F_t(\theta_t(\alpha))$.

Under the condition that the Hermite rank of the function G is equal to 1, we have the following central limit theorem:

Theorem 7.38 *Let $m = 1$.*

(a) *CLT for $\widehat{F}_{t_i}(y)$: Let $y \in \mathbb{R}$, $k \geq 1$ and $t_1^0 < t_2^0 < \cdots < t_k^0$ (with $t_i^0 \in (0, 1)$) be fixed. Define*

$$U_{i,n} = (nb)^{1-H} \frac{[\widehat{F}_{t_i}(y) - F_{t_i}(y) - b^2 A(t_i, y)]}{\sqrt{B(t_i, y)}}, \quad t_i = t_{i_n} = i_n/n$$

with $t_i \to t_i^0$ $(i = 1, 2, \ldots, k)$ *as* $n \to \infty$. *Then as* $n \to \infty$, *the random vector*

$$\mathbf{U}_n = (U_{1,n}, U_{2,n}, \ldots, U_{k,n})^T$$

converges in distribution to $\mathbf{Z}^u = (Z_1^u, Z_2^u, \ldots, Z_k^u)^T$ *where* Z_i^u, $i = 1, 2, \ldots, k$ *are independent and identically distributed standard normal random variables.*

(b) *CLT for* $\hat{\theta}_{t_i}(\alpha)$: *Let* $\alpha \in (0, 1)$ *and* $k \geq 1$ *be fixed, and* t_i^0 *as before. Define*

$$W_{i,n} = (nb)^{1-H} \frac{[\hat{\theta}_{t_i}(\alpha) - \theta_{t_i}(\alpha) - b^2 A(t_i, \theta_{t_i}(\alpha))/f_{t_i}(\theta_{t_i}(\alpha))]}{\sqrt{B(t_i, \theta_{t_i}(\alpha))}/f_{t_i}(\theta_{t_i}(\alpha))},$$

$$t_i = t_{i_n} = i_n/n$$

with t_{i_n} *as above. Then as* $n \to \infty$, *the random vector*

$$\mathbf{W}_n = (W_{1,n}, W_{2,n}, \ldots, W_{k,n})^T$$

converges in distribution to $\mathbf{Z}^w = (Z_1^w, Z_2^w, \ldots, Z_k^w)^T$ *where* Z_i^w, $i = 1, 2, \ldots, k$ *are independent and identically distributed standard normal random variables.*

Proof (a) Due to Theorem 7.36, as $n \to \infty$, for each $t \in (0, 1)$

$$(nb)^{1-H} \left| \widehat{F}_t(y) - F_t(y) - b^2 A(t, y) - R_n(t, y) \right| \to 0$$

in probability, where

$$R_n(t, y) = (nb)^{-1} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) c_1(t_i, y) Z_i.$$

Note that $(nb)^{1-H} R_n(t, y)$ has a normal distribution because it is a linear combination of standard normal random variables that are also jointly normal. Also, $cov((nb)^{1-H} \widehat{F}_t(y), (nb)^{1-H} \widehat{F}_s(y))$ for $t \neq s$ converges to zero in probability. The result follows by considering the sequence of random vectors $\mathbf{U}_n$ and Theorem 7.36(i) in Csörgő and Mielniczuk (1995a).

(b) The proof follows from (a) above and the arguments of Theorem 7.37(b). □

7.7 Partial Linear Models

A partial linear model is a semiparametric regression model containing a nonparametric as well as a linear parametric regression component. An example is as follows:

$$y(i) = \mathbf{x}^T(i)\beta + \mu(t_i) + \varepsilon(i)$$

where $y(i),\ i = 1, 2, \ldots, n$ is an observation on the dependent variable y, $\mathbf{x}^T(i)$ is a (row) vector of explanatory variables

$$\mathbf{x}^T(i) = \big(x_1(i), x_2(i), \ldots, x_p(i)\big), \quad p \geq 1,$$

β is a (column) vector of regression parameters

$$\beta^T = (\beta_1, \beta_2, \ldots, \beta_p)$$

and $t_i = i/n$ is rescaled time. The nonparametric component μ is an unknown but smooth function in $C^2[0, 1]$ whereas $\varepsilon(i)$ is the error term with zero mean. Of special interest is the case when $\varepsilon(i)$ is a stationary long-memory process. Specifically, let $\varepsilon(i)$ have a covariance function γ_ε and a spectral density f_ε

$$\gamma_\varepsilon(k) = Cov\big(\varepsilon(j), \varepsilon(j+k)\big) = \int_{-\pi}^{\pi} \exp(ik\lambda) f_\varepsilon(\lambda)\, d\lambda,$$
$$f_\varepsilon(\lambda) \sim c_\varepsilon |\lambda|^{-2d_\varepsilon} \quad \text{as } \lambda \to 0$$

where as usual $\sim$ means that the left-hand side divided by the right-hand side converges to one, c_ε is a positive constant and $0 \leq d_\varepsilon < \frac{1}{2}$. Let $E(\varepsilon\varepsilon^T) = \Gamma_{\varepsilon,n} = \Gamma_\varepsilon = [\gamma_\varepsilon(i-j)]_{i,j=1,2,\ldots,n}$. The uncorrelated case, namely when β and μ are unknown but the errors are uncorrelated, is considered in Speckman (1988). He suggests a $\sqrt{n}$-consistent estimator for β under the assumption that also the explanatory variables contain a rough component. Beran and Ghosh (1998) examine Speckman's method of estimation under long-memory in the errors. As it turns out, even under long-memory, a $\sqrt{n}$-rate of convergence of the slope estimates can be achieved. In this section, we take a closer look at some of these results.

To start with, we set our notations: we observe $(\mathbf{x}^T(i), y(i))$ at time points $i = 1, 2, \ldots, n$. Using vector notations, we define

$$\begin{aligned}
\mathbf{x}^T(i) &= \big(x_1(i), x_2(i), \ldots x_p(i)\big), \quad i = 1, 2, \ldots, n,\\
\mathbf{y}^T &= \big(y(1), y(2), \ldots, y(n)\big),\\
\mu^T &= \big(\mu(t_1), \mu(t_2), \ldots, \mu(t_n)\big), \quad t_i = i/n,\\
\varepsilon^T &= (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n).
\end{aligned}$$

Let the $n \times p$ full design matrix be

$$\mathbf{X} = \mathbf{M} + \eta$$

where M is a deterministic matrix of order $n \times p$ and η is a random matrix, its elements being zero mean random variables. The ith row of $\mathbf{X}$ is $\mathbf{x}^T(i)$, the columns of $\mathbf{M}$ are $(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_p)$,

$$\mathbf{m}_j^T = \big(m_j(t_1), m_j(t_2), \ldots, m_j(t_n)\big), \quad j = 1, 2, \ldots, p$$

whereas the ith row of $\mathbf{M}$ is

$$\big(m_1(t_i), m_2(t_i), \ldots, m_p(t_i)\big), \quad i = 1, 2, \ldots, n.$$

The functions $m_j(\cdot)$ are in $C^2[0, 1]$. The columns of the random matrix η are denoted by $\mathbf{e}_j$, i.e.

$$\eta = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$$

where

$$\mathbf{e}_j^T = \big(e_j(1), e_j(2), \ldots, e_j(n)\big), \quad j = 1, 2, \ldots, p,$$

rows are given by

$$\mathbf{e}^T(i) = \big(e_1(i), e_2(i), \ldots, e_p(i)\big).$$

The random "error" terms in $\mathbf{X}$ are assumed to have the following properties: η is independent of ε. As for the covariances,

$$\gamma_{e_j}(k) = Cov\big(e_j(s), e_j(s+k)\big) = \int_{-\pi}^{\pi} \exp(ik\lambda) f_{e_j}(\lambda)\, d\lambda,$$

$$f_{e_j}(\lambda) \sim c_{e_j} |\lambda|^{-2d_{e_j}} \quad \text{as } |\lambda| \to 0$$

where c_{e_j} is a positive constant and $0 \le d_{e_j} < \frac{1}{2}$. Let $\sigma_{\mathbf{e}}(j, l) = Cov(e_j(i), e_l(i))$ so that the $p \times p$ matrix of zero-lag cross-covariances is $E(\mathbf{e}(i)\mathbf{e}^T(i)) = \Gamma_{\mathbf{e}} = [\sigma_{\mathbf{e}}(j, l)]_{j,l=1,2,\ldots,p}$. The partial linear model is then of the form

$$\mathbf{y} = \mathbf{X}\beta + \mu + \varepsilon = \mathbf{M}\beta + \eta\beta + \mu + \varepsilon.$$

In the above formula, $\mathbf{M}\beta + \mu$ is deterministic whereas $\eta\beta + \varepsilon$ is random. The main idea is to smooth the values of $\mathbf{y}$ to obtain an estimate of the deterministic part and consequently an estimate of the error. Similarly, the error in $\mathbf{X}$ can be estimated by detrending the data series containing the values of the explanatory variables. These error estimates are then used in a regression model to recover β. For instance, consider the Nadaraya–Watson kernel (see Gasser et al. 1985)

$$K(t_i, t_j, n, b) = \frac{w(\frac{t_i - t_j}{b})}{n^{-1} \sum_{i=1}^{n} w(\frac{t_l}{b})}$$

and define the kernel matrix

$$\mathbf{K} = \big[K(t_i, t_j, n, b)\big]_{i,j=1,2,\ldots,n}.$$

Here b is a bandwidth satisfying in particular that as $n \to \infty$, $b \to 0$, $nb \to \infty$, and w is a bounded, non-negative, symmetric and piecewise continuous function with support $[-1, 1]$ such that $\int_{-1}^{1} w(s)\, ds = 1$. Additional conditions on b that are

used to prove the asymptotic results concerning the estimated slope are in Beran and Ghosh (1998).

Define the residuals

$$\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{K})\mathbf{X}, \qquad \tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{K})\mathbf{y}.$$

Then the semiparametric regression estimate of the slope parameter β can be given by

$$\hat{\beta} = \left(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}}.$$

In addition to the conditions stated earlier, let, as $n \to \infty$,

$$n\left(\eta^T \eta\right)^{-1} \eta^T \Sigma_\varepsilon \eta \left(\eta^T \eta\right)^{-1} \to \mathbf{A}$$

almost surely, and

$$\sqrt{n}\left(\eta^T \eta\right)^{-1} \eta^T \varepsilon \to N(0, \mathbf{A})$$

in distribution where $N(0, \mathbf{A})$ denotes a p-variate normal distribution with zero mean and covariance matrix $\mathbf{A}$. These conditions ensure that β can be estimated with $\sqrt{n}$-convergence. For sufficient conditions for these to hold, see Sect. 7.2 (and in particular Yajima 1991 and Künsch et al. 1993). Under the conditions stated above, the following asymptotic results can be derived.

Theorem 7.39 *Let* $d_0 = \max_{j=1,\ldots,p} d_{e_j}$. *Then as* $n \to \infty$, *conditionally on* $\mathbf{X}$,

$$E(\hat{\beta}|\mathbf{X}) - \beta = O\left(b^4\right) + O\left((nb)^{d_0 - \frac{1}{2}} b^2\right),$$

$$n \operatorname{Var}(\hat{\beta}|\mathbf{X}) \to \mathbf{A} \quad \textit{almost surely},$$

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, \mathbf{A}) \quad \textit{in distribution}.$$

Note in particular that asymptotically the bias is of a smaller order than the variance. For the proof of the theorem and additional technical conditions on the bandwidth, see Beran and Ghosh (1998). In applications, the covariance matrix $\mathbf{A}$ would have to be estimated. These authors recommend fitting a parametric model $f_\varepsilon(\lambda; \hat{\theta})$ for the spectral density to the residuals $\hat{\varepsilon}(i) = \tilde{y}(i) - \tilde{\mathbf{x}}^T(i)\beta$ and setting $\hat{\Gamma}_\varepsilon = \Gamma_\varepsilon(\hat{\theta})$. For an extension of these results to testing for partial linear models with long memory, see Aneiros-Pérez et al. (2004).

7.8 Inference for Locally Stationary Processes

7.8.1 Introduction

In this short section, we discuss estimation for locally stationary long-memory processes. In the context of weakly dependent processes, the mathematical background

stems from Dahlhaus (1997) (also see, e.g. Priestley 1981 for earlier references). In a long-memory setting, the general idea is that the long-memory parameter is treated as a smooth function of time (that is, the dependence parameter becomes a curve). Specifically, Whitcher and Jensen (2000) propose locally stationary ARFIMA processes. Ghosh et al. (1997) consider subordinated locally stationary Gaussian processes in the context of quantile estimation. Asymptotic theory for estimators of the "dependence curves" is presented in Beran (2009). The results use tools from kernel regression, as discussed before in Sect. 7.4. Roueff and von Sachs (2011) discuss estimation for locally stationary processes using wavelet methods.

The motivation for considering locally stationary processes is the observation that often time series appear to be stationary when one looks at short time periods; however, in the long run, the structure changes. If changes are not abrupt, then such data can be modelled by the so-called locally stationary processes. The general idea is that the probabilistic structure of the process changes smoothly in time such that locally the series are stationary in a first approximation. In engineering, this idea has been used long before exact mathematical definitions of local stationarity were introduced. A systematic mathematical approach was initiated by pioneering contributions of Subba Rao (1970), Hallin (1978) and Priestley (1981), followed by Dahlhaus (1997) who developed a general theory based on an exact definition of locally stationary processes in terms of their spectral representation $X_t = \int e^{it\lambda} A(e^{-i\lambda}; u_{t,n})\, dM_\varepsilon(\lambda)$ where M_ε is the spectral measure of white noise, $u_{t,n} = t/n$ and A depends (smoothly) on rescaled time $u_{t,n}$. More exactly, we have a sequence of processes

$$X_{t,n} = \int_{-\pi}^{\pi} e^{it\lambda} A^0_{t,n}\big(e^{-i\lambda}; \theta(u_{t,n})\big)\, dM_\varepsilon(\lambda) \tag{7.214}$$

with transfer functions $A^0_{t,n}(e^{-i\lambda}; \theta)$ such that

$$\sup_{\lambda\in[-\pi,\pi], t=1,2,\ldots,n} \big|A^0_{t,n}\big(e^{-i\lambda}; \theta(u_{t,n})\big) - A\big(e^{-i\lambda}; \theta(u_{t,n})\big)\big| \le Cn^{-1} \tag{7.215}$$

for all n, some constant C and a certain transfer function $A(e^{-i\lambda}; \theta)$. This definition allows for changes in the linear dependence structure. As an alternative definition that also includes the possibility of changes in the spectral measure $dM_\varepsilon(\cdot)$, Ghosh et al. (1997) and Ghosh and Draghicescu (2002a, 2002b) propose using the concept of subordination, defining $X_{t,n} = G(\zeta_t; u_n)$ where ζ_t is a stationary process and $G(\cdot; u)$ is a smooth function of u. In the following, we discuss inference for processes that are locally stationary in the sense of definition (7.214).

In the context of long-memory processes, changes in the long-memory parameter d are of particular interest. Numerous data examples are reported in the literature where d may be changing in time (see, e.g. Vesilo and Chan 1996; Whitcher and Jensen 2000; Whitcher et al. 2000, 2002; Lavielle and Ludena 2000; Ray and Tsay 2002; Granger and Hyung 2004; Falconer and Fernandez 2007). This motivated Whitcher and Jensen (2000) to consider locally stationary fractional ARIMA (FARIMA) processes. Optimal fitting of parameters in locally stationary

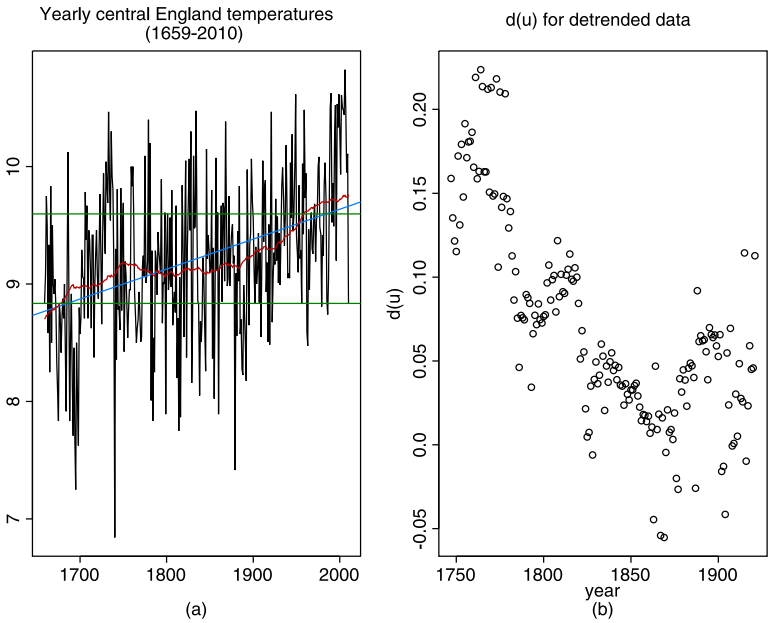

Fig. 7.15 (**a**) Central England temperature series with fitted linear and nonparametric trend function respectively; (**b**) local maximum likelihood estimates of d for detrended series, based on moving blocks of 176 years and a fractional ARIMA$(0, d, 0)$ model

long-memory processes is discussed in Beran (2009). An example is plotted in Figs. 7.15(a)–(b). After subtracting the nonparametric trend (see the nonlinear line in Fig. 7.15(a)), estimated values of d based on moving (overlapping) blocks of 175 years are plotted against the year in the middle of each block. The plot indicates that long memory is stronger for the initial measurements and then declines to a lower level.

7.8.2 Optimal Estimation for Locally Stationary Processes

In the following, we consider a locally stationary long-memory model of the following form. Define a sequence of processes $X_{t,n}$ with a time-varying infinite autoregressive representation given by

$$X_{t,n} = \sum_{j=1}^{\infty} b_{j,n} X_{t-j,n} + \varepsilon_t \tag{7.216}$$

where ε_t are i.i.d. zero-mean random variables with finite variance $\sigma_\varepsilon^2 = \sigma_\varepsilon^2(u_n)$ $(u_n = t/n)$ and coefficients $b_{j,n} = b_j(\theta(u_n))$. For fixed u, it is assumed that $d(u) \in$

$(0, \frac{1}{2})$ and the coefficients are such that

$$b_j\big(\theta(u)\big) \underset{j\to\infty}{\sim} c_b(u) j^{-d(u)-1} < \infty \tag{7.217}$$

$$\frac{\sigma_\varepsilon^2(u)}{2\pi}\left|1 - \sum_{j=1}^{\infty} b_j e^{-ij\lambda}\right|^{-2} \underset{|\lambda|\to 0}{\sim} c_f(u)|\lambda|^{-2d(u)} \tag{7.218}$$

where c_b, c_f are positive constants. Specifically, we may consider a locally stationary fractional ARIMA(p, d, q) process. Then $c_f(u) = \sigma_\varepsilon^2(u)/(2\pi)$ and for $z \in \mathbb{C}$, with $|z| \le 1$ and $z \ne 1$,

$$1 - \sum_{j=1}^{\infty} b_j\big(\theta(u)\big) z^j = \varphi(z; u)\psi^{-1}(z; u)(1 - z)^{d(u)} \tag{7.219}$$

where $\theta(u) = [d(u), \varphi_1(u), \ldots, \varphi_p(u), \psi_1(u), \ldots, \psi_q(u)]^T$,

$$\varphi(z; u) = 1 - \varphi_1(u)z - \cdots - \varphi_p(u)z^p \ne 0 \quad \big(|z| \le 1\big), \tag{7.220}$$

$$\psi(z; u) = 1 + \psi_1(u)z + \cdots + \psi_q(u)z^q \ne 0 \quad \big(|z| \le 1\big). \tag{7.221}$$

Separating σ_ε from the other parameters in the spectral representation, we can write

$$X_{t,n} = \sigma_\varepsilon(u_{t,n}) \int_{-\pi}^{\pi} e^{it\lambda} A_{t,n}^0\big(e^{-i\lambda}; \theta(u_{t,n})\big)\, dM_\varepsilon(\lambda) \tag{7.222}$$

with

$$A_{t,n}^0\big(z; \theta(u)\big) = \frac{\psi(z; u)}{\varphi(z; u)}(1 - z)^{-d(u)}. \tag{7.223}$$

Let $\theta^0(u)$ denote the true parameter function, and $X_{t,n}$ a locally stationary FARIMA process. In general, the shape of $\theta^0(\cdot)$ is unknown. Under smoothness conditions, estimation of $\theta^0(\cdot)$ can be done in a similar manner as regression smoothing. Suppose we would like to estimate θ^0 at a fixed rescaled time point $u_0 \in (0, 1)$. A natural approach is to apply quasi-maximum likelihood estimation based on time points in a small neighbourhood of u_0. Using the Gaussian likelihood, this is essentially equivalent to local minimization of the sum of squared residuals estimated from (7.216). Thus, let $t_0(n) = [nu_0]$, $u_{t_0,n} = t_0(n)/n$. Given a kernel function $K \ge 0$ with $K(-x) = K(x)$, $K(x) = 0$ $(|x| > 1)$ and $\int K(x)\, dx = 1$, a kernel estimate of $\theta^0(u_0)$ minimizes

$$\mathscr{L}_n(\theta) = \sum_{t=t_0-[nb]}^{t_0+[nb]} K\left(\frac{t_0(n) - t}{nb}\right) e_t^2(\theta) \tag{7.224}$$

or solves the equation

$$\dot{\mathscr{L}}_n(\hat{\theta}) = \sum_{t=t_0-[nb]}^{t_0+[nb]} K\left(\frac{t_0(n)-t}{nb}\right)\varepsilon_t^*(\hat{\theta})\dot{\varepsilon}_t^*(\hat{\theta}) = 0 \tag{7.225}$$

where

$$\varepsilon_t^*(\theta) = X_t - \sum_{j=1}^{t-1} b_j(\theta)X_{t-j}, \qquad \dot{\varepsilon}_t^*(\theta) = \frac{\partial}{\partial\theta}\varepsilon_t^*(\theta) = -\sum_{j=1}^{t-1}\dot{b}_j(\theta)X_{t-j} \tag{7.226}$$

are approximations of

$$\varepsilon_t(\theta) = X_t - \sum_{j=1}^{\infty} b_j(\theta)X_{t-j} \tag{7.227}$$

and

$$\dot{\varepsilon}_t(\theta) = -\sum_{j=1}^{\infty}\dot{b}_j(\theta)X_{t-j}, \tag{7.228}$$

respectively, and $\dot{b}_j = \partial/\partial\theta b_j \in \mathbb{R}^{p+q+1}$. The asymptotic distribution of $\hat{\theta}(u_0)$ was derived in Beran (2009) in an analogous manner as for stationary processes. The same result was later also shown to hold for the local Whittle estimator (Palma and Olea 2010).

Theorem 7.40 *Let $X_{t,n}$ be a locally stationary FARIMA process defined by* (7.222) *and* (7.223) *and let $u_0 \in (0,1)$. Moreover, assume that, as n tends to infinity, $b \to 0$ and $nb^3 \to \infty$. Then, under regularity assumptions and moment conditions (see Beran* 2009*), there is a sequence $\hat{\theta}_n$ such that $\dot{\mathscr{L}}_n(\hat{\theta}_n) = 0$ and $\hat{\theta}_n \to \theta^0(u_0)$ in probability. Moreover,*

$$\sqrt{nb}\big(\hat{\theta}_n - E(\hat{\theta}_n)\big) \to_d N(0, V) \tag{7.229}$$

where

$$V = J^{-1}\big(\theta^0\big)\int_{-1}^{1} K^2(x)\,dx \tag{7.230}$$

with

$$J\big(\theta^0\big) = \left[\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{\partial}{\partial\theta_r}\log g\big(\lambda;\theta^0\big)\frac{\partial}{\partial\theta_s}\log g\big(\lambda;\theta^0\big)\,d\lambda\right]_{r,s=1,\ldots,k} \tag{7.231}$$

and $g(\lambda;\theta(u_{t,n})) = |A^0_{t,n}(e^{-i\lambda};\theta(u_{t,n}))|^2$.

Once the estimate of $\theta^0(u^0)$ is given, $\sigma_\varepsilon^2(u_0)$ can be estimated by

$$\hat{\sigma}_\varepsilon^2(u_0) = \sum_{t=t_0-[nb]}^{t_0+[nb]} K\left(\frac{t_0(n)-t}{nb}\right)\left(\varepsilon_t^*(\hat{\theta})\right)^2. \tag{7.232}$$

As in the stationary case, $\hat{\sigma}_\varepsilon^2(u_0)$ is asymptotically independent of $\hat{\theta}$ and the asymptotic distribution of $\hat{\theta}$ does not depend on σ_ε^2.

Example 7.36 Let $X_{t,n}$ be a local fractional ARIMA$(0, d, 0)$ process. Then $J = \pi^2/6$ for any value of $\theta^0(u^0)$. The asymptotic variance of $\sqrt{nb}(\hat{d} - d^0(u_0))$ is therefore nuisance parameter free. If we use, for instance, the rectangular kernel $K(x) = \frac{1}{2}1\{|x| \le 1\}$, then $\int K^2(x)\,dx = \frac{1}{2}$ and

$$V = \frac{6}{\pi^2}\frac{1}{2} = \frac{3}{\pi^2} \approx 0.304. \tag{7.233}$$

The limit theorem cannot be used directly for inference about θ^0 because it refers to the deviation of $\hat{\theta}$ from its expected value. What we would need instead is a result for $\hat{\theta} - \theta^0$. As always in nonparametric smoothing, an asymptotic formula for the bias $E(\hat{\theta}) - \theta^0$ is required. Since the order of the bias is not influenced by the dependence structure, we have $E(\hat{\theta}) - \theta^0 = O(b^2)$. Moreover, in contrast to nonparametric regression smoothing with long-memory errors, the rate of convergence of $\hat{\theta} - E(\hat{\theta})$ is the same as under independence. Therefore, the mean squared error $E[\|\hat{\theta}(u_0) - \theta^0(u_0)\|^2]$ can be approximated by the sum of a bias term of order $O(b^4)$ and a variance term of order $O((nb)^{-1})$, and the optimal bandwidth is of the order $O(n^{-\frac{1}{5}})$.

More specifically, suppose, for instance, that $X_{t,n}$ is a locally stationary fractional ARIMA$(0, d, 0)$ process. Then the optimal choice of b can be based on the following result.

Theorem 7.41 *Let $d \in C^2[0, 1]$ and $d''(u_0) \neq 0$. Then under regularity and moment assumptions* (*see Beran* 2009), *we have, as $n \to \infty$,*

1. *Bias*:

$$E\left[\hat{d}(u_0)\right] - d^0(u_0) = b^2\frac{1}{2}d''(u_0)\int_{-1}^{1} K(x)x^2\,dx + o\left(b^2\right); \tag{7.234}$$

2. *Variance*:

$$\operatorname{var}\left[\hat{d}(u_0)\right] = (nb)^{-1}J^{-1}\int_{-1}^{1} K^2(x)\,dx + o\left((nb)^{-1}\right) \tag{7.235}$$

$$= (nb)^{-1}\frac{6}{\pi^2}\int_{-1}^{1} K^2(x)\,dx + o\left((nb)^{-1}\right); \tag{7.236}$$

3. *Mean squared error*:

$$MSE(\hat{d}) = E\big[\big(\hat{d} - d^0\big)^2\big] = b^4 C_1 + (nb)^{-1} C_2 + o\big\{\max\big(b^4, (nb)^{-1}\big)\big\} \quad (7.237)$$

with

$$C_1(u_0) = \left[\frac{1}{2} d''(u_0) \int_{-1}^{1} K(x) x^2 \, dx\right]^2 \quad (7.238)$$

and

$$C_2 = J^{-1} \int_{-1}^{1} K^2(x) \, dx = \frac{6}{\pi^2} \int_{-1}^{1} K^2(x) \, dx. \quad (7.239)$$

By minimizing the asymptotic expression (7.237) with respect to b, the asymptotically optimal bandwidth is of the form

$$b_{\text{opt}}(u_0) = C_{\text{opt}}(u_0) n^{-1/5} \quad (7.240)$$

with

$$C_{\text{opt}}(u_0) = \left[\frac{C_2}{4C_1(u_0)}\right]^{1/5}. \quad (7.241)$$

The resulting MSE is of the order $O(n^{-4/5})$. This result is analogous to nonparametric regression with uncorrelated residuals. The reason is the $\sqrt{n}$-rate of convergence of $\hat{\theta}$. The second derivative d'' of the estimated d-curve influences the constant C_{opt}. The stronger the curvature of $d(u)$ at the point u_0, the smaller the locally optimal bandwidth $b_{\text{opt}}(u_0)$. Similar results are derived in Dahlhaus and Giraitis (1998) for locally stationary $AR(p)$ processes. For practical purposes, one may prefer using a global bandwidth that minimizes the asymptotic *integrated* mean squared error. To avoid boundary effects, one may use the formula

$$IMSE = b^4 \int_{\delta}^{1-\delta} C_1(u) \, du + (nb)^{-1} \int_{\delta}^{1-\delta} C_2(u) \, du \quad (7.242)$$

where $0 < \delta < \frac{1}{2}$. The constant C_{opt} in (7.240) has to be adjusted accordingly.

If the optimal bandwidth or a bandwidth of the same order is used, then inference about the curve $d^0(u)$ has to take into account that the bias is of the same order as the standard deviation. This means that a bias correction has to be subtracted before using the bounds based on the CLT. An easier solution is to used a bandwidth that is of a slightly smaller order than $O(n^{-1/5})$. This way one can avoid a bias correction. Approximate $(1 - \alpha/2)$-confidence intervals can then be given by

$$\hat{d}(u_0) \pm z_{1-\alpha/2} \frac{\sqrt{6}}{\pi} \left(\int_{-1}^{1} K^2(x) \, dx\right)^{\frac{1}{2}} (nb)^{-\frac{1}{2}}.$$

In particular, for the rectangular kernel we have $\int K^2\,dx = \frac{1}{2}$, so that the interval reduces to

$$\hat{d}(u_0) \pm z_{1-\alpha/2}\frac{\sqrt{3}}{\pi}(nb)^{-\frac{1}{2}}.$$

Analogous formulas can be given for FARIMA(p, d, q) processes with p and q arbitrary. However, in general the optimal bandwidth and the confidence intervals are no longer parameter free.

7.8.3 Computational Issues

In practice, the involved parameters and hence also C_{opt} and b_{opt} are unknown and have to be estimated. In the context of nonparametric regression with i.i.d. errors, various data driven methods for bandwidth choice are known (see, e.g. Gasser et al. 1991; Herrmann et al. 1992). Similar algorithms may be applied here. A possible solution to this problem is an iterative plug-in algorithm where one obtains initial parameter estimates using a first bandwidth. This yields new estimates of b_{opt} so that one can again obtain new parameter estimates and so on. Beran (2009) suggests, for instance, the following algorithm for locally stationary fractional ARIMA$(0, d, 0)$ processes:

Algorithm 1

- Step 1: Set $j = 0$ and set b_j equal to an initial bandwidth.
- Step 2: Estimate $d(\cdot)$ using the bandwidth b_j.
- Step 3: For each u_o, fit a local polynomial regression $\beta_0(u_0) + \beta_1(u_0)(u - u_0) + \frac{1}{2}\beta_2(u_0)(u - u_0)^2$ directly to $\hat{d}(u)$ (plotted against u) using a suitable bandwidth b_2.
- Step 4: For each u_0, set $\hat{d}''(u_0) = 2\beta_2(u_0)$, and calculate an estimate of $C_{\text{opt}}(u_0)$ (or a global value C_{opt} minimizing the integrated mean squared error).
- Step 5: Set $j = j + 1$ and $b_j = C_{\text{opt}} n^{-1/5}$. If b_j and b_{j-1} are very similar (according to a specified criterion), go to Step 6. Otherwise go to Step 2.
- Step 6: Fit a kernel regression with kernel K and bandwidth b_j to $\hat{d}(u)$ directly.

Note that the only purpose of Step 6 is to obtain a somewhat smoother curve, without changing the order of the mean squared error. This step is, however, not necessary. The algorithm can easily be generalized to FARIMA(p, d, q) or more general processes. To do so, one needs to define a suitable mean square error criterion such as $E[\|\hat{\theta} - \theta\|^2]$ and plug-in $\hat{\theta}$ into the asymptotic expression of the criterion. A more complicated algorithm has to be designed, if one wants to combine optimal bandwidth selection with data driven choice of the AR- and MA-orders p and q. A proposal in the context of short-memory AR(p) processes is given in Van Bellegen and Dahlhaus (2006) under the assumption that p (which is unknown) remains constant. Note, however, that even in the AR(p) case the assumption that

p is constant may not be reasonable. In view of the fact that even for stationary fractional ARIMA(p, d, q) processes choosing p and q in a data adaptive way is not easy (see, e.g. Sect. 5.5.6), the problem of including unknown orders p and q (which may also change in time) is far from trivial in the context of locally stationary processes. Alternatively, if the interest lies solely in estimating the long-memory curve $d(u)$, a possibly more elegant solution is to apply a semiparametric method for estimating $d(u)$ locally. This approach is discussed in Roueff and von Sachs (2011) where results on local wavelet estimation of d are obtained.

7.9 Estimation and Testing for Change Points, Trends and Related Alternatives

7.9.1 Introduction

Modelling time series by locally stationary processes is closely related to change point detection and estimation. The main difference is that in change point analysis the emphasis is on abrupt changes. Changes can occur in any aspect of the probability distribution, but most frequently these are the expected value, the marginal distribution or the correlation structure. Here we consider such questions in the long-memory context. An additional issue is that sample paths of short-range dependent processes with change points may be almost indistinguishable from a stationary process with long-range dependence (see, e.g. Bhattacharya et al. 1983; Künsch 1986; Granger and Ding 1996; Teverovsky and Taqqu 1997; Hidalgo and Robinson 1996; Bai 1998; Krämer and Sibbertsen 2000; Mikosch and Starica 2000, 2004; Diebold and Inoue 2001; Granger and Hyung 2004; Davidson and Sibbertsen 2005, also see Sibbertsen 2004 and Banerjee and Urga 2005 and references therein). An important question is therefore how to distinguish "genuine" long memory from such models.

Change point analysis is a classical field of probability theory and statistics, and the literature is enormous (for an overview, see, e.g. Basseville and Nikiforov 1993; Csörgő and Horváth 1998 and references therein), even if we restrict attention to long-memory processes. In the following, some exemplary change point problems are discussed in the context of long-memory processes.

We start with change points in the mean. The standard approach is based on the so-called CUSUM statistics and the asymptotic results follow directly from the asymptotic behaviour of partial sums discussed in Sect. 4.2. In the long-memory context, CUSUM tests are discussed in Horváth and Kokoszka (1997).

Changes in the distribution are detected using empirical processes. In a weakly dependent situation, a sequential empirical process converges to a bivariate Gaussian process, the so-called Kiefer process. In the long-memory set-up the latter process has to be replaced by a process that is degenerate in one dimension and a fractional Brownian bridge in the other. Such results follow from Dehling and Taqqu (1989a, 1989b), see also Sect. 4.8.

Changes in the spectrum (i.e. in the linear dependence structure) are considered in Giraitis and Leipus (1992), Beran and Terrin (1994) and Horváth and Shao (1999), among others. In the last two papers, the dependence parameter before and after a potential change is estimated using Whittle's estimator. Hence, the asymptotic distribution under the "no-change" assumption follows from results for quadratic forms.

Tests that distinguish between changes in the mean (as null hypothesis) and stationary long memory. The best available results are obtained in Berkes et al. (2006), further improvements are suggested in Baek and Pipiras (2011).

Finally, this section is concluded with the question of detecting so-called rapid change points. This notion refers to smooth but very fast changes in the mean. Results in the long-memory context and applications to paleoclimatology are discussed in Menéndez et al. (2010).

7.9.2 Changes in the Mean Under Long Memory

Suppose we would like to test whether a process is stationary against the alternative that there may be changes in the expected value. If, under the alternative, the mean function $\mu(t) = E(X_t)$ is expected to follow certain regularity conditions such as differentiability or L^2-integrability, then we are back to the question of simultaneous modelling of trend functions and dependence structure. We refer to Sects. 7.1, 7.4 and 7.5 for a discussion of this topic. On the other hand, if abrupt changes are expected, then this leads to questions in the realm of change point detection and estimation. (Another situation that is somewhere between standard nonparametric trend estimation and change point analysis is the so-called rapid change point detection discussed in Sect. 7.10.)

Specifically, consider the null hypothesis

$$H_0 : Y_t = \mu + X_t$$

where X_t is a zero mean second-order stationary process against the alternative

$$H_1 : Y_t = \mu + \Delta \cdot 1\{t > t_0 + 1\} + X_t \quad (\Delta \neq 0)$$

where t_0 $(1 \leq t_0 < n)$ is an unknown change point. The best known approach is based on the CUSUM statistic (originally introduced by Page 1954 in the context of quality control; also see Barnard 1959) defined by

$$\begin{aligned} D_{1,n} &= \max_{1 \leq i \leq n} |V_i| \\ &\approx \sup_{0<u<1} \left| S_n(u) - u S_n(1) \right| \end{aligned}$$

where we use the notation

$$V_i = S_{1,i} - \frac{i}{n} S_{1,n}, \qquad S_{i,j} = \sum_{t=i}^{j} Y_t$$

and

$$S_n(u) = \sum_{t=1}^{[nu]} Y_t.$$

Note that $n^{-1}V_i$ can also be written as a weighted sum of the difference between the two sample means before and after i, namely

$$n^{-1}V_i = \frac{i}{n}\left(1 - \frac{i}{n}\right)\left(\frac{1}{i}S_{1,i} - \frac{1}{n-i}S_{i+1,n}\right).$$

In the classical change point analysis, the process X_t is assumed to be in the area of attraction of Brownian motion in the sense that $S_n(u)$, properly standardized, converges in the space of càdlàg functions $D[0,1]$ to a standard Brownian motion $B(u)$ ($u \in [0,1]$). This result usually applies to second-order stationary short-memory processes where $\text{var}(S_n(1)) \sim c_S n$. Thus, under H_0, we have a functional limit theorem with $\tilde{Z}_n(u) = (S_n(u) - uS_n(1))c_S^{-\frac{1}{2}} n^{-\frac{1}{2}}$ converging to a Brownian bridge $\tilde{B}(u) = B(u) - uB(1)$, and hence

$$c_S^{-\frac{1}{2}} n^{-\frac{1}{2}} D_{1,n} \underset{d}{\to} \sup_{u\in[0,1]} \left|\tilde{B}(u)\right|.$$

In view of the limit theorems discussed in Chap. 4, this result can be generalized quite easily to processes with long memory and antipersistence, respectively. Suppose that X_t is in the domain of attraction of fractional Brownian motion $B_H(u)$ (again in the sense of a functional limit theorem) with self-similarity parameter $H \in (0,1)$. The case of short memory is included here, with $H = \frac{1}{2}$, antipersistence corresponds to $H < \frac{1}{2}$ and long memory to $H > \frac{1}{2}$. Then, under the null hypothesis formulated above, the process

$$\tilde{Z}_n(u) \approx L_S^{-\frac{1}{2}}(n) n^{-H}\left(S_n(u) - uS_n(1)\right)$$

(with L_S a slowly varying function as defined in Sect. 4.2.2) converges to a fractional Brownian bridge $\tilde{B}_H(u) = B_H(u) - uB_H(1)$. For the standardized statistic, we then have

$$T = L_S^{-\frac{1}{2}}(n) n^{-H} D_{1,n} \underset{d}{\to} \sup_{u\in[0,1]} \left|\tilde{B}_H(u)\right|.$$

In contrast, under the alternative H_1 with a change point in $\mu(t) = E(Y_t)$, the expected value of $S_n(u) - uS_n(1)$ is of the order $n \gg n^H$ so that $T \to_p \infty$ (for further results and detailed regularity assumptions, see, e.g. Csörgő and Horváth 1998;

Berkes et al. 2006). Note that an analogous result can be obtained in principle for processes in the domain of attraction of a Hermite process of any order.

The standardization $L_S^{-\frac{1}{2}}(n)n^{-H}$ contains the unknown self-similarity parameter H and the slowly varying function L_S. Both have to be estimated from the observed data. For most practical purposes, it is sufficient to assume that L_S converges to a constant $c_S > 0$ so that $\mathrm{var}(S_n(1)) \sim c_S \cdot n^{2H}$ $(n \to \infty)$. In view of Sect. 1.3.1, a natural way of rewriting the standardization is

$$L_S^{\frac{1}{2}}(n)n^H = \sqrt{\nu(d)c_{f_X}}\,n^{d+\frac{1}{2}} = \sqrt{\nu(d)\,f_X\left(n^{-1}\right)}\,n^{\frac{1}{2}}$$

with $d = H - \frac{1}{2}$,

$$\nu(d) = \frac{2\sin\pi d}{d(2d+1)} \quad (d \neq 0),$$
$$\nu(0) = 2\pi$$

and c_{f_X} such that $f_X(\lambda) \sim c_{f_X}|\lambda|^{-2d}$ $(\lambda \to 0)$. In the classical change point analysis, H is assumed to be equal to $\frac{1}{2}$ a priori so that only the constant c_f, or equivalently $f_X(0)$, needs to be estimated (see, e.g. Csörgő and Horváth 1998 and references therein). However, if we calculate T under this assumption but the true value of H is actually larger than $\frac{1}{2}$, then the asymptotic rejection probability tends to one even if the null hypothesis is true (for a further discussion along this line, see, e.g. Horváth and Kokoszka 1997; Wright 1998; Krämer et al. 2002; Sibbertsen 2004; for extensions to linear regression, see, e.g. Krämer and Sibbertsen 2000). In other words, assuming independence or short-range dependence ultimately leads to the erroneous conclusion that the mean is not constant. The formal reason is that the standardization by $n^{\frac{1}{2}}$ is too small by a factor proportional to $n^{H-\frac{1}{2}} \to \infty$ so that T tends to infinity. The *intuitive* explanation is that long-range dependent series exhibit local spurious trends and tend to stay on one side of the expected value for a long time. This often looks as if the mean were changing occasionally.

If we are not assuming $H = \frac{1}{2}$ a priori, then both parameters, c_f and H, need to be estimated consistently. Given such estimates, we define the statistic

$$T = n^{-\hat{H}}\hat{\nu}^{-\frac{1}{2}}\hat{c}_{f_X}^{-\frac{1}{2}}D_{1,n}$$

with $\hat{H} = \hat{d} + \frac{1}{2}$ and $\hat{\nu} = \nu(\hat{d})$. The null hypothesis of no change point is rejected at the level of significance α, if $T > q_{1-\alpha}$ where $q_{1-\alpha}$ is defined by

$$P\Big(\sup_{u\in[0,1]}\left|\tilde{B}_{\hat{H}}(u)\right| > q_{1-\alpha}\Big) = \alpha.$$

(Note that here the probability is evaluated for a fractional Brownian bridge with $\hat{H}$ being fixed.)

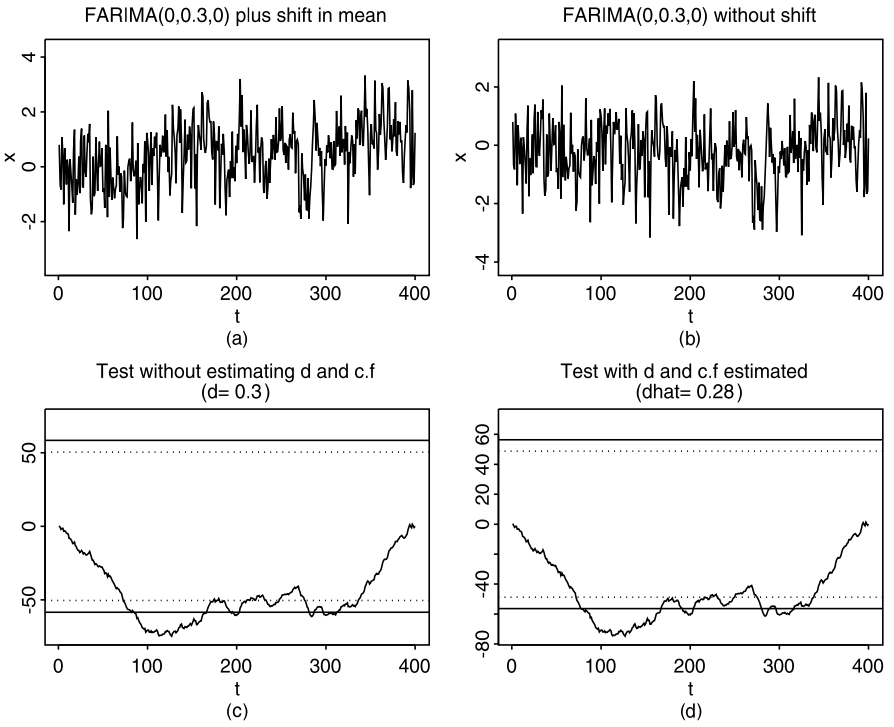

Fig. 7.16 Simulated sample paths of $Y_t = \Delta \cdot 1\{t \geq 120\} + X_t$ (**a**) and X_t (**b**) where X_t is a FARIMA$(0, 0.3, 0)$ process and $\Delta = 1$. The values of $V_i = S_{1,i} - (i/n)S_{1,n}$ are plotted against i in (**c**) and (**d**), with 5 %- and 10 %-critical values (*horizontal lines*) based on the true (**c**) and estimated parameters d and c_f (**d**), respectively

Example 7.37 Let X_t be generated by a fractional ARIMA$(0, d, 0)$ process with zero mean i.i.d. innovations ε_t. Then $c_f = \sigma_\varepsilon^2/(2\pi)$ and we may estimate $\theta = (\sigma_\varepsilon^2, d)$ by one of the (quasi-) maximum likelihood methods discussed in Sect. 5.5. The test statistic simplifies to

$$\tilde{T} = n^{-\frac{1}{2}-\hat{d}}\hat{\nu}^{-\frac{1}{2}}\sqrt{2\pi}\hat{\sigma}_\varepsilon^{-1}D_{1,n}.$$

Example 7.38 Figure 7.16(a) displays simulated sample paths of

$$Y_t = \Delta \cdot 1\{t \geq 120\} + X_t$$

$(t = 1, 2, \ldots, 400)$ with $\Delta = 1$ and 0, respectively, and X_t generated by a fractional ARIMA$(0, 0.3, 0)$ process. The shift is hardly visible by eye. Nevertheless, H_0 is rejected at the 5 %-level of significance. The fact that H and c_f have to be estimated does not make much of a difference. This can be seen from Figs. 7.16(c)–(d) where the values of $S_{1,i} - \frac{i}{n}S_{1,n}$ are plotted against i, together with critical 10 %- and 5 %-limits (horizontal lines) based on the true parameters (Fig. 7.16(c)) and the estimated parameters (Fig. 7.16(d)), respectively. The estimated value of H is 0.78.

Although in this example the estimation of d and c_f has almost no influence on the result, this may not always be the case. In fact, under the alternative, the observed process is no longer stationary. This may have undesirable effects on the estimates. Sometimes it may first be necessary to remove an estimated trend function $\hat{\mu}(t)$ before estimating d and c_f. This brings us back, however, to the question how to fit a trend function in the presence of dependent errors (see Sects. 7.1, 7.4 and 7.5). If a step function with a finite but unknown number of change points is expected under the alternative, then one may try, for instance, wavelet thresholding with Haar wavelets (see Sect. 7.5) or nonlinear regression with piecewise constant polynomials (see Sect. 7.3). Another possibility is to first calculate parameter estimates based on relatively short disjoint blocks of observations and then take their average. For quasi-maximum likelihood estimation, this can be done without any loss of asymptotic efficiency (Beran and Terrin 1996). This approach is illustrated in the following example.

Example 7.39 Figure 7.17(a) displays a sample path of $Y_t = \mu(t) + X_t$ where X_t is a FARIMA$(0, 0.1, 0)$ process and $\mu(t)$ has multiple change points with values switching between 0 and 1 as displayed in Fig. 7.17(b). The values of $V_i = S_{1,i} - (i/n)S_{1,n}$ are plotted in Figs. 7.17(c)–(d). In Fig. 7.17(c), the horizontal lines correspond to 10 %- and 5 %-critical values when using $\hat{d}$ and $\hat{c}_f$ estimated (by QMLE) from the complete series Y_t $(t = 1, 2, \ldots, n)$ directly, whereas in Fig. 7.17(d), the critical boundaries are based on averages of estimates $\hat{d}_j$ and $\hat{c}_{f,j}$ $(j = 1, 2, \ldots, 10)$ obtained from disjoint blocks $Y_{t+(j-1)100}, \ldots, Y_{j100}$ of length 100. In the first case, $d^0 = 0.1$ is overestimated by the amount of $\hat{d} - d^0 = 0.13$ whereas in the second case overestimation is less severe with $\hat{d} - d^0 = 0.06$. This leads to clear rejection of H_0 at the 5 %-level in the second case; however, no rejection in the first case.

The test statistics above do not take into account that the variance function of $\tilde{B}_H(u)$ is not constant. More specifically, we have

$$\begin{aligned}\operatorname{var}\big(\tilde{B}_H(u)\big) &= E\big[B_H^2(u)\big] + u^2 E\big[B_H^2(1)\big] - 2uE\big[B_H(u)B_H(1)\big] \\ &= u(1-u)\big[u^{2H-1} - 1 + (1-u)^{2H-1}\big] \\ &=: w_H(u).\end{aligned}$$

Since w_H is zero at both ends and achieves its maximum in the middle (see Fig. 7.18), the test based on T or $\tilde{T}$ may have little power when change points occur near the two ends. One therefore sometimes prefers to standardize by $\sqrt{w_H(u)}$ before taking the supremum. This means that one defines a test based on $D^*_{1,n} = \max |V_i|/\sqrt{w(\frac{i}{n})}$. The asymptotic distribution of $D^*_{1,n}$ is, however, more difficult to derive.

The statistics $w^{-\frac{1}{2}} V_i$ $(i = 2, \ldots, n-1)$ are also often used for estimating the change point t_0 itself, namely by choosing $\hat{t}_0 = i$ such that $|w^{-\frac{1}{2}} V_i|$ is minimal. For i.i.d. data, the asymptotic distribution of $\hat{t}_0$ has been derived by Antoch et al.

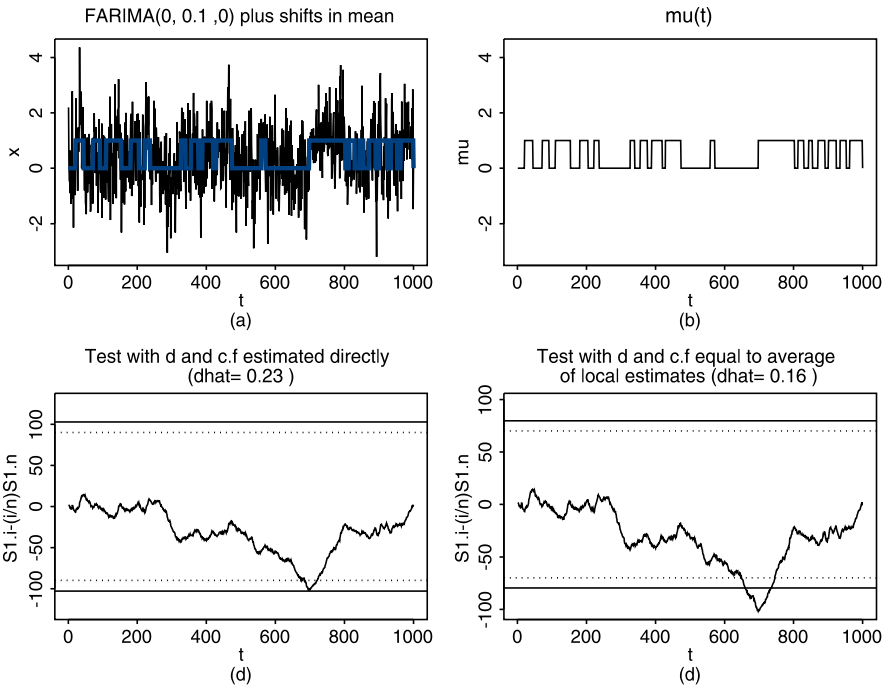

Fig. 7.17 Figure (**a**) shows a sample path of $Y_t = \mu(t) + X_t$ where X_t is a FARIMA$(0, 0.1, 0)$ process and $\mu(t)$ has multiple change points with values switching between 0 and 1 as displayed in (**b**). The values of $V_i = S_{1,i} - (i/n)S_{1,n}$ are plotted in (**c**) and (**d**). The *horizontal lines* correspond to 10 %- and 5 %-critical values using estimates of d and c_f. In (**c**), the estimates were based on Y_t $(t = 1, 2, \ldots, n)$, whereas in (**d**) these are averages of estimates $\hat{d}_j$ and $\hat{c}_{f,j}$ $(j = 1, 2, \ldots, 10)$ obtained from disjoint blocks $Y_{1+(j-1)100}, \ldots, Y_{j100}$ of length 100

(1995) (also see Hinkley 1970; Yao 1987 for earlier results). Similar results in the context of short-range dependence can be found, for instance, in Bagshaw and Johnson (1975), Davis et al. (1995), Horváth (1993), Johnson and Bagshaw (1974) and Tang and MacNeill (1993). Horváth and Kokoszka (1997) derive limit theorems for $\hat{t}_0$ under more general dependence assumptions in the domain of attraction of fractional Brownian motion with $H \in (0, 1)$, and also consider a more general class of estimators.

Change point estimation in the mean can be extended to the problem of structural breaks in regression models. Results along this line in the long-memory context can be found, for instance, in Wright (1998), Krämer and Sibbertsen (2003), Sibbertsen (2004), Lazarova (2005), Gil-Alana (2008). Also see Ben Hariz and Wylie (2005) and Ben Hariz et al. (2007) for general results. Change point estimation in the long-memory context based on the Wilcoxon two-sample test is considered in Dehling et al. (2013), rank tests are developed in Wang (2008).

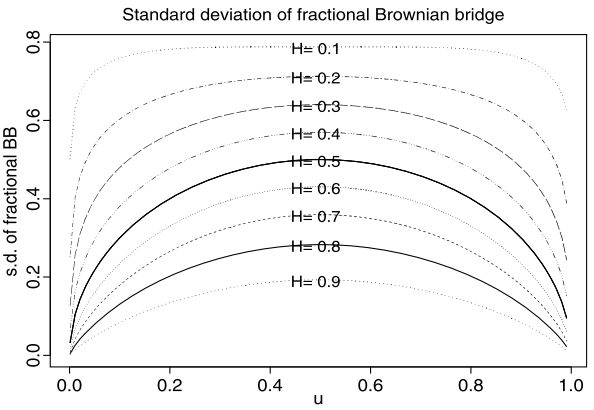

Fig. 7.18 Standard deviation of a fractional Brownian bridge $\tilde{B}_H(u)$

7.9.3 Changes in the Marginal Distribution

Instead of testing for changes in the mean, one may more generally test whether any changes in the marginal distribution occur. If we do not want to specify which features of the distribution may change, then we are led to nonparametric testing based on the empirical distribution function. This problem has been addressed, for instance, in Giraitis et al. (1996b) by studying a test based on the Kolmogorov–Smirnov statistic. In the i.i.d. and short memory context, such tests have been studied extensively (see, e.g. Picard 1985; Carlstein 1988; Leipus 1988; Dümbgen 1991; Ferger and Stute 1992; Carlstein and Lele 1993; Ferger 1994; also see Csörgő and Horváth 1988, 1998; Brodsky and Darkhovsky 1993 and references therein).

The essential probabilistic result one needs is the asymptotic distribution of the empirical process. More specifically, suppose we observe $Y_1, \ldots, Y_n$ generated by a stationary process with marginal distribution $F(y) = P(Y \le y)$. A natural statistic for testing for changes in the marginal distribution function can be constructed by comparing an estimated cumulative distribution of $Y_1, \ldots, Y_i$ with the corresponding estimate for $Y_{i+1}, \ldots, Y_n$. Let

$$F_{i,j}(y) = \frac{1}{(j-i+1)} \sum_{t=i}^{j} 1\{Y_t \le y\}$$

where $j \ge i$, and

$$F_{1,[nu]}(y) = F_{[nu]}(y)$$

with $u \in [0,1]$ and $[nu]$ denoting the largest integer not exceeding nu. Then we consider weighted differences

$$V_i(y) = \frac{i}{n}\left(1 - \frac{i}{n}\right)\left[F_{1,i}(y) - F_{i+1,n}(y)\right] \quad (i = 1, \ldots, n-1).$$

Let $u \in (0,1)$ and $i = [nu]$. Then we can rewrite $V_i(y)$ as

$$\begin{aligned} V_i(y) &= V_{[nu]}(y) \\ &= \frac{[nu]}{n}\left(1 - \frac{[nu]}{n}\right)\big[F_{1,[nu]}(y) - F_{[nu]+1,n}(y)\big] \\ &= \left(1 - \frac{[nu]}{n}\right)\left\{\frac{[nu]}{n}F_{[nu]}(y)\right\} - \frac{[nu]}{n}\left\{F_n(y) - \frac{[nu]}{n}F_{[nu]}(y)\right\} \\ &= F_{[nu]}(y) - \frac{[nu]}{n}F_n(y). \end{aligned}$$

This is analogous to the quantities used for the CUSUM statistic in the previous section. The only difference is that instead of the observations themselves we average the 0–1-variables $1\{Y_t \le y\}$. The CUSUM statistic is then of the form

$$\begin{aligned} D_{1,n} &= \sup_{\substack{1 \le i \le n-1 \\ y \in \mathbb{R}}} \big|V_i(y)\big| \\ &= \sup_{\substack{n^{-1} \le u \le 1-n^{-1} \\ y \in \mathbb{R}}} \left|\frac{[nu]}{n}\left(1 - \frac{[nu]}{n}\right)\big[F_{1,[nu]}(y) - F_{[nu]+1,n}(y)\big]\right| \\ &= \sup_{u,y}\left|F_{[nu]}(y) - \frac{[nu]}{n}F_n(y)\right| \end{aligned}$$

(see, e.g. Picard 1985). The asymptotic distribution of $D_{1,n}$ follows easily, once we have a suitable functional limit theorem for the difference $F_{[nu]}(y) - F(y)$, understood as a stochastic process in $(u, y) \in [0,1] \times [-\infty, \infty]$.

Suppose that there is a suitable sequence of numbers $v_n \to 0$ such that

$$v_n^{-\frac{1}{2}}\big[F_{[nu]}(y) - F(y)\big]$$

converges (weakly in a suitable manner) to a process $W(u, y)$. Then we define the test statistic

$$T = v_n^{-\frac{1}{2}} D_{1,n}.$$

Under the null hypothesis that the marginal distribution remains the same, we have

$$\begin{aligned} T &= \sup_{u,y}\left|v_n^{-\frac{1}{2}}\left\{F_{[nu]}(y) - \frac{[nu]}{n}F_n(y)\right\}\right| \\ &\underset{d}{=} \sup_{(u,y) \in [0,1] \times \mathbb{R}} \big|W(u,y) - uW(1,y)\big| + o_p(1). \end{aligned}$$

Thus, a rejection region at a level of significance α can be defined by $K_\alpha = \{T > q_{1-\alpha}\}$ where $q_{1-\alpha}$ are $(1-\alpha)$-quantiles defined by

$$P\Big(\sup_{(u,y)\in[0,1]\times\mathbb{R}} \big|W(u,y) - uW(1,y)\big| > q_{1-\alpha}\Big) = \alpha.$$

For i.i.d. observations, it is well known that the asymptotic limit of

$$W_n(u,y) = n^{\frac{1}{2}}\big[F_{[nu]}(y) - F(y)\big]$$

is a Kiefer process $W(u,y)$ where convergence is in the space $D([0,1] \times [-\infty,\infty])$. Recall that a Kiefer process is a Gaussian process (in (u,y)) with zero mean and covariance function

$$cov\big(W(u_1,y_1), W(u_2,y_2)\big) = \min\{u_1,u_2\} \cdot \big[F\big(\min(y_1,y_2)\big) - F(y_1)F(y_2)\big]$$

(see, e.g. Shorack and Wellner 1986 and references therein). This result can be generalized to standard short-memory conditions to obtain a Gaussian limiting process with covariance function

$$cov\big(W(u_1,y_1), W(u_2,y_2)\big) = \min\{u_1,u_2\} \cdot \sigma(y_1,y_2)$$

where

$$\sigma(y_1,y_2) = \sum_{t=-\infty}^{\infty} \big[P(Y_0 \le y_1, Y_t \le y_2) - P(Y_0 \le y_1)P(Y_t \le y_2)\big]$$

(see, e.g. Berkes and Philipp 1977). In contrast, under long memory the rate of convergence is slower and one obtains a degenerate limiting process (see Sect. 4.8). For instance, let $Y_t = G(Z_t)$ where Z_t is a zero mean Gaussian process with variance one, slowly decaying autocovariances $\gamma_Z(k) \sim L_\gamma(k)|k|^{2d-1}$ and assume that $1\{G(Z_t) \le y\}$ has Hermite rank $m = 1$. Then Dehling and Taqqu (1989b) showed that

$$W_{n,H}(u,y) = L_S^{-\frac{1}{2}}(n) n^{1-H}\big[F_{[nu]}(y) - F(y)\big]$$

(with $H = d + \frac{1}{2}$ and $L_S(n) = L_\gamma(n)(d(2d+1))^{-1}$, see Sect. 4.2.2) converges in $D([0,1] \times [-\infty,\infty])$ equipped with the sup-norm to a constant (depending on y) times a fractional Brownian motion B_H, or more specifically,

$$W(u,y) = W_H(u,y) = J_1(y)B_H(u)$$

where $J_1(y) = E[1\{G(Z) \le y\}Z]$. An analogous result holds for higher Hermite ranks with B_H replaced by the corresponding Hermite process of order m. This result is remarkable because along the y-axis, no stochasticity is involved. Once u is fixed and the random variable $B_H(u)$ is generated, the process evolves in y only

via multiplication by the deterministic function $J_1(y)$. The asymptotic distribution of $D_{1,n}$ is therefore much simpler than under short memory. Defining

$$T = L_S^{-\frac{1}{2}}(n)n^{-H}D_{1,n},$$

we obtain

$$T \underset{d}{=} \zeta + o_p(1)$$

with

$$\begin{aligned}\zeta &= \sup_{y\in\mathbb{R}}\big|J_1(y)\big| \cdot \sup_{u\in[0,1]}\big|B_H(u) - uB_H(1)\big| \\ &= \sup_{y\in\mathbb{R}}\big|J_1(y)\big| \cdot \sup_{u\in[0,1]}\big|\tilde{B}_H(u)\big|.\end{aligned}$$

The first factor is a deterministic constant that only depends on the transformation G. The second term is the usual supremum of a fractional Brownian bridge. Now we can calculate critical values for testing the null hypothesis that we observe a stationary process $Y_t = G(Z_t)$ with a certain (unknown) marginal distribution F against the alternative

$$H_1 : Y_t = X_{t,1} \quad (1 \le t \le t_0), \qquad Y_t = X_{t,2} \quad (t_0 < t \le n)$$

where $X_{t,1}$, $X_{t,2}$ are two stationary processes with marginal distributions $F_1 \neq F_2$ and t_0 is an unknown change point. A rejection region at level of significance α can be defined by

$$T > \sup_{y\in\mathbb{R}}\big|J_1(y)\big| \cdot q_{1-\alpha},$$

or equivalently,

$$D_{1,n} > L_S^{\frac{1}{2}}(n)n^H \cdot \sup_{y\in\mathbb{R}}\big|J_1(y)\big| \cdot q_{1-\alpha}$$

where $q_{1-\alpha}$ is defined by

$$P\Big(\sup_{u\in[0,1]}\big|\tilde{B}(u)\big| > q_{1-\alpha}\Big) = \alpha.$$

Example 7.40 Let Y_t be a Gaussian FARIMA$(0, d, 0)$ process with $\operatorname{var}(\varepsilon_t) = 1$. Then $Y_t = \sigma_Y Z_t$ with $\sigma_Y^2 = \operatorname{var}(Y_t) = \Gamma(1-2d)/\Gamma^2(1-d)$ and

$$J_1(y) = E\big[1\{\sigma_Y Z \le y\}Z\big] = \int_{-\infty}^{\sigma_Y^{-1}y} z\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\,dz = -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\sigma_Y^{-2}y^2}.$$

The supremum of $|J_1(y)|$ is $1/\sqrt{2\pi}$. Moreover,

$$L_\gamma(n) = \Gamma(1-2d)/\big[\Gamma(d)\Gamma(1-d)\big]$$

so that

$$L_S(n) = L_\gamma(n)\big(d(2d+1)\big)^{-1} = \frac{\Gamma(1-2d)}{\Gamma(1+d)\Gamma(1-d)(2d+1)}.$$

A critical region at level α is therefore given by

$$\left\{T > \frac{1}{\sqrt{2\pi}} \cdot q_{1-\alpha}\right\} = \left\{D_{1,n} > n^H \cdot \sqrt{\frac{\Gamma(1-2d)}{2\pi\Gamma(1+d)\Gamma(1-d)(2d+1)}} \cdot q_{1-\alpha}\right\}$$

where $H = d + \frac{1}{2}$.

7.9.4 Changes in the Linear Dependence Structure

Often the dependence structure in an observed time series is not constant. Slow changes can be captured by locally stationary processes. This has been discussed in Sect. 7.8. On the other hand, there are situations where the dependence structure changes suddenly. Such situations are in the realm of change point analysis. The null hypothesis we are testing is that the observed process Y_t is stationary with a fixed spectral distribution F_Y. The alternative is that there is a change point t_0 such that Y_t has the spectral distributions F_1 and F_2 for $t \le t_0$ and $t > t_0$, respectively, with $F_1 \neq F_2$. Note that here F denotes the *spectral* distribution, and not the marginal distribution.

A simple way of testing for change points in the correlation structure is considered in Beran and Terrin (1994). Suppose we have a parametric model with $\theta = (\sigma_\varepsilon^2, d, \dots)^T = (\sigma_\varepsilon^2, \eta)^T$ where the central limit theorem holds for quasi-maximum likelihood estimates as discussed in Sect. 5.5. For instance, we may assume a FARIMA(p, d, q) process with spectral density

$$f(\lambda;\theta) = \sigma_\varepsilon^2 \big|1 - \exp(-i\lambda)\big|^{-2d} \left|\frac{\psi(e^{-i\lambda})}{\phi(e^{-i\lambda})}\right|^2.$$

First, we divide the time axis into m blocks $I_1 = \{1, 2, \dots, n_1\}$, $I_2 = \{n_1 + 1, \dots, n_1 + n_2\}, \dots$ such that $\sum n_j = n$ and $n_j/n \to p_j \in (0, 1)$. For each block of observations Y_t $(t \in I_j)$ a quasi-MLE $\hat\eta_j$ is computed. Similar arguments as in Sect. 5.5 (Beran and Terrin 1994) show that, as $n \to \infty$, $Z_{j,n} = \sqrt{n_j}(\hat\eta_j - \eta)$ $(j = 1, 2, \dots, m)$ are asymptotically independent of each other, with limiting $N(0, \Sigma_j)$-distribution where $\Sigma_j = 4\pi V^{-1}$ and

$$V = \left\{\int \frac{\partial}{\partial\eta} \log f(\lambda;\theta) \left[\frac{\partial}{\partial\eta} \log f(\lambda;\theta)\right]^T d\lambda\right\}^{-1}.$$

This can be used for testing whether the parameter η remains constant over time. For simplicity suppose that we are only interested in changes of the long-memory

parameter d. Then the null hypothesis is that Y_t is stationary, which means in particular that d is constant. Denoting by d_j the long-memory parameter in block I_j $(j = 1, 2, \ldots, m)$, the null hypothesis implies $d_1 = \cdots = d_m = d$. The alternative is specified by the existence of at least one pair $j_1, j_2 \in \{1, 2, \ldots, m\}$ such that $d_{j_1} \neq d_{j_2}$. Suppose for simplicity that $n_1 = \cdots = n_m = nm^{-1}$ and denote by $v_{m,n} = 4\pi[V^{-1}]_{11}mn^{-1}$ the approximate variance of each $\hat{d}_j$. Using the notation $\bar{d} = m^{-1}\sum \hat{d}_j$, a simple test statistic of H_0 can be based on

$$\chi^2 = v_{m,n}^{-1}\sum_{j=1}^{m}(\hat{d}_j - \bar{d})^2$$

$$= \frac{1}{4\pi[V^{-1}]_{11}}\frac{n}{m}\sum_{j=1}^{m}(\hat{d}_j - \bar{d})^2.$$

Under H_0, the statistic is approximately χ^2_{m-1}-distributed. In contrast, under the alternative, $\sum(\hat{d}_j - \bar{d})^2$ converges in probability to $\sum_{j=1}^{m}(d_j - d)^2 > 0$ where $d = m^{-1}\sum d_j$ so that χ^2 diverges to infinity.

Example 7.41 Let Y_t be a FARIMA$(0, d, 0)$ process. Then $4\pi[V^{-1}]_{11} = 6/\pi^2$. The null hypothesis is rejected at the level of significance α, if

$$\frac{\pi^2}{6}\frac{n}{m}\sum_{j=1}^{m}(\hat{d}_j - \bar{d})^2 > \chi^2_{m-1;1-\alpha}$$

with $\chi^2_{m-1;1-\alpha}$ denoting the $(1-\alpha)$-quantile of a χ^2_{m-1}-distribution. We apply this test to the detrended central England temperatures displayed in Fig. 7.19(b). The sample size is $n = 352$. Using $m = 4$ blocks of length $n_j = 88$, and a FARIMA$(0, d, 0)$ fit for each block, the maximum likelihood estimates $\hat{d}_j$ $(j = 1, 2, 3, 4)$ are equal to 0.30, 0.07, 0.02 and 0.29, respectively. The value of the χ^2-statistic is about 9.15 which corresponds to a p-value (based on a χ^2_3-distribution) of 0.027. Thus, there is quite strong evidence for a change in d. This confirms the visual impression of the log–log-periodogram plots for the four blocks in Figs. 7.19(c)–(f), and also the impression obtained by fitting a locally stationary FARIMA$(0, d, 0)$ process in Sect. 7.8. (Note also that the FARIMA$(0, d, 0)$ model does indeed fit the data reasonably well, locally.)

In situations where the location of change points is unknown, one would prefer a method where one does not have to divide the time axis into blocks by hand. Assume again a parametric model with spectral density $f(\lambda; \theta)$ and a p-dimensional parameter $\theta = (\sigma_\varepsilon^2, d, \ldots)^T = (\sigma_\varepsilon^2, \eta)^T$. Suppose for simplicity of presentation that we are only interested in changes in the long-memory parameter d. A CUSUM type

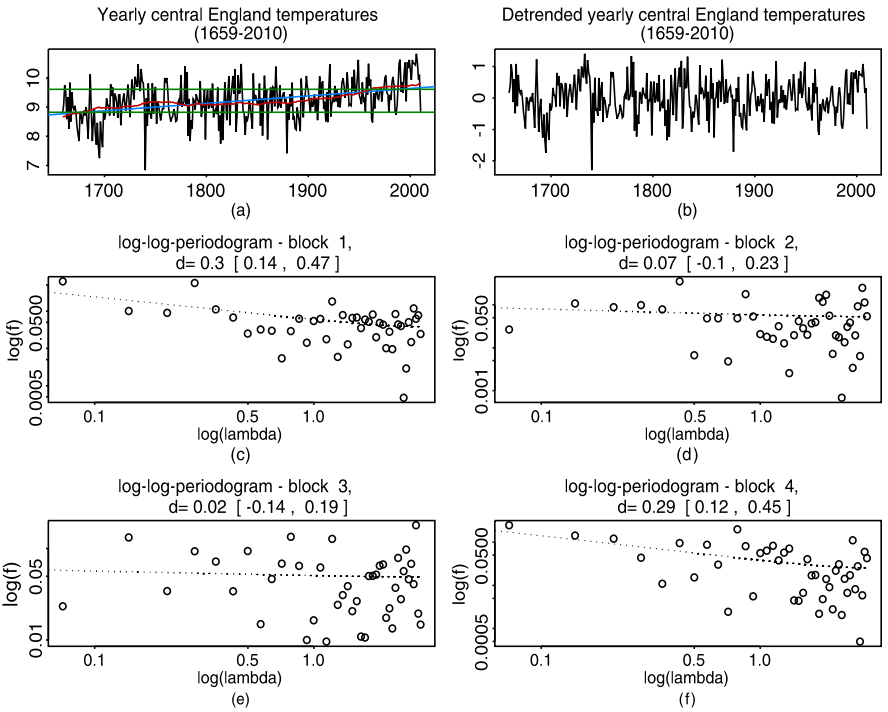

Fig. 7.19 Yearly Central England temperatures 1659–2010 (**a**) and the detrended series (**b**) after subtracting a nonparametric trend function. Also displayed are log–log-periodograms and FARIMA$(0, d, 0)$ spectral densities fitted to four disjoint blocks of length $n_j = 88$

statistic can be defined by

$$D_{1,n} = \max_{n_{\text{low}} \le i \le n_{\text{up}}} \left| \frac{i}{n} \left(1 - \frac{i}{n} \right) (\hat{d}_{1,i} - \hat{d}_{i+1,n}) \right|$$

with $\hat{d}_{1,i} = [\hat{\eta}_{1,i}]_1$, $\hat{d}_{i+1,n} = [\hat{\eta}_{i+1,n}]_1$ where $\hat{\eta}_{1,i}$ and $\hat{\eta}_{i+1,n}$ are estimates of $\eta = (d, \dots)^T$ based on $X_1, X_2, \dots, X_i$ and $X_{i+1}, \dots, X_n$, respectively. Note that, in contrast to the sample mean, the estimates require a certain minimal size of the sample. Therefore, in practice n_{low} has to be chosen larger than 1, and n_{up} smaller than n.

Suppose now that under the null hypothesis H_0 the observed time series Y_t $(t = 1, \dots, n)$ is generated by a stationary process in the parametric class with $\theta = \theta^0$. The alternative H_1 we would like to test against is that there is a change point $1 < t_0 < n$ such that the long-memory parameter is $d = d_1$ for $t \le t_0$ and $d = d_2 \neq d_1$ for $t > t_0$. To estimate θ^0 we use one of the approximate quasi-maximum likelihood estimators derived from the normal likelihood. Recall that under H_0, the central limit theorem holds for $\hat{\theta}$ with a $\sqrt{n}$-rate of convergence, and the scale estimator is asymptotically independent of $\hat{\eta}$. The proof of this result relies

either on a central limit theorem for quadratic forms or on an approximation by martingale differences (see Sect. 5.5). For instance, if we use the second approach, then $\hat{\eta}$ is defined by minimizing $\sum e_t^2(\eta)$ where $e_t(\eta) = \sum_{j=0}^{t-1} b_j(\eta) Y_{t-j}$ is an approximation of ε_t obtained from the autoregressive representation $\varepsilon_t = \sum_{j=0}^{\infty} b_j(\eta) Y_{t-j}$, and $\hat{\theta}_1 = \hat{\sigma}_\varepsilon^2$ is set equal to $n^{-1} \sum e_t^2(\hat{\eta})$. Then, based on n observations, we have the approximation

$$\hat{\eta} - \eta^0 = n^{-1} S_n + o_p\big(n^{-1}\big)$$

where

$$S_n = \big(S_n^1, \ldots, S^{p-1}\big)^T = M^{-1} \sum_{t=2}^{n} \dot{\varepsilon}_t\big(\eta^0\big) \varepsilon_t\big(\eta^0\big),$$

$M = E(\dot{\varepsilon}_t \dot{\varepsilon}_t^T)$ and $\dot{\varepsilon}_t = \partial / \partial \eta \varepsilon_t(\eta) \,|_{\eta=\eta^0} = \sum \dot{b}_j Y_{t-j}$. Using the notation

$$\zeta_t = \big(\zeta_t^1, \ldots, \zeta_t^{p-1}\big)^T = M^{-1} \dot{\varepsilon}_t\big(\eta^0\big) \varepsilon_t\big(\eta^0\big)$$

and

$$\zeta_t^j = \sum_{l=1}^{p-1} \tilde{m}_{jl} \left\{ \frac{\partial}{\partial \eta_l} \varepsilon_t\big(\eta^0\big) \varepsilon_t\big(\eta^0\big) \right\}$$

with $M^{-1} = [\tilde{m}_{jl}]_{j,l=1,\ldots,p-1}$, we can write $S_n = \sum_{t=2}^{n} \zeta_t$. Since we are only interested in d, the only relevant component of S_n is

$$S_n^1 = \sum_{t=2}^{n} \zeta_t^1.$$

This means that asymptotically $\hat{d} - d^0$ can be approximated by a sample mean, and $D_{1,n}$ can be written in the form of a usual CUSUM statistic with sample means. Furthermore, since $\dot{\varepsilon}_t(\eta^0)\varepsilon_t(\eta^0)$ is a martingale difference, we have, under suitable moment conditions, a functional limit theorem

$$n^{-\frac{1}{2}} S_n^1(u) = n^{-\frac{1}{2}} \sum_{t=2}^{[nu]} \zeta_t^1 \to \text{const} \cdot B(u)$$

where convergence is in $D[0,1]$ and $B(u)$ $(u \in [0,1])$ is a standard Brownian motion. Assuming that $n_{\text{low}}/n \to 0$ and $n_{\text{up}}/n \to 1$, we may therefore write

$$\begin{aligned} \sqrt{n} D_{1,n} &= \sqrt{n} \max_{n_{\text{low}} \le i \le n_{\text{up}}} \left| \frac{i}{n} \left(1 - \frac{i}{n}\right) (\hat{d}_{1,i} - \hat{d}_{i+1,n}) \right| \\ &= \sqrt{n} \max_{n_{\text{low}} \le i \le n_{\text{up}}} \left| \frac{i}{n} \left(1 - \frac{i}{n}\right) \big(i^{-1} S_i^1 - (n-i)^{-1} \big(S_n^1 - S_i^1\big)\big) \right| + o_p(1) \end{aligned}$$

$$= \max_{n_{\text{low}} \le i \le n_{\text{up}}} \left| n^{-\frac{1}{2}} \left(S_i^1 - \frac{i}{n} S_n^1 \right) \right| + o_p(1)$$

$$= \text{const} \cdot \sup_{0 \le u \le 1} \left| \tilde{B}(u) \right| + o_p(1)$$

with $\tilde{B}$ denoting a standard Brownian bridge. Analogous arguments can be carried out using a quasi-MLE based on quadratic forms. The derivation given here is, of course, purely heuristic, an exact proof is more difficult. For the approach based on quadratic forms, a complete proof can be found in Horváth and Shao (1999). Specifically, the following result is derived.

Theorem 7.42 *Consider a parametric family* $Y_t = \sum_{j=-\infty}^{\infty} a_j(\eta)\varepsilon_{t-j}$ *of second-order stationary linear processes with* $\theta = (\sigma_\varepsilon^2, \eta^T)^T = (\sigma_\varepsilon^2, d, \dots)^T \in \Theta \subseteq \mathbb{R}_+ \times (0, \frac{1}{2}) \times \mathbb{R}^{p-2}$. *Suppose that we observe* $Y_1, \dots, Y_n$ *with the true parameter* θ^0 *in the interior of* Θ^0. *Let* $\hat{d}_{1,i}$ *and* $\hat{d}_{i+1,n}$ *be the first components of* $\hat{\eta}_{1,i}$ *and* $\hat{\eta}_{i,n}$ *respectively obtained by Whittle estimation. Assume furthermore that the conditions in the central limit theorem for Whittle estimators given in Giraitis and Surgailis (1990) hold, and also* $E(\varepsilon_t^{4+r}) < \infty$ *for some* $r > 0$. *Denote by* $\Sigma_\eta = 4\pi V^{-1}$ *the asymptotic covariance matrix of* $\hat{\eta}$ *with*

$$V = \int \partial/\partial\eta \log f [\partial/\partial\eta \log f]^T \, d\lambda$$

and by $v_d = [\Sigma_\eta]_{11}$ *the asymptotic variance of* $\hat{d}$. *Then*

$$n^{\frac{1}{2}} u(1-u)(\hat{d}_{1,i} - \hat{d}_{i+1,n}) \to \sqrt{v_d}\, \tilde{B}(u)$$

where $\tilde{B}(u)$ *is a standard Brownian bridge.*

The theorem implies that under the null hypothesis

$$T = \sqrt{n} D_{1,n} = \sqrt{n} v_d^{-\frac{1}{2}} \max_{n_{\text{low}} \le i \le n_{\text{up}}} \left| \frac{i}{n} \left(1 - \frac{i}{n} \right) (\hat{d}_{1,i} - \hat{d}_{i+1,n}) \right| \underset{d}{\to} \sup_{u \in [0,1]} \left| \tilde{B}(u) \right|.$$

Thus, we reject H_0 at the level of significance α, if $T > q_{1-\alpha}$ where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup_{u \in [0,1]} |\tilde{B}(u)|$.

Example 7.42 Let Y_t be a FARIMA$(0, d, 0)$ process. Then $v_d = 6/\pi^2$ so that an approximate rejection region at level α is given by

$$T = \sqrt{n} \frac{\pi}{\sqrt{6}} \max_{n_{\text{low}} \le i \le n_{\text{up}}} \left| \frac{i}{n} \left(1 - \frac{i}{n} \right) (\hat{d}_{1,i} - \hat{d}_{i+1,n}) \right| > q_{1-\alpha}.$$

We apply this method to the detrended central England temperature series considered before. The practical difficulty one encounters is that it is not clear how

to choose n_{low} and n_{up}. Although the results in Horváth and Shao suggest that asymptotically one may choose $n_{\text{low}} = 1$ and $n_{\text{up}} = n$, this is not really true because the calculation of the MLE based on one (or a very small number of) observation is not meaningful; in fact, for very small samples, numerical optimization often fails to find a solution in the interior of the parameter space. Here, we chose $n_{\text{low}} = 100$ and $n_{\text{up}} = n - 100 = 252$. This means, however, that $u = n/n_{\text{low}} \approx 0.28$ and $u = n_{\text{up}}/n \approx 0.72$ are far from the left and right border of the interval $[0, 1]$. Instead of using quantiles of the supremum of $|\tilde{B}(u)|$ over the whole range of $u \in [0, 1]$ we therefore calculated quantiles of $\sup_{u\in[0.28,0.72]} |\tilde{B}(u)|$. The critical 5 %-level value is about 1.34. The observed value of T is 0.99 so that, in contrast to the simple χ^2-test calculated previously, H_0 is not rejected.

The failure to reject in this example may be due to the (conjectured) possibility that the potential change points are near the two borders of the observational period (recall that the estimates of d calculated for the four blocks were 0.30, 0.07, 0.02 and 0.29). The test based on T has little power when changes occur near the borders because the variance of $\tilde{B}(u)$ is equal to $u(1-u)$ and thus approaches zero at the two ends. One may increase the power by changing the standardization by the factor $[u(1-u)]^{-\frac{1}{2}}$ and hence using the statistic

$$\tilde{T} = \sqrt{n}\tilde{D}_{1,n} = \sqrt{n}v_d^{-\frac{1}{2}} \max_{n_{\text{low}}\le i\le n_{\text{up}}} \left|\sqrt{\frac{i}{n}\left(1-\frac{i}{n}\right)}(\hat{d}_{1,i} - \hat{d}_{i+1,n})\right|.$$

The derivation of the asymptotic distribution of $\tilde{T}$ is more involved, however, because convergence in $D[0, 1]$ no longer holds. The statistic $\tilde{T}$ was suggested in Beran and Terrin (1996), its asymptotic distribution was derived by Horváth and Shao (1999). Under additional regularity conditions, Horváth and Shao obtain the asymptotic expression

$$\lim_{n\to\infty} P\left\{\sqrt{2\log n}\sqrt{n}v_d^{-\frac{1}{2}} \max_{1\le i<n}\left|\sqrt{\frac{i}{n}\left(1-\frac{i}{n}\right)}(\hat{d}_{1,i} - \hat{d}_{i+1,n})\right| \le c(x)\right\}$$
$$= \exp\left(-2e^{-x}\right)$$

where

$$c(x) = x + 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi.$$

Thus, given a level of significance α, we first need to determine x_α such that $\exp(-2e^{-x_\alpha}) = 1 - \alpha$. We reject H_0 at the level of significance α, if

$$\tilde{T} > \frac{c(x_\alpha)}{\sqrt{2\log n}},$$

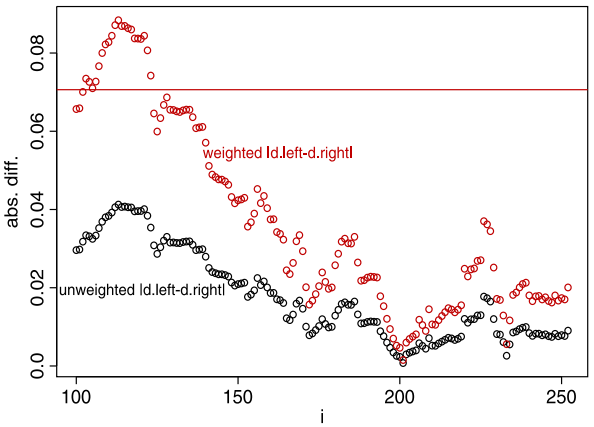

Fig. 7.20 Plot of $|\frac{i}{n}(1-\frac{i}{n})(\hat{d}_{1,i}-\hat{d}_{i+1,n})|$ and $|\sqrt{\frac{i}{n}(1-\frac{i}{n})}(\hat{d}_{1,i}-\hat{d}_{i+1,n})|$ against $i = 100, \ldots, 252$ for detrended yearly Central England temperatures. The *horizontal line* corresponds to the 5 %-critical value for the second statistic. The corresponding critical value for the first statistic is outside the plotted range

where

$$x_\alpha = -\log\log\frac{1}{\sqrt{1-\alpha}}.$$

For instance, for $\alpha = 0.05$ we have $x_\alpha = 3.66$ and $c(x_\alpha) = 5.82$.

Example 7.43 We apply the test based on $\tilde{T}$ to the detrended Central England series, using a FARIMA$(0, d, 0)$ model. For $\alpha = 0.01$ and 0.05 we have $c(x_\alpha)/\sqrt{2\log n} = 2.43$ and 1.70, respectively. The value of $\tilde{T}$ turns out to be 2.13. Thus, in contrast to the test based on T, we can reject H_0 at $\alpha = 0.05$. Figure 7.20 shows a comparison between $|i/n(1-i/n)(\hat{d}_{1,i}-\hat{d}_{i+1,n})|$ and $|\sqrt{i/n(1-i/n)}(\hat{d}_{1,i}-\hat{d}_{i+1,n})|$. Due to the new standardization, the second statistic is indeed much larger near the left border.

7.9.5 Changes in the Mean vs. Long-Range Dependence

One of the controversial issues in the applied literature is whether long-memory phenomena may not be caused by changes in parameters of a short-memory process rather than stationary long-range dependence (see, e.g. Klemes 1974; Boes and Salas 1978; Roughan and Veitch 1999; Veres and Boda 2000; Karagiannis et al. 2004; Diebold and Inoue 2001; Granger and Hyung 2004; Mikosch and Starica 2004; Charfeddine and Guegan 2009; Mills 2007). One way to answer this is the pragmatic view that in situations where the data were actually generated by a more complex short-memory mechanism, stationary processes with long-range dependence often provide a convenient parsimonious model (by including just one additional parameter d or H). Nevertheless, one would at least like to be able to distinguish long memory from certain simple alternatives. Among the most important competitors are short-memory processes with changes in the expected value. Essentially, we may distinguish two situations: (a) $E(Y_t)$ changes gradually; (b) $E(Y_t)$

changes abruptly. In the first case, the standard nonparametric approach is to consider a sequence of models $Y_{t,n} = m(t/n) + X_t$ where X_t is a zero mean stationary process and $m : [0, 1] \to \mathbb{R}$ satisfies certain regularity conditions such as $m \in C[0, 1]$ or $L^2[0, 1]$. This leads back to the question of estimating a deterministic trend function m and parameters describing the stochastic dependence structure simultaneously. This topic is discussed in Sects. 7.4 and 7.5. (Note, in particular, that wavelet thresholding provides a way of distinguishing m from the dependence structure of X_t even if m is not smooth, which is the case under alternatives in change point analysis.)

In this section, we turn to scenario (b) where changes in the expected value are abrupt. The fundamental difficulty of distinguishing between a stationary long-memory process and a short-memory process with change points can be illustrated by the following example. Suppose that X_t are i.i.d. with zero mean. We observe $Y_t = \mu(t) + X_t$ with $\mu(t) = \mu(t; \omega) \in \{0, 1\}$ generated by an ON–OFF process that is independent of X_t and has long memory. In other words,

$$\mu(t; \omega) = W(t) = \sum_{j=-\infty}^{\infty} 1\{\tau_{j-1} \le t < \tau_{j-1} + T_{j,\mathrm{on}}\},$$

with $T_j = \tau_j - \tau_{j-1} = T_{j,\mathrm{on}} + T_{j,\mathrm{off}}$ as defined in Sect. 2.2.3 (there we used the notation $X_{j,\mathrm{on}}$, $X_{j,\mathrm{off}}$ instead of $T_{j,\mathrm{on}}$, $T_{j,\mathrm{off}}$). The distributions of the ON and OFF intervals are such that $P(T_{j,\mathrm{on}} > x) \sim C_{\mathrm{on}} x^{-\alpha_{\mathrm{on}}}$ and $P(T_{j,\mathrm{off}} > x) \sim C_{\mathrm{off}} x^{-\alpha_{\mathrm{off}}}$ with $1 < \alpha_{\mathrm{on}} < \alpha_{\mathrm{off}} < 2$. Then $cov(\mu(t), \mu(t + k)) \sim \mathrm{const} \cdot |k|^{-(\alpha_{\mathrm{on}} - 1)}$. This means that $\mu(t)$ and hence also Y_t has long-range dependence. On the other hand, *conditionally* on $\mu(t; \omega)$ the observations Y_t $(t = 1, 2, \ldots, n)$ are independent. Figures 7.21(a)–(f) show simulated sample paths of $\mu(t; \omega)$, X_t and Y_t, respectively, and the corresponding empirical correlograms. Here, $T_{j,\mathrm{on}}$ and $T_{j,\mathrm{off}}$ are equal to 10 times standard Pareto-distributed variables with $\alpha_{\mathrm{on}} = 1.1$ and $\alpha_{\mathrm{off}} = 1.2$, respectively, i.e. $P(T_{i,\mathrm{off}} > x) = (x/10)^{-1.1}$ and $P(T_{i,\mathrm{off}} > x) = (x/10)^{-1.2}$ (for $x \ge 10$). The correlogram of X_t—which is the same as the *conditional* correlogram of Y_t given $\mu(t; \omega)$—does not show any dependence, whereas in the (unconditional) correlogram of Y_t the long memory of μ leaks in. If we observe one sample path of the process Y_t only, then in principle we are not able to tell whether $\mu(t)$ has been generated randomly or if it is deterministic, unless we know or assume a priori that the class of possible deterministic functions has certain properties that make them distinguishable asymptotically from typical sample paths of the long-memory ON–OFF process. If, however, no assumptions are imposed on the function $E(Y_t)$, then one realization of the process Y_t with μ generated by the ON–OFF process can also be interpreted as a series of independent observations with deterministic shifts in the expected value. More generally, one can say that the question whether we have stationarity with long memory or short memory with shifts in the mean function is ill-posed, unless one specifies a priori some detailed properties of the shifts in $E(Y_t)$. Such restrictions may be, for example, the maximal number, the frequency, the location, the spacing, integrability or the size of shifts.

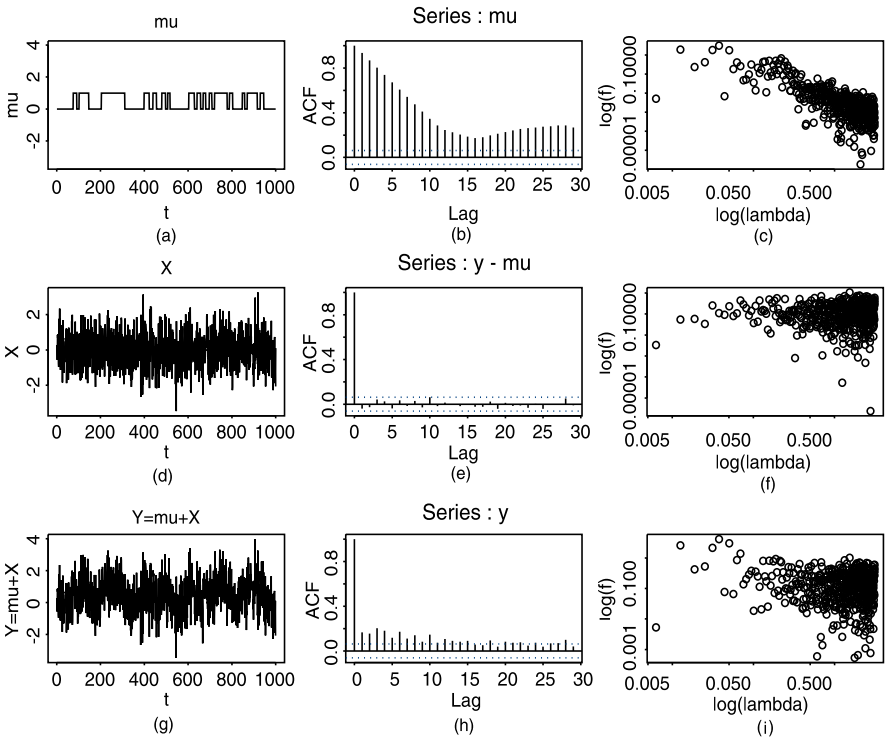

Fig. 7.21 Figure (**g**) shows a simulated sample path of $Y_t = \mu(t/n) + X_t$ where X_t are i.i.d. $N(0, 1)$-variables and $\mu(u)$ $(u \in [0, 1])$ is generated by an ON–OFF-process with long-range dependence. The ON–OFF-process is displayed in (**a**), the residual process X_t in (**d**). Also shown are the corresponding correlograms ((**b**), (**e**) and (**h**)) and log–log-periodograms ((**c**), (**f**) and (**i**))

Once we have decided on what type of change point models we would like to compare with, an appropriate statistical test can be set up. Depending on the application, the assumption of stationarity with long memory can be assigned to the null hypothesis H_0 or to the alternative H_1. The former is considered, for instance, in Ohanissian et al. (2008), Müller and Watson (2008), Qu (2010), Kuswanto (2011), the latter in Berkes et al. (2006), Jach and Kokoszka (2008) and Baek and Pipiras (2011).

As an example, we discuss the method proposed by Berkes et al. (2006). The idea is to start with testing

$$H_0 : Y_t = \mu + \Delta \cdot 1\{t > t_0 + 1\} + X_t \quad (\Delta \neq 0)$$

where $1 \leq t_0 < n$ and X_t is a fourth-order stationary zero mean *short-memory* process with absolutely summable autocovariances $\gamma_X(k)$ in the domain of attraction of a Brownian motion. The alternative is

$$H_1 : Y_t = \mu + X_t$$

where X_t is a fourth-order stationary zero mean *long-memory* process with autocovariances $\gamma_X(k) \sim c_\gamma |k|^{2d-1}$ ($|k| \to \infty$) for some $0 < d < \frac{1}{2}$, in the domain of attraction of a *fractional* Brownian motion. An additional technical assumption is that under H_0 the fourth-order cumulants

$$\begin{aligned}\kappa(k_1, k_2, k_3) &= cum(X_t, X_{t+k_1}, X_{t+k_2}, X_{t+k_3}) \\ &= E(X_t X_{t+k_1} X_{t+k_2} X_{t+k_3}) \\ &\quad - \big(\gamma_X(k_1)\gamma_X(k_2 - k_3) + \gamma_X(k_2)\gamma_X(k_1 - k_3) + \gamma_X(k_3)\gamma_X(k_1 - k_2)\big)\end{aligned}$$

are such that

$$\sup_{k_1} \sum_{k_2,k_3=-\infty}^{\infty} \big|\kappa(k_1, k_2, k_3)\big| < \infty.$$

Under H_1, the fourth-order cumulants are assumed to be such that

$$\sup_{k_1} \sum_{k_2,k_3=-n}^{n} \big|\kappa(k_1, k_2, k_3)\big| = O\big(n^{2d}\big).$$

The idea of the test proposed in Berkes et al. (2006) is to use a CUSUM statistic with a standardization of the order $O(\sqrt{n})$ that leads to a well known limiting distribution under H_0, but to divergence under H_1 because there dividing by $n^{\frac{1}{2}}$ is not enough. The distribution of CUSUM statistics is well known under the assumption of no change in the mean. Under the null hypothesis considered here, we have *one* change point. If we knew the change point t_0, then we could consider a CUSUM statistic for $Y_1, \ldots, Y_{t_0}$ and another CUSUM statistic for $Y_{t_0+1}, \ldots, Y_n$ separately. For each statistic, the asymptotic distribution could be calculated using the supremum of a Brownian bridge. A natural approach to testing H_0 is therefore to first estimate the change point t_0, and then to consider the two CUSUM statistics for Y_t ($t \le \hat{t}_0$) and Y_t ($t \ge \hat{t}_0 + 1$). Estimation of t_0 can also be done by means of a CUSUM statistic. Thus, we define

$$\hat{t}_0 = \min\Big\{i : |V_i| = \max_{1 \le i \le n} |V_i|\Big\}$$

where

$$V_i = S_{1,i} - \frac{i}{n} S_{1,n}.$$

Given $\hat{t}_0$, we consider

$$D_{1,\hat{t}_0} = \max_{1 \le i \le \hat{t}_0} \left| S_{1,i} - \frac{i}{\hat{t}_0} S_{1,\hat{t}_0} \right|$$

and

$$D_{\hat{t}_0+1,n} = \max_{\hat{t}_0+1 \le i \le n} \left| S_{\hat{t}_0+1,i} - \frac{i - \hat{t}_0}{n - \hat{t}_0} S_{\hat{t}_0+1,n} \right|.$$

Note that in both cases, the location parameter is removed automatically. The essential part is therefore the standardization of $D_{1,\hat{t}_0}$ and $D_{\hat{t}_0+1,n}$. To obtain a standardization that corresponds to $\sqrt{\mathrm{var}(S_{1,t_0})}$ and $\sqrt{\mathrm{var}(S_{t_0+1,n})}$ asymptotically under H_0, but remains of the order $O(\sqrt{n})$ under H_1, Berkes et al. (2006) propose Bartlett estimators defined by

$$v_{1,\hat{t}_0} = \sum_{u=-(m_{\hat{t}_0}-1)}^{m_{\hat{t}_0}-1} \left(1 - \frac{|u|}{m_{\hat{t}_0}}\right) \hat{\gamma}_{1,\hat{t}_0}(u),$$

$$v_{\hat{t}_0+1,n} = \sum_{u=-(m_{n-\hat{t}_0}-1)}^{m_{n-\hat{t}_0}-1} \left(1 - \frac{|u|}{m_{n-\hat{t}_0}}\right) \hat{\gamma}_{\hat{t}_0+1,n}(u)$$

where $m_{\hat{t}_0}$ and $m_{n-\hat{t}_0}$ tend to infinity at a slower rate than n. Here we use the notation

$$\hat{\gamma}_{i,j}(u) = \frac{1}{n_{i,j}} \sum_{t=i}^{j-|u|} (Y_t - \bar{y}_{i,j})(Y_{t+|u|} - \bar{y}_{i,j})$$

for the sample autocovariance at lag u (where $j > i$), based on observations $Y_i, Y_{i+1}, \ldots, Y_j$, with $n_{i,j} = j - i + 1$ and $\bar{y}_{i,j} = n_{i,j}^{-1} S_{i,j}$. If it is assumed that under H_0 the change point $\hat{t}_0$ is asymptotically proportional (but not equal) to n, then $v_{1,\hat{t}_0}$ and $v_{\hat{t}_0+1,n}$ both converge in probability to $\sum_{u=-\infty}^{\infty} \gamma_X(u) = 2\pi f_X(0)$. This is the asymptotic variance of a standardized sum since $\mathrm{var}(S_{1,n}) \sim 2\pi f_X(0) n$. On the other hand, under H_1, $\mathrm{var}(S_{1,n}) \sim c_S n^{2d}$, but $v_{1,\hat{t}_0}$ and $v_{\hat{t}_0+1,n}$ diverge to infinity at a slower rate than n^{2d}. This essentially follows from $\sum_{k=1}^{m} k^{2d-1} \sim \mathrm{const} \cdot m^{2d} = o(n^{2d})$. Thus we obtain the desired asymptotic properties for the test statistics

$$T_{1,\hat{t}_0} = \hat{t}_0^{-\frac{1}{2}} v_{1,\hat{t}_0}^{-\frac{1}{2}} D_{1,\hat{t}_0}$$

and

$$T_{\hat{t}_0+1,n} = (n - \hat{t}_0)^{-\frac{1}{2}} v_{\hat{t}_0+1,n}^{-\frac{1}{2}} D_{\hat{t}_0+1,n}.$$

More specifically, Berkes et al. (2006) use following additional conditions:

$$t_0 = [n\vartheta] \quad \text{for some } 0 < \vartheta < 1,$$

$$\Delta \to 0, \qquad n\Delta^2 \to \infty, \qquad m_n \Delta^2 = O(1),$$

and

$$\Delta^2 |\hat{t}_0 - t_0| = O_p(1).$$

The joint distribution of the two statistics under H_0 is given by

Theorem 7.43 *Suppose H_0 holds, and m_n is nondecreasing, $m_n \to \infty$ and such that*

$$\sup_{k\geq 0} \frac{m_{2^{k+1}}}{m_{2^k}} < \infty, \qquad m_n(\log n)^4 = O(n).$$

Then, under the conditions above,

$$(T_{1,\hat{t}_0}, T_{\hat{t}_0+1,n}) \underset{d}{\to} \Big(\sup_{0\leq u\leq 1} \big|\tilde{B}^{(1)}(u)\big|, \sup_{0\leq u\leq 1} \big|\tilde{B}^{(2)}(u)\big| \Big)$$

where $\tilde{B}^{(1)}$, $\tilde{B}^{(2)}$ are two independent Brownian bridges, i.e. $\tilde{B}^{(i)}(u) = B^{(i)}(u) - uB^{(i)}(1)$ with $B^{(i)}$ $(i = 1, 2)$ two independent standard Brownian motions.

In contrast, under the alternative, we have long-range dependence so that the rate of convergence of sums is slower, the two statistics are no longer asymptotically independent and their distribution can be expressed in terms of *one* common fractional Brownian motion:

Theorem 7.44 *Suppose that H_1 holds, and m_n is nondecreasing, $m_n \to \infty$ and such that*

$$\sup_{k\geq 0} \frac{m_{2^{k+1}}}{m_{2^k}} < \infty, \qquad m_n(\log n)^{\frac{7}{2-4d}} = O(n).$$

Then, under the conditions above,

$$\left(\left(\frac{m_{\hat{t}_0}}{n}\right)^d T_{1,\hat{t}_0}, \left(\frac{m_{n-\hat{t}_0}}{n}\right)^d T_{\hat{t}_0+1,n} \right) \underset{d}{\to} (Z_1, Z_2)$$

where

$$Z_1 = \tau^{-\frac{1}{2}} \sup_{0\leq u\leq \tau} \left| B_H(u) - \frac{u}{\tau} B_H(\tau) \right|,$$

$$Z_2 = (1-\tau)^{-\frac{1}{2}} \sup_{\tau\leq u\leq 1} \left| B_H(u) - B_H(\tau) - \frac{u-\tau}{1-\tau}\big(B_H(1) - B_H(\tau)\big) \right|,$$

B_H is a fractional Brownian motion with self-similarity parameter $H = d + \frac{1}{2}$ and

$$\tau = \inf\Big\{ t \geq 0 : \big|B_H(t)\big| = \sup_{0\leq u\leq 1} \big|B_H(u)\big| \Big\}.$$

By assumption $m_{\hat{t}_0}/n$ and $m_{n-\hat{t}_0}/n$ converge to zero so that, under H_1, the vector $(T_{1,\hat{t}_0}, T_{\hat{t}_0+1,n})$ diverges to (∞, ∞) in probability. Defining

$$T = \max\{T_{1,\hat{t}_0}, T_{\hat{t}_0+1,n}\},$$

we have

$$T \underset{d}{\to} \max\Big\{ \sup_{0\leq u\leq 1} \big|\tilde{B}^{(1)}(u)\big|, \sup_{0\leq u\leq 1} \big|\tilde{B}^{(2)}(u)\big| \Big\},$$

under H_0 whereas under H_1 the statistic diverges to infinity. The results can be extended to H_0 including several shifts in the mean.

An essential element in the test procedure by Berkes et al. (2006) is the Bartlett estimator based on sample autocovariances. Apart from the difficulty of choosing appropriate sequences $m_{\hat{t}_0}$ and $m_{n-\hat{t}_0}$, more efficient estimators of the asymptotic values of $\gamma_X(k)$ exist because $\gamma_X(k) \sim c_\gamma |k|^{2d-1}$ is characterized by two parameters only. A test where all autocovariances are estimated by the sample autocovariance is likely to have relatively low power. Baek and Pipiras (2011) therefore suggest a more powerful test procedure where the hyperbolic shape of the autocovariances and the spectral density is exploited more directly. As before, in a first step $\hat{t}_0$ is calculated. In a second step, the data are centred using $\hat{t}_0$ by defining

$$\hat{X}_t = Y_t - \bar{y}_{1,\hat{t}_0} \quad (1 \le t \le \hat{t}_0),$$
$$\hat{X}_t = Y_t - \bar{y}_{\hat{t}_0+1,n} \quad (\hat{t}_0 + 1 \le t \le n).$$

The third step is to estimate the long-memory parameter from $\hat{X}_1, \ldots, \hat{X}_n$. If $\hat{t}_0$ converges to t_0 fast enough, then $\hat{d}$ converges to the true value d_0 under H_0 and under H_1. Thus, if we are able to establish that under H_0 a standardized statistic $n^\beta(\hat{d} - d^0)$ converges to a nondegenerate random variable ζ, then we may use the test statistic $T^* = |n^\beta(\hat{d} - \frac{1}{2})|$. Under H_0, T^* converges in distribution to $|\zeta|$ whereas under H_1 the statistic diverges to infinity because the true value of d is not $\frac{1}{2}$. For instance, Baek and Pipiras (2011) show the following result for the local Whittle estimator.

Theorem 7.45 *Let $\hat{d}$ be a local Whittle estimator based on $\hat{X}_t$ using m Fourier frequencies $\lambda_j = 2\pi j/n$ $(j = 1, 2, \ldots, m)$. Suppose that conditions used in the theorems above as well as regularity conditions needed for the Whittle estimator (see Theorem 2 in Robinson 1995b; also see Chap. 5) hold. Furthermore, assume*

$$\frac{m \log^2 m}{n\Delta^2} \to 0.$$

Then, under H_0,

$$\sqrt{m}\left(\hat{d} - \frac{1}{2}\right) \underset{d}{\to} \zeta \sim N\left(0, \frac{1}{4}\right),$$

whereas under H_1 with $d^0 \in (0, \frac{1}{2})$,

$$\hat{d} \underset{d}{\to} d^0.$$

For exact regularity conditions and detailed proofs, see Baek and Pipiras (2011). Note that Δ is even allowed to tend to zero; however, at a slower rate than $\log m\sqrt{m/n}$. The theorem essentially says that estimation of t_0 does not change the asymptotic distribution of the local Whittle estimator under H_0, and under H_1 the

estimator remains consistent. We may therefore reject H_0 at the level of significance α if

$$T^* = \left| \sqrt{m} \left(\hat{d} - \frac{1}{2} \right) \right| > \frac{1}{2} z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution.

7.10 Estimation of Rapid Change Points in the Trend Function

In this section, we address rapid change point detection in a nonparametric regression function where the regression residuals are Gaussian subordinated via an unknown function (see Sect. 7.6) with long-memory. Due to a specific application that we have in mind, we base our estimation procedure on time series observed at unevenly spaced time points. In fact, this type of problem tends to occur in palaeoclimatic research where in order to answer questions concerning past environmental changes, one may analyse environmental proxies such as pollens, oxygen and other gas isotopes that are found in ice or sediment samples. Such environmental proxies give rise to time series data, where the successive observations are unevenly spaced in time. One important topic is rapid climate change where one is concerned with identification of rapid change points in the trend function; see Ammann et al. (2000) for background information on palaeoclimatic research. Most of the material covered in this section can be found in Menéndez et al. (2010); also see Menéndez (2009) and Menéndez et al. (2012). We start by introducing a continuous time stationary Gaussian process $Z(u)$ $(u \in \mathbb{R})$ with $E[Z(u)] = 0$, $\mathrm{var}(Z) = 1$ and

$$\gamma_Z(v) = cov\big(Z(u), Z(u+v)\big) \sim C_Z v^{2H-2}$$

as $v \to \infty$ where $H \in (0,1)$. Here "$\sim$" means that the ratio of the left and right hand side tends to one. The observed time series $Y_1, \ldots, Y_n$ is assumed to be generated by a nonparametric regression model of the form

$$Y_i = m(t_i) + \varepsilon_i$$

where $\varepsilon_i = G(Z(T_i), t_i)$, $T_i \in \mathbb{R}_+$, $T_1 \le T_2 \le \cdots \le T_n$, $t_i = T_i / T_n \in [0,1]$ and $m(\cdot)$ is a smooth function. For each fixed $t \in [0,1]$ the function $G(\cdot, t)$ is assumed to be in the L^2-space of functions (on $\mathbb{R}$) with $E[G(Z,t)] = (2\pi)^{-\frac{1}{2}} \int G(z,t) \exp(-z^2/2)\, dz = 0$ and $\|G\|^2 = E[G^2(Z,t)] < \infty$. This implies a convergent L^2-expansion

$$G(Z_i, t_i) = \sum_{k=q}^{\infty} \frac{c_k(t_i)}{k!} H_k(Z_i)$$

where $H_k(\cdot)$ are Hermite polynomials and $q \ge 1$ is the Hermite rank. The function G provides the possibility of having non-Gaussian residuals with a changing marginal

distribution (see Sect. 7.6). The spacings between the successive time points is arbitrary except for some technical conditions (similar in spirit as the equidistant case, where $T_i = iT_n/n$ and $t_i = i/n$).

Rapid change is defined in terms of derivatives of the trend function. Such a change may be rapid but it is a continuous change in the trend function m. More specifically, rapid change is said to occur whenever the absolute value of the first derivative of m has a local maximum and exceeds a certain threshold. Let $m^{(i)}(t)$ denote the ith derivative of m with respect to t. We shall follow this definition of a rapid change point considered in Müller and Wang (1994) in the context of hazard rate estimation:

Definition 7.9 Given a threshold $\eta > 0$, the p time points $\{\tau_1, \tau_2, \ldots, \tau_p\} \in (0, 1)$ are rapid change points of the trend function m if

$$\begin{aligned}
&\left|m^{(1)}(\tau_1)\right| \geq \left|m^{(1)}(\tau_2)\right| \geq \cdots \geq \left|m^{(1)}(\tau_p)\right| \geq \eta, \\
&m^{(2)}(\tau_i) = 0, \quad i = 1, \ldots, p \quad \text{and} \\
&0 < \left|m^{(3)}(\tau_i)\right| < \infty.
\end{aligned}$$

In applications, the trend derivatives will have to be estimated. Thus consider the non-parametric curve estimates using Priestley–Chao type kernel estimator

$$\hat{m}^{(\nu)}(t) = \frac{(-1)^\nu}{b^{\nu+1}} \sum_{i=1}^{n} (t_i - t_{i-1}) K^{(\nu)}\left(\frac{t_i - t}{b}\right) Y_i$$

where $\nu = 0, 1, 2, \ldots, t_0 = 0$ and the kernel K satisfies the following conditions (Gasser and Müller 1984):

(i) $K \in C^{\nu+1}[-1, 1]$;
(ii) $K(x) \geq 0$, $K(x) = 0$ $(|x| > 1)$, $\int_{-1}^{1} K(x)\,dx = 1$;
(iii) $\forall x, y \in [-1, 1]$, $|K^{(\nu)}(x) - K^{(\nu)}(y)| \leq L_0|x - y|$ where $L_0 \in \mathbb{R}^+$ is a constant;
(iv) K is of order (ν, k), $\nu \leq k - 2$, where k is a positive integer, i.e.

$$\int_{-1}^{1} K^{(\nu)}(x) x^j\,dx = \begin{cases} (-1)^\nu \nu!, & j = \nu, \\ 0, & j = 0, \ldots, \nu - 1, \nu + 1, \ldots, k - 1, \\ \theta, & j = k \end{cases}$$

where $\theta \neq 0$ is a constant;
(v) $K^{(j)}(1) = K^{(j)}(-1) = 0$ for all $j = 0, 1, \ldots, \nu - 1$.

It turns out that by Lemma 1 in Gasser and Müller (1984) one can also write

$$\int_{-1}^{1} K(x) x^j\,dx = \begin{cases} 1, & j = 0, \\ 0, & j = 1, \ldots, k - \nu - 1, \\ (-1)^\nu \theta \frac{(k-\nu)!}{k!}, & j = k - \nu. \end{cases}$$

For a given sample and a fixed value of the first derivative threshold η, the number of change points $\hat{p}$ where $\hat{m}^{(2)}$ is zero is random whereas the true number of change points p is unknown. However, as the sample size increases, under suitable regularity conditions on m, consistency of $\hat{m}$ and $\hat{p}$ follows. The following technical conditions are used to prove the consistency result in the theorem below:

(A1) The coefficients $c_k(t) = E[G(Z,t)H_k(Z)]$ in the Hermite expansion of $G(Z,t)$ are continuously differentiable with respect to $t \in [0,1]$;
(A2) $1-(2q)^{-1} < H < 1$;
(A3) $m \in C^{\nu+1}[0,1]$;
(A4) $0 \le T_1 \le T_2 \le \cdots \le T_n$, $t_i = T_i/T_n \in [0,1]$;
(A5) $\alpha_n^{-1} \le t_j - t_{j-1} \le \beta_n^{-1}$ where $\alpha_n \ge \beta_n > 0$ and $\beta_n \to \infty$;
(A6) $b \to 0$, $b^{2\nu}(T_n b)^{(2-2H)q} \to \infty$, and $b\beta_n \to \infty$;
(A7) $\lim_{n\to\infty}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2} = 0$;
(A8) $K \in C^{\nu+1}[0,1]$ with $0 < c_{\nu+1} = \sup_{u\in[0,1]}|K^{(\nu+1)}(u)| < \infty$.

The following observations can be made. (A1) implies a slowly changing marginal distribution of the regression residuals. This may be understood as a type of local-stationarity. Due to (A2), the long-memory property of Z_i is inherited by the subordinated error process. (A5) ensures that no repeated time points and, more generally, no extreme clustering of the time points occurs. A special case is when the observations are available at equidistant time points (set $\alpha_n = \beta_n = n$). The first condition in (A6) is needed to avoid an asymptotic bias in $\hat{m}^{(\nu)}(t)$ whereas the second and the third conditions ensure convergence of the asymptotic expression for the variance of $\hat{m}^{(\nu)}(t)$ to zero. (A7) is needed for the asymptotic approximation of the mean squared error. Due to (A2), $(2-2H)q < 1$ so that (A7) is possible although $\alpha_n \ge \beta_n$. For additional discussions and related results, specifically for monotone transforms G and slightly different conditions on the spacings between successive observations $T_i - T_{i-1}$, see Menéndez et al. (2012).

Theorem 7.46 *Under the assumptions stated earlier in this section and* (A1)–(A7), *we have for* $t \in (0,1)$:

$$\begin{aligned}
\mathrm{Bias}\big(\hat{m}^{(\nu)}(t)\big) &= E\big[\hat{m}^{(\nu)}(t)\big] - m^{(\nu)}(t) = b^{k-\nu}J_{\nu,k} + o\big(b^{k-\nu}\big),\\
\mathrm{Var}\big(\hat{m}^{(\nu)}(t)\big) &= b^{-2\nu}(T_n b)^{(2H-2)q} I_q(t) + o\big(b^{-2\nu}(T_n b)^{(2H-2)q}\big),\\
MSE\big(\hat{m}^{(\nu)}(t)\big) &= E\big[\big(\hat{m}^{(\nu)}(t) - m^{(\nu)}(t)\big)^2\big]\\
&= b^{2(k-\nu)}J_{\nu,k}^2(t) + b^{-2\nu}(T_n b)^{(2H-2)q} I_q(t)\\
&\quad + o\big(\max\big(b^{2(k-\nu)}, b^{-2\nu}(T_n b)^{(2H-2)q}\big)\big)
\end{aligned}$$

where

$$I_q(t) = \frac{c_q^2(t)}{q!} C_Z^q \int_{-1}^{1}\int_{-1}^{1} K^{(\nu)}(u)K^{(\nu)}(v)|u-v|^{(2H-2)q}\,du\,dv$$

and

$$J_{v,k}(t) = \frac{m^{(k)}(t)}{k!} \int_{-1}^{1} K^{(v)}(u) u^{k-v}\, du.$$

Proof Let $t \in (0,1)$ be a scalar. The expression for the bias follows from a Taylor series expansion of m and properties of the kernel. To prove the result for the variance, note that

$$\begin{aligned} &b^{2v}(T_n b)^{(2-2H)q} \operatorname{Var}\big(\hat{m}^{(v)}(t)\big) \\ &\quad = b^{-2}(T_n b)^{(2-2H)q} \sum_{i,j=1}^{n} (t_i - t_{i-1})(t_j - t_{j-1}) K^{(v)}\left(\frac{t - t_i}{b}\right) K^{(v)}\left(\frac{t - t_j}{b}\right) V_{i,j} \end{aligned}$$

where

$$V_{i,j} = Cov(Y_i, Y_j) = \sum_{l=q}^{n} \frac{c_l(t_i) c_l(t_j)}{l!} \gamma_Z^l(T_i - T_j).$$

Recalling

$$\gamma_Z(T_i - T_j) \sim C_Z |T_i - T_j|^{2H-2}$$

and $-1 < (2H-2)q < 0$, we have

$$Cov(Y_i, Y_j) \sim \frac{c_q^2(t)}{q!} \gamma_Z^q(T_i - T_j)$$

for $i, j \in U_b(t)$ with $U_b = \{k \in \mathbb{N} : |t - T_k / T_n| \le b\}$. It is then sufficient to consider

$$\begin{aligned} S_n = b^{-2}(T_n b)^{(2-2H)q} &\sum_{i \neq j} (t_i - t_{i-1})(t_j - t_{j-1}) K^{(v)}\left(\frac{t_i - t}{b}\right) K^{(v)} \\ &\times \left(\frac{t_j - t}{b}\right) |T_i - T_j|^{(2H-2)q}. \end{aligned}$$

Since $K(u) = 0$ for $|u| > 1$, we have

$$S_n = \sum_{i : |T_i - t T_n| \le T_n b} K^{(v)}\left(\frac{t_i - t}{b}\right) \frac{t_i - t_{i-1}}{b} [S_{i,1} + S_{i,2}]$$

where

$$S_{i,1} = \sum_{j \in A_i} K^{(v)}\left(\frac{t_j - t}{b}\right) \cdot \left(\frac{t_i - t_j}{b}\right)^{(2H-2)q} \frac{t_j - t_{j-1}}{b},$$

$$S_{i,2} = \sum_{j \in B_i} K^{(v)}\left(\frac{t_j - t}{b}\right) \cdot \left(\frac{t_i - t_j}{b}\right)^{(2H-2)q} \frac{t_j - t_{j-1}}{b},$$

$$A_i = \{j \in \mathbb{N} : 1 \le j \le i-1, |T_i - tT_n| \le T_n b\} \quad \text{and}$$
$$B_i = \{j \in \mathbb{N} : i+1 \le j \le n, |T_i - tT_n| \le T_n b\}.$$

Setting

$$h_n(x) = K^{(\nu)}\left(x - \frac{t}{b}\right) \times \left(\frac{t_i}{b} - x\right)^{(2H-2)q},$$

we have

$$S_{i,1} = \int_{t_1/b}^{t_{i-1}/b} h_n(x)\,dx + \sum_{j \in A_i} h_n'(x_j)\left(\frac{t_j - t_{j-1}}{b}\right)^2 = \int_{t_1/b}^{t_{i-1}/b} h_n(x)\,dx + r_{n,i,1}$$

and an analogous expression for $S_{i,2}$ where $t_{j-1}/b \le x_j \le t_j/b$ and $h_n'(x) = g_{n,1}(x) + g_{n,2}(x)$ with

$$g_{n,1}(x) = K^{(\nu+1)}\left(x - \frac{t}{b}\right) \times \left(\frac{t_i}{b} - x\right)^{(2H-2)q} \quad \text{and}$$
$$g_{n,2}(x) = K^{(\nu)}\left(x - \frac{t}{b}\right) \times \left(\frac{t_i}{b} - x\right)^{(2H-2)q-1} \times (2-2H)q.$$

By assumption we have $\alpha_n^{-1} \le |t_j - t_{j-1}| \le \beta_n^{-1}$, $-1 < (2H-2)q < 0$ and

$$0 \le \sup_{u \in [0,1]} \left|K^{(\nu+1)}(u)\right| = c_{\nu+1} < \infty.$$

Also note that the assumption $b\beta_n \to \infty$ implies $b\alpha_n \to \infty$. Using the notation $j_1 = [\alpha_n(t-b)]$ and $j_2 = [\alpha_n(t+b)]$, an upper bound can be given by

$$\left|\sum_{j \in A_i} g_{n,1}(x_j)\left(\frac{t_j - t_{j-1}}{b}\right)^2\right| \le c_{\nu+1} b^{-2}\beta_n^{-2} \sum_{j=j_1}^{j_2} \left(\frac{t_i - t_j}{b}\right)^{(2H-2)q}$$
$$\le c_{\nu+1} b^{-2}\beta_n^{-2} \sum_{j=1}^{[2b\alpha_n]} \left(\frac{j}{b\alpha_n}\right)^{(2H-2)q}$$
$$= c_{\nu+1} b^{-1}\alpha_n\beta_n^{-2} \sum_{j=1}^{[2b\alpha_n]} \left(\frac{j}{b\alpha_n}\right)^{(2H-2)q} \frac{1}{b\alpha_n}$$
$$\le c_{\nu+1} b^{-1}\alpha_n\beta_n^{-2} \int_o^2 x^{(2H-2)q}\,dx.$$

Thus if $(2H-2)q > -1$ and $\lim_{n\to\infty} b^{-1}\alpha_n\beta_n^{-2} = 0$ there is a uniform (in i) upper bound on the remainder term $r_{n,i,1}$. Note that $1 + (2-2H)q > 1$ and $b\alpha_n \to \infty$ so that $\lim_{n\to\infty} b\alpha_n(b\beta_n)^{-2} = 0$ follows from the assumption that

$\lim_{n\to\infty}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2} = 0$. Similarly, considering the remainder term $r_{n,,i,2}$ for $g_{n,2}$, we have

$$\begin{aligned}
\left|\sum_{j\in A_i} g_{n,2}(x_j)\left(\frac{t_j - t_{j-1}}{b}\right)^2\right| &\le c_{\nu+1}(b\beta_n)^{-2}\sum_{j=j_1}^{j_2}\left(\frac{t_i - t_j}{b}\right)^{(2H-2)q-1} \\
&\le c_{\nu+1}(b\beta_n)^{-2}\sum_{j=1}^{[2b\alpha_n]}\left(\frac{j}{b\alpha_n}\right)^{(2H-2)q-1} \\
&= c_{\nu+1}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2}\sum_{j=1}^{[2b\alpha_n]} j^{(2H-2)q-1} \\
&\le c_{\nu+1}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2}\sum_{j=1}^{\infty} j^{(2H-2)q-1}
\end{aligned}$$

so that, under the assumption that $H < 1$ and $\lim_{n\to\infty}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2} = 0$, there is a uniform (in i) upper bound on the remainder term $r_{n,i,1}$. Analogous arguments apply to $S_{i,2}$ so that the sum S_n converges to the corresponding double integral and $c_q^2(t)/q!C_Z$ times S_n converges to the asymptotic variance as given in the theorem. □

The asymptotic formula for the mean squared error stated above implies an asymptotically optimal bandwidth of the form

$$b_{\text{opt}} = \left[\frac{2\nu + (2-2H)q}{2(k-\nu)}\frac{I_q}{J_{\nu,k}^2}\right]^{\frac{1}{2k+(2-2H)q}} T_n^{\frac{(2H-2)q}{2k+(2-2H)q}}.$$

The central limit theorem in the corollary below states that if the Hermite rank q equals 1, the limiting distribution of $\hat{m}^{(\nu)}(t)$ is normal and the estimates at different fixed values $t_1, \ldots, t_k$ are asymptotically independent. If, however, $q \ge 2$, a similar limit theorem can be derived but with a non-normal asymptotic distribution which would correspond to the marginal distribution of a Hermite process of order q.

Corollary 7.4 *Suppose that the Hermite rank q of G is one. Let $\mathbf{t} = (t_1, \ldots, t_k)'$, $\hat{\mathbf{m}}^{(\nu)}(\mathbf{t}) = [\hat{m}^{(\nu)}(t_1), \ldots, \hat{m}^{(\nu)}(t_k)]'$ and define the $k \times k$ diagonal matrix*

$$\mathbf{D} = \operatorname{diag}\left(\sqrt{I_1(t_1)}, \ldots, \sqrt{I_1(t_k)}\right).$$

Then, under the assumptions of Theorem 7.46, we have, as n tends to infinity,

$$b^{\nu}(T_n b)^{1-H} D^{-1}\left\{\hat{\mathbf{m}}^{(\nu)}(\mathbf{t}) - E\left[\hat{\mathbf{m}}^{(\nu)}(\mathbf{t})\right]\right\} \underset{d}{\to} (\zeta_1, \ldots, \zeta_k)'$$

where ζ_i are i.i.d. standard normal variables.

Proof The result follows from the previous theorem and the fact that asymptotically the distribution of

$$\Delta_n = (T_n b)^{(1-H)q}\left\{\hat{m}^{(2)}(\tau_i) - E\left[\hat{m}^{(2)}(\tau_i)\right]\right\}$$

is equivalent to the asymptotic distribution of

$$\tilde{\Delta}_n = (T_n b)^{(1-H)q}\frac{(-1)^\nu}{nb^{\nu+1}}\sum_{j=1}^{n} K^{(\nu)}\left(\frac{t_j-\tau_i}{b}\right)\frac{c_q(\tau_i)}{q!}H_q(Z_j)$$

$$= (T_n b)^{1-H}\frac{(-1)^\nu}{nb^{\nu+1}}\sum_{j=1}^{n} K^{(\nu)}\left(\frac{t_j-\tau_i}{b}\right)c_1(\tau_1)Z_j$$

which is a sequence of normal variables. Asymptotic independence of $\hat{m}^{(\nu)}(t)$ and $\hat{m}^{(\nu)}(s)$ for $t \neq s$ follows by analogous arguments as in the proof of the last theorem, along the lines of Csörgő and Mielniczuk (1995b). □

Note that the estimate of the change points will involve estimates of the trend derivatives, which in turn will depend on the respective bandwidths. As we have seen in the theorem earlier, if b is too large, and in particular if $b^{-2\nu}(T_n b)^{(2H-2)q}$ is of smaller order than $b^{2(k-\nu)}$, then the bias of $\hat{\tau}_n$ will dominate the mean squared error and no reasonable confidence interval for τ can be given. Consider, however, (i) $b^{2k} = o((T_n b)^{(2H-2)q})$ which allows the bias to be asymptotically negligible, or (ii) $b^{2k} \sim C \cdot (T_n b)^{(2H-2)q}$ which makes the asymptotic contribution of both bias and variance of the same order. For these cases, if the Hermite rank of G is one, asymptotic normality of $\hat{\tau}_n$ follows.

Theorem 7.47 *Let $\tau = (\tau_1, \tau_2, \ldots, \tau_p)'$ be the points of rapid change of m, and suppose that the assumptions of the corollary to the last theorem hold. Then there is a sequence $\hat{\tau}_n = (\hat{\tau}_{n;1}, \hat{\tau}_{n;2}, \ldots, \hat{\tau}_{n;p})'$ such that $\hat{m}^{(2)}(\hat{\tau}_{n;i}) = 0$ $(1 \le i \le p)$ and $\hat{\tau}_n \to_p \tau$. Moreover, define the $p \times p$ diagonal matrix*

$$\tilde{\mathbf{D}} = \operatorname{diag}\left(\sqrt{I_1(\tau_1)}\Big/\left|m^{(3)}(\tau_1)\right|, \ldots, \sqrt{I_1(\tau_p)}\Big/\left|m^{(3)}(\tau_p)\right|\right).$$

Then the asymptotic distribution of $\hat{\tau}_n$ is given as follows:

(i) *If $b^{2k} = o((T_n b)^{2H-2})$ then $(T_n b)^{1-H}\tilde{\mathbf{D}}^{-1}(\hat{\tau}_n - \tau) \underset{d}{\to} (\zeta_1, \ldots, \zeta_p)'$ where ζ_i are i.i.d. standard normal variables*;

(ii) *If $b^{2k} \sim C \cdot (T_n b)^{2H-2}$ then $(T_n b)^{1-H}\tilde{\mathbf{D}}^{-1}(\hat{\tau}_n - \tau) \underset{d}{\to} (\mu_1 + \zeta_1, \ldots, \mu_p + \zeta_p)'$ where ζ_i are as in* (i) *and*

$$\mu_i = \left[\frac{m^{(k)}(\tau_i)}{k!}\int_{-1}^{1} K^{(\nu)}(u)u^{k-\nu}\,du\right]\Big/ m^{(3)}(\tau_i).$$

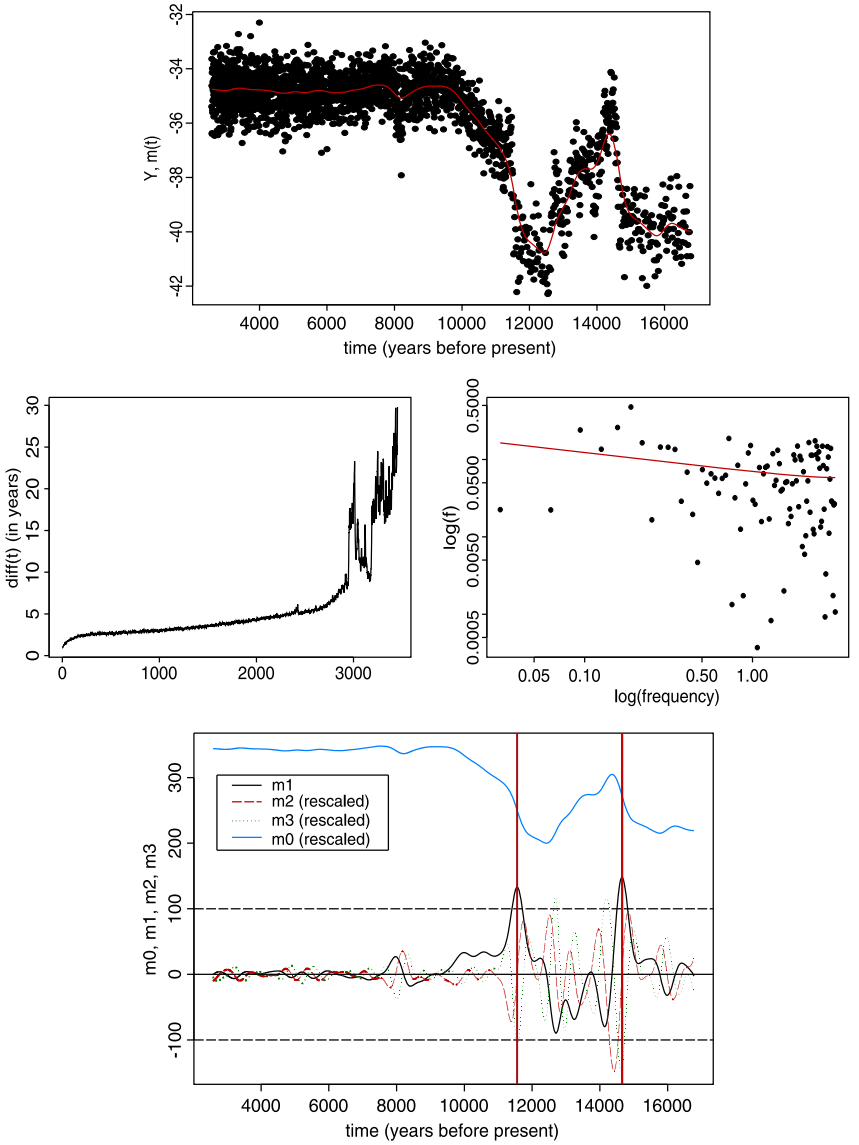

Fig. 7.22 *Top*: Oxygen isotope values plotted against age (years before present or 1989) and an estimated trend curve. *Left middle*: Distance between successive time points. *Right middle*: Periodogram of residuals and fitted spectral density in log-log coordinates. *Bottom*: Estimated trend derivatives $\widehat{m^{(\nu)}}$ ($\nu = 0, 1, 2, 3$). The curve estimates are rescaled for better visibility. The *two vertical lines* mark rapid climate change points where the threshold for the speed of change is set at $\eta = 100$. The two main points of rapid climate change points are estimated to be at around 11,560 and 14'658 years before 1989. The asymptotic 95 %-confidence intervals for the change points (in years before 1989) ignoring bias in estimation are $(11,554; 11,566)$ and $(14,646; 14,670)$, respectively. *Data source*: Greenland Ice Core Project dataset, Johnsen et al. 1997. *The figure is reproduced from the Journal of Statistical Planning and Inference* (2010), vol. **40**, 3343–3354

Proof Consistency follows from $m(t) \in C^{\nu+1}[0, 1]$ and the consistency of $\hat{m}^{(2)}(t)$. For the asymptotic distribution of $\hat{\tau}_n$, we have by Taylor expansion

$$\hat{\tau}_{n:i} - E(\hat{\tau}_{n:i}) = -\hat{m}^{(2)}(\tau_i)\big[m^{(3)}(\tau_i)\big]^{-1} + o_p\big(b^{-2}(T_n b)^{H-1}\big).$$

Since the Hermite rank q of G is equal to one, the limiting behaviour given in (i) and (ii) then follows from the last theorem and its corollary. □

Note that, a similar non-Gaussian limit theorem can be derived for $q \geq 2$. By analogous arguments as above, it can be shown that the number of zeros of $\hat{m}^{(2)}$ with $|\hat{m}^{(2)}| > \eta$ converges to p in probability, so that when n is sufficiently large, p can be estimated with arbitrary precision and in particular, the estimate of p can be plugged-in for computing confidence intervals for the change points.

The example below is concerned with evidence of rapid climate changes in the northern hemisphere approximately 20,000 years before present ('present' being set at 1989). The observations are oxygen isotope ratio measurements from a Greenland ice core (Johnsen et al. 1997) resulting in unevenly spaced time series observations, so that a continuous time process is appropriate for modelling the regression errors. The data are analysed and rapid change points in the trend functions are identified by using the methods described in this section. For curve estimation, the Gaussian kernel and its derivatives with support $\mathbb{R}$ were used which gave very smooth curve estimates. This is appropriate in the current example. The regression residuals are estimated by detrending the data series locally, using an optimal bandwidth (formula given in the text above). The distribution of the residuals turned out to be very close to normal so that one may assume $q = 1$ and $c_1^2(t_i) \approx \mathrm{var}(Y_i)$. On the original time scale in years (before 1989) the method identifies the main points of rapid change around the epoch known as the *Younger Dryas* at about 11,560 and 14,658 years before 1989 (see Fig. 7.22). For further details of the data analysis, see Menéndez et al. (2010).

Chapter 8
Forecasting

8.1 Forecasting for Linear Processes

8.1.1 Introduction

Here we briefly recall some basic results from forecasting. For details, see standard time series books such as Priestley (1981) and Brockwell and Davis (1991).

8.1.1.1 Prediction Given the Infinite Past

Suppose we observe $X_1, \ldots, X_n$ generated by a stationary process with Wold decomposition

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = A(B)\varepsilon_t$$

where

$$A(z) = \sum_{j=0}^{\infty} a_j z^j,$$

ε_t are identically distributed uncorrelated zero mean variables with variance $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t)$, $\sum a_j^2 < \infty$ and $B^j \varepsilon_t = \varepsilon_{t-j}$. We would like to predict X_{n+k} for some $k \geq 1$.

Before we focus on long-memory processes, we recall some basic facts from time series analysis (see, e.g. Priestley 1981; Brockwell and Davis 1991). The simplest formulas can be obtained for linear prediction based on the infinite past X_t $(t \leq n)$,

$$\hat{X}_{n+k} = \sum_{j=0}^{\infty} \beta_{j,k} X_{n-j} \tag{8.1}$$

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_8, © Springer-Verlag Berlin Heidelberg 2013

with suitably chosen weights. If X_t is invertible, then

$$\sum_{j=0}^{\infty} b_j X_{t-j} = \varepsilon_t$$

with

$$\sum_{j=0}^{\infty} b_j z^j = A^{-1}(z) = \left(\sum_{j=0}^{\infty} a_j z^j\right)^{-1} \quad (|z| \le 1)$$

and the σ-algebra generated by X_t $(t \le n)$ is the same as the one generated by ε_t $(t \le n)$. Therefore, $\hat{X}_{n+k}$ can also be written as

$$\hat{X}_{n+k} = \sum_{j=0}^{\infty} \alpha_{j,k} \varepsilon_{n-j}. \tag{8.2}$$

To judge the performance of the prediction, we use the mean squared prediction error

$$MSE(k) = E\big[(\hat{X}_{t+k} - X_{t+k})^2\big].$$

The best linear predictor minimizes the $MSE(k)$. Since ε_t are uncorrelated, we have from (8.2)

$$\begin{aligned} X_{n+k} - \hat{X}_{n+k} &= \sum_{j=0}^{\infty} a_j \varepsilon_{n+k-j} - \sum_{j=0}^{\infty} \alpha_{j,k} \varepsilon_{n-j} \\ &= \sum_{j=0}^{\infty} [a_{j+k} - \alpha_{j,k}] \varepsilon_{n-j} + \sum_{j=0}^{k-1} a_j \varepsilon_{n+k-j}. \end{aligned}$$

Hence,

$$MSE(k) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} [a_{j+k} - \alpha_{j,k}]^2 + \sigma_\varepsilon^2 \sum_{j=0}^{k-1} a_j^2.$$

The second term on the right-hand side does not depend on our choice of $\alpha_{j,k}$. The minimum is therefore achieved for

$$\alpha_{j,k} = a_{j+k} \quad (j = 0, 1, 2, \dots).$$

We thus obtain the optimal linear predictor

$$\hat{X}_{n+k} = \sum_{j=0}^{\infty} a_{j+k} \varepsilon_{n-j} \tag{8.3}$$

and the optimal mean squared error

$$MSE(k) = \sigma_\varepsilon^2 \sum_{j=0}^{k-1} a_j^2.$$

In particular, for the one-step prediction $\hat{X}_{n+1}$ we obtain

$$MSE(1) = \sigma_\varepsilon^2.$$

On the other hand, for predictions far into the future we have

$$\lim_{k\to\infty} \hat{X}_{n+k} = 0 = E(X_t)$$

(with convergence in the sense of mean squared error, i.e. in $L^2(\Omega)$) and

$$\lim_{k\to\infty} MSE(k) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} a_j^2 = \mathrm{var}(X_t).$$

More generally, if

$$X_t = \mu + \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$$

then

$$\lim_{k\to\infty} \hat{X}_{n+k} = \mu,$$

i.e. we predict the infinitely remote observation by $\mu = E(X_t)$, and the asymptotic prediction error is the variance of X_t. The proportion of the variability explained by past observations can be measured by

$$R^2(k) = \frac{MSE(k)}{MSE(\infty)} = \frac{\sum_{j=k}^{\infty} a_j^2}{\sum_{j=0}^{\infty} a_j^2}.$$

Formula (8.3) is not computable directly because the innovations ε_s, $s \le n$ are not observable. Since X_t was assumed to be invertible, the optimal weights in (8.1) can be obtained from

$$A^{(k)}(z)A^{-1}(z) = \sum_{j=0}^{\infty} \beta_{j,k} z^k \quad (|z| \le 1) \tag{8.4}$$

where

$$A^{(k)}(z) = \sum_{j=0}^{\infty} a_{j+k} z^j.$$

Prediction intervals follow directly from expressions for $MSE(k)$, at the least if X_t is a Gaussian series. In this case, a prediction interval with confidence level $1-\alpha$ is given by

$$I_\alpha(k) = \left[-z_{1-\alpha/2}\sigma_\varepsilon \sqrt{\sum_{j=0}^{k-1} a_j^2},\ z_{1-\alpha/2}\sigma_\varepsilon \sqrt{\sum_{j=0}^{k-1} a_j^2} \right].$$

8.1.1.2 Construction of the Wold Decomposition from the Spectral Density

Some models are given in terms of their spectral density f_X so that the coefficients in the Wold decomposition need to be calculated based on that information. The solution is due to a classical result by Whittle (1962). If $\int_{-\pi}^{\pi} \log f_X(\lambda)\, d\lambda > -\infty$ and the autocovariance generating function

$$G(z) = \sum_{k=-\infty}^{\infty} \gamma_X(k) z^k$$

is such that $L(z) = \log(G(z))$ is holomorphic in a ring $r^{-1} < |z| < r$ for some $r > 1$, then X_t has the Wold representation

$$X_t = A(B)\varepsilon_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$$

with

$$A(z) = \sum_{j=0}^{\infty} a_j z^j = 1 + \sum_{j=1}^{\infty} a_j z^j = \exp\left(\sum_{j=1}^{\infty} \alpha_j z^j \right), \tag{8.5}$$

$$\alpha_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ij\lambda} \log f_X(\lambda)\, d\lambda, \tag{8.6}$$

and the stationary uncorrelated zero-mean process ε_t defined by

$$\varepsilon_t = \int e^{it\lambda} \frac{1}{A(e^{-i\lambda})}\, dM_X(\lambda;\omega) = \sum_{j=0}^{\infty} b_j X_{t-j}.$$

The coefficients a_j can be obtained by

$$a_j = \frac{\frac{d^j}{dz^j} A(z)\,|_{z=0}}{j!}. \tag{8.7}$$

Similarly, the coefficients b_j are obtained from

$$\frac{1}{A(z)} = \sum_{j=0}^{\infty} b_j z^j = \exp\left(-\sum_{j=1}^{\infty} \alpha_j z^j\right)$$

by writing down the left-hand side as a power series and comparing the coefficients on both sides. Furthermore,

$$f_X(\lambda) = e^{\alpha_o} \left|A\left(e^{-i\lambda}\right)\right|^2 = \frac{\sigma_\varepsilon^2}{2\pi} \left|A\left(e^{-i\lambda}\right)\right|^2$$

and

$$\begin{aligned}\sigma_\varepsilon^2 &= \operatorname{var}(\varepsilon_t) = 2\pi \exp(\alpha_o) \\ &= 2\pi \exp\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda\right).\end{aligned}$$

Note that f_X can also be written as

$$f_X(\lambda) = \exp\left(\sum_{j=-\infty}^{\infty} \alpha_j e^{-i\lambda}\right) = \frac{\sigma_\varepsilon^2}{2\pi} \exp\left(\sum_{j=1}^{\infty} \tilde{\alpha}_j \cos j\lambda\right) \tag{8.8}$$

with

$$\tilde{\alpha}_j = 2\alpha_j \quad (j \geq 1).$$

8.1.1.3 Prediction Based on the Finite Past

Optimal linear prediction given the finite past, i.e. observations $X_1, \ldots, X_n$, is of the form

$$\hat{X}_{n+k} = \sum_{j=1}^{n} \varphi_{n,j}(k) X_{n-j+1} = \left[\varphi(n;k)\right]^T X(n)$$

with $X(n) = (X_n, X_{n-1}, \ldots, X_1)^T$ and $\varphi(n;k) = (\varphi_{n1}(k), \varphi_{n2}(k), \ldots, \varphi_{nn}(k))^T$ such that $MSE(k) = E[(X_{n+k} - \hat{X}_{n+k})^2]$ is minimized. By orthogonal projection on the L^2-closure of the span of $X_1, \ldots, X_n$ (see, e.g. Brockwell and Davis 1991) it follows that the optimal coefficients φ_{nj} can be obtained from the autocovariances by

$$\gamma_X(k+s-1) = \varphi_{n,1}(k)\gamma_X(s-1) + \cdots + \varphi_{n,n}(k)\gamma_X(s-n) \quad (s = 1, 2, \ldots, n).$$

In matrix form, with

$$\begin{aligned}\gamma_X(n;k) &= \left(\gamma_X(k), \gamma_X(k+1), \ldots, \gamma_X(k+n-1)\right)^T, \\ \Sigma_n &= \left[\gamma_X(i-j)\right]_{i,j=1,2,\ldots,n},\end{aligned}$$

this can be written as

$$\varphi(n;k) = \Sigma_n^{-1}\gamma_X(n;k).$$

The forecast is then

$$\hat{X}_{n+k} = \big[\varphi(n;k)\big]^T X(n) = \big[\gamma_X(n;k)\big]^T \Sigma_n^{-1} X(n),$$

and the mean squared prediction error is

$$\begin{aligned} MSE(k) &= E\big[(X_{n+k} - \hat{X}_{n+k})^2\big] \\ &= \gamma_X(0) - \gamma_X^T(n;k)\Sigma_n^{-1}\gamma_X(n;k). \end{aligned}$$

Forecast intervals are calculated as before, but with this formula for the k-step mean squared error $MSE(k)$.

Another important notion is partial correlation. If $\hat{X}_{n+1}(2,n)$ denotes the best linear prediction of X_{n+1} given $X_2,\ldots,X_n$ and $\hat{X}_1(2,n)$ the best linear prediction of X_1 given $X_2,\ldots,X_n$, then the partial correlation (pacf) at lag n is defined as $corr(X_{n+1} - \hat{X}_{n+1}(2,n), X_1 - \hat{X}_1(2,n))$ and turns out to be equal to the coefficient of X_1 in the forecast of X_{n+1}, i.e.

$$corr\big(X_{n+1} - \hat{X}_{n+1}(2,n), X_1 - \hat{X}_1(2,n)\big) = \varphi_{nn}(1).$$

The coefficients of $\varphi(n;1)$ can be calculated recursively, for instance, by the Durbin–Levinson algorithm (see Brockwell and Davis 1991). The coefficients $\varphi(n;k)$ for $k \geq 2$ can then be obtained recursively by repeated conditioning and insertion of corresponding one-step forecasts.

8.1.2 Forecasting for FARIMA Processes

Fractional ARIMA processes are very convenient when it comes to linear forecasting because they are defined in terms of difference equations. This makes the calculation of optimal prediction coefficients and prediction errors relatively easy. Explicit and recursive formulas are available. There is an extended, mainly applied, literature on forecasting with FARIMA and related processes (see, e.g. Reinsel and Lewis 1987; Peiris and Pereira 1988; Smith and Yadav 1994; Crato and Ray 1996; Palma and Chan 1997; Beran and Ocker 1999; Brodsky and Hurvich 1999; Hauser and Kunst 2001; Baillie and Chung 2002; Bos et al. 2002; Hidalgo and Yajima 2002; Ramjee et al. 2002; Ravishanker and Ray 2002; Bhansali and Kokoszka 2003; Bhardwaj and Swanson 2006; Man and Tiao 2006). Here we focus on the main basic formulas.

A FARIMA(p,d,q) process with $-\frac{1}{2} < d < \frac{1}{2}$ has the Wold decomposition

$$X_t = A(B)\varepsilon_t = (1-B)^{-d}\frac{\psi(B)}{\varphi(B)}\varepsilon_t,$$

with

$$A(z) = \sum_{j=0}^{\infty} a_j z^j = \sum_{j=0}^{\infty} \binom{-d}{j} (-1)^j \frac{1 + \psi_1 z + \cdots + \psi_q z^q}{1 - \varphi_1 z - \cdots - \varphi_p z^p} \tag{8.9}$$

(see Sect. 2.1.1.4). If $d \neq 0$, then $a_j \sim c_a j^{d-1}$. For instance, for a FARIMA$(0, d, 0)$ process $c_a = 1/\Gamma(d)$. Thus,

$$\sum_{j=k}^{\infty} a_j^2 \sim c_a^2 \int_1^{\infty} x^{2d-2}\, dx \cdot k^{2d-1} = \frac{c_a^2}{1-2d} k^{2d-1}.$$

Hence

$$R^2(k) = \frac{\sum_{j=k}^{\infty} a_j^2}{\sum_{j=0}^{\infty} a_j^2} \sim \text{const} \cdot k^{2d-1}.$$

In contrast, if $d = 0$, then we have a short-memory ARMA(p, q) process with an exponential decay $|a_j| \leq O(c^j)$ for some $0 \leq c < 1$ so that

$$R^2(k) = \frac{\sum_{j=k}^{\infty} a_j^2}{\sum_{j=0}^{\infty} a_j^2} = O\big(c^k\big).$$

In other words, under short memory the explanatory power of past observations decays at an exponential rate, whereas for long-memory and antipersistent processes the decay is much slower. This essentially means that for short-memory processes accurate forecasts cannot be made too far into the future whereas the "forecastable" time horizon is much longer for long-memory processes. This is illustrated in Fig. 8.1 with a comparison of $R^2(k)$ for FARIMA$(0, d, 0)$ models and AR(1) processes with parameters chosen such that the variance and the lag-one correlation is the same for both processes. The difference between short and long memory can also be seen by looking at the length $l_\alpha(k)$ of prediction intervals which is proportional to $(\sum_{j=0}^{k-1} a_j^2)^{1/2}$, compared to the length of the interval for the infinite time horizon. The monotonically nondecreasing ratio $l_\alpha(k)/l_\alpha(\infty) = \sqrt{1 - R^2(k)}$ always converges to 1, but for long-memory processes the convergence is rather slow (namely hyperbolic).

The optimal coefficients $\beta_{j,k}$ in the optimal forecast (8.1) are defined by

$$\begin{aligned} A^{(k)}(z) A^{-1}(z) &= \sum_{j=0}^{\infty} \beta_{j,k} z^j \\ &= \sum_{j=0}^{\infty} a_{j+k} z^j \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j z^j \frac{1 - \varphi_1 z - \cdots - \varphi_p z^p}{1 + \psi_1 z + \cdots + \psi_q z^q}. \end{aligned}$$

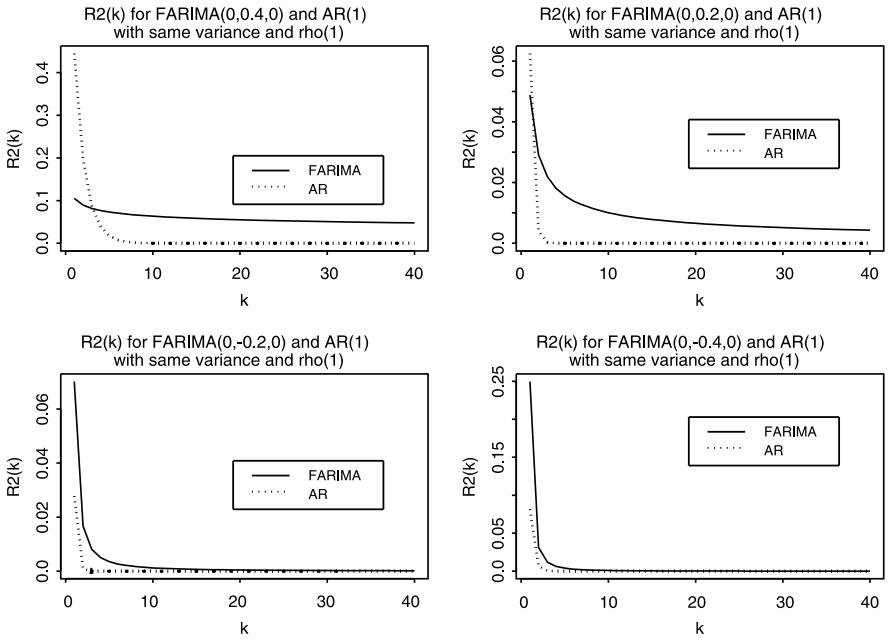

Fig. 8.1 $R^2(k)$ for FARIMA$(0, d, 0)$ processes with $d = 0.4, 0.2, -0.2$ and -0.4, respectively, together with corresponding values for AR(1) processes with the same variance and lag-one correlation

Using the notation

$$\alpha_k(B) = \sum_{j=0}^{k-1} a_j B^j,$$

alternative expressions for $\beta_{j,k}$ can be obtained from

$$\begin{aligned}\hat{X}_{n+k} &= \sum_{j=k}^{\infty} a_j \varepsilon_{n+k-j} = X_{n+k} - \sum_{j=0}^{k-1} a_j \varepsilon_{n+k-j} \\ &= B^{-k}\left\{1 - \alpha_k(B)\frac{\varphi(B)}{\psi(B)}(1-B)^d\right\} X_n \\ &= \sum_{j=0}^{\infty} \beta_{j,k} X_{n-j} = \beta_k(B) X_n\end{aligned}$$

which implies

$$\sum_{j=0}^{\infty} \beta_{j,k} z^j = z^{-k}\left\{1 - z^{-k}\sum_{j=0}^{k-1} a_j z^j \sum_{j=0}^{\infty} b_j z^j\right\}$$

(see also Bisaglia and Bordignon 2002). Multiplying out yields the relationship

$$\beta_{j,k} = \sum_{i=0}^{k-1} b_i a_{k+j-i}.$$

For predictions based on the finite past, one can apply the usual Durbin–Levinson algorithm to obtain $\varphi(n;k)$. This is particularly simple for FARIMA$(0,d,0)$ processes because there one has explicit formulas for $\gamma_X(k)$. In particular, Hosking (1981) showed the partial autocorrelation of a FARIMA$(0,d,0)$ process to be equal to

$$\varphi_{n,n}(1) = \frac{d}{n-d}.$$

It is interesting that $\varphi_{n,n}(1)$ is proportional to n^{-1} and this rate does not depend on d. The other coefficients do depend on d, however, with

$$\varphi_{n,j}(1) = -\binom{n}{j} \frac{\Gamma(j-d)\Gamma(n-j-d+1)}{\Gamma(-d)\Gamma(n-d+1)} \sim -\frac{1}{\Gamma(-d)} j^{-d-1}$$

with the last equivalence under the assumption that $j, n \to \infty$, $j/n \to 0$. Note that the hyperbolic decay and the very slow rate of $\varphi_{n,n}(1)$ are again in contrast to ARMA processes with $d = 0$ where we have an exponential bound. Also, the very slow rate n^{-1} of $\varphi_{n,n}(1)$ can be understood in the sense that the additional information from the past that is encoded in the earliest available observation X_1 is highly relevant for the future observation X_{n+1}.

Example 8.1 The variety of possible forecast intervals one may obtain with FARIMA(p,d,q) models is displayed in Fig. 8.2.

8.1.3 Forecasting for FEXP Processes

FEXP processes (Beran 1993; Robinson 1994a) are defined in terms of their spectral density. Therefore the *MA*-coefficients a_j in the Wold representation and the *AR*-coefficients b_j have to be calculated from f_X. Consider, for instance, cosine-based models. A cosine based FEXP(p) process has the spectral density

$$\begin{aligned} f_X(\lambda) &= \frac{\sigma_\varepsilon^2}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d} \exp\left(\sum_{j=1}^{p} \vartheta_j \cos j\lambda\right) \\ &= \left|1 - e^{-i\lambda}\right|^{-2d} f_{\mathrm{EXP}}(\lambda) \end{aligned}$$

where f_{EXP} is the spectral density of an FEXP(p) model with $d = 0$, which is also called an EXP process Bloomfield (1973). Thus X_t can also be written as

$$X_t = (1-B)^{-d} Z_t$$

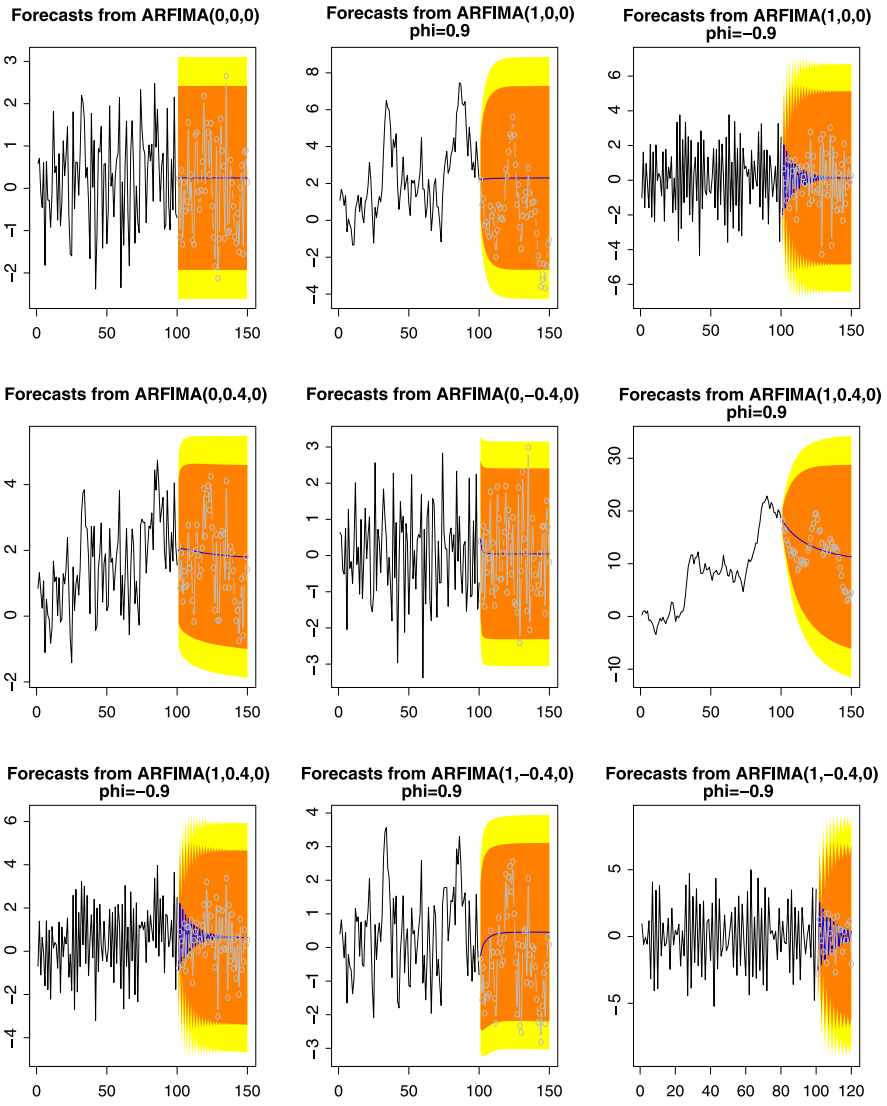

Fig. 8.2 Simulated FARIMA(p, d, q) series for various parameter settings. In each picture, $n = 100$ observations were simulated and the optimal k-step forecast was calculated ($k = 1, 2, \ldots, 20$) together with 95 %- and 99 %-prediction intervals. The forecasts (*full line*) as well the actual simulated observations $X_{n+1}, \ldots, X_{n+20}$ (*circles*) are also displayed

where Z_t is an EXP(p) process with short memory. Since the coefficients in the linear filters $(1 - B)^{-d}$ and $(1 - B)^d$ are known, a natural approach to obtain *MA*- and *AR*-coefficients of X_t is to obtain the *MA*-filter of Z_t first and multiply it by the fractional differencing filter $(1 - B)^{-d}$ and $(1 - B)^d$, respectively Hurvich (2002). The coefficients based on f_{EXP} were derived by Bloomfield (1973). From (8.5) we

have

$$A_{\mathrm{EXP}}(z) = \sum_{j=0}^{\infty} a_j z^j = \exp\left(\frac{1}{2}\sum_{j=1}^{p} \vartheta_j z^j\right) = \exp\big(h(z)\big)$$

so that

$$\frac{d^j}{dz^j} A_{\mathrm{EXP}}(z) \mid_{z=0} = \sum_{s=1}^{j} \binom{j-1}{s-1} h^{(s)}(0).$$

Equation (8.7) then implies $a_{0,\mathrm{EXP}} = 1$ and

$$a_{j,\mathrm{EXP}} = \frac{\frac{d^j}{dz^j} A_{\mathrm{EXP}}(z) \mid_{z=0}}{j!} = \frac{1}{2j} \sum_{s=1}^{j} s\vartheta_s a_{j-s} \quad (j > 0)$$

(with $\vartheta_s := 0$ for $s > p$). Moreover, the *AR*-coefficients are obtained from $\exp(-h(z))$ so that one obtains $b_{0,\mathrm{EXP}} = 1$ and

$$b_{j,\mathrm{EXP}} = -\frac{1}{2j} \sum_{s=1}^{j} s\vartheta_s b_{j-s} \quad (s > 0).$$

The *MA*-coefficients of X_t are then obtained by

$$\sum_{j=0}^{\infty} a_j z^j = (1-z)^{-d} \sum_{j=0}^{\infty} a_{j,\mathrm{EXP}} z^j.$$

Comparing powers leads to

$$a_j = \sum_{s=0}^{j} \binom{-d}{s} (-1)^s a_{j-s,\mathrm{EXP}}.$$

Similarly, for the *AR*-coefficients one has

$$b_j = \sum_{s=0}^{j} \binom{d}{s} (-1)^s b_{j-s,\mathrm{EXP}}.$$

For a detailed derivation of these formulas, see Bloomfield (1973) and Hurvich (2002).

8.2 Forecasting for Nonstationary Processes

Suppose we observe an integrated process Y_t $(t = 1, 2, \ldots, n)$ such that

$$Y_t - Y_{t-1} = X_t$$

where $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ is a stationary linear process as before with $a_j \sim c_a j^{d-1}$ $(0 < d < \frac{1}{2})$. Then an observation k steps ahead is of the form

$$Y_{n+k} = Y_n + U_{n+k}$$

with

$$U_{n+k} = \sum_{j=1}^{k} X_{n+j}.$$

Note that $X_2, \ldots, X_n$ can be reconstructed from $Y_1, \ldots, Y_n$ by differencing. Considering linear prediction of Y_{n+k}, we have

$$\hat{Y}_{n+k} = Y_n + \sum_{j=1}^{n-1} \beta_j(k) X_{n-j+1} = Y_n + \beta^T X_{n:2}$$

with $X_{2:n} = (X_n, X_{n-1}, \ldots, X_2)^T$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_{n-1})^T$. The mean squared error is

$$\begin{aligned} MSE(k) &= E\big[(Y_{n+k} - \hat{Y}_{n+k})^2\big] \\ &= \sum_{s=-(k-1)}^{k-1} \big(k - |s|\big)\gamma_X(s) - 2\sum_{i=1}^{k}\sum_{j=1}^{n-1} \beta_j E[X_{n+i} X_{n-j+1}] + \beta^T \Sigma_n \beta \\ &= \sum_{s=-(k-1)}^{k-1} \big(k - |s|\big)\gamma_X(s) - 2\tilde{\gamma}^T \beta + \beta^T \Sigma_n \beta \end{aligned}$$

where

$$\tilde{\gamma} = \tilde{\gamma}(n;k) = \sum_{i=1}^{k} \gamma_X(n-1;i),$$

and the vectors

$$\gamma_X(n-1;i) = \big(\gamma_X(i), \gamma_X(i+1), \ldots, \gamma_X(i+n-2)\big)^T$$

are defined as in the stationary case. Minimizing with respect to β then leads to the optimal solution

$$\beta_{\text{opt}} = \Sigma_n^{-1} \tilde{\gamma}.$$

The optimal MSE is given by

$$MSE_{\text{opt}}(k) = \sum_{s=-(k-1)}^{k-1} \big(k - |s|\big)\gamma_X(s) - \tilde{\gamma}^T \Sigma_n^{-1} \tilde{\gamma}.$$

In contrast to the stationary case, the MSE diverges to infinity as $k \to \infty$ because the first term is simply the variance of the sum of k observations $X_{n+1}, \ldots, X_{n+k}$. More specifically, we have

$$MSE(k) \sim c_f \nu(d) k^{2d-1}.$$

For more details and examples, together with extensions to forecasting in the presence of trends, see Beran and Ocker (1999).

8.3 Forecasting for Nonlinear Processes

Prediction for nonlinear processes can differ quite substantially from the case of linear processes. There is an enormous number of possibilities for nonlinear behaviour (see, e.g. the classical book by Tong 1993). Here, we focus on volatility models because most nonlinear processes with long-range dependence known so far fall into this category. We give a brief account of some basic problems.

Consider a volatility model $X_t = \sigma_t \varepsilon_t$ ($t \in \mathbb{Z}$) with ε_t i.i.d., independent of the past, $E(\varepsilon_t) = 0$, $\mathrm{var}(\varepsilon_t) = 1$, and σ_t a function of X_s ($s \le t-1$). The best linear prediction of X_{n+k} given X_t ($t \le n$) is $E(X) = 0$ because the observations are uncorrelated. However, in contrast to linear processes, the conditional quadratic forecasting error

$$MSE_{\mathrm{cond}}(k) = E\big[(\hat{X}_{n+k} - X_{n+k})^2 \,\big|\, X_t, t \le n-1\big] = E\big[X_{n+k}^2 \,\big|\, X_t, t \le n-1\big]$$

can be quite different from the unconditional error

$$MSE(k) = E\big[(\hat{X}_{n+k} - X_{n+k})^2\big] = E\big[X_{n+k}^2\big].$$

This is in particular true for processes with *long memory* in volatility. Moreover, for the purpose of forecasting one actually needs the standard deviation rather than the mean squared error. However, in general $\sqrt{MSE(k)}$ is not equal to $\tilde{\sigma} = E[|\sigma_n(X_{n-1}, X_{n-2}, \ldots)|]$, and the latter quantity is difficult to calculate (but can be evaluated approximately by simulations). In contrast, the conditional standard deviation is readily available due to the definition of $\sigma_t(X_s, s \le t-1)$.

We illustrate this by considering a LARCH process with weights $b_j \sim \mathrm{const} \cdot j^{d-1}$ ($j \to \infty$) for some $0 < d < \frac{1}{2}$, and $\sum b_j^2 < 1$. The conditional value of σ_n,

$$E[\sigma_n | X_t, t \le n-1] = \sigma_n(X_{n-1}, X_{n-2}, \ldots) = b_0 + \sum_{j=1}^{\infty} b_j X_{n-j}, \tag{8.10}$$

can be calculated directly. For the unconditional expected value, we know that $E(\sigma_n) = 0$ and

$$E\big(\sigma_n^2\big) = \mathrm{var}\big(\sigma_n^2\big) = b_0^2 \frac{\|b\|_2^2}{1 - \|b\|_2^2}$$

where we use the notation $\|b\|_2^2 = \sum_{j=1}^{\infty} b_j^2$ (see Sect. 2.1.3.6). However, there is no closed form formula for $\tilde{\sigma} = E(|\sigma_n|)$. Thus, $\tilde{\sigma}$ has to be evaluated by simulation.

Given X_t $(t \le n)$, X_{n+1} is distributed like $\sigma_{n+1}\varepsilon_{n+1}$ where $\sigma_{n+1} = \sigma_{n+1}(X_t, t \le n)$ is a fixed number defined by (8.10). Suppose that ε_t are symmetrically distributed. Since $E(X_{n+1} \mid X_t, t \le n) = 0$, a conditional $(1-\alpha)$-prediction interval can be given by

$$\begin{aligned} I_{\text{cond}}(\alpha) &= \left[\hat{X}_{n+1} - |\sigma_{n+1}| q_{\varepsilon;\frac{\alpha}{2}}, \hat{X}_{n+1} + |\sigma_{n+1}| q_{\varepsilon;1-\frac{\alpha}{2}}\right] \\ &= \left[-|\sigma_{n+1}| q_{\varepsilon;\frac{\alpha}{2}}, |\sigma_{n+1}| q_{\varepsilon;1-\frac{\alpha}{2}}\right] \end{aligned}$$

where $q_{\varepsilon;\frac{\alpha}{2}}$ and $q_{\varepsilon;1-\frac{\alpha}{2}}$ are the $\frac{\alpha}{2}$- and $(1-\frac{\alpha}{2})$-quantiles of ε_t. In particular, if ε_t are standard normal variables, then

$$I_{\text{cond}}(\alpha) = \pm\sigma_{n+1} z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. To calculate an unconditional prediction interval, one would have to evaluate the corresponding unconditional quantiles of $X_t = \sigma_t \varepsilon_t$. Although ε_t is independent of σ_t, calculating such quantiles is quite difficult due to the complicated distribution of σ_t. Moreover, conditional prediction intervals are more accurate because they have the correct coverage probability even if one looks at the conditional distribution. In other words,

$$P\big(X_{n+1} \notin I_{\text{cond}}(\alpha) \mid X_t, t \le n\big) = \alpha.$$

If we use an unconditional prediction interval $I_{\text{uncond}}(\alpha) = [-c, c]$, then the situation changes. Even though the constant c is chosen such that $P(X_{n+1} \notin I_{\text{uncond}}(\alpha)) = \alpha$, the conditional probability $P(X_{n+1} \notin I_{\text{uncond}}(\alpha) \mid X_t, t \le n)$ is a nondegenerate random variable. If the innovations ε_t are continuous, then the conditional coverage probability of unconditional prediction intervals is almost surely wrong. Thus, in summary, we may conclude that it is advisable to use conditional prediction intervals.

8.4 Nonparametric Prediction of Exceedance Probabilities

In Sect. 7.4, we have considered nonparametric estimation of time dependent distribution functions and quantiles when the time series observations are Gaussian subordinated via an unknown function. The aim of this section is a slight extension of that and, namely, nonparametric forecasting of exceedance (or, non-exceedance) probabilities, a topic with major practical significance in many fields. Needless to say, this approach can also be applied to other trend functions, i.e. to means of stochastic processes. To start with we recall our model for the observations $Y_1, Y_2, \ldots, Y_n$ namely,

$$Y_j = G(Z_j, t_j),$$

$t_j = j/n$ being rescaled times and $\{Z_j,\ j = 1, 2, \ldots\}$ is an unobserved zero mean stationary Gaussian process with long-memory. The unknown function $G(x, \cdot)$ is a Lebesgue measurable function that is square integrable with respect to the normal density. This assumption will allow us to use Hermite polynomial expansions. The Gaussian process Z_j has correlations that decay slowly, i.e. it has long-memory, and in particular

$$Cov(Z_j, Z_{j+u}) = \gamma\big(|u|\big) \sim C \times |u|^{2H-2}, \quad \text{as } |u| \to \infty$$

with $1/2 < H < 1$ and $C > 0$. We want to forecast values of the non-exceedance probability $F_t(y)$ for a prespecified value of $y \in \mathbb{R}$ where

$$F_{t_j}(y) = P(Y_j \leq y)$$

or of the level-crossing probability $1 - F_t(y)$ at time t. We let F_t, $t \in (0, 1)$ be continuous, finitely differentiable, as many times as is required, with respect to t and y, and we denote the probability density function at time t by,

$$f_t(y) = \frac{\partial}{\partial y} F_t(y).$$

Finally, under some suitable conditions, we will also derive CLT's facilitating construction of asymptotic prediction bands for $F_t(y)$ where y is a prespecified real number. The main ideas covered here are in Ghosh and Draghicescu (2002a, 2002b) and Beran and Ocker (1999). Additional important information concerning the asymptotic properties is in Beran (1986), Dehling and Taqqu (1989a, 1989b), Csörgő and Mielniczuk (1995a); also relevant is Leadbetter et al. (1983).

The time dependent Gaussian subordination model considered here is appropriate for processes for which the marginal distribution function at any point of time can have an arbitrary shape and an adequate description by means of a parametric family becomes difficult. This property then demands that the method of estimation and prediction be sufficiently flexible. For nonparametric function prediction of a function $\mu(\cdot)$ at a future time $t_n + \delta$ based on data until time t_n, the main idea is to use a Taylor series expansion of $\mu(t_n + \delta)$ around $\mu(t_n)$ and then to plug in estimates of the various derivatives. For instance, a k-step forecast based on n data points would correspond to $t_n + \delta = (n + k)/n$. We sketch the main ideas here, but there are various other factors that affect the quality of prediction. One such factor is the issue of boundary bias when $\mu(t)$ and its derivatives where t close to 1 is being estimated from the data. For illustration, for estimating $\mu(t)$ or its derivatives at $t \in (0, 1)$, we use kernel smoothing for estimating derivatives. Needless to say, local polynomials of a suitable degree would be more appropriate when t reaches the boundary.

Figure 8.3 illustrates the time series of the mean daily precipitation totals (mm) in the 1900s from Grand St. Bernard, Switzerland and the estimated probability (as a function of time) that the precipitation level stays below 5 mm. The figure shows estimated probabilities for 1901–1970 with prediction for the next 20 years. For comparison, estimates are also shown for the entire series.

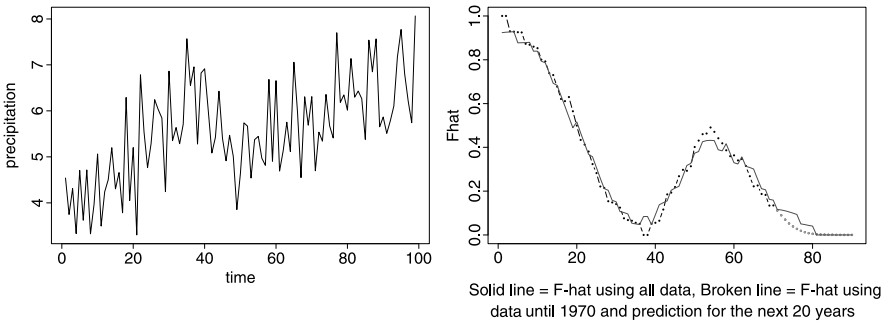

Fig. 8.3 Mean daily precipitation in Grand St. Bernard (mm) in the 1900s: (*left*) precipitation time series, (*right*) estimated (*solid lines*) and predicted (*broken lines*) probability function. *Data source: MeteoSwiss, Switzerland. The figures have been reproduced from International Journal of Forecasting* (*2002*), *Vol.* ***18***, *pp. 283–290*

At the beginning of the century, an average daily precipitation value of 5 mm is near the right tail of the distribution. However, after some 100 years, the precipitation distribution had shifted to the right, well beyond this daily average.

Using a symmetric probability density function $K(u)$, $u \in (-1, 1)$ and a sequence of bandwidths $b_n = b$ for which $b \to 0$ and $nb \to \infty$ as $n \to \infty$, we start by defining the estimate of the ith derivative $F_t^{(i)}(y) = \frac{\partial^i}{\partial t^i} F_t(y)$ as follows: let $K_i(\cdot)$ be a kernel of order $i + 2$ (see Gasser and Müller 1984 and Eubank 1987). Define the Priestley–Chao estimator of $F_t^{(i)}(y)$ as

$$\widehat{F}_t^{(i)}(y) = \frac{(-1)^i}{nb^{i+1}} \sum_{j=1}^{n} K_i\left(\frac{t_j - t}{b}\right) I_j(y)$$

where

$$I_j(y) = 1 \quad \text{if } Y_j \le y \quad \text{and} \quad I_j(y) = 0 \quad \text{otherwise.}$$

When $i = 0$, the above estimator is simply the usual Priestley–Chao estimator of $F_t(y)$.

Since the indicator function $I_j(y)$ is a function of Y_j, it is also Gaussian subordinated. We assume that the following Hermite polynomial expansion holds

$$W(t_j, y) = I_j(y) - P(Y_j \le y) = \sum_{l=m}^{\infty} \frac{c_l(t_j, y)}{l!} H_l(Z_j)$$

where $m \ge 1$ is the Hermite rank, c_l are the Hermite coefficients, and H_l denotes the Hermite polynomial of degree l. Suitable regularity conditions will be assumed for the Hermite coefficients. For instance, due to the orthogonality of the Hermite

polynomials, and since $\text{var}(H_l(Z_j)) = l!$,

$$\text{var}\big(W(t, y)\big) = \sum_{l=m}^{\infty} \frac{c_l^2(t, y)}{l!}$$

which implies

$$\sum_{l=m}^{\infty} \frac{c_l^2(t, y)}{l!} < \infty$$

where $y \in \mathbb{R}$ and $t \in (0, 1)$. Furthermore, we will assume that

$$\frac{\partial^2}{\partial t^2} \text{var}\big(W(t, y)\big) < \infty$$

for $t \in (0, 1)$ and $y \in \mathbb{R}$. These conditions essentially imply smoothness of changes in the indicator function W (Draghicescu 2002). We also assume that the long-memory parameter $H > 1 - 1/(2m)$, in which case, $W(t_j, y)$, $j = 1, 2, \ldots$ will have long-memory.

For prediction of $F_t(y)$ to a future point $t + \delta$, it is convenient to consider the logistic transformation

$$V_t(y) = \log\left(\frac{F_t(y)}{1 - F_t(y)}\right)$$

and define

$$U_{t,\delta}(y) = V_t(y) + \sum_{j=1}^{k} \frac{\delta^j}{j!} \times \left(\frac{\partial^j}{\partial t^j} V_t(y)\right).$$

Then by Taylor series expansion around $\delta = 0$,

$$V_{t+\delta}(y) = U_{t,\delta}(y) + \big(\delta^{k+1}/(k+1)!\big) R_{\tilde{t}}(y)$$

where

$$R_t(y) = \frac{\partial^{k+1}}{\partial t^{k+1}} V_t(y)$$

and $t < \tilde{t} < t + \delta$. For convenience, we assume that $V_t(y)$ has k continuous derivatives with respect to t in $[0, 1]$ and a finite $(k+1)$st derivative in $(0, 1)$. Specifically, there exist a measurable function $M_t(y)$ defined on $[0, 1] \times \mathbb{R} \to \mathbb{R}$ and a constant a such that $|R_t(y)| < M_t(y)$ and $E[M_t(Y)] < a < \infty$ (see, e.g. Rao 1973 and Serfling 1980).

The above discussion shows that when δ converges to zero, the difference between $V_{t+\delta}(y)$ and $U_{t,\delta}(y)$ diminishes. Since $V_{t+\delta}(y)$ is the logistic transformation of a cumulative probability distribution function at time $t + \delta$, it is monotone in y,

justifying the use of $U_{t,\delta}(y)$ for a 'δ *steps ahead of* t' prediction. In particular, when δ is small, the inverse logistic transform

$$\frac{\exp(U_{t,\delta}(y))}{1+\exp(U_{t,\delta}(y))}$$

is a valid probability distribution function. In particular, in that case, the predicted value of the α-quantile ($0<\alpha<1$) defined as any value $\theta_t(\alpha)$ for which

$$\theta_t(\alpha)=\inf_y\big\{y|F_t(y)\geq\alpha\big\},$$

can then be obtained by simply inverting the predicted distribution function. For convenience, we will assume that $\theta_t(\alpha)$ is unique.

For illustration, consider $k=2$, in which case the expression for $U_{t,\delta}(y)$ becomes

$$\begin{aligned}U_{t,\delta}(y)&=\log\left(\frac{F_t(y)}{1-F_t(y)}\right)+\delta\frac{\frac{\partial}{\partial t}F_t(y)}{F_t(y)(1-F_t(y))}\\&\quad+\frac{\delta^2}{2!}\left[\frac{\frac{\partial^2}{\partial t^2}F_t(y)}{F_t(y)(1-F_t(y))}-\left\{\frac{\frac{\partial}{\partial t}F_t(y)}{F_t(y)(1-F_t(y))}\right\}^2\right.\\&\quad\left.+2F_t(y)\left\{\frac{\frac{\partial}{\partial t}F_t(y)}{F_t(y)(1-F_t(y))}\right\}^2\right]\\&=\psi(F_0,F_1,F_2),\quad\text{say,}\end{aligned}$$

where $F_i=\frac{\partial^i}{\partial t^i}F_t(y)$, $i=0,1,2$. Similarly, define $\hat{U}_{t,\delta}(y)$ by substituting the estimates $\hat{F}_i$, $i=0,1,2$.

Fix t and y and consider

$$a_i=a_i(t,\delta,y)=\frac{\partial}{\partial F_i}\psi(F_0,F_1,F_2),\quad i=0,1,2,$$

where

$$\begin{aligned}a_0&=\frac{1}{F_0(1-F_0)}+\frac{\delta F_1(2F_0-1)}{(F_0(1-F_0))^2}+\frac{\delta^2}{2}\left[\frac{F_2(2F_0-1)}{(F_0(1-F_0))^2}+\frac{2F_1^2(4F_0-3F_0^2-1)}{F_0^2(1-F_0)^4}\right.\\&\quad\left.-\frac{2F_1^2(2F_0-1)}{(F_0(1-F_0))^3}\right],\\a_1&=\frac{\delta}{F_0(1-F_0)}+\frac{\delta^2F_1}{F_0(1-F_0)^2}-\frac{\delta^2F_1}{(F_0(1-F_0))^2},\\a_2&=\frac{\delta^2}{2F_0(1-F_0)}.\end{aligned}$$

Then,

$$\hat{U}_{t,\delta}(y) = U_{t,\delta}(y) + \sum_{i=0}^{2} a_i(\hat{F}_i - F_i) + R_n(t, \delta, y)$$

where

$$R_n(t, \delta, y) = o_p\big(\max(\hat{F}_0 - F_0, \hat{F}_1 - F_1, \hat{F}_2 - F_2)\big).$$

For simplicity, let $b_0 = b_1 = b_2 = b$. Let as $n \to \infty$, $b \to 0$ and $nb \to \infty$. We have

Theorem 8.1 *As* $n \to \infty$,

(a) *Bias*: $E(\hat{U}_{t,\delta}(y)) - U_{t,\delta}(y) = O(b^2)$;
(b) *Variance*: $\text{var}(\hat{U}_{t,\delta}(y)) = V_n(t, \delta, y; m, b, H) + o(b^{-4}(nb)^{m(2H-2)})$

where

$$V_n(t, \delta, y; m, b, H) = \sum_{i=0}^{2}\sum_{j=0}^{2} a_i(t, \delta, y)a_j(t, \delta, y)B_{n,i,j}(t, y; m, b, H)$$

and

$$B_{n,i,j}(t, y; m, b, H) = \frac{(nb)^{m(2H-2)}}{b^{i+j}} \frac{C^m}{m!} c_m^2(t, y) \int_{-1}^{1}\int_{-1}^{1} K_i(u)K_j(v)|u - v|^{m(2H-2)}\, du\, dv.$$

Proof The proof follows using arguments of Sect. 7.4. □

Theorem 8.2 *Let* $m = 1$. *We assume the regularity conditions on* $V_t(y)$ *mentioned above. Then under the conditions stated above and if* $b^{-4}(nb)^{m(2H-2)} \to 0$ *as* $n \to \infty$, *for every fixed* $t \in (0, 1)$ *and* $y \in \mathbb{R}$,

$$\big(\hat{U}_{t,\delta}(y) - V_{t+\delta}(y)\big)\Big/\sqrt{V_n(t, \delta, y; m, b, H)}$$

converges to a standard normal variable.

Proof Note that due to the previous theorem, for every fixed $t \in (0, 1)$ and $y \in \mathbb{R}$, the mean squared error of $\hat{U}_{t,\delta}(y)$, i.e.

$$E\big(\hat{U}_{t,\delta}(y) - V_{t+\delta}(y)\big)^2 = O\big(b^{-4}(nb)^{m(2H-2)}\big) + O\big(b^4\big),$$

so that if $b^{-4}(nb)^{m(2H-2)} \to 0$ as $n \to \infty$, by Chebyshev's inequality, $|\hat{U}_{t,\delta}(y) - V_{t+\delta}(y)|$ converges to zero in probability. Normality follows from Dehling and Taqqu (1989a, 1989b). □

By using inverse logistic transformation, for $0 < \alpha < 1$ and $y \in \mathbb{R}$, an approximate $100(1-\alpha)$ %-prediction interval can now be given by $(\hat{F}^{(l)}_{t+\delta}(y), \hat{F}^{(u)}_{t+\delta}(y))$ where

$$\hat{F}^{(l)}_{t+\delta}(y) = \frac{e^{\hat{U}_{t,\delta}(y) - z_{\alpha/2}\sqrt{W_n(t,\delta,y;m,b,H)}}}{1 + e^{\hat{U}_{t,\delta}(y) - z_{\alpha/2}\sqrt{V_n(t,\delta,y;m,b,H)}}},$$

$$\hat{F}^{(u)}_{t+\delta}(y) = \frac{e^{\hat{U}_{t,\delta}(y) + z_{\alpha/2}\sqrt{W_n(t,\delta,y;m,b,H)}}}{1 + e^{\hat{U}_{t,\delta}(y) + z_{\alpha/2}\sqrt{V_n(t,\delta,y;m,b,H)}}}$$

and $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

Chapter 9
Spatial and Space-Time Processes

9.1 Spatial Models on $\mathbb{Z}^k$

Spatial data play an important role in many areas such as ecology, biology, environmental monitoring, agronomy, remote sensing, geology, to name a few. Sometimes observations are obtained on a regular lattice (see, e.g. Whittle 1962; Bartlett 1974; Besag 1974; Cressie 1993; Christakos 1992, 2000; Benson et al. 2006; Sain and Cressie 2007; and references therein). This leads to considering spatial processes on a grid, or more specifically, random fields X_t with index $t \in \mathbb{Z}^2$. On the other hand, if the spatial points are not on a regular grid, then random fields with $t \in \mathbb{R}^2$ are used.

Suppose now that our data can be modelled by a stationary random field $X_t \in \mathbb{R}$ ($t \in \mathbb{Z}^2$) on a regular grid. In general, the autocovariances $\gamma(k) = cov(X_t, X_{t+k})$ (where $t = (t_1, t_2)$, $k = (k_1, k_2)$) are a function of the two lags k_1 and k_2 and may differ from $\gamma(\tilde{k}_1, \tilde{k}_2)$ even if $\|k\|^2 = k_1^2 + k_2^2$ is equal to $\|\tilde{k}\|^2 = \tilde{k}_1^2 + \tilde{k}_2^2$. If this is the case for at least one pair of (two-dimensional) lags $k, \tilde{k}$, then X_t is called anisotropic. Otherwise, X_t is isotropic and we can write (in a slight misuse of notation) $\gamma(k) = \gamma(\|k\|)$, i.e. the autocovariance function depends on the Euclidian distance only and not on the direction. The random field is said to have long memory if

$$\sum_{k \in \mathbb{Z}^2} \big|\gamma(k)\big| = \infty$$

(see, e.g. Lavancier 2006, 2007). Most work in the literature focusses on long-range dependent random fields with a hyperbolic decay of γ of the form

$$\gamma(k) \sim L\big(\|k\|\big) h\left(\frac{k}{\|k\|}\right) \|k\|^{-\alpha} = L\big(\|k\|\big) h\left(\frac{k}{\|k\|}\right) \|k\|^{2d-2} \quad (0 < \alpha < 2)$$

as $\|k\| \to \infty$, where L is slowly varying at infinity and h is a continuous function on the unit sphere in $\mathbb{R}^2$. Note that $d = 1 - \frac{1}{2}\alpha \in (0, 1)$. Also note that this definition can be generalized to random fields on a grid of arbitrary dimension m, i.e. $t \in \mathbb{Z}^m$.

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_9, © Springer-Verlag Berlin Heidelberg 2013

Since $h(u)$ ($u = (u_1, u_2) = k/\|k\|$) is an arbitrary function of u_1, u_2, this definition includes isotropic as well anisotropic fields. However, since asymptotically the essential part is $\|k\|^{-\alpha}$, Lavancier (2006, 2007) suggests to call such long-memory fields isotropic, even if h depends on the direction. In a similar way as for time series, the definition of long memory based on the autocovariance function is equivalent under suitable regularity conditions to a pole of the spatial spectral density of the following form:

Definition 9.1 A stationary random field X_t ($t \in \mathbb{Z}^2$) is said to have long-range dependence or long memory if it has a spectral distribution that is continuous everywhere except at zero, where it has a pole of the form

$$f(\lambda) \sim L\big(\|\lambda\|\big)h\left(\frac{\lambda}{\|\lambda\|}\right)\|\lambda\|^{\alpha-2} = L\big(\|\lambda\|\big)h\left(\frac{\lambda}{\|\lambda\|}\right)\|\lambda\|^{-2d} \quad (0 < \alpha < 2)$$

with L slowly varying at zero, h continuous on the unit sphere and $\lambda \in [-\pi, \pi]^2$.

This definition implies that the long-memory property, characterized by the parameter d, is the same in all directions. In some applications, this assumption is too restrictive. For instance, in ground water flow and contaminant transport studies, it is common practice to model physical properties by scalar fields with stronger long memory in the direction of the flow (Guo et al. 2009). Ponson et al. (2005) conjecture the existence of universal anisotropic long-memory exponents in fracture surfaces that correspond to certain physical properties of the material such as roughness, growth and the so-called dynamic exponents. Also see, e.g. Makse et al. (1996), Elliott et al. (1997), Hristopulos (2002), Kelbert et al. (2005) for other examples from physics, geophysics, etc. An overview of recent results on anisotropic random fields can be found in Lavancier (2006, 2007) (also see Matheron 1973; Mandelbrot 1983; Solo 1992; Heyde and Gay 1993; Anh et al. 1999; Angulo et al. 2000; Ruiz-Medina et al. 2003; Chan and Wood 2004; Fernández-Pascual et al. 2006).

More generally, we may thus extend the definition of long memory to spectral densities that are unbounded on at least one line $\lambda_2 + \beta\lambda_1 = 0$ or $\lambda_1 + \beta\lambda_2 = 0$, or even more generally, on a one-dimensional set (or curve) $A_{\text{pole}} \subseteq [-\pi, \pi]^2$ of positive one-dimensional Lebesgue measure (i.e. of positive length). The meaning of $\lambda_2 + \beta\lambda_1 = 0$ is that long-range dependence in the sense of time series is present when following a transect in the direction $k = (k_1, k_2)$ with $k_2 = \beta k_1$. This can be seen by considering the autocovariance function along this line, namely

$$\gamma(k) = \gamma(k_1, \beta_1 k_1) = \iint e^{ik_1(\lambda_1+\beta_1\lambda_2)} f(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2 =: \gamma^*(k_1).$$

Assuming a pole of the form $f(\lambda) \sim L(|\lambda_2 + \beta\lambda_1|)|\lambda_2 + \beta\lambda_1|^{-2d}$ with $0 < d < \frac{1}{2}$, analogous arguments as in Sect. 1.3.1 lead to $\gamma^*(k_1) \sim L^*(|k_1|)|k_1|^{2d-1}$ as $|k_1| \to \infty$. Note that here d is limited to the range $(0, 1/2)$ because the pole is directional.

9.2 Spatial FARIMA Processes

A simple model that follows this definition and can accommodate anisotropic long memory can be obtained, for instance, by extending the fractional ARIMA process to space as follows (Beran et al. 2009). Define polynomials

$$\begin{aligned} \varphi_1(z) &= 1 - \sum_{j=1}^{p_1} \varphi_{1j} z^j, \qquad & \varphi_2(z) &= 1 - \sum_{j=1}^{p_2} \varphi_{2j} z^j, \\ \psi_1(z) &= 1 + \sum_{j=1}^{q_1} \psi_{1j} z^j, \qquad & \psi_2(z) &= 1 + \sum_{j=1}^{q_2} \psi_{2j} z^j \end{aligned} \tag{9.1}$$

with no roots for $|z| \le 1$, and let ε_{rs} $(r, s \in \mathbb{Z})$ be i.i.d. random variables with $E(\varepsilon_{rs}) = 0$ and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_{rs}) < \infty$. Denoting by B_1 and B_2 backshift operators in the horizontal and vertical direction, respectively (i.e. $B_1\varepsilon_{rs} = \varepsilon_{r,s-1}$, $B_2\varepsilon_{rs} = \varepsilon_{r-1,s}$) we define "vertical" and "horizontal" linear filters $\Lambda_1(B_1) = \varphi_1^{-1}(B_1)\psi_1(B_1)$ and $\Lambda_2(B_2) = \varphi_2^{-1}(B_2)\psi_2(B_2)$, and the product

$$\Lambda(B_1, B_2) = \Lambda_1(B_1)\Lambda_2(B_2).$$

To include the possibility of long memory and antipersistence, we define further for $d_1, d_2 \in (-\frac{1}{2}, \frac{1}{2})$ the fractional differencing operators $(1 - B_1)^{d_1}$ and $(1 - B_2)^{d_2}$, respectively. Then a process X_{rs} $(r, s \in \mathbb{Z})$ that solves

$$X_{rs} = (1 - B_1)^{-d_1}(1 - B_2)^{-d_2}\Lambda(B_1, B_2)\varepsilon_{rs} = \Psi_1(B_1)\Psi_2(B_2)\varepsilon_{rs} \tag{9.2}$$

is called a spatial fractional ARIMA process, or ARFIMA($\mathbf{p}, \mathbf{d}, \mathbf{q}$) process where $\mathbf{p} = (p_1, p_2)$, $\mathbf{d} = (d_1, d_2)$ and $\mathbf{q} = (q_1, q_2)$. The idea of the model is that there is a vertical and a horizontal dependence structure in form of a fractional ARIMA model. From the definition, it follows that the spectral density of X_{rs} is equal to

$$f(\lambda_1, \lambda_2) = \frac{\sigma_\varepsilon^2}{4\pi^2}\left|1 - e^{-i\lambda_1}\right|^{-2d_1}\left|1 - e^{-i\lambda_2}\right|^{-2d_2}\left|\frac{\psi_1(e^{-i\lambda_1})}{\varphi_1(e^{-i\lambda_1})}\right|^2\left|\frac{\psi_2(e^{-i\lambda_2})}{\varphi_2(e^{-i\lambda_2})}\right|^2 \tag{9.3}$$

$$= \sigma_\varepsilon^2 f_1(\lambda_1) f_2(\lambda_2) \tag{9.4}$$

where f_i $(i = 1, 2)$ are the spectral densities of a fractional ARIMA model with innovation variance one, $p = p_i$, $d = d_i$, $q = q_i$, and the MA- and AR-polynomials $\varphi(z) = \varphi_i(z)$, $\psi(z) = \psi_i(z)$, respectively. Note that here

$$\begin{aligned} A_{\text{pole}} &= \left\{\lambda \in [-\pi, \pi]^2 : \lambda_1 = 0, \lambda_2 \in [-\pi, \pi]\right\} \cup \left\{\lambda \in \mathbb{R}^2 : \lambda_1 \in [-\pi, \pi], \lambda_2 = 0\right\} \\ &= \{\lambda : \lambda_1 + \beta\lambda_2 = 0 \text{ with } \beta = 0\} \cup \{\lambda : \lambda_2 + \beta\lambda_1 = 0 \text{ with } \beta = 0\}, \end{aligned}$$

provided that $d_1, d_2 > 0$. The short-memory version of this model (i.e. $d_1 = d_2 = 0$) was introduced in Martin (1979). Sethuraman and Basawa (1995) considered a version with $d_2 = 0$, $d_1 > 0$ and Ψ_1 finite. The fully spatial version was introduced in

Beran et al. (2009). Note furthermore that we also may obtain directional antipersistence for negative values d_1 or d_2, respectively.

It is worth noting that for spatial data, there are many other ways of generating long-range dependence. Generally speaking, we may start with the spectral representation of an uncorrelated (i.e. white noise) spatial process

$$\varepsilon_t = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{i\langle t,\lambda\rangle}\, dM_\varepsilon(\lambda_1, \lambda_2)$$

where $M_\varepsilon(\lambda_1, \lambda_2) = M_\varepsilon(\lambda_1, \lambda_2; \omega)$ is the spectral measure of ε_t and $\langle t, \lambda\rangle = t_1\lambda_1 + t_2\lambda_2$. Given a function $a(e^{-i\lambda}) = a(e^{-i\lambda_1}, e^{-i\lambda_2}) \in L^2([-\pi, \pi]^2)$, the process

$$X_t = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{i\langle t,\lambda\rangle} a\big(e^{-i\lambda}\big)\, dM(\lambda_1, \lambda_2)$$

is well defined with spectral density

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{(2\pi)^2} \big|a\big(e^{-i\lambda}\big)\big|^2 \quad \big(\lambda \in [-\pi, \pi]^2\big).$$

Long-range dependence is achieved whenever $a(\lambda)$ is unbounded on a set A_{pole}. For instance, consider

$$X_t = \big(1 - B_1 B_2^m\big)^{-d} \varepsilon_t$$

with $0 < d < \frac{1}{2}$ and $m \geq 2$ (see Lavancier 2006, 2007). Although there is only one fractional differencing parameter d, long memory is not at all isotropic. The spectral density is equal to

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{4\pi^2} \big|1 - e^{i(\lambda_1 + m\lambda_2)}\big|^{-2d}$$

so that

$$\gamma_X(k_1, mk_1) \sim \text{const} \cdot |k_1|^{2d-1}$$

as $|k_1| \to \infty$, whereas $\gamma_X(k) = 0$ for all other directions.

9.3 Maximum Likelihood Estimation

For a linear process on a regular grid, a straightforward approach to maximum likelihood estimation can be obtained from the conditional representation of X_t ($t \in \mathbb{Z}^2$) given all other observations. The reason is that for an invertible linear process, $E(X_t \mid X_s, s \neq t)$ is linear in X_s ($s \neq t$) and the residuals $\varepsilon_t = X_t - E(X_t \mid X_s, s \neq t)$ are i.i.d. variables with distribution F_ε. In particular, if ε_t are Gaussian and $E(X_t \mid X_s, s \neq t)$ is characterized by a finite dimensional parameter $\vartheta^0 = (\sigma^2_{\varepsilon;0}, \theta^0)$,

then maximum likelihood estimation of ϑ^0 can usually be approximated asymptotically by minimizing the residual sum of squares $SSE = \sum \varepsilon_t^2(\theta)$ with respect to θ and setting $\hat{\sigma}_\varepsilon^2 = n^{-1}SSE(\hat{\theta})$. As in the time series context, there are two main issues that have to be addressed to prove that this estimator has asymptotically the same distribution as the exact MLE: First of all, in the exact MLE, conditioning can be done on the observed values X_s only (i.e. a finite number of X_s-values). The second problem which is related to the first one is that none of the ε_t can be evaluated exactly because only a finite number of X_s-values are known. More generally, the same estimator can be used for linear processes with an arbitrary distribution of ε_t (satisfying certain moment conditions). In the general case, the method no longer approximates the MLE, but usually shares the same or similar limiting properties. Approximations of the likelihood function of short-memory Gaussian lattice processes are discussed, for instance, in Besag (1974), Tjostheim (1978), Martin (1979), Guyon (1982, 1995), Kashyap (1984), Dahlhaus and Künsch (1987), Huang and Anh (1992). In the long-memory case, the issue of obtaining a good approximation is more delicate because of the farther reaching dependence structure.

Consider, for instance, the fractional ARIMA lattice model introduced above (Sect. 9.2). If $0 < d_1, d_2 < \frac{1}{2}$ and the roots of the polynomials

$$\begin{aligned} \varphi_1(z) &= 1 - \sum_{j=1}^{p_1} \varphi_{1j} z^j, \qquad \varphi_2(z) = 1 - \sum_{j=1}^{p_2} \varphi_{2j} z^j, \\ \psi_1(z) &= 1 + \sum_{j=1}^{q_1} \psi_{1j} z^j, \qquad \psi_2(z) = 1 + \sum_{j=1}^{q_2} \psi_{2j} z^j \end{aligned} \tag{9.5}$$

are outside the unit circle, then X_{rs} is stationary and invertible so that we have the representation

$$\varepsilon_{rs} = (1 - B_1)^{d_1}(1 - B_2)^{d_2} \Lambda^{-1}(B_1, B_2) X_{rs} \tag{9.6}$$

$$= \sum_{j,l=0}^{\infty} b_j(\theta_{\text{col}}) b_l(\theta_{\text{row}}) X_{r-j,s-l}. \tag{9.7}$$

The unknown parameter vector is $\vartheta = (\sigma_\varepsilon^2, \theta^T)$, where $\sigma_\varepsilon^2 = \text{var}(\varepsilon_{rs}) > 0$, and $\theta = (\theta_{\text{row}}^T, \theta_{\text{col}}^T) \in \Theta \subseteq \mathbb{R}^{p_1+q_1+1} \times \mathbb{R}^{p_2+q_2+1}$ with

$$\theta_{\text{row}} = \left(d_1, \varphi_1^T, \psi_1^T\right)^T, \qquad \theta_{\text{col}} = \left(d_2, \varphi_2^T, \psi_2^T\right)^T, \tag{9.8}$$

and

$$\varphi_i = (\varphi_{i1}, \ldots, \varphi_{ip_i})^T, \qquad \psi_i = (\psi_{i1}, \ldots, \psi_{iq_i})^T \quad (i = 1, 2). \tag{9.9}$$

For convenience, we use the same notation φ_i and ψ_i for the polynomials and the corresponding parameter vectors, respectively. Approximate MLE as above is discussed in Beran et al. (2009). Related results for special cases are discussed in

Boissy et al. (2005) and Sethuraman and Basawa (1995). For local Whittle estimation based on low frequencies of the spatial periodogram, see Guo et al. (2009). Due to the factorized form of the linear filter, it is possible to calculate approximate values of ε_{rs} even if we observe X_{rs} on an irregularly shaped area $(r, s) \in A \subseteq \mathbb{Z}^2$. To obtain asymptotic results, A has to grow with increasing sample size n. Thus, we assume that observations consist of

$$X_{rs}, (r, s) \in A_n \tag{9.10}$$

with

$$A_n = \left\{(r, s) \in \mathbb{N}_+^2 : m_{\text{row},L} \le r \le m_{\text{row},U}, m_{\text{col},L}(r) \le s \le m_{\text{col},U}(r)\right\} \tag{9.11}$$

where $m_{\text{col},L}(\cdot)$ and $m_{\text{col},U}(\cdot)$ are functions with finite support mapping $\mathbb{N}_+$ (the set of positive integers excluding zero) to $\mathbb{N}$. This definition includes quite general shapes. For instance, an L-shaped area can be defined by setting $m_{\text{row},L} = 1$, $m_{\text{row},U} = n$, $m_{\text{col},L}(r) \equiv 1$, $m_{\text{col},U}(r) = n - [n/2] \cdot 1\{r > [n/2]\}$. A rectangular area with side lengths n and $[na_0]$ is obtained by setting $m_{\text{row},L} = 1$, $m_{\text{row},U} = n$, $m_{\text{col},L}(r) \equiv 1$, and $m_{\text{col},U}(r) \equiv [na_0]$. The computable approximation of $\varepsilon_{rs}(\theta)$ is

$$e_{rs}(\theta) = \sum_{j,l \in B_n(r,s)} b_j(\theta_{\text{col}}) b_l(\theta_{\text{row}}) X_{r-j,s-l} \quad (\theta \in \Theta), \tag{9.12}$$

with $B_n(r, s) = \{j, l \ge 0 : (r - j, s - l) \in A_n\}$. The estimate of θ is set equal to

$$\hat{\theta} = \arg\min \sum_{(r,s) \in A_n} e_{rs}^2(\theta). \tag{9.13}$$

In the following, we use the notation

$$\dot{e}_{rs}(\theta) = [\dot{e}_{rs;1}, \ldots, \dot{e}_{rs;p_1+q_1+p_2+q_2+2}]^T \tag{9.14}$$

with

$$\dot{e}_{rs;j} = \frac{\partial}{\partial \theta_j} e_{rs}(\theta) \tag{9.15}$$

and

$$\tilde{S}_n(\theta) = \sum_{(r,s) \in A_n} \dot{e}_{rs}(\theta) e_{rs}(\theta). \tag{9.16}$$

The estimator $\hat{\theta}$ can also be defined as the solution of

$$\tilde{S}_n(\hat{\theta}) = 0. \tag{9.17}$$

This definition is useful for deriving the asymptotic distribution. Beran et al. (2009) use the following assumptions:

- (A1) Let $\varepsilon_{rs} = \varepsilon_{rs}(\theta^0)$ be i.i.d. zero mean random variables with finite variance and denote by

$$\begin{aligned}\dot{\varepsilon}_{rs}\big(\theta^0\big) &= \sum_{j,l=0}^{\infty} \frac{\partial}{\partial\theta}\big[b_j(\theta_{\text{col}})b_l(\theta_{\text{row}})\big]\big|_{\theta=\theta^0} X_{r-j,s-l} \\ &= \big[\dot{\varepsilon}_{rs;1}\big(\theta^0\big), \ldots, \dot{\varepsilon}_{rs;p_1+q_1+p_2+q_2+2}\big(\theta^0\big)\big]^T \end{aligned} \tag{9.18}$$

  the derivative of $\varepsilon_{rs}(\theta)$ at $\theta = \theta^0$. Then, as $n \to \infty$,

$$n^{-1} \max_{1\le r,s\le n} \big\|\varepsilon_{rs}\big(\theta^0\big)\dot{\varepsilon}_{rs}\big(\theta^0\big)\big\|^2 = o_p(1) \tag{9.19}$$

  where $\|\cdot\|$ denotes the Euclidean norm and $o_p(1)$ means that the sequence of random variables converges to zero in probability as n tends to infinity.
- (A2) With the same notation as in (A1)

$$\lim_{n\to\infty} n^{-1} E\Big\{ \max_{1\le r,s\le n} \big\|\varepsilon_{rs}\big(\theta^0\big)\dot{\varepsilon}_{rs}\big(\theta^0\big)\big\|^2 \Big\} = 0. \tag{9.20}$$

- (A3) Θ is compact and $\theta \in \Theta^0$ where Θ^0 denotes the interior of Θ.
- (A4)

$$X_{rs}, (r,s) \in A_n, \tag{9.21}$$

  with A_n defined in (11) and such that there exist constants $0 < \kappa_{\text{row}}, \kappa_{\text{col}} \le 1$, $0 \le a < b \le 1$, with

$$\lim_{n\to\infty} n^{-1}[m_{\text{row},U} - m_{\text{row},L}] = \kappa_{\text{row}} \tag{9.22}$$

  and

$$\lim_{n\to\infty} n^{-1} \min_{a\le r\le nb} \big[m_{\text{col},U}(r) - m_{\text{col},L}(r)\big] = \kappa_{\text{col}}. \tag{9.23}$$

  Moreover, the number of points in A_n is such that

$$0 < \lim_{n\to\infty} n^{-2}|A_n| = A < \infty. \tag{9.24}$$

The intuitive meaning of these assumptions can be explained as follows. The first two assumptions, (A1) and (A2), depend on extreme value properties of ε_{rs}. These conditions hold, for instance, if ε_{rs} are in the maximum domain of attraction of the Gumbel distribution (see Embrechts et al. 1997). Assumption (A3) is standard in the context of parameter estimation. It makes sure that asymptotically the solution of the estimating equation is bounded away from the border of the parameter space. Finally, (A4) provides an asymptotic characterization of the observational area A_n. For example, for a rectangle with sides of length 1 and a we have $\kappa_{\text{row}} = 1$, $\kappa_{\text{col}} = a$ and $A = a$, whereas for the L-shaped area described previously we have $\kappa_{\text{row}} = 1$,

$A = \frac{3}{4}$, $a = 0$, $b = 1$, $\kappa_{\text{col}} = \frac{1}{2}$. Under (A1)–(A4), the asymptotic distribution of $\hat{\theta}$ can be derived by defining

$$S_n(\theta) = \sum_{(r,s)\in A_n} \dot{\varepsilon}_{rs}(\theta)\varepsilon_{rs}(\theta) \tag{9.25}$$

and essentially showing

$$\lim_{n\to\infty} |A_n|^{-1} E\big[\big\| S_n\big(\theta^0\big) - \tilde{S}_n\big(\theta^0\big)\big\|^2\big] = 0$$

and

$$|A_n|^{-\frac{1}{2}} S_n\big(\theta^0\big) \underset{d}{\to} Z \sim N\big(0, \sigma_\varepsilon^2 V\big(\theta^0\big)\big), \tag{9.26}$$

where

$$V\big(\theta^0\big) = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} \tag{9.27}$$

with the $(p_1 + q_1 + 1) \times (p_1 + q_1 + 1)$ matrix

$$V_1 = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_{\text{row}}} \log f_1(\lambda) \left[\frac{\partial}{\partial \theta_{\text{row}}} \log f_1(\lambda) \right]^T d\lambda$$

and the $(p_2 + q_2 + 1) \times (p_2 + q_2 + 1)$ matrix

$$V_2 = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_{\text{col}}} \log f_2(\lambda) \left[\frac{\partial}{\partial \theta_{\text{col}}} \log f_2(\lambda) \right]^T d\lambda.$$

Note that the convergence of $|A_n|^{-\frac{1}{2}} S_n(\theta^0)$ follows from a martingale property analogous to the case of a time series. Conditioning is, however, more complex since it has to be done in space. More specifically, define

$$Z_{t,n} = n^{-\frac{1}{2}} S_t\big(\theta^0\big) \tag{9.28}$$

$$= \sum_{u=1}^{t} \xi_{u,n} \tag{9.29}$$

with

$$\xi_{u,n} = n^{-\frac{1}{2}} \left[\sum_{j=1}^{u} \dot{\varepsilon}_{uj}\big(\theta^0\big)\varepsilon_{uj}\big(\theta^0\big) + \sum_{j=1}^{u-1} \dot{\varepsilon}_{ju}\big(\theta^0\big)\varepsilon_{ju}\big(\theta^0\big) \right]$$

$$= n^{-\frac{1}{2}} \left[\sum_{j=1}^{u} \eta_{uj} + \sum_{j=1}^{u-1} \zeta_{uj} \right].$$

Defining the array of σ-algebras $\mathscr{F}_{t,n} = \sigma(Z_{u,n}, u \le t)$, we have

$$E[\xi_{t,n}|\mathscr{F}_{t-1,n}] = 0$$

and

$$E[Z_{t,n}|\mathscr{F}_{t-1,n}] = Z_{t-1,n}.$$

Thus, $\xi_{t,n}$ is an array of martingale differences and $Z_{t,n}$ an array of martingales. The central limit theorem for

$$n^{-1} S_n(\theta^0) = n^{-\frac{1}{2}} Z_{n,n}$$

then follows from Theorem 3.2 in Hall and Heyde (1980) since their sufficient conditions

$$n^{-1} \max_{1 \le u \le n} \xi_{u,n}^2 = o_p(1) \tag{9.30}$$

and

$$\lim_{n \to \infty} n^{-1} E\left\{ \max_{1 \le u \le n} \xi_{u,n}^2 \right\} = 0 \tag{9.31}$$

turn out to hold under the given assumptions. The final result for $\hat{\theta}_n$ can be stated as follows:

Theorem 9.1 *Under* (A1)–(A4), *there exists a sequence* $\hat{\theta}_n$ *such that* (9.17) *holds*, $\hat{\theta}_n$ *converges to* θ^0 *in probability and*

$$|A_n|^{\frac{1}{2}} (\hat{\theta}_n - \theta^0) \underset{d}{\to} Z \sim N(0, V^{-1}(\theta^0)). \tag{9.32}$$

This result holds even if $d_i = 0$ or $d_i < 0$ (as long as d_1 and d_2 are estimated). It is interesting to note that the shape of V implies that $\hat{\theta}_{\text{row}}$ and $\hat{\theta}_{\text{col}}$ are asymptotically independent. This can be used to obtain a simple test of isotropy. Under the null hypothesis $H_0 : d_1 = d_2$, the statistic

$$T = |A_n|^{\frac{1}{2}} \frac{\hat{d}_1 - \hat{d}_2}{\sqrt{\sigma_1^2 + \sigma_2^2}},$$

with σ_i^2 $(i = 1, 2)$ equal to the asymptotic variances of $\hat{d}_i$, is approximately standard normal. In general, σ_i^2 $(i = 1, 2)$ have to be replaced by estimates obtained from $\hat{\theta}$. The only exception is the case with $p_i = q_i = 0$ where $\sigma_i^2 = 6/\pi^2 \approx 0.608$. Beran et al. (2009) apply this test to show that, for the ozone data introduced in Sect. 1.2 (Fig. 1.19), there is evidence for stronger long memory in the north–south direction.

9.4 Latent Spatial Processes: An Example from Ecology

In the context of species diversity assessment, one is often interested in estimating the total number of unseen species which have not been discovered during a survey. Consider, for example, the problem of estimating the number of plant species in a very large landscape. The usual method consists of counting the total number of species in an area and then extrapolating this value to a much larger area. Extrapolation takes place after one has fitted a suitable curve to the data that are obtained from surveying areas of increasing size. Such a curve is known as the species–area curve. It is popular to fit a linear model to the scatter plot of the (x, y) observations in log–log coordinates where x denotes area and y is species count. Since the number of species typically increases with increasing area, this approach leads to a positive slope, implying a predicted value of an infinite number of species when area tends to infinity. This disturbing outcome leads one to postulate a spatial model that is decisive of species occurrence and examining regularity conditions that would lead to a finite predicted value for the total species number.

To proceed with our argument, we assume a systematic survey design, a lattice $\mathbf{u} = (i, j)$, $i, j = 1, 2, \ldots, n$ such that $k = n^2$ denotes the total number of sites where species have been counted. Species occurrence is assumed to depend on a background process $X(\mathbf{u})$ which could be, for instance, moisture, soil quality, exposure to sunlight, and so on. This process that is decisive of species occurrence will be assumed to be of the form

$$X(\mathbf{u}) = G_X\big(Z(\mathbf{u})\big).$$

Here, $Z(\mathbf{u})$ is a zero mean, unit variance Gaussian spatial process with covariance function

$$cov\big(Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})\big) = \gamma_Z\big(|\mathbf{h}|\big),$$

$\mathbf{h} = (l_1, l_2)$, $l_1, l_2 = 0, \pm 1, \pm 2, \ldots$ and G_X is an arbitrary Lebesgue-measurable L^2 function with respect to the standard normal density. Note that by definition, $X(\mathbf{u})$ is also stationary but it need not be Gaussian. Stationarity of the background process $X(\mathbf{u})$ implies that we are concerned with a large area where apart from stochastic variations ecologically similar conditions prevail, so that the same species–area law can be postulated. Let X be univariate, although generalizations can be carried out. We consider two correlation types:

Type 1: Short memory: $\sum_{\mathbf{h}} |\gamma_Z(\mathbf{h})|^m < \infty$ and,
Type 2: Long memory: $\gamma_Z(\mathbf{h}) \sim |\mathbf{h}|^{-2\alpha}$, as $|\mathbf{h}| \to \infty$, where $0 < \alpha < 1/m$ and $m \geq 1$ is a positive integer. In this case, $\sum_{\mathbf{h}} |\gamma_Z(\mathbf{h})|^m = \infty$.

Let there be S different species in the landscape, which have been assigned serial numbers $1, 2, \ldots, S$ and suppose that $X(\mathbf{u}) \in A_s$ where A_s is an interval on the real line suitable for species occurrence so that we may define the spatial process Y_s that

takes binary values on the lattice as follows:

$$\begin{aligned} Y_s(\mathbf{u}) &= 1 \quad \text{if } X(\mathbf{u}) \in A_s, \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

having expected value $E[Y_s(\mathbf{u})] = P(X(\mathbf{u}) \in A_s) = p_s(\mathbf{u}) = p_s$ which does not depend on $\mathbf{u}$ due to stationarity of X. In addition, suppose that the species have been ordered such that $s_1 < s_2$ implies $p_{s_1} \leq p_{s_2}$. As S is assumed to be large, we will assume $p_s = p(x_s)$ where $x_s = s/S$ and p is a sufficiently regular function on $[0, 1]$.

Note that by definition, Y_s is also a transformation of the same Gaussian spatial process Z, and we may write

$$Y_s(\mathbf{u}) - p(x_s) = G_Y\big(Z(\mathbf{u}), x_s\big),$$

for an appropriately defined G_Y. We assume that G_Y admits the following Hermite-polynomial expansion:

$$Y_s(\mathbf{u}) - E\big[Y_s(\mathbf{u})\big] = \sum_{l=m}^{\infty} \frac{c_l(x_s)}{l!} H_l\big(Z(\mathbf{u})\big)$$

having a Hermite rank $m \geq 1$. The Hermite coefficients $c_l(x)$ are continuous functions, and as before, $H_l(x)$ is the Hermite polynomial of degree l in $x \in \mathbb{R}$. Summing the values of the indicator process Y_s over k plots, one arrives at the number of plots where species s occurs. Let this number be denoted by $N_{k,s} = \sum_{i=1}^{k} Y_s(\mathbf{u_i})$ where $\mathbf{u_i} = (i_1, i_2)$, $i_1, i_2 = 1, 2, \ldots, n$, $i = 1, 2, \ldots, k$.

Theorem 9.2 *Under the assumptions stated above, as $n \to \infty$ and $n^2 = k$,*

$$\operatorname{var}(N_{k,s}) \sim Ck^{\delta}$$

for some constant $0 < C < \infty$, where $\delta = 1$ in case of short memory and $\delta = 2 - m\alpha$, $0 < \alpha < 1/m$, when $Z(\mathbf{u})$ has long-memory correlations and the Hermite rank of G_Y is $m \geq 1$.

Proof We have

$$\begin{aligned} \operatorname{var}(N_{k,s}) = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} cov\Bigg\{ & \sum_{l_1=m}^{\infty} \frac{c_{l_1}(x_s)}{l_1!} H_{l_1}\big(Z(i_1, i_2)\big), \\ & \sum_{l_2=m}^{\infty} \frac{c_{l_2}(x_s)}{l_2!} H_{l_2}\big(Z(j_1, j_2)\big)\Bigg\} \\ = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\sum_{l=m}^{\infty} & \frac{c_l^2(x_s)}{(l!)^2} cov\big\{H_l\big(Z(i_1, i_2)\big), H_l\big(Z(j_1, j_2)\big)\big\} \end{aligned}$$

$$= \sum_{v_1=-(n-1)}^{n-1} \sum_{v_2=-(n-1)}^{n-1} \sum_{l=m}^{\infty} \frac{c_l^2(x_s)}{l!}\big(n-|v_1|\big)\big(n-|v_2|\big)\gamma^l\big(\big|\sqrt{{v_1}^2+{v_2}^2}\big|\big)$$

$$= n^2 \sum_{l=m}^{\infty} \frac{c_l^2(x_s)}{l!} \sum_{v_1=-(n-1)}^{n-1} \sum_{v_2=-(n-1)}^{n-1} \gamma^l\big(\big|\sqrt{{v_1}^2+{v_2}^2}\big|\big)\big[1+O(1)\big].$$

Case 1. Short memory: In this case, the auto-covariances are summable. Moreover, $\mathrm{var}(H_l(Z(\mathbf{u}))) = l!$, so that, due to the orthogonality of the Hermite polynomials, $\mathrm{var}(Y_s(\mathbf{u})) = \sum_{l=m}^{\infty} \frac{c_l^2(x_s)}{l!} < \infty$, uniformly in s. Thus, substituting $k = n^2$, $\mathrm{var}(N_{k,s}) \sim O(k)$.

Case 2. Long memory: In this case,

$$cov\big(Y_s(\mathbf{u}), Y_s(\mathbf{v})\big) = \sum_{l=m}^{\infty} \frac{c_l^2(x_s)}{l!}\gamma_Z^l\big(|\mathbf{u}-\mathbf{v}|\big) \sim \sum_{l=m}^{\infty} \frac{c_l^2(x_s)}{l!}|\mathbf{u}-\mathbf{v}|^{-2l\alpha}$$

as $|\mathbf{u}-\mathbf{v}| \to \infty$. The result follows by noting that

$$\mathrm{var}(N_{k,s}) \sim n^2 \frac{c_m^2(x_s)}{m!} \sum_{v_1=-(n-1)}^{n-1} \sum_{v_2=-(n-1)}^{n-1} \big\{{v_1}^2+{v_2}^2\big\}^{-m\alpha}$$

$$= O\big(n^{4-2m\alpha}\big) = O\big(k^{2-m(2-2H)}\big), \quad 1-1/(2m) < H < 1,$$

$H = 1-\alpha/2$ being the Hurst parameter. □

Thus in case of short memory, $\mathrm{var}(N_{k,s}) = \sigma_k^2(x_s) = O(k)$, whereas when $m = 1$ and $Z(\mathbf{u})$ has long-memory $\sigma_k^2(x_s) = O(k^{2H})$, $0.5 < H < 1$.

Figure 9.1 shows two typical examples of $Y_s(\mathbf{u})$ (right) for one species (i.e. s fixed) generated by a background process $X(\mathbf{u})$. In the upper left figure, we see an image plot of a simulated process $X(\mathbf{u})$ with moderate long-range dependence, and on the right next to it, the corresponding spatial distribution of $Y_s(\mathbf{u})$ (white corresponds to $Y_s(\mathbf{u}) = 1$). The two lower panels display the same pictures for strong long-range dependence. Figure 9.2(a) shows typical species–area curves, i.e. plots of the total number of species, say ξ_k (see (9.34) below), observed in k plots, against k, as well as a plot of bootstrap averages of $\xi_k - \xi_{k-1}$ against k in log–log-coordinates. The different paths of ξ_k were obtained by a specific bootstrap procedure developed in Ghosh (2009).

Define the sum

$$W_{s,k} = N_{k,s}/\sqrt{\mathrm{var}(N_{k,s})} = \sum_{i=1}^{k}\big[Y_s(\mathbf{u_i}) - p(x_s)\big]\big/\sigma_k(x_s).$$

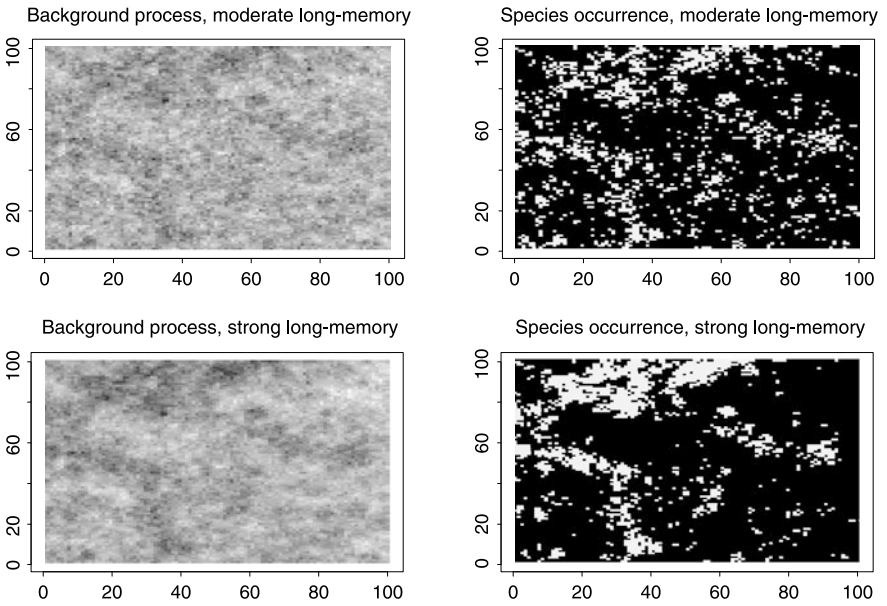

Fig. 9.1 Spatial distribution of one species (*right*) generated by a background process with long-range dependence (*left*): the *upper two panels* correspond to moderate long-range dependence, the *lower two* to strong long-range dependence in the background process. *White patches* in the two right panels indicate presence of the species

We will assume that for every fixed s, $W_{s,k}$ has an asymptotic cumulative probability distribution function $F_s(x)$ with an exponentially decaying tail

$$\log\bigl(F_s(-x)\bigr) \sim -\frac{a_s x^{\theta}}{2}\bigl[1+o(1)\bigr], \quad x \to \infty, \tag{9.33}$$

for some $a_s, \theta > 0$. For instance, for the standard normal distribution Φ, we have:

Lemma 9.1 *Let $F_s = \Phi$. Then*

$$\log\bigl(F_s(-x)\bigr) \sim -\frac{x^2}{2}\bigl[1+o(1)\bigr], \quad x \to \infty.$$

Proof The reader may also refer to Feller (1971) for an outline of the proof. Let $f(x) = \phi(x)/x$ where $\phi(x)$ is the pdf of the standard normal distribution. Since $f'(x) = -\phi(x)\{1+\frac{1}{x^2}\}$, we have an alternative expression

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-y^2/2}\left\{1+\frac{1}{y^2}\right\} dy.$$

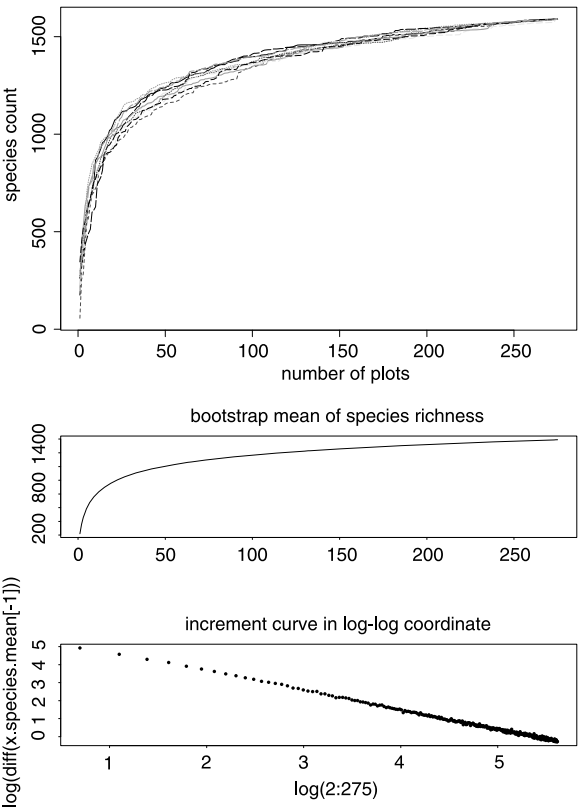

Fig. 9.2 Species–area curves for vascular plant species-richness data from Switzerland. (**a**) *Top*: Ten randomly selected species–area curves: species counts plotted against the number of surveyed plots. (**b**) *Middle*: Bootstrap average of species counts from 1000 simulations plotted against the number of plots; (**c**) *Bottom*: first finite difference of the average species counts plotted against the number of plots in log–log coordinates. Data source: Federal Office of the Environment (FOEN), Switzerland and Hintermann & Weber, AG, Switzerland. The figures are reproduced from Sankhya, B (2009) with the permission of the Indian Statistical Institute, Kolkata, India

Since, however, for every real y, $e^{-y^2/2} < e^{-y^2/2}\{1 + \frac{1}{y^2}\}$, we have

$$\int_x^\infty \phi(y)\,dy < \phi(x)/x.$$

Also,

$$\frac{d}{dx}\big[\phi(x)\big\{1/x - 1/x^3\big\}\big] = \phi(x)\big\{3/x^4 - 1\big\}$$

which implies

$$\phi(x)\big\{1/x - 1/x^3\big\} = \int_x^\infty \phi(x)\big\{1 - 3/x^4\big\}\,dx < \int_x^\infty \phi(x)\,dx = 1 - \Phi(x)$$

where $\Phi(x) = \int_{-\infty}^x \phi(y)\,dy$. Combining these results, for large x,

$$1 - \Phi(x) \approx \phi(x)\frac{1}{x}.$$

Taking logarithm of both sides,

$$\begin{aligned}\log\{1-\Phi(x)\} &\approx -\frac{x^2}{2} - \log\left(\sqrt{2\pi}\cdot x\right)\\ &= -\frac{x^2}{2}\left\{1+\frac{2(\log(x)+\log(\sqrt{2\pi}))}{x^2}\right\}\\ &= -\frac{x^2}{2}\{1+o(1)\}, \quad \text{as } x\to\infty.\end{aligned}$$

The result follows. □

In what follows, we assume that the Hermite rank m equals 1, which holds if the first Hermite coefficient is non-zero, i.e.

$$c_1(x_s) = E\big[Z(\mathbf{u})\{Y_s(\mathbf{u}) - p(x_s)\}\big] = \int_{I_s} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\, dz \neq 0,$$

where $I_s = \{z \mid X(\mathbf{u}) = G_X(z) \in A_s\}$. Note that $c_1(x_s)$ will be equal to zero if I_s is exactly symmetric around zero, an unlikely situation in the present example. If $N_{k,s} > 0$, species s occurs in at least one of the k plots which have been surveyed. So, considering the new indicator process $T_{k,s}$ which takes the value 1 if $N_{k,s} > 0$, and it equals zero otherwise, the sum

$$\xi_k = T_{k,1} + T_{k,2} + \cdots + T_{k,S} \tag{9.34}$$

is the total number of (different) species in the k plots. Thus the mean number of species in the first k plots is

$$\begin{aligned}E[\xi_k] &= \sum_{s=1}^{S} P\{S_{k,s} > 0\} = \sum_{s=1}^{S} P\left\{\sum_{j=1}^{k} Y_s(\mathbf{u_j}) > 0\right\}\\ &= \sum_{s=1}^{S}\left[1 - P\left(\sum_{j=1}^{k} Y_s(\mathbf{u_j}) - kp(x_s) \le -kp(x_s)\right)\right].\end{aligned}$$

Now, we let S and k tend to ∞ and $\frac{k}{S} \to 0$ so that the number of species is much larger than the number of plots on the sampling grid. This takes us to an asymptotic expression for the number of unseen species, i.e. species which have not been discovered in the first k plots. Under suitable conditions (which have to take into account that we are taking simultaneous limits in S and k) an approximation of the

following form can be obtained (using (9.33)):

$$S - E(\xi_k) = \sum_{s=1}^{S} P\left(\frac{\sum_{j=1}^{k} Y_s(\mathbf{u_j}) - kp(x_s)}{\sigma_k(x_s)} \leq -\frac{kp(x_s)}{\sigma_k(x_s)} \right)$$
$$\sim \sum_{s=1}^{S} \exp\left\{ -k^\theta p^\theta(x_s)/\sigma_k^\theta(x_s) \cdot \frac{a(x_s)}{2} \right\}, \quad \text{as } k \to \infty.$$

We summarize this results in the following theorem (Ghosh 2009):

Theorem 9.3 *Under the assumptions given above, the difference between S and the expected number of observed species $E(\xi_k)$ in k plots can be approximated by*

$$S - E(\xi_k) \approx S \cdot \int_o^1 \exp\{-p^\theta(x) b_k(x)\}\, dx$$

where

$$b_k(x) = \frac{1}{2} \frac{a(x) k^\theta}{\sigma_k^\theta(x)}.$$

Moreover, for the increments, we obtain the approximation

$$E(\xi_k) - E(\xi_{k-1}) \sim S \cdot \int_o^1 \left[\exp\{-p^\theta(x) b_k(x)\} - \exp\{-p^\theta(x) b_{k-1}(x)\} \right] dx.$$

The most common example is $\theta = 2$, i.e. all $W_{s,k}$ are asymptotically standard normal. Let $\sigma_k(x_s) = k^\beta$ with $0.5 \leq \beta < 1$. In particular, if $p^2(x)a(x) = c^2 \cdot x^2$ for some constant $c \neq 0$, we have

$$S - E(\xi_k) \sim S \cdot \int_o^1 \exp\left\{ -\frac{1}{2} c^2 k^{2-2\beta} x^2 \right\} dx$$
$$= S \cdot \sqrt{2\pi} c^{-1} k^{\beta-1} \frac{1}{\sqrt{2\pi c^{-2} k^{2\beta-2}}} \int_o^1 \exp\left\{ -\frac{x^2}{2(c^{-1}k^{\beta-1})^2} \right\} dx$$
$$= S \cdot \sqrt{2\pi} c^{-1} k^{\beta-1} \left[\Phi\left(ck^{1-\beta}\right) - \frac{1}{2} \right] \approx S \cdot \sqrt{\frac{\pi}{2}} c^{-1} k^{\beta-1}.$$

For large S and $k \to \infty$, we therefore may use the following approximation:

Corollary 9.1 *Under the conditions stated above, for $S \to \infty$, $k/S \to 1$, the number of unseen species $S - E(\xi_k)$ can be approximated by*

$$S - E(\xi_k) \sim S \cdot \sqrt{\frac{\pi}{2}} c^{-1} k^{\beta-1}$$

where $\beta \in [0.5, 1)$. *Moreover, for the increments, we have the approximation*

$$E(\xi_k) - E(\xi_{k-1}) \sim S\sqrt{\frac{\pi}{2}}c^{-1}\left[k^{\beta-1} - (k-1)^{\beta-1}\right]$$
$$\sim S\sqrt{\frac{\pi}{2}}c^{-1}(\beta-1)k^{\beta-2}.$$

In other words, if $\delta_k = \xi_k - \xi_{k-1} =$ *increment in number of species from* $k-1$ *plots to* k *plots*, then in log–log coordinates,

$$\log(\delta_k) \sim \text{constant} + (\beta - 2)\log(k).$$

Thus the increment in the species–area curve will be hyperbolic with an exponent equal to $\beta - 2 < -1$, so that the predicted number of species in the infinitely large landscape governed by this background process will be finite. Also, the number of unseen species, i.e. the difference from the total number S will also decrease hyperbolically, with an exponent equal to $\beta - 1 < 0$.

Chapter 10
Resampling

10.1 General Introduction

Resampling or bootstrap methods refer to techniques where statistical inference is based on a simulated distribution of a statistic T_n obtained by resampling from an observed sample $X_1, \ldots, X_n$. Inference of this type is always conditional on the sample. In the most general version, no model assumptions are used except for global conditions such as stationarity, existence of some moments, etc. In the most restricted version, a parametric model is specified and resampling is used only as a simple way of obtaining an approximate distribution of T_n. Note that different terms such as 'bootstrap', 'resampling', 'subsampling', etc. are used in the literature for different variations of the same general idea. Since there does not seem to be a unified terminology, we use 'resampling' and 'bootstrap' as synonyms.

The original bootstrap (Efron 1979) was developed for i.i.d. data. Under the i.i.d. assumption, only the marginal distribution is unknown. Suppose, for instance, that we are interested in inference about the location parameter μ, given the observed data $\mathscr{Y}_n = (Y_1, \ldots, Y_n)$ where $Y_j = \mu + X_j \in \mathbb{R}$ and X_j are i.i.d. with distribution F_X. If we estimate μ by the sample mean $T_n = \bar{y}$, then we can write T_n as a functional $T_n(F_n) = \int u \, dF_n(u)$ of the empirical distribution function $F_n(x) = n^{-1} \sum 1\{Y_j \le x\}$. If the distribution function $F_Y(x) = F_X(x - \mu)$ of Y were known, then, in principle, the distribution of T_n could be calculated exactly by evaluating the n-dimensional integral $F_{T_n}(x) = P(T_n \le x) = \int_A dF_Y(y_1)\, dF_Y(y_2) \cdots dF_Y(y_n)$ where $A = \{y \in \mathbb{R}^n : y_1 + \cdots + y_n \le nx\}$. Usually, F_Y is unknown and is therefore replaced by an estimate $\hat{F}_Y$. One then has to evaluate $\hat{F}_{T_n}(x) = \hat{P}(T_n \le x) = \int_A d\hat{F}_Y(y_1)\, d\hat{F}_Y(y_2) \cdots d\hat{F}_Y(y_n)$. In most cases, the numerical evaluation of high dimensional integrals is difficult. The easiest alternative is Monte Carlo approximation which means that we approximate $\hat{F}_{T_n}$ by a simulated distribution, say $\hat{F}^*_{T_n}$, based on a sufficiently large sample of i.i.d. values $T^*_{n,1}, \ldots, T^*_{n,N}$ with $T^*_{n,j} \sim \hat{F}_{T_n}$. This can be done without actually computing $\hat{F}_{T_n}$ directly (after all that is what we wanted to avoid), namely by resampling. Independent samples $\mathscr{Y}^*_{n,j} = \{Y^*_{1,j}, \ldots, Y^*_{n,j}\}$ $(j = 1, 2, \ldots, N)$ are simulated and the

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_10, © Springer-Verlag Berlin Heidelberg 2013

sample means $T^*_{n,j} = n^{-1}\sum_{i=1}^n Y^*_{i,j} = T_n(F^*_{n,j})$ (with $F^*_{n,j}$ denoting the empirical distribution function of $Y^*_{1,j}, \ldots, Y^*_{n,j}$) are computed. For each j, the values $Y^*_{i,j}$ $(i = 1, 2, \ldots, n)$ are obtained by simulating n independent realizations of a random variable $Y^* \sim \hat{F}_Y$. If $\hat{F}_Y$ is equal to the empirical distribution function F_n, then this is the same as drawing $Y^*_{i,j}$ $(i = 1, 2, \ldots, n)$ randomly with replacement (and equal probability n^{-1}) from the original set of observations $\{Y_1, \ldots, Y_n\}$.

Resampling procedures can thus be considered as a simulation device to obtain an approximate distribution function of a statistic T_n. It should be noted here that the sample mean is a relatively simple statistic because it can be expressed explicitly as a function of $Y_1, \ldots, Y_n$. Many estimators in statistics are defined by equations that do not lead to an explicit expression for T_n and F_{T_n}. For example, most non-Gaussian maximum likelihood estimators, M-estimators or minimum contrast estimators are defined as solutions of nonlinear equations for which no explicit solution exists. This makes resampling procedures even more useful because explicit expressions are not required.

The obvious question is how accurate a bootstrap approximation $\hat{F}^*_{T_n}$ of F_{T_n} is and, in fact, whether it works at all. Usually, if T_n is an appropriately standardized statistic, then it converges in distribution to a certain nondegenerate random variable $Z \sim F_Z$. For instance, in the i.i.d. example above, we may redefine T_n as $T_n = \sqrt{n}(\bar{y}_n - \mu)/\sigma$ which converges to a standard normal variable, provided that $\sigma^2 = \mathrm{var}(X_j)$ is finite. The asymptotic distribution F_Z is a natural competitor of the bootstrap approximation $\hat{F}^*_{T_n}$. Since F_Z is exactly correct asymptotically, the first requirement is that the same is true for $\hat{F}^*_{T_n}$. This is also called 'validity' of the bootstrap procedure. Thus, one needs to prove that $\hat{F}^*_{T_n}$ converges to F_Z as n tends to infinity. Once validity is shown, the next question is why we should prefer to use $\hat{F}^*_{T_n}$ instead of the asymptotic distribution F_Z. There are at least two possible reasons: (i) F_Z may be complicated or unknown, (ii) $\hat{F}^*_{T_n}$ may be more accurate than the asymptotic distribution F_Z.

The first reason is certainly relevant in the context of long-range dependence. For instance, under Gaussian subordination with Hermite rank two or higher, asymptotic distributions of normalized sums are marginals of non-Gaussian Hermite processes. These distributions are rather complicated and, in practice, we actually do not even know which one applies because the Hermite rank is an unknown quantity (in fact, we do not even know whether Gaussian subordination applies). Also, even in the case of a Gaussian limit (i.e. Hermite rank one), the exponent of n in the standardization is unknown and the normalizing constant (or even a slowly varying function) may be complicated. Resampling procedures based on self-normalized statistics that avoid explicit estimation of this exponent (and the constant or slowly varying function) provide a simple alternative to more explicit model based approaches. Other examples where F_Z may be complicated are encountered in the context of stable laws (see below).

To justify the second reason for using $\hat{F}^*_{T_n}$, namely improved accuracy, more refined asymptotic results are required since convergence of $\hat{F}^*_{T_n}$ to F_Z (which is

a basic prerequisite for considering $\hat{F}^*_{T_n}$ at all) does not automatically imply that, compared to F_Z, $\hat{F}^*_{T_n}$ is closer to the true finite sample distribution F_{T_n}. Suppose that $F_{T_n}(x) = F_Z(x) + a_n(x) + o(a_n)$ (with $a_n = o(1)$) and $F_{T_n}(x) = \hat{F}^*_{T_n}(x) + b_n(x) + o_p(b_n)$. (Note that in contrast to $a_n(x)$, $b_n(x)$ is random because $\hat{F}^*_{T_n}(x)$ is calculated conditionally on the observed sample.) For *validity* it is sufficient to show that $\tilde{b}_n = \sup_x |b_n(x)| = o_p(1)$. To prove that $\hat{F}^*_{T_n}$ is *more accurate* than F_Z, one needs to make a second order comparison. Such comparisons are usually based on Edgeworth expansions (see, e.g. Hall 1992). In many situations, it is indeed possible to show that $\tilde{b}_n = o_p(a_n)$ which means that the bootstrap error is of a smaller order than the one of the asymptotic approximation. The implications of such an improvement are often clearly visible. For instance, if F_X in the i.i.d. example above is highly skewed, then the distribution of $T_n = \sqrt{n}(\bar{y}_n - \mu)/\sigma$ can be highly skewed too, even for relatively large sample sizes. In such a case, an approximation by the standard normal distribution F_Z is inappropriate whereas a bootstrap distribution tends to mimic the asymmetry of F_{T_n} rather well.

The validity and accuracy of resampling techniques is fairly well understood in the i.i.d. case (see, e.g. Hall 1992; Politis et al. 1999; Lahiri 2003, and references therein). Once the assumption of independence is abandoned, further complications arise because the marginal distribution is not the only unknown quantity. In full generality, a statistic T_n is a functional of the complete joint n-dimensional distribution $F_{\mathscr{Y}_n}(y_1, \ldots, y_n) = P(Y_1 \le y_1, \ldots, Y_n \le y_n)$. The question how to resample from an observed series $\mathscr{Y}_n = (Y_1, \ldots, Y_n)$ is therefore much more difficult. First of all, we have one observation only (namely $\mathscr{Y}_n$ itself) from the n-dimensional distribution $F_{\mathscr{Y}_n}$ so that no consistent estimate of $F_{\mathscr{Y}_n}$ is available, unless certain assumptions are imposed. This is, of course, a general problem of statistical inference for stochastic processes, and led, already in the early days of time series analysis, to the introduction of properties such as stationarity and ergodicity. Most of the resampling theory for stochastic processes is concerned with the question under what kind of general conditions bootstrap works, which modifications are required to ensure validity and how to improve the second-order error. The original approach of drawing individual observations $Y^*_{i,j}$ $(i = 1, 2, \ldots, n)$ independently with replacement from $\{Y_1, \ldots, Y_n\}$ does not provide valid results in general because the dependence structure is removed completely by the resampling scheme.

There are two main ideas how to solve this problem. The first approach is to resample whole blocks $B_r = (Y_r, \ldots, Y_{r+l-1})$ of adjacent observations instead of individual values. By letting the block length l tend to infinity such that at the same time $l/n \to 0$, an infinite time horizon is captured ultimately within each block while at the same time the number of blocks (and thus the number of items to resample from) also tends to infinity. Methods of this type are also called block or blockwise bootstrap or subsampling. The problem is, of course, that in general F_{T_n} depends on the complete n-dimensional distribution $F_{\mathscr{Y}_n}$ whereas the subsampling procedure essentially relies on estimating the lower-dimensional probability function $F_{\mathscr{Y}_l}$. Although l tends to infinity, we also have $l = o(n)$. It is therefore not clear

a priori whether information about the dependence structure beyond lag l is asymptotically negligible when characterizing the distribution of $T_n(F_{\mathscr{Y}_n})$, and in how far it matters that $F_{\mathscr{Y}_l}$ actually has to be estimated as well. As it turns out, the main dividing line is between short and long memory. The validity and second-order accuracy of relatively simple versions of blockwise subsampling can be established under short-memory assumptions (Carlstein 1986; Künsch 1989; Politis and Romano 1993). This is not the case in general for long-memory processes although some modifications of blockwise resampling work under certain specific assumptions (see below).

A second approach to adapting bootstrap to dependent data consists of removing all or some of the dependence *before* applying resampling. Resampling methods based on this principle are subsumed under the name 'sieve bootstrap'. For instance, under the assumption that a causal linear process $Y_t = \sum a_j \varepsilon_{t-j}$ (with ε_t i.i.d.) is observed, one may use a sequence of autoregressive filters $\Phi_n(B) = 1 - \varphi_{1,n}B - \cdots - \varphi_{p_n,n}B^{p_n}$ with $p_n \to \infty$ and $\varphi_{j,n}$ estimated by minimizing the least squares criterium $\sum(Y_t - \Phi_n(B)Y_t)^2$. Resampling is then applied to the residual process $e_{t,n} = \Phi_n(B)Y_t$. Under suitable short-memory conditions, it can be shown that with $p_n \to \infty$ it is possible to approximate the actual i.i.d. residuals ε_t with sufficient accuracy (for early literature on autoregressive fitting with $p_n \to \infty$, see, e.g. Parzen 1974; Berk 1974; Hannan and Deistler 1988; also see Shibata 1980 for the connection to optimal prediction and Akaike's information criterion). Note that, if the order p_n is kept fixed, then we are relying on the stronger assumption that Y_t is generated by a finite-order autoregressive process. This is a special case of a 'parametric bootstrap'. Validity and second-order accuracy of the sieve bootstrap have been established under short-memory conditions (see, e.g. Bühlmann 1997, 2002, and references therein). In general, sieve methods rely on more restrictive assumptions than blockwise bootstrap because the choice of the preprocessing device has to be appropriate. On the other hand, if the assumptions are correct, then the sieve bootstrap tends to provide more accurate approximations (see, e.g. Choi and Hall 2000).

While both approaches (blockwise and sieve) are quite well understood under short-memory conditions, the situation is more difficult in the presence of long memory. Generally, the validity of standard blockwise methods no longer holds, unless specific modifications are applied (see, e.g. Lahiri 1993, 2003; Hall et al. 1998; Nordman et al. 2006). The easiest situation is encountered for the parametric bootstrap where not only validity but also improved second-order accuracy has been established for certain classes of estimators under long-memory conditions (see, e.g. Andrews et al. 2006; Andrews and Lieberman 2005). Similar results are available for the sieve bootstrap based on autoregressive fitting as above with $p_n \to \infty$ such that $n^{\frac{1}{2}-d}(\log n)^{\frac{1}{2}-d}p_n \to 0$ (Poskitt 2007a, 2007b). Note that the results in Poskitt (2007a, 2007b) are also interesting from the point of view of parameter estimation for a long-memory process because it is shown that the fitted AR-coefficients $\varphi_{j,n}$ converge to the coefficients a_j in the Wold representation with a simultaneous bound on the estimation error $|\varphi_{j,n} - a_j|$ $(j = 1, 2, \ldots, p_n)$. This is achieved without using fractional differencing or direct estimation of the fractional differencing parameter

d (in contrast to comparable AR-fitting methods such as Bhansali et al. 2006; see Sect. 5.9.3).

In the following sections, a few selected resampling methods will be discussed in more detail in the context of long-range dependence. For further literature on resampling methods and Edgeworth expansions for long-memory processes, see, e.g. Lahiri (2003), and references given in Lieberman et al. (2001, 2003), Giraitis and Robinson (2003), Faÿ et al. (2004), Lieberman and Phillips (2004), Andrews and Lieberman (2005), Nordman and Lahiri (2005), Andrews et al. (2006), McElroy and Politis (2007), Poskitt (2007a, 2007b), Jach et al. (2012), Kim and Nordman (2011).

10.2 Some Basics on Bootstrap for i.i.d. Data

Let $\mathscr{Y}_n = \{Y_1, \ldots, Y_n\}$ be a sample from the distribution F. Note that at this moment we do not assume any particular dependence structure of the original sequence Y_j ($j \in \mathbb{N}$), except that the marginal distribution is the same. The simplest bootstrap procedure starts with drawing a sample $Y_1^*, \ldots, Y_n^*$ with replacement from $\mathscr{Y}_n$. Conditionally on $\mathscr{Y}_n$, the random variables $Y_1^*, \ldots, Y_n^*$ are i.i.d., no matter what the original model is. Moreover,

$$P_*\big(Y_1^* = Y_j\big) := P\big(Y_1^* = Y_j | \mathscr{Y}_n\big) = 1/n, \quad j = 1, \ldots, n,$$

which means that the common (random) distribution function of Y_j^* ($j = 1, 2, \ldots, n$) is equal to the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1\{Y_j \le x\}.$$

To keep things simple, we consider estimation of the expected value $\mu = E(Y_1)$ by the sample mean $\bar{Y}_n$. Denote by $\bar{Y}_n^* = n^{-1} \sum_{j=1}^n Y_j^*$ the bootstrap sample mean. Also, let E_* be the expectation w.r.t. P_*. We have the following moment properties:

$$E_*\big(Y_i^*\big) = \int x \, dF_n(x) = \frac{1}{n} \sum_{j=1}^{n} Y_i = \bar{Y}_n,$$

$$E_*\big(\bar{Y}_n^*\big) = E\big(\bar{Y}_n^* | \mathscr{Y}_n\big) = \frac{1}{n} \sum_{j=1}^{n} E(Y_j | \mathscr{Y}_n) = \bar{Y}_n,$$

$$E\big(\bar{Y}_n^*\big) = E\big[E\big(\bar{Y}_n^* | \mathscr{Y}_n\big)\big] = E(\bar{Y}_n) = E(Y),$$

$$\mathrm{var}_*\big(Y_i^*\big) = \int x^2 \, dF_n(x) - \left(\int x \, dF_n(x)\right)^2 = \frac{1}{n} \sum_{j=1}^{n} Y_j^2 - \left(\frac{1}{n} \sum_{j=1}^{n} Y_j\right)^2 =: s^2,$$

and recalling that Y_j^* are conditionally independent,

$$\operatorname{var}_*\big(\bar{Y}_n^*\big) = \frac{1}{n}\operatorname{var}_*\big(Y_1^*\big) = \frac{s^2}{n}. \tag{10.1}$$

Let us now focus on the case where $Y_1, \ldots, Y_n$ are i.i.d. observations with a finite variance. The standardized sample mean is asymptotically standard normal, i.e.

$$T_n = \frac{\bar{Y}_n - \mu}{\sqrt{\operatorname{var}(\bar{Y}_n)}} = \sqrt{n}\frac{\bar{Y}_n - \mu}{\sqrt{\operatorname{var}(Y_1)}} \xrightarrow{\mathrm{d}} N(0,1). \tag{10.2}$$

In the bootstrap approach, the initial population one sampled from is replaced by $\mathscr{Y}_n$. Thus, the bootstrap version of T_n is obtained by replacing $\bar{Y}_n$ by the bootstrap sample mean $\bar{Y}_n^*$, the population mean μ by the bootstrap population mean $E_*(Y_1^*) = \bar{Y}_n$, and the population variance $\operatorname{var}(Y_1)$ by the bootstrap population variance $\operatorname{var}_*(Y_1^*) = s^2$. The bootstrap version of T_n is therefore given by

$$T_n^* = \frac{\bar{Y}_n^* - E_*(Y_1^*)}{\sqrt{\operatorname{var}_*(\bar{Y}_n^*)}} = \sqrt{n}\frac{\bar{Y}_n^* - E_*(Y_1^*)}{\sqrt{\operatorname{var}_*(Y_1^*)}} = \sqrt{n}\frac{\bar{Y}_n^* - \bar{Y}_n}{s}. \tag{10.3}$$

Since $\bar{Y}_n$ converges in probability to μ and the denominator converges in probability to $\sqrt{\operatorname{var}(Y)}$, T_n^* has the same behaviour as T_n asymptotically. More specifically, the following lemma justifies validity of the bootstrap for i.i.d. data with a finite variance (see, e.g. Lahiri 2003, Theorem 2.1).

Lemma 10.1 *Assume that $Y_1, \ldots, Y_n$ are i.i.d. with $\operatorname{var}(Y_i) < \infty$. Then*

$$\sup_x \big| P_*\big(T_n^* \le x\big) - \Phi(x) \big| = o_p(1),$$

where $\Phi(x)$ is the standard normal distribution.

10.3 Self-normalization

Consider $Y_j = \mu + X_j$ ($j \in \mathbb{N}$) with X_j a stationary zero-mean sequence and assume that after suitable standardization the sample mean converges to a nondegenerate random variable Z, or in other words,

$$T_n := \frac{\sum_{j=1}^n Y_j - n\mu}{v_n} = \frac{n}{v_n}(\bar{Y}_n - \mu) \xrightarrow{\mathrm{d}} Z \sim F_Z \tag{10.4}$$

where F_Z is a nondegenerate distribution. Usually, the choice of v_n is $v_n^2 = \operatorname{var}(\sum_{j=1}^n Y_j)$, provided that this quantity exists. In the i.i.d. case with finite variance, we have $v_n^2 = n \cdot \operatorname{var}(X_1)$ and Z standard normal. Usually, v_n has to be estimated. In some situations, v_n is not even computable or requires an additional

estimation step. For example, if the random variables X_j are i.i.d. with a regularly varying distribution with index $-\alpha$ ($\alpha \in (0,2)$), then $v_n = n^{1/\alpha} L(n)$ where $L(n)$ is a slowly varying function, and Z is a stable random variable. Thus, in principle, we would need to estimate α (and even the slowly varying function L) before computing T_n. Often, it is possible to replace v_n by a data-based normalizer V_n without explicit estimation of model specific quantities, such as α or L. For example, for i.i.d. data (both with finite and infinite variance), we can replace v_n by the square root of $V_n^2 = n^{-1} \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$. Given a data-based normalizer V_n we then consider the 'self-normalized' statistic

$$U_n := \frac{\sum_{j=1}^n Y_j - n\mu}{V_n} = \frac{n}{V_n}(\bar{Y}_n - \mu). \tag{10.5}$$

The choice of the normalizer V_n has to be modified for dependent sequences to guarantee that V_n/v_n converges to one in probability.

Denote by Z_0 the limit of U_n. If Z in (10.4) is normal, then Z_0 is also a standard normal variable. In general, however, the distributions of Z and Z_0 can be quite complicated, and may even differ. For example, if the data are i.i.d. with infinite variance, then Z is a stable random variable, but Z_0 is different. To see this, assume that X_j ($j \in \mathbb{N}$) are i.i.d. and regularly varying with index $-\alpha$. Consider

$$W_n := \frac{n}{V_n}(\bar{Y}_n - \mu) \tag{10.6}$$

where

$$V_n^2 = \sum_{j=1}^n (Y_j - \bar{Y}_n)^2.$$

We note that the random variables Y_j^2 ($j \in \mathbb{N}$) are regularly varying with index $-\alpha/2$ and thus have an infinite mean. In particular, $n^{-2/\alpha} \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$, and hence $n^{-1/\alpha} V_n$ converges to a stable random variable. This implies that

$$W_n = \frac{n^{-1/\alpha} \sum_{j=1}^n (Y_j - \mu)}{n^{-1/\alpha} V_n}$$

converges to a ratio R of two dependent stable random variables. In principle, we may use this information to construct confidence intervals for μ of the form

$$\left[\bar{Y}_n - z_{1-\frac{1}{2}p_0} n^{-1} V_n, \bar{Y}_n - z_{\frac{1}{2}p_0} n^{-1} V_n\right],$$

where z_p denotes the $(100p)$th percentile of R. However, these percentiles may not be easily computable. Resampling methods are useful to overcome this problem.

10.4 The Moving Block Bootstrap (MBB)

Lemma 10.1 provides validity of the bootstrap procedure in the case of i.i.d. data with existing second moments. Now we turn our attention to the case of dependent data. Assume that $Y_j = \mu + X_j$ ($j \in \mathbb{N}$) is a stationary sequence of random variables with short memory and $\sigma^2 := \mathrm{var}(Y) < \infty$. Then convergence (10.2) has to be replaced by

$$\sqrt{n}\frac{\bar{Y}_n - \mu}{\sigma_0} \xrightarrow{\mathrm{d}} N(0,1), \tag{10.7}$$

with

$$\sigma_0^2 = \mathrm{var}(Y) + 2\sum_{k=1}^{\infty} cov(Y_0, Y_k). \tag{10.8}$$

However, as mentioned above, sampling with replacement from $\mathscr{Y}_n$ produces conditionally independent random variables. Therefore, if we use T_n^* defined in (10.3), then the result in Lemma 10.1 still applies. This contradicts (10.7) so that the bootstrap procedure is no longer valid (except in the special case of uncorrelated observations). The asymptotic variance of bootstrap replicates is wrong by the factor $(\sigma_0/\sigma)^2$. The reason is that the bootstrap procedure cannot recreate $\mathrm{var}(\bar{Y}_n)$. More exactly, recall that $\mathrm{var}_*(\bar{Y}_n^*) = s^2/n$ (10.1). The expected value of the conditional variance is then equal to

$$E\big[\mathrm{var}_*\big(\bar{Y}_n^*\big)\big] = \frac{1}{n}\left\{E\big(Y^2\big) - \frac{1}{n^2}\sum_{j,j'=1}^{n} E(Y_j Y_{j'})\right\} = \frac{1}{n}\big\{\mathrm{var}(Y) - \mathrm{var}(\bar{X}_n)\big\}. \tag{10.9}$$

Since $\mathrm{var}(\bar{X}_n) \to 0$ (except for degenerate cases that are not of interest here), the expected variance is approximately equal to

$$E\big[\mathrm{var}_*\big(\bar{Y}_n^*\big)\big] \sim \frac{1}{n}\mathrm{var}(Y) = \frac{\sigma^2}{n}.$$

This is in contrast to

$$\mathrm{var}(\bar{Y}_n) \sim \frac{\sigma_0^2}{n}.$$

To obtain a valid bootstrap procedure, a suitable modification is required. One of the possible solutions is the so-called *Moving Block Bootstrap* (*MBB*) (Carlstein 1986; Künsch 1989). To preserve most of the dependence structure, we sample (with replacement) blocks $B_1^*, \ldots, B_k^*$ from the set of all available blocks $B_r = (Y_r, \ldots, Y_{r+l-1})$ ($r = 1, \ldots, N_b$; $N_b = n - l + 1$) instead of sampling single observations. A bootstrapped sample $Y_1^*, \ldots, Y_n^*$ is generated by pasting $k = [n/l]$ sampled blocks $B_1^*, \ldots, B_k^*$ next to each other. Note that, by definition, $B_r^* = (Y_{(r-1)l+1}^*, \ldots, Y_{rl}^*)$ ($r = 1, \ldots, k$). For example, if $k = 2$ and blocks, say, B_1 and

B_3 are selected, then the bootstrap sample is

$$\left(Y_1^*, \ldots, Y_l^*, Y_{l+1}^*, \ldots, Y_{2l}^*\right) = (Y_1, \ldots, Y_l, Y_3, \ldots, Y_{l+2}).$$

Also note that the actual length of the bootstrapped series is $\tilde{n} = kl = [n/l]l$ (where $[n/l]$ denotes the largest integer not exceeding n/l), but the difference between $\tilde{n}$ and n is negligible asymptotically. We will therefore write $n = kl$ for simplicity. Denote by

$$\zeta_r = \zeta_{r,l} = \sum_{j \in B_r} Y_j = \sum_{j=r}^{r+l-1} Y_j$$

$(r = 1, 2, \ldots, N_b)$ the block sums and by

$$\zeta_r^* = \zeta_{r,l}^* = \sum_{j \in B_r^*} Y_j^* = \sum_{j=(r-1)l+1}^{rl} Y_j^*$$

the corresponding bootstrapped quantities (the index l will be dropped unless it needs to be emphasized). The bootstrap mean is given by

$$\bar{Y}_n^* = n^{-1} \sum_{j=1}^{n} Y_j^* = \frac{1}{k} \sum_{r=1}^{k} \frac{1}{l} \zeta_{r,l}^* = \frac{1}{k} \sum_{r=1}^{k} \left(\frac{1}{l} \sum_{j=(r-1)l+1}^{rl} Y_j^* \right).$$

When drawing block B_r^*, each of the blocks B_s $(s = 1, \ldots, N_b)$ has the same probability of being chosen. Thus, for any $r \in \{1, \ldots, k\}$,

$$P_*\left(B_r^* = B_s\right) = \frac{1}{N_b} \quad (s = 1, \ldots, N_b) \tag{10.10}$$

so that

$$\begin{aligned} E_*\left(\bar{Y}_n^*\right) &= E_*\left[\frac{1}{k} \sum_{r=1}^{k} \left(\frac{1}{l} \sum_{Y_j^* \in B_r^*} Y_j^* \right) \right] \\ &= \frac{1}{N_b} \sum_{r=1}^{N_b} \left(\frac{1}{l} \sum_{Y_j \in B_r} Y_j \right) = \frac{1}{N_b l} \sum_{r=1}^{N_b} \sum_{j=r}^{r+l-1} Y_j \\ &= \frac{1}{N_b} \sum_{r=1}^{N_b} \frac{1}{l} \zeta_{r,l}. \end{aligned}$$

Note that, if $l/n \to 0$ fast enough, then $E_*(\bar{Y}_n^*)$ may be approximated by the sample mean $\bar{Y}_n$ because all variables Y_j occur in the sum l times except for l observations on the left and right border, respectively.

Now, recalling that the blocks are conditionally independent, the conditional variance of the bootstrap mean is

$$\begin{aligned}\operatorname{var}_*\big(\bar{Y}_n^*\big) &= \operatorname{var}_*\left(\frac{1}{k}\sum_{r=1}^{k}\frac{1}{l}\zeta_r^*\right)\\ &= \frac{k}{(kl)^2}\operatorname{var}_*\big(\zeta_r^*\big) = \frac{k}{(kl)^2}\operatorname{var}_*\left(\sum_{j=1}^{l} Y_j^*\right)\\ &= \frac{k}{n^2}\left\{\frac{1}{N_b}\sum_{r=1}^{N_b}\left(\sum_{j=r}^{r+l-1} Y_j\right)^2 - \left(\frac{1}{N_b}\sum_{r=1}^{N_b}\sum_{j=r}^{r+l-1} Y_j\right)^2\right\}.\end{aligned}$$

For the unconditional expected value of the variance, we may assume, without loss of generality, that $\mu = 0$. Then the second term does not contribute asymptotically, and we obtain

$$E\big[\operatorname{var}_*\big(\bar{Y}_n^*\big)\big] \sim \frac{k}{n^2}E\left[\left(\sum_{j=1}^{l} Y_j\right)^2\right] = \frac{1}{nl}\operatorname{var}\left(\sum_{j=1}^{l} Y_j\right).$$

If the stationary sequence Y_j has short memory and $n, l \to \infty$ such that $l/n \to 0$, this leads to

$$E\big[\operatorname{var}_*\big(\bar{Y}_n^*\big)\big] \sim \frac{1}{nl}\sigma_0^2 l = \frac{\sigma_0^2}{n},$$

where σ_0 is given in (10.8). Therefore, the bootstrap variance of the bootstrap mean is asymptotically the same as $\operatorname{var}(\bar{Y}_n)$ and the MBB bootstrap statistic

$$T_n^* = \frac{\bar{Y}_n^* - E_*(Y_1^*)}{\sqrt{\operatorname{var}_*(\bar{Y}_n^*)}}$$

has the same asymptotic distribution as $T_n = (\bar{Y}_n - \mu)/\sigma_0$.

However, if the random variables X_j ($j \in \mathbb{N}$) are Gaussian with autocovariance function $\gamma_X(k) \sim L_\gamma k^{2d-1}$ ($0 < d < \frac{1}{2}$), then

$$\operatorname{var}(\bar{Y}_n) \sim n^{-2}v_n^2$$

and

$$T_n = \frac{n(\bar{Y}_n - \mu)}{v_n} \xrightarrow{\mathrm{d}} N(0,1)$$

where $v_n^2 = n^{2d+1} L_S$ with $L_S = C_1 L_\gamma$ (Sect. 4.2.2). On the other hand,

$$E\big[\mathrm{var}_*(\bar{Y}_n^*)\big] \sim \frac{k}{n^2} E\left(\sum_{j=1}^{l} Y_j\right)^2 = \frac{1}{nl}\,\mathrm{var}\left(\sum_{j=1}^{l} Y_j\right)$$
$$\sim C\frac{1}{nl} l^{2d+1} = C\frac{l^{2d}}{n}.$$

Thus

$$\frac{E[\mathrm{var}_*(\bar{Y}_n^*)]}{\mathrm{var}(\bar{Y}_n)} \sim \mathrm{const}\left(\frac{l}{n}\right)^{2d} \to 0$$

and

$$\frac{\bar{Y}_n^* - E_*(\bar{Y}_n^*)}{\sqrt{\mathrm{var}(\bar{Y}_n)}} = \frac{\bar{Y}_n^* - E_*(\bar{Y}_n^*)}{\sqrt{\mathrm{var}_*(\bar{Y}_n^*)}} \sqrt{\frac{\mathrm{var}_*(\bar{Y}_n^*)}{\mathrm{var}(\bar{Y}_n)}} = T_n^* \sqrt{\frac{\mathrm{var}_*(\bar{Y}_n^*)}{\mathrm{var}(\bar{Y}_n)}} \to 0.$$

This means that the MBB bootstrap heavily underestimates the variability of the sample mean $\bar{Y}_n$ such that the asymptotic coverage probabilities of bootstrap confidence intervals for μ are zero. The reason is that too much of the long-memory property is lost by pasting together independent blocks. In the short-memory case, the rate of $\sum_{t=1}^{n} Y_t$ is $O_p(\sqrt{n})$ which is the same as for i.i.d. data, and therefore also the same as for $\sum_{r=1}^{k} \zeta_r^* = O_p(\sqrt{kl})$ with $kl = n$. The error in the standardization is only a multiplicative constant that can be made arbitrarily small by letting l tend to infinity. This is no longer the case under long memory because independent sampling of blocks changes the *rate* of the original sum $\sum_{t=1}^{n} Y_t = O_p(n^d \cdot n^{\frac{1}{2}})$ to the smaller rate of the bootstrapped sum given by $\sum_{r=1}^{k} \zeta_r^* = O_p(k^{\frac{1}{2}} l^{d+\frac{1}{2}}) = O_p(l^d \cdot n^{\frac{1}{2}})$.

A simple remedy to make the MBB bootstrap work in the long-memory context is suggested in Lahiri (1993). Instead of using the sample mean directly, we consider a statistic that takes into account independence introduced by blockwise resampling. This can be done by adjusting the standardization accordingly. As before $k = [n/l]$ blocks $B_1^*, \ldots, B_k^*$ are sampled independently with replacement, but we now consider the correctly standardized statistic

$$\tilde{T}_n^* = k^{-\frac{1}{2}} \sum_{r=1}^{k} \frac{\zeta_r^* - l \cdot E_*(Y_1^*)}{v_l}$$
$$= k^{-\frac{1}{2}} \sum_{r=1}^{k} l^{-d-\frac{1}{2}} \frac{\zeta_r^* - l \cdot E_*(Y_1^*)}{\sqrt{C_1 L_\gamma}},$$

or

$$\tilde{T}_n^* = k^{-\frac{1}{2}} \sum_{r=1}^{k} l^{-d-\frac{1}{2}} \frac{\zeta_r^* - l \cdot \bar{Y}_n}{\sqrt{C_1 L_\gamma}}.$$

Since $\tilde{T}_n^*$ is equal to $k^{-\frac{1}{2}}$ times a sum of k independent equally distributed standardized variables, the central limit theorem holds and one can even show uniform convergence (Lahiri 1993)

$$\sup_{x \in \mathbb{R}} \left| P_* \left(\tilde{T}_n^* \leq x \right) - \Phi(x) \right| = o_p(1).$$

This result has to be interpreted with care, however, because we are dealing with the case of long memory. For instance, consider the Gaussian subordination model $Y_j = \mu + G(X_j)$ where X_j is a stationary Gaussian process with $E(X_j) = 0$, $\mathrm{var}(X_j) = 1$, $E[G(X_j)] = 0$ and autocovariance function $\gamma_X(k) \sim L_\gamma(k)|k|^{2d_X - 1}$ (as $k \to \infty$) for some $0 < d_X < \frac{1}{2}$. If G has Hermite rank one, then the standardized sample mean converges to a standard normal variable and the standardization is the same as in $\tilde{T}_n^*$. In this sense, validity of the modified MBB procedure is established. However, if G has a Hermite rank m higher than one and $d_X > \frac{1}{2}(1 - m^{-1})$, then the asymptotic limit of the standardized sample mean is non-Gaussian. This means that the modified MBB is no longer valid. The question then arises why the modified MBB should be used at all. The reason is obviously not a complicated asymptotic distribution since validity holds only in the case where the asymptotic distribution is normal. As discussed previously, another possible motivation for using resampling is a better approximation of finite sample distributions. In how far the conditional distribution of $\tilde{T}_n^*$ does indeed provide a better approximation of the distribution of T_n has not yet been fully explored in the long-memory context. However, the idea of a modified MBB can be extended to other problems where the definition of a bootstrap based statistic with known asymptotic distribution is useful in its own right. For instance, Beran and Shumeyko (2012b) develop an MBB based test of the null hypothesis that a nonparametric trend function is continuous (see Sect. 10.7.2 below).

10.5 The Sampling Window Bootstrap (SWB)

As we saw above, the modified MBB is not valid under Gaussian subordination unless the Hermite rank of G is one. The reason is that independent sampling of blocks automatically entails the central limit theorem, independently of the Hermite rank. A natural idea to solve this problem is to avoid independent resampling. In the so-called sampling window (SW) approach, independent sampling of blocks is replaced by including all available blocks with equal weight in an empirical distribution function.

To be specific, we consider as before estimation of μ for the process $Y_j = \mu + G(X_j)$ where G has Hermite rank m, and X_j ($j \in \mathbb{N}$) is a stationary Gaussian

sequence with $E(X_j) = 0$, $\mathrm{var}(X_j) = 1$ and autocovariances $\gamma_X(k) \sim L_\gamma(k)k^{2d_X-1}$ with $L_\gamma(k) = c_\gamma > 0$, $\frac{1}{2}(1-m^{-1}) < d < \frac{1}{2}$. From Theorem 4.4 we have

$$n^{-(1-m(\frac{1}{2}-d))}L_S^{-1/2}\left(\sum_{j=1}^{n} Y_j - n\mu\right) \xrightarrow{\mathrm{d}} \frac{J(m)}{m!}Z_{m,H}(1), \tag{10.11}$$

where $L_S = J^2(m)/m!C_m c_\gamma^m$, $v_n = n^{1-m(\frac{1}{2}-d)}L_S^{1/2}$ and $Z = Z_{m,H}(1)$. As before, the replicates $T^*_{n,1}, \ldots, T^*_{n,N_b}$ are based on standardized sums over blocks $B_r = (Y_r, \ldots, Y_{r+l-1})$ $(r = 1, 2, \ldots, N_b)$ of length l. However, instead of resampling blocks independently and pasting them together, we use all N_b (partially overlapping) blocks to obtain the empirical distribution function

$$F^*_{T_n}(x) = \frac{1}{N_b}\sum_{r=1}^{N_b} 1\{T^*_{n,r} \le x\} = \frac{1}{N_b}\sum_{r=1}^{N_b} 1\left\{\frac{S_{n,l,r} - l\bar{Y}_n}{v_l} \le x\right\}$$

with

$$T^*_{n,r} := T^*_{n,l,r} := \frac{\sum_{j=r}^{r+l-1} Y_j - l\bar{Y}_n}{v_l} = \frac{S_{n,l,r} - l\bar{Y}_n}{v_l} \quad (r = 1, 2, \ldots, N_b).$$

By assigning equal weights to all available blocks and avoiding any kind of random reshuffling of the sequence, the complete dependence structure can essentially be preserved. Why this is so can be seen in more detail as follows. Recall that, as $n \to \infty$, $F_{T_n}(x) = P(T_n \le x) \to F_Z(x) := P(Z_{m,H}(1) \le x)$ for all $x \in \mathbb{R}$ and note that $E[F^*_{T_n}(x)] = P(T^*_{n,l,1} \le x)$. We will prove (Hall et al. 1998):

Theorem 10.1 *Let X_j be as defined above, and $l, n \to \infty$ such that $l/n \to 0$. Then*

$$\sup_{x\in\mathbb{R}} \left|F^*_{T_n}(x) - F_{T_n}(x)\right| \xrightarrow{\mathrm{p}} 0. \tag{10.12}$$

Proof In the first step, we will replace $\bar{Y}_n$ by μ in the definition of $F^*_{T_n}(x)$. To justify this, we note that with $\tilde{T}_{n,l,r} = (S_{n,l,r} - l\mu)/v_l$ we have

$$\tilde{T}_{n,l,r} - T^*_{n,l,r} = \frac{l}{v_l}(\bar{Y}_n - \mu) = \frac{lv_n}{nv_l}T_n.$$

On account of (10.11), T_n converges in distribution to the finite random variable $Z = Z_{m,H}(1)$. Furthermore,

$$\frac{lv_n}{nv_l} \to 0$$

since it was assumed that $l, n \to \infty$ and $l/n \to 0$.

The next useful fact is that both $F_{T_n}(x)$ and $F_{T_l}(x)$ converge to $F_Z(x)$ as $n, l \to \infty$. It is therefore sufficient to prove

$$\sup_{x\in\mathbb{R}} \big|\tilde{F}_{T_n}(x) - F_{T_l}(x)\big| \xrightarrow{\mathrm{p}} 0,$$

where

$$\tilde{F}_{T_n}(x) = \frac{1}{N_b}\sum_{r=1}^{N_b} 1\{\tilde{T}_{n,l,r} \le x\} = \frac{1}{N_b}\sum_{r=1}^{N_b} 1\big\{(S_{n,l,r} - l\mu)/v_l \le x\big\}.$$

We note that $E[\tilde{F}_{T_n}(x)] = P(T_l \le x) = F_{T_l}(x)$. Therefore,

$$\begin{aligned} E\big[\big(\tilde{F}_{T_n}(x) - F_{T_l}(x)\big)^2\big] &= \mathrm{var}\big(\tilde{F}_{T_n}(x)\big) = \frac{1}{N_b}\,\mathrm{var}\big(1\{\tilde{T}_{n,l,1} \le x\}\big) \\ &\quad + \frac{2}{N_b}\sum_{r=2}^{l} cov\big(1\{\tilde{T}_{n,l,1} \le x\}, 1\{\tilde{T}_{n,l,r} \le x\}\big) \\ &\quad + \frac{2}{N_b}\sum_{r=l+1}^{N_b} cov\big(1\{\tilde{T}_{n,l,1} \le x\}, 1\{\tilde{T}_{n,l,r} \le x\}\big) \\ &\le \frac{1}{N_b} + \frac{2l}{N_b} + \frac{2}{N_b}\sum_{r=l+1}^{N_b} cov\big(1\{\tilde{T}_{n,l,1} \le x\}, 1\{\tilde{T}_{n,l,r} \le x\}\big). \end{aligned}$$

Now, let us consider the case $m = 1$ only, so that v_l^2 is proportional to l^{2d_X+1}. Then the random variables $\tilde{T}_{n,l,r}$, $r = l+1, \ldots, N_b$, are centred Gaussian and w.l.o.g. we can assume that they have unit variance (formally, $\mathrm{var}(\tilde{T}_{n,l,r}) \sim 1$ as $l \to \infty$). Note that for a standardized bivariate normal vector $Z = (Z_1, Z_2)$ we have

$$\big|cov(Z_1, Z_2)\big| = \big|corr(Z_1, Z_2)\big| \ge \big|cov\big(1\{Z_1 \le x\}, 1\{Z_2 \le x\}\big)\big|.$$

Moreover, the separation between blocks B_1 and B_r is $r - l$. Therefore,

$$\begin{aligned} &\frac{2}{N_b}\sum_{r=l+1}^{N_b} Cov\big(1\{\tilde{T}_{n,l,1} \le x\}, 1\{\tilde{T}_{n,l,r} \le x\}\big) \\ &\quad \le \frac{2}{N_b}\sum_{r=l+1}^{N_b}\sum_{j=1}^{l}\sum_{j'=r}^{r+l-1} \gamma_X(j' - j) \\ &\quad \le \frac{2l^2}{N_b v_l^2}\sum_{r=l+1}^{N_b} \gamma_X(r-l) \sim C\frac{2l^2}{N_b v_l^2} N_b^{2d} \sim C\frac{l^{1-2d}}{N_b^{1-2d}} = C\left(\frac{l}{n-l+1}\right)^{1-2d} \to 0 \end{aligned}$$

as $l, n \to \infty$ such that $l/n \to 0$.

The arguments for $m > 1$ are analogous, but covariances between Hermite polynomials of higher order have to be considered. □

We conclude that the empirical distribution $F^*_{T_n}(x)$ is a consistent estimator of the limiting distribution $F_Z(x)$ so that the SW bootstrap is a valid procedure under Gaussian subordination with arbitrary Hermite rank. This is in contrast to the MBB bootstrap which is valid for Hermite rank one only. Since the SW approach preserves non-Gaussianity, one may also hope that it will provide better finite sample approximations even in the case of a Gaussian limit. Some examples in the next section illustrate this conjecture.

Remark 10.1 This theorem is adapted from Hall et al. (1998); see also Lahiri (2003, Theorem 10.4). We note that the authors consider a general form of $\gamma_X(k)$ with a possible slowly varying function. It requires slightly modified assumptions on the length l of the blocks. Furthermore, Theorem 2.4 in Hall et al. (1998) implies that it is enough to prove (10.12) for a fixed x.

Remark 10.2 The proof above also works for weakly dependent random variables (informally, when $d = 0$), and under Gaussian subordination with $0 < d_X < \frac{1}{2}(1 - m^{-1})$.

So far, we assumed that the standardization sequence v_n is known. In practice, this is, of course, not the case because $v_n = n^{(1-m(\frac{1}{2}-d_X))} L_S^{1/2}$ depends on the long-memory parameter d_X and the constant $L_\gamma(n) \equiv c_\gamma$. There are at least two possible solutions to this problem. The first one is to estimate the parameters d_X and c_γ directly by fitting a parametric or semiparametric model (see Sects. 5.5, 5.6, 5.7, 5.8 and 5.9). The standardization v_n is then replaced by $\hat{v}_n = n^{\hat{d}_X + \frac{1}{2}} \hat{L}_S^{1/2}$. Note, however, that in general the true Hermite rank m is not known. Nevertheless, if m is larger than one, then the exponent of n can also be estimated by the same methods. The difference is that we are then not estimating d_X but rather $\tilde{d} = (1 - m(\frac{1}{2} - d_X)) - \frac{1}{2}$. The other solution is to replace v_n by a direct fully nonparametric estimate V_n. Thus, we consider the statistics

$$U_n := \frac{\sum_{j=1}^n Y_j - n\mu}{V_n} = \frac{n(\bar{Y}_n - \mu)}{V_n},$$

and, with the blocks defined as before,

$$U^*_{n,r} := U^*_{n,l,r} := \frac{\sum_{j=r}^{r+l-1} Y_j - l\bar{Y}_n}{V_l} = \frac{S_{n,l,r} - l\bar{Y}_n}{V_l} \quad (r = 1, \ldots, N_b).$$

Note that, compared to the previous parametric or semiparametric estimation of v_n, direct estimators of v_n are more general, but at the same also less efficient, if the model assumptions needed for estimating d and L_γ by parametric or semiparametric

methods hold. A possible, though somewhat arbitrary, choice is, for instance,

$$V_l^2 = V_{n,l}^2 = \frac{E_{n,l,m_1}^4}{E_{n,l,m_2}^2}$$

where

$$E_{n,l,m_i}^2 = \frac{1}{l - m_i + 1} \sum_{j=1}^{l-m_i+1} (S_{n,m_i,j} - m_i \bar{Y}_n)^2$$

and

$$S_{n,m_i,j} = \frac{1}{m_i} \sum_{h=j}^{j+m_i-1} Y_h.$$

The crucial part of this construction is that

$$\frac{V_n^2}{\operatorname{var}(\sum_{j=1}^n Y_j)} = \frac{E_{n,n,m_1}^4}{v_n^2 E_{n,n,m_2}^2} \xrightarrow{\text{p}} 1$$

as $n \to \infty$. Therefore, the limiting distribution of $U_n = nV_n^{-1}(\bar{Y}_n - \mu)$ is the same as that of $T_n = nv_n^{-1}(\bar{Y}_n - \mu)$, namely $F_Z(x) = P(Z_{m,H}(1) \le x)$. We state the following result without proof (see Hall et al. 1998 or Lahiri 2003, Theorem 10.5).

Theorem 10.2 *Assume that X_j ($j \in \mathbb{N}$) is a stationary sequence of standard normal random variables, such that $\gamma_X(k) \sim L_\gamma k^{2d-1}$, $d \in (0, 1/2)$. Let*

$$F_{U_n}^*(x) = \frac{1}{N} \sum_{r=1}^{N} 1\{U_{n,l,r}^* \le x\}$$

and $F_{U_n}(x) = P(U_n \le x)$. If $l, n \to \infty$ such that $l/n \to 0$, then, as $n \to \infty$,

$$V_n^2 / \operatorname{var}\left(\sum_{j=1}^n Y_j\right) = V_n^2 / v_n \xrightarrow{\text{p}} 1$$

and

$$\sup_{x \in \mathbb{R}} \left| F_{U_n}^*(x) - F_{U_n}(x) \right| \xrightarrow{\text{p}} 0. \tag{10.13}$$

Combining Theorems 10.1 and 10.2 implies that the empirical distribution function $F_{U_n}^*(x)$ approximates $F_{U_n}(x)$ which in turn approximates $F_Z(x) = \lim_{n\to\infty} P(T_n \le x)$. Thus, validity of the SW bootstrap based on U_n is also established.

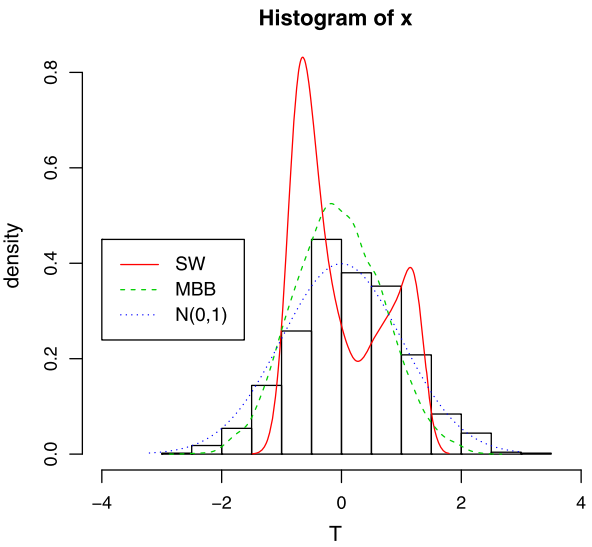

Fig. 10.1 Histogram of a simulated series $Y_t = G(X_t)$ of length $n = 1000$, where X_t is a FARIMA$(0, 0.4, 0)$ process with variance one and $G(x) = x + 0.005(x^3 - x)$. Also plotted are distributions obtained by blockwise bootstrap with block length $l = 177$, and by an analogous SW bootstrap

10.6 Some Practical Issues

The main practical problem with the bootstrap procedures above is that it is not clear how to choose the tuning parameters for an observed data set with a finite number of observations and unknown data generating process. For both bootstrap procedures, the block length is to be chosen such that l tends to infinity at a slower rate than n. Even if we restrict attention to block lengths proportional to $n^{1-\varepsilon}$ for some $0 < \varepsilon < 1$, one needs to specify ε and the proportionality constant. For the block bootstrap, there is an additional tuning parameter k.

As a general rule, the block length should be neither too small nor too large, compared to n. If l is very small, then the computed statistics fail to capture the asymptotic effect of long-range dependence. On the other hand, if l is too large, then the number of blocks to choose from is small so that there is not enough variability among the (highly dependent) block statistics, and the results may heavily depend on spurious features of the observed series. The latter problem is more likely to occur for the SW bootstrap because there the whole shape of the sample path plays a role. This is illustrated in Figs. 10.1, 10.2 and 10.3. The figures are based on a simulated series of the process $Y_t = G(X_t)$ where X_t is a FARIMA$(0, 0.4, 0)$ process with variance one and $G(x) = x + 0.005(x^3 - x)$. Since the Hermite rank of G is one, both bootstrap procedures are valid. Given the dominant linear part and the relatively large sample size of $n = 1000$, one would expect a good approximation by any reasonable bootstrap method. In Fig. 10.1, l is chosen to be equal to $n^{1-\varepsilon}$ with $\varepsilon = \frac{1}{4}$ so that $l = 177$. While the block bootstrap and even the asymptotic standard normal approximation are close to the simulated histogram, the SW bootstrap yields a completely wrong bimodal distribution. The reason for the bimodal shape can be seen in Figs. 10.2(a)–(d). Due to strong long memory (with $d = 0.4$), the simulated sample path stays below zero for a relatively long time in the beginning and towards

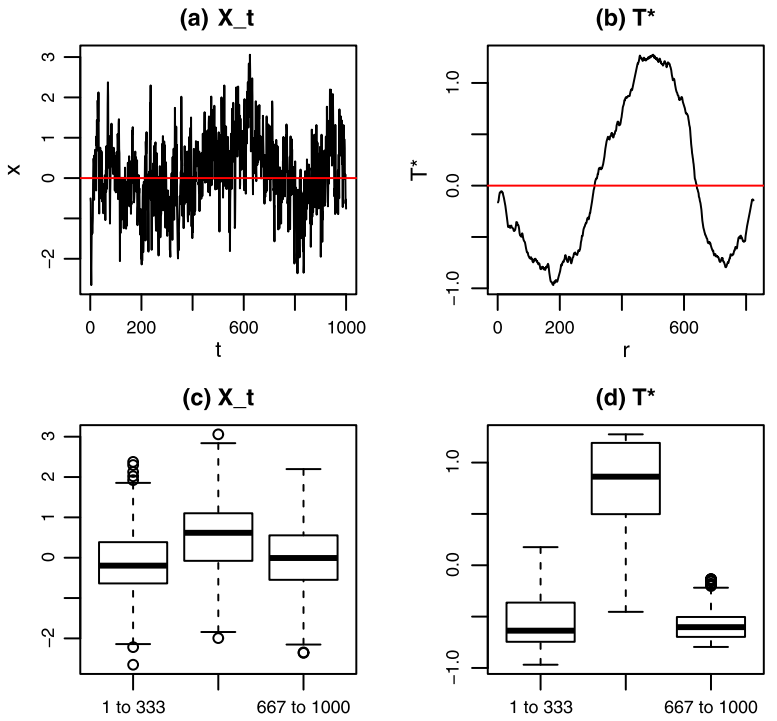

Fig. 10.2 Same simulated series as for the histogram in Fig. 10.1 (**a**), together with values of $T^*_{n,r}$ for blocks moving from left to right (**b**), and boxplots of X_t and $T^*_{n,r}$ for three different regions ((**c**) and (**d**))

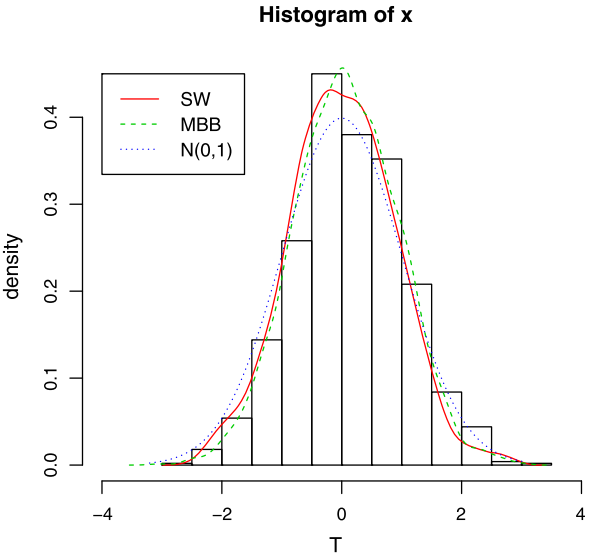

Fig. 10.3 Histogram of a simulated series $Y_t = G(X_t)$ of length $n = 1000$, where X_t is a FARIMA$(0, 0.4, 0)$ process with variance one and $G(x) = x + 0.005(x^3 - x)$. Also plotted are distributions obtained by blockwise bootstrap with block length $l = 5$, and by an analogous SW bootstrap

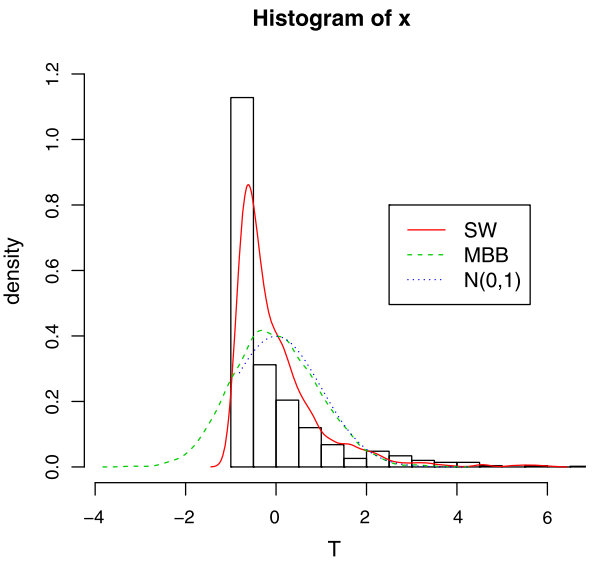

Fig. 10.4 Histogram of a simulated series $Y_t = G(X_t)$ of length $n = 1000$, where X_t is a FARIMA(0, 0.4, 0) process with variance one and $G(X_t) = H_2(X_t) = X_t^2 - 1$. Also plotted are distributions obtained by blockwise bootstrap with block length $l = 5$, and by an analogous SW bootstrap

the end whereas it is above zero most of the time in the middle period. As a result, conditionally on the observed sample path, block sums and hence the values of $T^*_{n,r}$ exhibit a bimodal distribution (Figs. 10.1 and 10.2(b), (d)). In contrast, for the block bootstrap the long wave in the observed series does not influence the result because blocks are resampled randomly. The dependence of the SW bootstrap on spurious features can be alleviated by choosing a smaller block length. This illustrated Fig. 10.3 where $\varepsilon = \frac{3}{4}$ and hence $l = 5$ was used.

Figure 10.4 shows an example where only the SW bootstrap is a valid resampling procedure. The simulated series is $Y_t = G(X_t)$ with $G(X_t) = H_2(X_t) = X_t^2 - 1$. Since the Hermite rank is two, the asymptotic distribution is given by the marginal of the Hermite–Rosenblatt process. This distribution is skewed to the right. The simulated histogram of T_n with $n = 1000$ is indeed highly skewed. In contrast, the distribution obtained by the MBB is symmetric and very close to the standard normal density. The SW bootstrap provides a much better approximation with a skewed shape. As before, however, the concrete choice of the block length is crucial. The good approximation in Fig. 10.4 with $l = 5$ ($\varepsilon = \frac{1}{4}$) is in sharp contrast to the disastrous result in Fig. 10.5 with $l = 177$ ($\varepsilon = \frac{3}{4}$).

Generally, one may conclude that the SW method is quite flexible since it is able to capture non-Gaussian limits. This is very useful even for large sample sizes because the distribution of Hermite processes is rather complicated except for Hermite rank one. On the other hand, the flexibility of the SW method comes at a price. Since almost the complete dependence structure of the observed series is preserved, results may heavily depend on the particular sample path. This lack of 'robustness' can lead to artefacts. A good choice of the block length l plays an important role. On the one hand, l needs to be large enough to come as close as possible to the situation with n observations. On the other hand, if l is too large, then some spurious properties

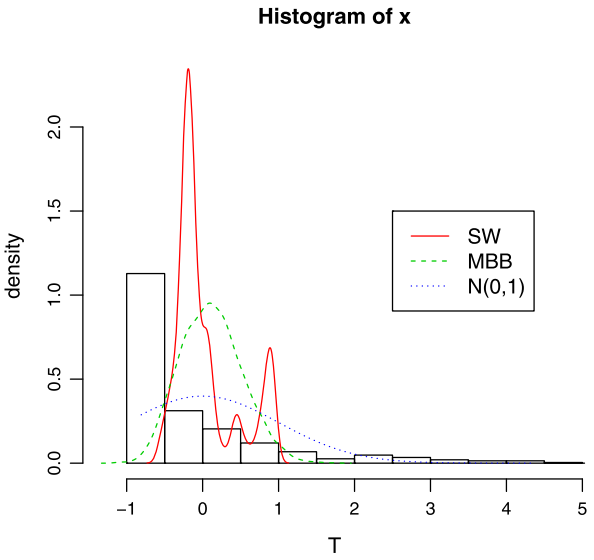

Fig. 10.5 Histogram of a simulated series $Y_t = G(X_t)$ of length $n = 1000$, where X_t is a FARIMA$(0, 0.4, 0)$ process with variance one and $G(X_t) = H_2(X_t) = X_t^2 - 1$. Also plotted are distributions obtained by blockwise bootstrap with block length $l = 177$, and by an analogous SW bootstrap

of the observed sample path may have an undue influence on the result (see, e.g. Fig. 10.1). Thus, as is so often in nonparametric statistics, a suitable balance has to be achieved between two conflicting aims.

10.7 More Complex Models

10.7.1 Bootstrap for the Heavy-Tailed SV Model

10.7.1.1 The HTLM Model

We consider a stochastic volatility model $X_t = \xi_t \sigma_t$, where the random variables ξ_t are i.i.d., strictly positive and regularly varying with index $-\alpha$, $\alpha \in (1, 2)$, that is,

$$P(\xi_1 > x) \sim Ax^{-\alpha}.$$

The sequence $\sigma_t = \exp(\zeta_t)$ is stationary and ergodic, and independent of the sequence ξ_t. Furthermore, ζ_t is a Gaussian long-memory process with parameter d. Suppose that $E[\xi_1] \neq 0$. We saw in Example 4.17 that, if $1/2 + d < 1/\alpha$, then

$$n^{-1/\alpha} S_n(u) \Rightarrow A^{1/\alpha} C_\alpha^{-1/\alpha} \left(E\left[\sigma_1^\alpha\right]\right)^{1/\alpha} \tilde{Z}_\alpha(u), \tag{10.14}$$

where $\tilde{Z}_\alpha(\cdot)$ is an α-stable Lévy process such that $\tilde{Z}_\alpha(1) \stackrel{\mathrm{d}}{=} S_\alpha(1, 1, 0)$. On the other hand, if $1/2 + d > 1/\alpha$, then

$$n^{-(1/2+d)} L_1^{-1/2}(n) S_n(u) \Rightarrow J(1) E[\xi_1] B_H(u)\,, \tag{10.15}$$

where $B_H(\cdot)$ is a fractional Brownian motion, $H = d + \frac{1}{2}$, $L_1(n) = C_1 L_\gamma(n)$ and $J(1) = E(\zeta_1 \exp(\zeta_1))$. In Example 4.17, we called this model LMSD. A very similar model was considered in McElroy and Politis (2007). There, $X_t = \xi_t \sigma_t$ with $\sigma_t = \sigma(\zeta_t)$. The function $\sigma(\cdot)$ is supposed to have Hermite rank 1, and furthermore $E[\sigma_1] = 0$. For this model, we have the same dichotomy as in (10.14)–(10.15), only the constants of the limiting distributions change. McElroy and Politis coined the term "HTLM (Heavy Tailed with Long Memory)".

10.7.1.2 Subsampling for the HTLM Model

We consider $Y_t = \mu + X_t$, where X_t ($t \in \mathbb{N}$) is the HTLM model described above, with $E[\xi] \neq 0$ (but $E[X_t] = 0$ since the subordinated Gaussian sequence σ_t is centred). We noted above that the limiting distribution F is either stable or normal. Furthermore, the scaling v_n is the maximum of $n^{1/\alpha}$ and $n^{d+1/2}L(n)$, where $L(n)$ is a slowly varying function.

Recall the self-normalized statistics W_n from (10.6). Since our data are dependent, we have to change the self-normalizer. It can be constructed as

$$V_n^2 = \sum_{j=1}^{n} (Y_j - \bar{Y}_n)^2 + nLM_n(\rho),$$

where

$$LM_n(\rho) = \left| \sum_{|k|=1}^{[n^\rho]} \frac{1}{n - |k|} \sum_{j=1}^{n-k} \left(Y_j Y_{j+k} - \bar{Y}_n^2 \right) \right|^{1/\rho}, \quad \rho \in (0, 1).$$

To get an idea about the behaviour of V_n^2, we note that Y_j^2 ($j \in \mathbb{N}$) are regularly varying with index $-\alpha/2$ and thus they have an infinite mean. This implies that the behaviour of Y_j^2 is free of long memory. In particular, $\sum_{j=1}^n Y_j^2$ grows at rate $n^{2/\alpha}$, $n^{-2/\alpha} \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$ converges to a stable random variable and $n^{-(2d+1)} L^{-2}(n) \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$ converges in probability to 0. As for $LM_n(\rho)$, we recognize $(n - |h|)^{-1} \sum_{j=1}^{n-h} (Y_j Y_{j+k} - \bar{Y}_n^2)$ as the sample covariance at lag k associated with the sequence Y_j ($j \in \mathbb{N}$) which is the same as the sample covariance of X_j ($j \in \mathbb{N}$). We expect that they converge in probability to $\gamma_X(k) = E^2[\xi_0] E[\sigma_0 \sigma_k]$. If we assume that $\gamma_X(k) \sim L_\gamma k^{2d-1}$, $d \in (0, 1/2)$, then we expect $LM_n(\rho)$ to grow at the rate

$$C \left| \sum_{|k|=1}^{[n^\rho]} k^{2d-1} \right|^{1/\rho} \approx C n^{2d},$$

since the Hermite rank is one. Thus, with $v_n := \max\{n^{1/\alpha}, n^{d+1/2}L(n)\}$ we may conclude that

$$W_n := \frac{n(\bar{Y}_n - \mu)}{V_n} = \frac{v_n^{-1}\sum_{j=1}^n (Y_j - \mu)}{v_n^{-1} V_n}$$

converges to a non-degenerate random variable.

Now, using the blocks $B_r = (Y_r, \ldots, Y_{r+l-1})$, $r = 1, \ldots, N_b$, we construct replicates of W_n as

$$W^*_{n,l,r} = l\frac{\bar{Y}_{n,l,r} - \bar{Y}_n}{V_{n,l,r}}, \quad r = 1, \ldots, N_b,$$

where

$$\bar{Y}_{n,l,r} = \frac{1}{l}\sum_{j=r}^{r+l-1} Y_j$$

and

$$LM_{n,l,r}(\rho) = \left| \sum_{|k|=1}^{[l^\rho]} \frac{1}{l-|k|} \sum_{j=r}^{r+l-1-|k|} \left(Y_j Y_{j+k} - \bar{Y}^2_{n,l,r}\right) \right|^{1/\rho}.$$

A $(1-\theta)$-confidence interval can be constructed as

$$[\bar{Y}_n - z_{1-\frac{\theta}{2}} V_n, \bar{Y}_n - z_{\frac{\theta}{2}} V_n],$$

where $z_{\frac{\theta}{2}}$ is the $(1-\theta)$-percentile of the empirical distribution function

$$F^*_n(x) = \frac{1}{n-l+1}\sum_{r=1}^{n-l+1} 1\{W^*_{n,l,r} \le x\}.$$

For details, we refer to McElroy and Politis (2007) and Jach et al. (2012).

10.7.2 Testing for Jumps in a Trend Function

In some situations, the modified MBB approach can be useful for defining test statistics whose distribution under the null hypothesis is asymptotically normal due to the resampling device. For instance, consider a model with a nonparametric trend function given by

$$Y_i = m(t_i) + e_i \tag{10.16}$$

where $m \in L^2[0,1]$ and e_i a Gaussian process with autocovariance function $\gamma(k) \sim L_\gamma |k|^{-\alpha}$ for some $\alpha = 2d - 1 \in (0,1)$. Beran and Shumeyko (2012b) derive an MBB-based test for

$$H_0 : m \in C[0,1]$$

against the alternative H_1 that m has at least one isolated jump. The idea is to use the wavelet estimator

$$\hat{m}(t) = \hat{m}_{\text{low}}(t) + \hat{m}_{\text{high}}(t)$$

given in Sect. 7.5. The low resolution component $\hat{m}_{\text{low}}(t)$ is an optimal estimator of m, if m is continuous whereas the high resolution part $\hat{m}_{\text{high}}(t)$ captures departures from continuity. A natural idea is therefore to test H_0 against H_1 by designing a test statistic that compares two types of residuals, $\hat{e}_i = Y_i - \hat{m}(t) = Y_i - \hat{m}_{\text{low}}(t) - \hat{m}_{\text{high}}(t)$ and $\hat{e}_{i,\text{low}} = Y_i - \hat{m}(t) = Y_i - \hat{m}_{\text{low}}(t)$. This can be done, for instance, as follows. For a given block size l, define block sums

$$\zeta_r = \hat{e}_r + \cdots + \hat{e}_{r+l-1} = \sum_{\hat{e}_j \in B_r} \hat{e}_j$$

and

$$\zeta_{r,\text{low}} = \hat{e}_{r,\text{low}} + \cdots + \hat{e}_{r+l-1,\text{low}} = \sum_{\hat{e}_j \in B_r} \hat{e}_{j,\text{low}}$$

($1 \le r \le N_b = n - l + 1$). Then, k blocks $B_1^*, \ldots, B_k^*$ are sampled independently with replacement and bootstrap samples $\zeta_1^*, \ldots, \zeta_k^*$ and $\zeta_{1,\text{low}}^*, \ldots, \zeta_{k,\text{low}}^*$ are computed. The corresponding bootstrap statistics are

$$T_{kl}^* = k^{-1/2} \sum_{r=1}^{k} \frac{\zeta_r^*}{v_l}, \qquad T_{kl,\text{low}}^* = k^{-1/2} \sum_{r=1}^{k} \frac{\zeta_{r,\text{low}}^*}{v_l}$$

with $v_l = L_\gamma^{1/2} l^{d+\frac{1}{2}}$. Extending the proofs in Lahiri (1993) and Beran and Shumeyko (2012a), the following result can be derived (Beran and Shumeyko 2012b):

Theorem 10.3 *Suppose that $m \in L^2[0,1]$, m' exists except for a finite set $\mathcal{N} \subset [0,1]$ and is piecewise continuous outside of $\mathcal{N}$. Moreover, let*

$$l = O\left(n^{\delta}\right)$$

where

$$\frac{1}{2r+\alpha} < \delta < \frac{2}{2r+\alpha}$$

and define $\tilde{\sigma}^2 = 2\sigma^2(1-\alpha)^{-1}(2-\alpha)^{-1}$ where $\sigma^2 = \text{var}(e_t)$. Then, under $H_0 : m \in C[0,1]$, we have

$$E_*\left(T_{kl,\text{low}}^*\right) = E_*\left(T_{kl}^*\right) + o_p\left(n^{0.5\alpha\delta - \ln n}\right) = o_p(1),$$

$$\text{Var}_*\left(T_{kl,\text{low}}^*\right) = \text{Var}_*\left(T_{kl}^*\right) + o_p\left(n^{\alpha\delta - 2\ln n}\right) = \tilde{\sigma}^2 + o_p(1),$$

$$T_{kl,\text{low}}^* = T_{kl}^* + O_p\left(n^{0.5\alpha\delta - \ln n}\right)$$

and

$$\sup_{x\in\mathbb{R}}\left|P_*\big(T^*_{kl,\mathrm{low}}\le x\big)-\Phi\left(\frac{x}{\tilde\sigma}\right)\right|=o_p(1),$$

$$\sup_{x\in\mathbb{R}}\left|P_*\big(T^*_{kl}\le x\big)-\Phi\left(\frac{x}{\tilde\sigma}\right)\right|=o_p(1),$$

$$\sup_{x\in\mathbb{R}}\big|P_*\big(T^*_{kl}\le x\big)-P_*\big(T^*_{kl,\mathrm{low}}\le x\big)\big|=o_p(1).$$

Thus, under H_0, the two statistics are asymptotically equivalent and converge uniformly in distribution to the $N(0,\tilde\sigma^2)$ distribution. This is no longer the case under H_1:

Theorem 10.4 *Suppose that the same assumptions as in the previous theorem hold except that m has at least one isolated jump. Then the first two moments and the distribution of T^*_{kl} as well as $E_*(T^*_{kl,\mathrm{low}})$ are the same asymptotically as under H_0. However,*

$$\mathrm{Var}_*\big(T^*_{kl,\mathrm{low}}\big)=\tilde\sigma^2+w_n+o_p(1)$$

where

$$w_n=C^*n^\beta$$

with

$$0<\beta=\alpha\delta-\frac{\alpha}{2r+\alpha}<\frac{\alpha}{2r+\alpha}.$$

Moreover,

$$\sup_{x\in\mathbb{R}}\left|P_*\big(T^*_{kl,\mathrm{low}}\le x\big)-\Phi\left(\frac{x}{\sqrt{\tilde\sigma^2+w_n}}\right)\right|=o_p(1),$$

$$\sup_{x\in\mathbb{R}}\left|P_*\big(T^*_{kl}\le x\big)-\Phi\left(\frac{x}{\tilde\sigma}\right)\right|=o_p(1).$$

Note in particular that under H_1 the ratio of the variances $\mathrm{var}(T^*_{kl,\mathrm{low}})/\mathrm{var}(T^*_{kl})$ diverges to infinity. We may therefore test

$$H_0:\mathrm{var}_*\big(T^*_{kl,\mathrm{low}}\big)=\mathrm{var}\big(T^*_{kl}\big)$$

against

$$H_1:\mathrm{var}_*\big(T^*_{kl,\mathrm{low}}\big)>\mathrm{var}\big(T^*_{kl}\big).$$

Repeating the bootstrap procedure described so far, say N_T times, we calculate

$$W_{\mathrm{low}}=\tilde\sigma^{-2}\sum_{i=1}^{N_T}\big(T^{*(i)}_{kl,\mathrm{low}}-\bar T^*_{kl,\mathrm{low}}\big)^2$$

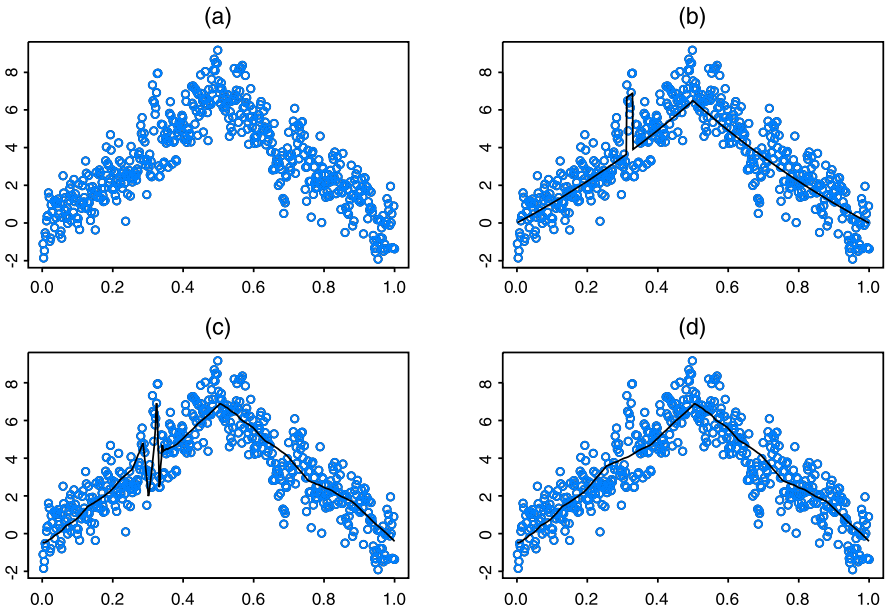

Fig. 10.6 (**a**) Simulated series $Y_i = m(t_i) + e_i$ with e_i generated by a FARIMA process and (**b**) a trend function with a local jump. The wavelet estimate of $m(t_i)$ is shown in (**c**), a kernel estimate in (**d**). The bootstrap based test (using the trend estimate in (**c**)) detects the jump at the 5 %-level of significance

and reject H_0, if W_{low} is too large. Conditionally on the sample, the simulated statistics $T^{*(i)}_{kl,\text{low}}$ $(i = 1, 2, \ldots, N_T)$ are independent. Moreover, under H_0 they are asymptotically $N(0, \tilde{\sigma}^2)$-distributed so that W_{low} is approximately $\chi^2_{N_T-1}$-distributed. Approximate critical values for W_{low} are therefore given by corresponding quantiles of the $\chi^2_{N_T-1}$-distribution. To obtain more exact finite sample quantiles, one can instead simulate the distribution of

$$W = \tilde{\sigma}^{-2} \sum_{i=1}^{N_T} \left(T^{*(i)}_{kl} - \bar{T}^{*}_{kl}\right)^2$$

via resampling. This approach is adopted in Beran and Shumeyko (2012b).

Figure 10.6 shows a typical example where the wavelet decomposition and the test based on W_{low} enables us to detect a very local discontinuity in the trend function. In spite of the presence of local spurious trends caused by strong long memory in the residuals, the local disturbance in the trend function (Fig. 10.6(b)) is captured by the high resolution component (Fig. 10.6(c)). This is in contrast to other nonparametric regression methods such as kernel or local polynomial regression (Fig. 10.6(d)).

Appendix A
Function Spaces

A.1 Convergence of Functions and Basic Definitions

In what follows, E denotes a measurable space.

- A sequence of functions $f_n : E \to \mathbb{R}$ converges pointwise to a function f if

$$\lim_{n\to\infty} f_n(x) = f(x) \tag{A.1}$$

  for each $x \in E$.
- A sequence of functions $f_n : E \to \mathbb{R}$ converges uniformly on $A \subseteq \mathbb{R}$ if

$$\sup_{x\in A} \left| f_n(x) - f(x) \right| \to 0. \tag{A.2}$$

- We have *local* uniform convergence if (A.2) holds for any compact interval A.
- For two functions, $g(x) \sim h(x)$ $(x \to x_0)$ means that $g(x)/h(x)$ converges to one as x tends to x_0.

Definition A.1 A function $f : \mathbb{R} \to \mathbb{R}$ has bounded variation if

$$\sup \sum \left| f(x_i) - f(x_{i-1}) \right| < \infty,$$

where the supremum is taken over all possible partitions of $\mathbb{R}$.

A.2 L Spaces

Throughout the book we use several function spaces:

- Let (E, ν) be a measurable space. Then $L^p(E, \nu)$ denotes the space of functions $f : E \to \mathbb{R}$ such that

$$\int_E \left| f(x) \right|^p d\nu(x) < \infty.$$

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

In particular, $L^2(\mathbb{R}, \text{Leb})$ is the space of functions $f : \mathbb{R} \to \mathbb{R}$ that are square integrable with respect to the Lebesgue measure.

- $L^2(\Omega)$ is the space of random variables X (or, equivalently, the space of distribution functions F) such

$$\|X\|_2^2 = E\big(X^2\big) = \int x^2 \, dF(x) < \infty.$$

A.3 The Spaces C and D

We denote by $C[0, M]$ the space of continuous real-valued functions $x : [0, M] \to \mathbb{R}$. The uniform metric on $C[0, M]$ is defined by

$$d_M\big(x(\cdot), y(\cdot)\big) = \sup_{t \in [0,M]} \big|x(t) - y(t)\big| =: \|x - y\|_M.$$

The uniform metric on $C[0, \infty)$ is defined by

$$d_\infty\big(x(\cdot), y(\cdot)\big) = \sum_{M=1}^{\infty} \frac{\min\{d_M(x(\cdot), y(\cdot)), 1\}}{2^M}.$$

Two functions are close in the uniform topology if their graphs are close.

We denote by $D[0, 1]$ the space of real-valued functions that are right-continuous functions on $[0, 1)$, with finite left limits on $(0, 1]$. Let

$$\Lambda = \big\{\lambda : [0, 1] \to [0, 1] : \lambda(0) = 0, \lambda(1) = 1; \lambda \text{ continuous, strictly increasing}\big\}.$$

The Skorokhod J_1 metric is defined by

$$d\big(x(\cdot), y(\cdot)\big) = \inf_{\lambda \in \Lambda} \max\big(\|\lambda - I\|_1, \|x - y \circ \lambda\|_1\big),$$

where I is the identity function. In other words, two functions are close in the Skorokhod topology if there exists a strictly increasing transformation mapping one into another. A typical example is given by

$$x_n(u) = 1\left\{0 \le u \le \frac{1}{2} + \frac{1}{n}\right\}, \qquad x(u) = 1\left\{0 \le u < \frac{1}{2}\right\}(u).$$

These two functions are close to each other in the J_1 topology, however, the uniform distance is 1.

We refer to Whitt (2002) for more details on different Skorokhod metrics.

Appendix B
Regularly Varying Functions

In this section we collect several results on regularly varying and slowly varying functions that are used in the book. The main references for this material are Bingham et al. (1987) and Resnick (2007).

Definition B.1 A measurable function $L : (c, \infty) \to \mathbb{R}$ $(c \geq 0)$ is called slowly varying at infinity in Karamata's sense if it is positive for x large enough and such that for any $u > 0$,

$$L(ux) \sim L(x) \quad (x \to \infty).$$

The function is called slowly varying at infinity in Zygmund's sense if for x large enough, it is positive and for any $\delta > 0$, there exists a finite number $x_0(\delta) > 0$ such that for $x > x_0(\delta)$, both functions $p_1(x) = x^{\delta} L(x)$ and $p_2(x) = x^{-\delta} L(x)$ are monotone.

Similarly, L is called slowly varying at the origin if $\tilde{L}(x) = L(x^{-1})$ is slowly varying at infinity.

Definition B.2 A measurable function $g : \mathbb{R}_+ \to \mathbb{R}$ is called regularly varying (at infinity) with exponent α if $g(x) \neq 0$ for large x and for any $u > 0$,

$$\lim_{x \to \infty} \frac{g(ux)}{g(x)} = u^{\alpha}.$$

The class of such functions is denoted by $Re(\alpha)$.

Similarly, a function g is called regularly varying at the origin with exponent α if $\tilde{g}(x) = g(x^{-1}) \in Re(-\alpha)$. We will denote this class by $Re_0(\alpha)$.

Lemma B.1 (Karamata Theorem) *Let $g \in Re(\alpha)$ with $\alpha > -1$ and integrable on $(0, a)$ for any $a > 0$. Then $\int_0^x g(t)\, dt \in Re(\alpha + 1)$ and*

$$\int_0^x g(t)\, dt \sim \frac{x g(x)}{\alpha + 1} \quad (x \to \infty).$$

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

Lemma B.2 *Let $g \in Re(\alpha)$ with $\alpha < -1$ and integrable on (a, b) for any $0 < a \leq b < \infty$. Then $\int_x^\infty g(t)dt \in Re(\alpha + 1)$ and*

$$\int_x^\infty g(t)\, dt \sim -\frac{x g(x)}{\alpha + 1} \quad (x \to \infty).$$

Lemma B.3 (Potter's Bound) *Assume that $g(t)$ is regularly varying with index ρ. Let $\delta > 0$. Then there exists t_0 such that for all $x > 0$ and $t > t_0$,*

$$\frac{g(tx)}{g(t)} < (1 + \delta)\big(\max\{x, 1\}\big)^{\rho+\delta}.$$

Appendix C
Vague Convergence

We collect some notions and results on vague convergence. This concept is used to prove convergence of point processes. For details, the reader is referred to standard literature, such as Kallenberg (1997) and Resnick (2007).

Let ν be a measure on E. It is called a Radon measure if $\nu(K) < \infty$ for all relatively compact sets $K \subseteq E$. If $E = \mathbb{R}^m$, then K is called relatively compact if it is bounded away from 0. For example, if $m = 1$, then $K \subset (0, \infty)$ or $K \subset (-\infty, 0)$. If $m = 2$, relative compactness means that K does not contain $(0, 0)$. We denote by $M_+(E)$ the set of all nonnegative Radon measures on E.

The simplest example of a Radon measure is the Dirac measure $\delta_x(\cdot)$: $\delta_x(K) = 1$ if $x \in K$ and 0 otherwise. Furthermore, a point measure $\sum_i \delta_{x_i}$ $(x_i \in E)$ is a nonnegative Radon measure. The set $M_p(E)$ consists of all Radon point measures of the form $\sum_i \delta_{x_i}$.

Let $C_K^+(E)$ be a set of all continuous functions $f : E \to \mathbb{R}_+$ with compact support. We say that a sequence ν_n of measures converges vaguely to ν, denoted by $\nu_n \xrightarrow{v} \nu$, if

$$\int f(x)\nu_n(dx) \to \int f(x)\nu(dx)$$

for all $f \in C_K^+(E)$. There is a close link between regular variation and vague convergence. Assume that the distribution F of a nonnegative random variable X is regularly varying. Let a_n be a sequence of constants such that $n\bar{F}(a_n x) \to x^{-\alpha}$ as $n \to \infty$. Define $\nu_n(K) = nP(a_n^{-1}X \in K)$. Then $\nu_n \xrightarrow{v} \nu$, where $\nu(x, \infty] = x^{-\alpha}$ in $M_+[0, \infty)$.

A sequence N_n of point processes converges weakly in $M_p(E)$ to N if for all sets $A_1, \ldots, A_m \subseteq E$ and all integers $n_1, \ldots, n_m$, we have

$$P\big(N_n(A_1) = n_1, \ldots, N_n(A_m) = n_m\big) \to P\big(N(A_1) = n_1, \ldots, N(A_m) = n_m\big)$$

as $n \to \infty$.

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

Appendix D
Some Useful Integrals

We collect several formulas for integrals that are used extensively in the book. We start with definitions of beta and gamma functions.

Definition D.1 (Gamma Function)

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx. \tag{D.1}$$

Definition D.2 (Beta Function)

$$B(a,b) = \int_0^\infty x^{a-1}(1+x)^{-(a+b)}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

$$\int_0^\infty x^{-1} \sin x\, dx = \pi/2, \tag{D.2}$$

$$\int_0^\infty x^{-\alpha} \sin x\, dx = \frac{\Gamma(2-\alpha)\cos(\pi\alpha/2)}{1-\alpha} \quad (\alpha \neq 1), \tag{D.3}$$

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin \pi z}, \tag{D.4}$$

$$\int_0^\infty \left(x + x^2\right)^{H_0 - \frac{3}{2}} dx = B\left(H_0 - \frac{1}{2}, 2 - 2H_0\right) = \frac{\Gamma(H_0 - \frac{1}{2})\Gamma(2-2H_0)}{\Gamma(\frac{3}{2} - H_0)}$$

$$= \frac{\sin \pi(H_0 - \frac{1}{2})}{\pi} \Gamma(2-2H_0). \tag{D.5}$$

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

Glossary

Notation

$\overset{\mathrm{d}}{=}$	Equality in distribution
$\overset{\mathrm{d}}{\to}$	Convergence in distribution
$\Rightarrow$	Weak convergence
$\overset{\mathrm{p}}{\to}$	Convergence in probability
$\overset{\mathrm{f.d.}}{\to}$	Finite-dimensional convergence
$B(\cdot)$	Brownian Motion
$B_H(\cdot)$	Fractional Brownian Motion (fBm) with Hurst parameter H
$Z_{m,H}$	Hermite–Rosenblatt process
$S\alpha S$	Symmetric α-Stable random variable
$Z_\alpha(\cdot)$	Stable Lévy process
$\tilde{Z}_{H,\alpha}(\cdot)$	Linear Fractional Stable Motion (LFSM)

Abbreviations

ARMA	Autoregressive Moving Average
FARIMA (ARFIMA)	Fractionally Integrated ARMA
fBm	Fractional Brownian Motion
fGn	Fractional Gaussian Noise
(G)ARCH	(Generalized) Autoregressive Conditionally Heteroscedastic
I(G)ARCH	Integrated GARCH
EGARCH	Exponential GARCH
FIGARCH	Fractionally Integrated GARCH
FIEGARCH	Fractionally Integrated Exponential GARCH
LARCH	Linear ARCH
SSSI	Self Similar with Stationary Increments
SV	Stochastic Volatility
LMSV	Long-Memory Stochastic Volatility
LMSD	Long-Memory Stochastic Duration

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

References

Abadir, K., & Talmain, G. (2002). Aggregation, persistence and volatility in a macro model. *Review of Economic Studies*, *69*(4), 749–779.

Abadir, K. M., Distaso, W., & Giraitis, L. (2007). Nonstationarity-extended local Whittle estimation. *Journal of Econometrics*, *141*(2), 1353–1384.

Abete, T., de Candia, A., Lairez, D., & Coniglio, A. (2004). Percolation model for enzyme gel degradation. *Physical Review Letters*, *93*, 228301.

Abramovich, F., Sapatinas, T., & Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society*, *60*(4), 725–749.

Abramowitz, M. & Stegun, I. A. (Eds.) (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (p. 773). New York: Dover.

Abry, P., & Flandrin, P. (1994). On the initialization of the discrete wavelet transform. *IEEE Signal Processing Letters*, *SPL-1*(2), 32–34.

Abry, P., & Veitch, D. V. (1998). Wavelet analysis of long-range-dependent traffic. *IEEE Transactions on Information Theory*, *44*(1), 2–15.

Abry, P., Goncalves, P., & Flandrin, P. (1995). Wavelets, spectrum analysis and $1/f$ processes. In A. Antoniadis & G. Oppenheim (Eds.), *Lecture notes in statistics: Vol. 103. Wavelets and statistics*. New York: Springer.

Abry, P., Veitch, D., & Flandrin, P. (1998). Long-range dependence: revisiting aggregation with wavelets. *Journal of Time Series Analysis*, *19*(3), 253–266.

Abry, P., Flandrin, P., Taqqu, M. S., & Veitch, D. (2003). Self-similarity and long-range dependence through the wavelet lens. In P. Doukhan, G. Openheim, & M. S. Taqqu (Eds.), *In, long-range dependence: theory and applications* (pp. 527–556). Boston: Birkhäuser.

Adams, R. A., & Fournier, J. J. F. (2003). *Sobolev spaces*. Amsterdam: Academic Press.

Adenstedt, R. K. (1974). On large-sample estimation for the mean of a stationary random sequence. *The Annals of Statistics*, *2*(6), 1095–1107.

Adler, R. J. (1981). *The geometry of random fields*. New York: Wiley.

Aharoni, A., & Feder, J. (1990). *Fractals in physics*. Amsterdam: North-Holland.

Ahmad, Z., Bhartia, P. K., & Krotkov, N. (2004). Spectral properties of backscattered UV radiation in cloudy atmospheres. *Journal of Geophysical Research*, *109*, D01201. doi:10.1029/2003JD003395.

Akaike, H. (1969). Power spectrum estimation through autoregressive model fitting. *Annals of the Institute of Statistical Mathematics*, *21*, 407–419.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov (Ed.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

Akonom, J., & Gourieroux, C. (1987). A functional central limit theorem for fractional processes. Preprint, CEREMAP, Paris.

Allan, D. W. (1966). Statistics of atomic frequency clocks. *Proceedings of the IEEE*, *54*, 221–230.

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, *85*, 749–759.

Ammann, B., Birks, H. J. B., Brooks, S. J., Eicher, U., von Grafenstein, U., Hofmann, W., Lemdahl, G., Schwander, J., Tobolski, K., & Wick, L. (2000). Quantification of biotic responses to rapid climatic changes around the Younger Dryas—a synthesis. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *159*, 313–347.

Andersen, T. G., & Bollerslev, T. (1997a). Heterogeneous information arrivals and return volatility dynamics: uncovering the long run in high frequency returns. *Journal of Finance*, *52*, 975–1005.

Andersen, T. G., & Bollerslev, T. (1997b). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, *4*, 115–158.

Anderson, Ch. A. (1967). Some properties of Appell-like polynomials. *Journal of Mathematical Analysis and Applications*, *19*, 475–491.

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley. xiv+704 pp.

Andrews, D. W. K., & Guggenberger, P. (2003). A bias-reduced log-periodogram regression estimator for the long memory parameter. *Econometrica*, *71*, 675–712.

Andrews, D., & Lieberman, O. (2005). Valid Edgeworth expansions for the Whittle maximum likelihood estimator for stationary long-memory Gaussian time series. *Econometric Theory*, *21*, 710–734.

Andrews, D. W. K., & Sun, Y. (2004). Adaptive local polynomial Whittle estimation of long-range dependence. *Econometrica*, *72*(2), 569–614.

Andrews, W. K., Lieberman, O., & Marmer, V. (2006). Higher-order improvements of the parametric bootstrap for long-memory Gaussian processes. *Journal of Econometrics*, *133*, 673–702.

Aneiros-Pérez, G., Gonzáles-Manteiga, W., & Vieu, P. (2004). Estimation and testing in a partial linear regression model under long-memory dependence. *Bernoulli*, *10*(1), 49–78.

Angulo, J. M., Ruiz-Medina, M. D., & Anh, V. V. (2000). Estimation and filtering of fractional generalised random fields. *Journal of the Australian Mathematical Society A*, *69*, 1–26.

Anh, V. V., Angulo, J. M., & Ruiz-Medina, M. D. (1999). Possible long-range dependence in fractional random fields. *Journal of Statistical Planning and Inference*, *80*(1–2), 95–110.

Anh, V. V., Leonenko, N. N., & Shieh, N.-R. (2009). Multifractal products of stationary diffusion processes. *Stochastic Analysis and Applications*, *27*(3), 475–499.

Annis, A. A., & Lloyd, E. H. (1976). The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika*, *63*, 111–116.

Antoch, J., Huskova, M., & Veraverbeke, N. (1995). Change-point problem and bootstrap. *Journal of Nonparametric Statistics*, *5*, 123–144.

Antoniadis, A. & Oppenheim, G. (Eds.) (1995). *Wavelets and statistics*. *Lecture notes in statistics* (Vol. 103). Heidelberg: Springer.

Appell, P. (1880). Sur une classe de polynômes. *Annales Scientifiques de l'École Normale Supérieure*, *2e série, 9*, 119–144.

Appell, P. (1881). Sur des polynômes de deux variables analogues aux polynômes de Jacobi. *Archiv der Mathematik und Physik*, *66*, 238–245.

Arcones, M. A. (1994). Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *Annals of Probability*, *22*(4), 2242–2274.

Arteche, J. (2004). Gaussian semiparametric estimation in long memory in stochastic volatility and signal plus noise models. *Journal of Econometrics*, *119*(1), 131–154.

Arteche, J. (2006). Semiparametric estimation in perturbed long memory series. *Computational Statistics & Data Analysis*, *51*, 2118–2141.

Arteche, J., & Robinson, P. M. (2000). Semiparametric inference in seasonal and cyclical long memory processes. *Journal of Time Series Analysis*, *21*, 1–25.

Arteche, J., & Velasco, C. (2005). Trimming and tapering semi-parametric estimates in asymmetric long memory time series. *Journal of Time Series Analysis*, *26*(4), 581–611.

Astrauskas, A. (1983). Limit theorems for forms of linear processes. *Litovskij Matematičeskij Sbornik*, *23*(4), 3–11.

Astrauskas, A., Levy, J., & Taqqu, M. S. (1991). The asymptotic dependence structure of the linear fractional Lévy motion. *Lietuvos Matematikos Rinkinys (Lithuanian Mathematical Journal)*, *31*, 1–28.

Avarucci, M., & Velasco, C. (2009). A Wald test for the cointegration rank in nonstationary fractional systems. *Journal of Econometrics*, *151*(2), 178–189.

Avnir, D. (Ed.) (1989). *The fractal approach to heterogeneous chemistry*. New York: Wiley.

Avram, F. (1988). On bilinear forms in Gaussian random variables and Toeplitz matrices. *Probability Theory and Related Fields*, *79*(1), 37–45.

Avram, F., & Taqqu, M. S. (1986). Weak convergence of moving averages with infinite variance. In E. Eberlein & M. S. Taqqu (Eds.), *Dependence in probability and statistics* (pp. 399–415). Boston: Birkhäuser.

Avram, F., & Taqqu, M. S. (1987). Noncentral limit theorems and Appell polynomials. *Annals of Probability*, *15*(2), 767–775.

Avram, F., & Taqqu, M. S. (1992). Weak convergence of sums of moving averages in the α-stable domain of attraction. *Annals of Probability*, *20*(1), 483–503.

Baek, C., & Pipiras, V. (2011). Statistical tests for a single change in mean against long-range dependence. *Journal of Time Series Analysis*, *33*, 131–151.

Bagshaw, M., & Johnson, R. A. (1975). The effect of serial correlation on the performance of CUSUM tests II. *Technometrics*, *17*, 73–80.

Bai, J. (1998). A note on spurious break. *Econometric Theory*, *14*, 663–669.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, *73*, 5–59.

Baillie, R., & Bollerslev, T. (1994). Cointegration, fractional cointegration, and exchange rate dynamics. *Journal of Finance*, *49*, 737–745.

Baillie, R. T., & Chung, S.-K. (2002). Modeling and forecasting from trend-stationary long memory models with applications to climatology. *International Journal of Forecasting*, *18*(2), 215–226.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996a). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*(1), 3–30.

Baillie, R. T., Chung, C.-F., & Tieslau, M. A. (1996b). Analysing inflation by the fractionally integrated ARFIMA-GARCH model. *Journal of Applied Econometrics*, *11*, 23–40.

Baillie, R. T., Kapetanios, G., & Papailias, F. (2012). Modified information criteria and selection of long memory time series models. Preprint.

Bajšanski, B., & Karamata, J. (1968/1969). *Regularly varying functions and the principle of equicontinuity* (pp. 235–242). Publ. Ramanujan Inst., 1.

Bak, P. (1996). *How nature works: the science of self-organised criticality*. New York: Copernicus Press.

Banach, S. (1948). Kurs funktsionalnogo analiza (A course of functional analysis). Kiev.

Banerjee, A., & Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics*, *129*, 1–34.

Bardet, J. M., Lang, G., Moulines, E., & Soulier, P. (2000). Wavelet estimator of long-range dependent processes. *Statistical Inference for Stochastic Processes*, *3*(1–2), 85–99 (English summary). 19th "Rencontres Franco–Belges de Statisticiens" (Marseille, 1998).

Baringhaus, L., & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Annals of the Institute of Statistical Mathematics*, *43*, 551–564.

Baringhaus, L., & Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statistics & Probability Letters*, *13*(4), 269–274.

Barnard, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society B*, *21*, 239–271.

Barndorff-Nielsen, O. E., & Leonenko, N. N. (2005). Burgers' turbulence problem with linear or quadratic external potential. *Journal of Applied Probability*, *42*(2), 550–565.

Barndorff-Nielsen, O., & Pedersen, B. V. (1979). The bivariate Hermite polynomials up to order six. *Scandinavian Journal of Statistics*, *6*(3), 127–128.

Barndorff-Nielsen, O. E., & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *63*, 167–241.

Barndorff-Nielsen, O. E., & Stelzer, R. (2011a). The multivariate supOU stochastic volatility model. *Mathematical Finance* doi:10.1111/j.1467-9965.2011.00494.x.

Barndorff-Nielsen, O. E., & Stelzer, R. (2011b). Multivariate supOU processes. *The Annals of Applied Probability*, *21*(1), 140–182.

Barndorff-Nielsen, O. E., Jensen, J. L., & Sørensen, M. M. (1998). Some stationary processes in discrete and continuous time. *Advances in Applied Probability*, *30*(4), 989–1007.

Barnsley, M. F. (1993). *Fractals everywhere* (2nd ed.). Boston: Academic Press.

Bartkiewicz, K., Jakubowski, A., Mikosch, T., & Wintenberger, O. (2011). Stable limits for sums of dependent infinite variance random variables. *Probability Theory and Related Fields*, *150*(3–4), 337–372 (English summary).

Bartlett, M. S. (1974). The statistical analysis of spatial pattern. *Advances in Applied Probability*, *6*(2), 336–358.

Basrak, B., Davis, R. A., & Mikosch, T. (1999). The sample ACF of a simple bilinear process. *Stochastic Processes and Their Applications*, *83*(1), 1–14.

Basrak, B., Davis, R. A., & Mikosch, T. (2002). Regular variation of GARCH processes. *Stochastic Processes and Their Applications*, *99*(1), 95–115.

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*. Englewood Cliffs: Prentice Hall.

Bassingthwaighte, J. B., & Raymond, G. M. (1994). Evaluating rescaled range analysis for time series. *Annals of Biomedical Engineering*, *22*, 432–444.

Batchelor, G. K. (1953). *The theory of homogeneous turbulence*. *Cambridge science classics*.

Bateman, G., & Erdelyi, A. (1974). *Higher transcendental functions: Bessel functions, parabolic-cylinder functions, orthogonal polynomials*. Moscow: Nauka. (In Russian).

Bauwens, L., & Hautsch, N. (2009). Modelling financial high frequency data using point processes. In T. Mikosch, J.-P. Kreiss, R. A. Davis, & T. G. Andersen (Eds.), *Handbook of financial time series* (pp. 953–979). Berlin: Springer.

Bauwens, L., Pohlmeier, W., & Veredas, D. (2008). Editor's introduction: recent developments in high frequency financial econometrics. In L. Bauwens, W. Pohlmeier, & D. Veredas (Eds.), *High frequency financial econometrics*. *Studies in empirical economics* (pp. 1–5).

Becker, K.-H., & Dörfler, M. (1989). *Dynamical systems and fractals*. Cambridge: Cambridge University Press.

Beirlant, J., Bouquiaux, C., & Werker, B. J. M. (2006). Semiparametric lower bounds for tail index estimation. *Journal of Statistical Planning and Inference*, *136*(3), 705–729.

Ben Hariz, S., & Wylie, J. J. (2005). Rates of convergence for the change-point estimator for long-range dependent sequences. *Statistics & Probability Letters*, *73*, 155–164.

Ben Hariz, S., Wylie, J., & Zhang, L. (2007). Optimal rate of convergence for nonparametric change-point estimators for non-stationary sequences. *The Annals of Statistics*, *35*, 1802–1826.

Benassi, A., Cohen, S., & Istas, J. (2002). Identification and properties of real harmonizable fractional Levy motions. *Bernoulli*, *8*(1), 97–115.

Bender, C. (2003a). An Itô formula for generalized functionals of a fractional Brownian motion with arbitrary Hurst parameter. *Stochastic Processes and Their Applications*, *104*(1), 81–106.

Bender, C. (2003b). An S-transform approach to integration with respect to a fractional Brownian motion. *Bernoulli*, *9*(6), 955–983.

Bender, C., Sottinen, T., & Valkeila, E. (2007). Arbitrage with fractional Brownian motion? *Theory of Stochastic Processes*, *13*(29), No. 1–2, 23–34.

Benedetti, J. K. (1977). On the nonparametric regression of regression functions. *Journal of the Royal Statistical Society. Series B*, *39*, 248–253.

Benhenni, K., Hedli-Griche, S., Rachdi, M., & Vieu, P. (2008). Consistency of the regression estimator with functional data under long memory conditions. *Statistics & Probability Letters*, *78*(8), 1043–1049.

Benson, D., Meerschaert, M. M., Bauemer, B., & Scheffler, H. P. (2006). Aquifer operator-scaling and the effect on solute mixing and dispersion. *Water Resources Research*, *42*(W01415), 1–18.

Beran, J. (1986). *Estimation, testing and prediction for self-similar and related processes*. Ph.D. Thesis, ETH Zürich.

Beran, J. (1989). A test of location for data with slowly decaying serial correlations. *Biometrika*, *76*, 261–269.

Beran, J. (1991). M-estimators of location for data with slowly decaying serial correlations. *Journal of the American Statistical Association*, *86*, 704–708.

Beran, J. (1992). A goodness-of-fit test for time series with long range dependence. *Journal of the Royal Statistical Society. Series B*, *54*(3), 749–760.

Beran, J. (1993). Fitting long-memory models by generalized linear regression. *Biometrika*, *80*, 817–822.

Beran, J. (1994a). *Statistics for long-memory processes. Monographs on statistics and applied probability* (Vol. 61). New York: Chapman and Hall/CRC.

Beran, J. (1994b). On a class of M-estimators for long-memory Gaussian models. *Biometrika*, *81*(4), 755–766.

Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short- and long-memory ARIMA models. *Journal of the Royal Statistical Society. Series B*, *57*, 659–672.

Beran, J. (1997). On heavy tail modeling and teletraffic data by S.I. Resnick. Invited discussion. *The Annals of Statistics*, *25*(5), 1852–1855.

Beran, J. (1999). *SEMIFAR Models: A semiparametric framework for modelling trends, long-range dependence, and nonstationarity*. CoFE discussion paper 99/16, University of Konstanz.

Beran, J. (2006). On location estimation for LARCH processes. *Journal of Multivariate Analysis*, *97*(8), 1766–1782.

Beran, J. (2007a). On M-estimation under long-range dependence in volatility. *Journal of Time Series Analysis*, *28*(1), 138–153.

Beran, J. (2007b). Systematic vs. random development, long-range dependence and nonstationarity. In F. Kienast, O. Wildi, & S. Ghosh (Eds.), *Landscape series: Vol. 8. A changing world—challenges for landscape research*. Berlin: Springer.

Beran, J. (2009). On parametric estimation for locally stationary long-memory processes. *Journal of Statistical Planning and Inference*, *139*(3), 900–915.

Beran, J., & Feng, Y. (2001a). Local polynomial estimation with a FARIMA-GARCH error process. *Bernoulli*, *7*(5), 733–750.

Beran, J., & Feng, Y. (2001b). A semiparametric fractional autoregressive model. *Statistical Review (Revista de Estatistica)*, *2*, 125–128.

Beran, J., & Feng, Y. (2002a). SEMIFAR models—a semiparametric framework for modelling trends, long-range dependence and nonstationarity. *Computational Statistics & Data Analysis*, *40*(2), 393–419.

Beran, J., & Feng, Y. (2002b). Iterative plug-in algorithms for SEMIFAR models—definition, convergence, and asymptotic properties. *Journal of Computational and Graphical Statistics*, *11*(3), 690–713.

Beran, J., & Feng, Y. (2002c). Local polynomial fitting with long memory, short memory and antipersistent errors. *Annals of the Institute of Statistical Mathematics*, *54*(2), 291–311.

Beran, J., & Feng, Y. (2007). Weighted averages and local polynomial estimation for fractional linear ARCH processes. *Journal of Statistical Theory and Practice*, *1*(2), 149–166.

Beran, J., & Ghosh, S. (1990). Goodness of t-tests and long-range dependence. In W. Stahel & S. Weisberg (Eds.), *IMA volumes in mathematics and its applications: Vol. 33. Directions in robust statistics and diagnostics, Part I* (pp. 21–33). New York: Springer.

Beran, J., & Ghosh, S. (1991). Slowly decaying correlations, testing normality, nuisance parameters. *Journal of the American Statistical Association*, *86*(415), 785–791.

Beran, J., & Ghosh, S. (1998). Root-n-consistent estimation in partial linear models with long-memory errors. *Scandinavian Journal of Statistics*, *25*, 345–357.
Beran, J., & Ghosh, S. (2000). Estimation of the dominating frequency for stationary and nonstationary fractional autoregressive processes. *Journal of Time Series Analysis*, *21*(5), 513–533.
Beran, J., & Künsch, H. (1985). *Location estimators for processes with long-range dependence* (Research Report No. 40). Seminar für Statistik, ETH, Zurich.
Beran, J., & Ocker, D. (1999). SEMIFAR forecasts, with applications to foreign exchange rates. *Journal of Statistical Planning and Inference*, *80*, 137–153.
Beran, J., & Ocker, D. (2000). *Temporal aggregation of stationary and nonstationary FARIMA*$(p, d, 0)$ *models* (COFE Working Paper). University of Konstanz. http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-5300.
Beran, J., & Ocker, D. (2001). Volatility of stock market indices—an analysis based on SEMIFAR models. *Journal of Business & Economic Statistics*, *19*(1), 103–116.
Beran, J., & Schell, D. (2010). On robust tail estimation. *Computational Statistics and Data Analysis*, *56*(11), 3430–3443.
Beran, J., & Schützner, M. (2008). The effect of long memory in volatility on location estimation. *Sankhya, Series B*, *70*(1), 84–112.
Beran, J., & Schützner, M. (2009). On approximate pseudo maximum likelihood estimation for LARCH-processes. *Bernoulli*, *15*(4), 1057–1081.
Beran, J., & Shumeyko, Y. (2012a). On asymptotically optimal wavelet estimation of trend functions under long-range dependence. *Bernoulli*, *189*(1), 137–176.
Beran, J., & Shumeyko, Y. (2012b). Bootstrap testing for discontinuities under long-range dependence. *Journal of Multivariate Analysis*, *105*(1), 322–347.
Beran, J., & Terrin, N. (1994). Estimation of the long-memory parameter, based on a multivariate central limit theorem. *Journal of Time Series Analysis*, *15*(3), 269–278.
Beran, J., & Terrin, N. (1996). Testing for a change of the long-memory parameter. *Biometrika*, *83*(3), 627–638.
Beran, J., & Weiershäuser, A. (2011). On spline regression under Gaussian subordination with long memory. *Journal of Multivariate Analysis*, *102*(2), 315–335.
Beran, J., Sherman, R., Taqqu, M. S., & Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, *43*(234), 1566–1579.
Beran, J., Bhansali, R. J., & Ocker, D. (1998). On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes. *Biometrika*, *85*(4), 921–934.
Beran, J., Feng, Y., Ghosh, S., & Sibbertsen, P. (2002). On robust local polynomial estimation with long-memory errors. *International Journal of Forecasting*, *18*, 227–241.
Beran, J., Ghosh, S., & Sibbertsen, P. (2003). Nonparametric M-estimation with long-memory errors. *Journal of Statistical Planning and Inference*, *17*, 199–206.
Beran, J., Ghosh, S., & Schell, D. (2009). On least squares estimation for long-memory lattice processes. *Journal of Multivariate Analysis*, *100*(10), 2178–2194.
Beran, J., Schützner, M., & Ghosh, S. (2010). From short to long memory: aggregation and estimation. *Computational Statistics & Data Analysis*, *54*(11), 2432–2442.
Beran, J., Das, B., & Schell, D. (2012). On robust tail index estimation for linear long-memory processes. *Journal of Time Series Analysis*, *33*(3), 406–423.
Beran, J., Weiershäuser, A., Galizia, C. G., Rein, J., Smith, B. H., & Strauch, M. (2013, in press). On piecewise polynomial regression under general dependence conditions, with an application to calcium-imaging data. *Sankhya Series B*.
Berche, B., Henkel, M., & Kenna, R. (2009). Critical phenomena: 150 years since Cagniard de la Tour. arXiv:0905.1886v1.
Berger, N. (2002). Transience, recurrence and critical behavior for long-range percolation. *Communications in Mathematical Physics*, *236*, 531–558.
Berger, D., Chaboud, A., & Hjalmarsson, E. (2009). What drives volatility persistence in the foreign exchange market? *Journal of Financial Economics*, *94*(2), 192–213.
Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics*, *2*, 489–502.

Berkes, I., & Horváth, L. (2003). Asymptotic results for long memory LARCH sequences. *The Annals of Applied Probability*, *13*, 641–668.
Berkes, I., & Horváth, L. (2004). The efficiency of the estimators of the parameters in GARCH processes. *The Annals of Statistics*, *32*, 633–655.
Berkes, I., & Philipp, W. (1977). An almost sure invariance principle for the empirical distribution function of mixing random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *41*, 115–137.
Berkes, I., Horváth, L., & Kokoszka, P. (2003). GARCH processes: structure and estimation. *Bernoulli*, *9*, 201–228.
Berkes, I., Horváth, L., Kokoszka, P., & Shao, Q.-M. (2006). On discriminating between long-range dependence and changes in mean. *The Annals of Statistics*, *34*(3), 1140–1165.
Berman, S. M. (1964). Limit theorems for the maximum term in stationary sequences. *The Annals of Mathematical Statistics*, *35*, 502–516.
Berman, S. M. (1971). Maxima and high level excursions of stationary Gaussian processes. *Transactions of the American Mathematical Society*, *160*, 65–85.
Bertail, P., Doukhan, P., & Soulier, P. (Eds.) (2006). *Lecture notes in statistics: Vol. 187. Dependence in probability and statistics*. New York: Springer.
Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, *36*(2), 192–236.
Besicovitch, A. S. (1928). On the fundamental geometrical properties of linearly measurable plain sets of points. *Mathematische Annalen*, *98*, 422–464.
Besicovitch, A. S. (1929). On linear sets of points of fractional dimension. *Mathematische Annalen*, *101*(1), 161–193.
Besicovitch, A. S., & Ursell, H. D. (1937). Sets of fractional dimensions. *Journal of the London Mathematical Society*, *12*(1), 18–25.
Bhansali, R. J. (1978). Linear prediction by autoregressive model fitting in the time domain. *The Annals of Statistics*, *6*(1), 224–231.
Bhansali, R. J. (1980). Autoregressive and window estimates of the inverse correlation function. *Biometrika*, *67*, 551–566.
Bhansali, R. J. (1986). A derivation of the information criterion for selecting autoregressive models. *Advances in Applied Probability*, *18*, 360–387.
Bhansali, R. J. (1993). Order selection for linear time series models: a review. In T. Subba Rao (Ed.), *Developments in time series analysis* (pp. 55–56). London: Chapman & Hall.
Bhansali, R. J., Giraitis, L., & Kokoszka, P. (1997). Approximations and limit theory for quadratic forms of linear processes. *Stochastic Processes and Their Applications*, *117*(1), 71–95.
Bhansali, R. J., & Kokoszka, P. S. (2003). Prediction of long-memory time series. In Doukhan et al. (Eds.), *Theory and applications of long-range dependence*, Basel: Birkhäuser.
Bhansali, R. J., Giraitis, L., & Kokoszka, P. S. (2006). Estimation of the memory parameter by fitting fractionally differenced autoregressive models. *Journal of Multivariate Analysis*, *97*(10), 2101–2130.
Bhardwaj, G., & Swanson, N. R. (2006). An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series. *Journal of Econometrics*, *131*(1–2), 539–578.
Bhattacharya, R. N., Gupta, V. K., & Waymire, E. (1983). The Hurst effect under trends. *Journal of Applied Probability*, *20*(3), 649–662.
Biagini, F., Hu, Y., Øksendal, B., & Zhang, T. (2008). *Stochastic calculus for fractional Brownian motion and applications*. Berlin: Springer.
Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.
Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1987). *Regular variation*. Cambridge: Cambridge University Press.
Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1989). *Regular variation*. Cambridge: Cambridge University Press.
Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, *17*, 656–660.

Bisaglia, L., & Bordignon, S. (2002). Mean square prediction error for long-memory processes. *Statistical Papers*, *43*, 161–175.

Biskup, M. (2004). On the scaling of the chemical distance in long range percolation models. *Annals of Probability*, *32*, 2938–2977.

Biskup, M. (2011). Graph diameter in long-range percolation. *Random Structures & Algorithms*, *39*(2), 210–227.

Blanke, D. (2004). Local Hölder exponent estimation for multivariate continuous time processes. *Journal of Nonparametric Statistics*, *16*(1–2), 227–244.

Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika*, *60*(2), 217–226.

Boas, R. P. (1954). *Entire functions*. New York: Academic Press.

Boas, R. P., & Buck, R. C. (1964). *Polynomial expansions of analytic functions*. New York: Academic Press.

Boes, D. C., & Salas, J. D. (1978). Nonstationarity of the mean and the Hurst phenomenon. *Water Resourses Research*, *14*, 135–143.

Boissy, Y., Bhattacharyya, B. B., Li, X., & Richardson, G. D. (2005). Parameter estimates for fractional autoregressive spatial processes. *The Annals of Statistics*, *33*(6), 2553–2567.

Bojdecki, T., Gorostiza, L. G., & Talarczyk, A. (2007). A long range dependence stable process and an infinite variance branching system. *Annals of Probability*, *35*(2), 500–527.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Bollerslev, T., & Mikkelsen, H. O. (1996). Modeling and pricing long memory in stock market volatility. *Journal of Econometrics*, *73*(1), 151–184.

Bollerslev, T., & Mikkelsen, H. O. (1999). Long-term equity anticipation securities and stock market volatility dynamics. *Journal of Econometrics*, *92*(1), 75–99.

Bollerslev, T., Sizova, N., & Tauchen, G. (2012). Volatility in equilibrium: asymmetries and dynamic dependencies. *Review of Finance*, *16*(1), 31–80.

Bolthausen, E., Deuschel, J.-D., & Zeitouni, O. (1995). Entropic repulsion of the lattice free field. *Communications in Mathematical Physics*, *170*(2), 417–443.

Bos, C. S., Franses, P. H., & Ooms, M. (2002). Inflation, forecast intervals and long memory regression models. *International Journal of Forecasting*, *18*, 243–264.

Bougerol, P., & Picard, N. (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics*, *52*, 115–127.

Bourbaki, N. (1976). *Elements of mathematics, functions of a real variable*. Reading: Addison-Wesley. (Translated from French).

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*, 353–360.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: forecasting and control*. San Francisco: Holden Day.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association*, *65*, 1509–1526.

Bramson, M., Cox, J. T., & Durrett, R. (1996). Spatial models for species area curves. *Annals of Probability*, *24*(4), 1727–1751.

Breidt, F. J., & Davis, R. A. (1998). Extremes of stochastic volatility models. *The Annals of Applied Probability*, *8*(3), 664–675.

Breidt, F. J., Crato, N., & de Lima, P. (1998). On the detection and estimation of long memory in stochastic volatility. *Journal of Econometrics*, *83*, 325–348.

Breiman, L. (1965). On some limit theorems similar to the arc-sine law. *Teoriâ Veroâtnostej I Ee Primeneniâ*, *10*, 351–360.

Breiman, L. (1992). *Probability*. Philadelphia: SIAM.

Breitung, J., & Hassler, U. (2002). Inference on the cointegration rank in fractionally integrated processes. *Journal of Econometrics*, *110*, 167–185.

Breitung, J., & Hassler, U. (2006). A residual-based LM type test against fractional cointegration. *Econometric Theory*, *22*, 1091–1111.

Breuer, P., & Major, P. (1983). Central limit theorems for nonlinear functionals of Gaussian fields. *Journal of Multivariate Analysis*, *13*(3), 425–441.

Brillinger, D. R. (1969). The canonical analysis of stationary time series. In *Multivariate analysis, II (Proc. second internat. sympos., Dayton, Ohio, 1968)* (pp. 331–350). New York: Academic Press.

Brillinger, D. (1994). Uses of cumulants in wavelet analysis. *Proceedings of SPIE, Advanced Signal Processing*, *2296*, 2–18.

Brillinger, D. (1996). Some uses of cumulants in wavelet analysis. *Nonparametric Statistics*, *6*, 93–114.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: theory and methods*. *Springer series in statistics*. New York: Springer.

Brockwell, P. J., & Marquardt, T. (2005). Lévy-driven and fractionally integrated ARMA processes with continuous-time parameter. *Statistica Sinica*, *15*, 477–494.

Brodsky, B. E., & Darkhovsky, B. S. (1993). *Mathematics and its applications: Vol. 243*. *Nonparametric methods in change-point problems*. Dordrecht: Kluwer Academic.

Brodsky, J., & Hurvich, C. M. (1999). Multi-step forecasting for long-memory processes. *Journal of Forecasting*, *18*(1), 59–75.

Brody, D. C., Syroka, J., & Zervos, M. (2002). Dynamical pricing of weather derivatives. *Quantitative Finance*, *2*(3), 189–198.

Bryk, A., & Mielniczuk, J. (2008). Randomized fixed design regression under long-range-dependent errors. *Communications in Statistics, Theory and Methods*, *37*(4), 520–531.

Buchmann, B., & Chan, N. H. (2007). Asymptotic theory of least squares estimators for nearly unstable processes under strong dependence. *The Annals of Statistics*, *35*(5), 2001–2017.

Buchmann, B., & Klüppelberg, C. (2005). Maxima of stochastic processes driven by fractional Brownian motion. *Advances in Applied Probability*, *37*(3), 743–764.

Buchmann, B., & Klüppelberg, C. (2006). Fractional integral equations and state space transforms. *Bernoulli*, *12*(3), 431–456.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, *3*, 123–148.

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, *17*, 52–72.

Bunde, A. & Havlin, S. (Eds.) (1995). *Fractals and disordered systems* (2nd ed.). Berlin: Springer.

Cambanis, S., & Houdré, C. (1995). On the continuous wavelet transform of second-order random processes. *IEEE Transactions on Information Theory*, *41*(3), 628–642.

Cantor, G. (1883). Über unendliche, lineare Punktmannigfaltigkeiten V. *Mathematische Annalen*, *51*, 545–591.

Cappé, O., Moulines, E., Pesquet, J.-C., Petropulu, A., & Yang, X. (2002). Long-range dependence and heavy-tail modeling for teletraffic data. *IEEE Signal Processing Magazine*, *19*(3), 14–27.

Carlstein, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *The Annals of Statistics*, *14*, 1171–1179.

Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, *16*, 188–197.

Carlstein, E., & Lele, S. (1993). Nonparametric change-point estimation for data from an ergodic sequence. *Theory of Probability and Its Applications*, *38*, 910–917.

Cassandro, M., & Jona-Lasinio, G. (1978). Critical point behaviour and probability theory. *Advances in Physics*, *27*(6), 913–941.

Chakrabarti, B. K., Bardhan, K. K., & Sen, A. K. (2009). *Lecture notes in physics: Vol. 762*. *Quantum and semiclassical percolation and breakdown in disordered solids*. Berlin: Springer.

Chambers, M. J. (1996). The estimation of continuous parameter long-memory time series models. *Econometric Theory*, *12*, 374–390.

Chambers, M. J. (1998). Long memory and aggregation in macroeconometric time series. *International Econometric Review*, *39*, 1053–1072.

Chan, N. H., & Ling, S. (2008). Residual empirical processes for long and short memory time series. *Annals of Statistics*, *36*(5), 2453–2470.

Chan, N. H., & Palma, W. (1998). State space modeling of long-memory processes. *The Annals of Statistics*, *26*(2), 719–740.
Chan, N. H., & Terrin, N. (1995). Inference for unstable long-memory processes with applications to fractional unit root autoregressions. *The Annals of Statistics*, *23*, 1662–1683.
Chan, G., & Wood, A. T. A. (1997). Increment-based estimators of fractal dimension for two-dimensional surface data. *Statistica Sinica*, *10*, 343–376.
Chan, G., & Wood, A. T. A. (2004). Estimation of fractal dimension for a class of non-Gaussian stationary processes and fields. *The Annals of Statistics*, *32*(3), 1222–1260.
Charfeddine, L., & Guegan, D. M. (2009). *Breaks or long memory behaviour: an empirical investigation* (CES Working Papers). Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne.
Chen, W. W., & Deo, R. S. (2006). Estimation of mis-specified long memory models. *Journal of Econometrics*, *134*(1), 257–281.
Chen, W. W., & Hurvich, C. M. (2003a). Estimating fractional cointegration in the presence of polynomial trends. *Journal of Econometrics*, *117*, 95–121.
Chen, W. W., & Hurvich, C. M. (2003b). Semiparametric estimation of multivariate fractional cointegration. *Journal of the American Statistical Association*, *98*, 629–642.
Chen, W. W., & Hurvich, C. M. (2006). Semiparametric estimation of fractional cointegrating subspaces. *The Annals of Statistics*, *34*, 2939–2979.
Chen, W. W., & Hurvich, C. M. (2009). Fractional cointegration. In T. Mikosch, J. P. Kreiß, R. A. Davis, & T. G. Andersen (Eds.), *Handbook of financial time series* (pp. 709–726). Berlin: Springer.
Cheng, B., & Robinson, P. M. (1991). Density estimation in strongly dependent nonlinear time series. *Statistica Sinica*, *1*(2), 335–359.
Cheng, B., & Robinson, P. M. (1994). Semiparametric estimation from time series with long-range dependence. *Journal of Econometrics*, *64*, 335–354.
Cheridito, P., Kawaguchi, H., & Maejima, M. (2003). Fractional Ornstein–Uhlenbeck processes. *Electronic Journal of Probability*, *8*(3), 1–14.
Cheung, Y.-W., & Diebold, F. X. (1994). On maximum likelihood estimation of the differencing parameter of fractionally-integrated noise with unknown mean. *Journal of Econometrics*, *62*, 301–316.
Cheung, Y.-W., & Lai, K. (1993). A fractional cointegration analysis of purchasing power parity. *Journal of Business and Economics Statistics*, *11*, 93–101.
Chihara, T. S. (1978). *An introduction to orthogonal polynomials*. New York: Gordon and Breach.
Chiriac, R., & Voev, V. (2010). Modelling and forecasting multivariate realized volatility. *Journal of Applied Econometrics*, *36*(6), 922–947.
Choi, E., & Hall, P. G. (2000). Bootstrap confidence regions from autoregressions of arbitrary order. *Journal of the Royal Statistical Society B*, *62*, 461–477.
Choy, K., & Taniguchi, M. (2001). Stochastic regression model with dependent disturbances. *Journal of Time Series Analysis*, *22*(2), 175–196.
Christakos, G. (1992). *Random field models in Earth sciences*. San Diego: Academic Press.
Christakos, G. (2000). *Modern spatiotemporal geostatistics*. New York: Oxford University Press.
Christensen, B. J., & Nielsen, M. (2006). Asymptotic normality of narrow-band least squares in the stationary fractional cointegration model and volatility forecasting. *Journal of Econometrics*, *133*, 343–371.
Chung, C. F. (1996a). Estimating a generalized long memory process. *Journal of Econometrics*, *73*, 237–259.
Chung, C.-F. (1996b). A generalized fractionally integrated ARMA process. *Journal of Time Series Analysis*, *17*(2), 111–140.
Chung, C.-F. (2002). Sample means, sample autocovariances, and linear regression of stationary multivariate long memory processes. *Econometric Theory*, *18*(1), 51–78.
Claeskens, G., & Hall, P. (2002). Effect of dependence on stochastic measures of accuracy of density estimators. *The Annals of Statistics*, *30*(2), 431–454.
Clark, R. M. (1975). A calibration curve for radiocarbon dates. *Antiquity*, *49*, 251–266.

Cline, D. B. H. (1983). Estimation and linear prediction for regression, autoregression and ARMA with infinite variance data. Thesis Ph.D., Colorado State University. ProQuest LLC, Ann Arbor, MI, 1983. 136 pp.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, *17*, 164–177.

Coeurjolly, J.-F. (2008a). Bahadur representation of sample quantiles for functional of Gaussian dependent sequences under a minimal assumption. *Statistics & Probability Letters*, *78*(15), 2485–2489.

Coeurjolly, J.-F. (2008b). Hurst exponent estimation of locally self-similar Gaussian processes using sample quantiles. *The Annals of Statistics*, *36*(3), 1404–1434.

Cohen, A., & Ryan, R. (1995). *Wavelets and multiscale signal processing*. London: Chapman & Hall.

Comte, F. (1996). Simulation and estimation of long memory continuous-time models. *Journal of Time Series Analysis*, *17*(1), 19–36.

Comte, F., & Renault, E. (1996). Long memory continuous-time models. *Journal of Econometrics*, *73*, 101–149.

Constantine, A. G., & Hall, P. (1994). Characterizing surface smoothness via estimation of effective fractal dimension. *Journal of the Royal Statistical Society, Series B*, *56*, 97–113.

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, *19*, 297–301.

Coppersmith, D., Gamarnik, D., & Sviridenko, M. (2002). The diameter of a long-range percolation graph. *Random Structures & Algorithms*, *21*, 1–13.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Econometrics*, *7*(2), 174–196.

Cox, D. R., & Isham, V. (1980). *Point processes*. London: Chapman and Hall.

Crato, N., & de Lima, P. J. (1993). Long-range dependence in the conditional variance of stock returns. *Economics Letters*, *25*, 281–285.

Crato, N., & Ray, B. K. (1996). Model selection and forecasting for long range dependent processes. *Journal of Forecasting*, *15*, 107–125.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, *31*, 377–403.

Crawford, N., & Sly, A. (2011). Heat-kernel upper bounds on long-range percolation cluster. Preprint. arXiv:0907.2434.

Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.

Crovella, M. E., & Bestavros, A. (1997). Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, *5*(6), 835–846.

Csörgő, S. (1981). Limit behaviour of the empirical characteristic function. *The Annals of Probability*, *9*, 130–144.

Csörgő, S. (1982). The empirical moment generating function. In B. V. Gnedenko, M. L. Puri, & I. Vincze (Eds.), *Nonparametric statistical inference* (pp. 139–150). London: North-Holland.

Csörgő, S. (1986). Testing for normality in arbitrary dimension. *The Annals of Statistics*, *14*, 708–723.

Csörgő, M., & Horváth, L. (1988). Nonparametric methods for the change-point problems. In P. R. Krishnaiah & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 7, pp. 403–425). Amsterdam: Elsevier.

Csörgő, S., & Horváth, L. (1998). *Limit theorems in change-point analysis*. New York: Wiley.

Csörgő, M., & Kulik, R. (2008a). Reduction principles for quantile and Bahadur–Kiefer processes of long-range dependent linear sequences. *Probability Theory and Related Fields*, *142*(3–4), 339–366.

Csörgő, M., & Kulik, R. (2008b). Weak convergence of Vervaat and Vervaat error processes of long-range dependent sequences. *Journal of Theoretical Probability*, *21*(3), 672–686.

Csörgő, S., & Mason, D. M. (1985). Central limit theorems for sums of extreme values. *Mathematical Proceedings of the Cambridge Philosophical Society*, *98*(3), 547–558.

Csörgő, S., & Mielniczuk, J. (1995a). Density estimation under long-range dependence. *The Annals of Statistics*, *23*(3), 990–999.
Csörgő, S., & Mielniczuk, J. (1995b). Nonparametric regression under long-range dependent normal errors. *The Annals of Statistics*, *23*(3), 1000–1014.
Csörgő, S., & Mielniczuk, J. (1999). Random-design regression under long-range dependent errors. *Bernoulli*, *5*(2), 209–224.
Csörgő, S., & Mielniczuk, J. (2000). The smoothing dichotomy in random design regression with long-memory errors based on moving averages. *Statistica Sinica*, *10*, 771–787.
Csörgő, S., & Viharos, L. (1997). Asymptotic normality of least-squares estimators of tail indices. *Bernoulli*, *3*(3), 351–370.
Csörgő, S., Deheuvels, P., & Mason, D. M. (1985). Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, *13*, 1050–1077.
Csörgő, M., Szyszkowicz, B., & Wang, L. (2006). Strong invariance principles for sequential Bahadur–Kiefer and Vervaat error processes of long-range dependent sequences. *The Annals of Statistics*, *34*(2), 1013–1044.
Dabrowski, A. R., & Jakubowski, A. (1994). Stable limits for associated random variables. *Annals of Probability*, *22*(1), 1–16.
Dacorogna, M., Muller, U., Nagler, R., Olsen, R., & Pictet, O. (1993). A geographical model for the daily and weekly seasoned volatility in the FX market. *Journal of International Money and Finance*, *12*, 413–438.
Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *The Annals of Statistics*, *17*(4), 1749–1766.
Dahlhaus, R. (1995). Efficient location and regression estimation for long range dependent regression models. *The Annals of Statistics*, *23*, 1029–1047.
Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, *25*, 1–37.
Dahlhaus, R., & Giraitis, L. (1998). On the optimal segment length for estimates for locally stationary time series. *Journal of Time Series Analysis*, *19*, 629–636.
Dahlhaus, R., & Künsch, H. (1987). Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, *74*, 877–882.
Daley, D. J. (1999). The Hurst index of long-range dependent renewal processes. *The Annals of Probability*, *27*(4), 2035–2041.
Daley, D. J., & Vere-Jones, D. (1988). *An introduction to the theory of point processes* (1st ed.). New York: Springer.
Daley, D. J., & Vere-Jones, D. (2007). *An introduction to the theory of point processes* (2nd ed.). New York: Springer.
Daley, D. J., & Vesilo, R. (1997). Long range dependence of point process with queueing examples. *Stochastic Processes and Their Applications*, *70*, 265–282.
Daley, D. J., Rolski, T., & Vesilo, R. (2000). Long-range dependent point processes and their Palm–Khinchin distributions. *Advances in Applied Probability*, *32*(4), 1051–1063.
Dalla, V., Giraitis, L., & Hidalgo, J. (2006). Consistent estimation of the memory parameter for nonlinear time series. *Journal of Time Series Analysis*, *27*, 211–251.
Daubechies, I. (1992). *CBMS-NSF regional conference series in applied mathematics: Vol. 61. Ten lectures on wavelets*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM). xx+357 pp.
Davidson, J. (2002). A model of fractional cointegration, and tests for cointegration using the bootstrap. *Journal of Econometrics*, *110*(2), 187–212.
Davidson, J. (2006). Alternative bootstrap procedures for testing cointegration in fractionally integrated processes. *Journal of Econometrics*, *133*, 741–777.
Davidson, J., & Hashimzade, N. (2009a). Type I and type II fractional Brownian motions: a reconsideration. *Computational Statistics & Data Analysis*, *53*, 2089–2106.
Davidson, J., & Hashimzade, N. (2009b). Representation and weak convergence of stochastic integrals with fractional integrator processes. *Econometric Theory*, *25*(6), 1589–1624.

Davidson, J., & Sibbertsen, P. (2005). Generating schemes for longmemory processes: regimes, aggregation and linearity. *Journal of Econometrics*, *128*(2), 253–282.

Davidson, J. E. H., Hendry, D. F., Srba, F., & Yeo, S. (1978). Econometric modelling of the aggregate time series relationship between consumer's expenditure and income in the United Kingdom. *The Economic Journal*, *88*, 661–692.

Davies, S., & Hall, P. (1999). Fractal analysis of surface roughness by using spatial data. *Journal of the Royal Statistical Society, Series B*, *61*, 3–37.

Davis, R. A. (1983). Stable limits for partial sums of dependent random variables. *Annals of Probability*, *11*(2), 262–269.

Davis, R. A., & Hsing, T. (1995). Point process and partial sum convergence for weakly dependent random variables with infinite variance. *Annals of Probability*, *23*(2), 879–917.

Davis, R. A., & Mikosch, T. (1998). The sample autocorrelations of heavy-tailed processes with applications to ARCH. *The Annals of Statistics*, *26*(5), 2049–2080.

Davis, R. A., & Mikosch, T. (2001). Point process convergence of stochastic volatility processes with application to sample autocorrelations. *Journal of Applied Probability*, *38A*, 93–104.

Davis, R. A., & Resnick, S. (1984). Tail estimates motivated by the extreme value theory. *The Annals of Statistics*, *12*, 1467–1487.

Davis, R., & Resnick, S. (1985). Limit theory for moving averages of random variables with regularly varying tail probabilities. *Annals of Probability*, *13*(1), 179–195.

Davis, R. A., & Resnick, S. I. (1986). Limit theory for the sample covariance and correlation functions of moving averages. *The Annals of Statistics*, *14*(2), 533–558.

Davis, R. A., & Resnick, S. I. (1988). Extremes of moving averages of random variables from the domain of attraction of the double exponential distribution. *Stochastic Processes and Their Applications*, *30*(1), 41–68.

Davis, R. A., & Resnick, S. I. (1996). Limit theory for bilinear processes with heavy-tailed noise. *The Annals of Applied Probability*, *6*(4), 1191–1210 (English summary).

Davis, R. A., Huang, D. D., & Yao, Y. C. (1995). Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, *23*, 282–304.

Davydov, Yu. A. (1970a). The invariance principle for stationary processes. *Teoriâ Veroâtnostej I Ee Primeneniâ*, *15*, 498–509 (Russian).

Davydov, Ju. A. (1970b). The invariance principle for stationary processes. *Theory of Probability and Its Applications*, *15*, 487–498.

De Haan, L., & Ferreira, A. (2006). *Extreme value theory. An introduction*. *Springer series in operations research and financial engineering*. New York: Springer.

De Haan, L., & Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica*, *52*, 60–70.

De Haan, L., & Resnick, S. (1998). On asymptotic normality of the Hill estimator. *Stochastic Models*, *14*, 849–867.

de Lima, B. N. B., & Sapozhnikov, A. (2008). On the truncated long-range percolation on $\mathbb{Z}^d$. *Journal of Applied Probability*, *45*(1), 287–291.

Decreusefond, L., & Üstünel, A. S. (1999). Stochastic analysis of the fractional Brownian motion. *Potential Analysis*, *10*, 177–214.

Deheuvels, P., Haeusler, E., & Mason, D. (1988). Almost sure convergence of the Hill estimator. *Mathematical Proceedings of the Cambridge Philosophical Society*, *104*(2), 371–381.

Dehling, H., & Taqqu, M. S. (1989a). The functional law of the iterated logarithm for the empirical process of some long-range dependent sequences. *Statistics & Probability Letters*, *7*(1), 81–85.

Dehling, H., & Taqqu, M. S. (1989b). The empirical process of some long-range dependent sequences with an application to U-statistics. *The Annals of Statistics*, *17*(4), 1767–1783.

Dehling, H., & Taqqu, M. S. (1991). Bivariate symmetric statistics of long-range dependent observations. *Journal of Statistical Planning and Inference*, *28*(2), 153–165.

Dehling, H., Mikosch, T., & Sörensen, M. (Eds.) (2002). *Empirical process techniques for dependent data*.

Dehling, H., Rooch, A., & Taqqu, M. S. (2013). Nonparametric change-point tests for long-range dependent data. *Scandinavian Journal of Statistics*, *40*, 153–173.

Dekkers, A. L. M., & de Haan, L. (1989). On the estimation of the extreme value index and large quantile estimation. *The Annals of Statistics*, *17*(4), 1795–1832.

Delgado, M. A., Hidalgo, J., & Velasco, C. (2005). Distribution free goodness-of-fit tests for linear processes. *The Annals of Statistics*, *33*(6), 2568–2609.

Denker, M., & Jakubowski, A. (1989). Stable limit distributions for strongly mixing sequences. *Statistics & Probability Letters*, *8*(5), 477–483.

Deo, R. S. (1997). Asymptotic theory for certain regression models with long memory errors. *Journal of Time Series Analysis*, *18*(4), 385–393.

Deo, R. S., & Chen, W. W. (2000). On the integral of the squared periodogram. *Stochastic Processes and Their Applications*, *85*(1), 159–176.

Deo, R., & Hurvich, C. M. (2001). On the log periodogram regression estimator of the memory parameter in long memory stochastic volatility models. *Econometric Theory*, *17*(4), 686–710.

Deo, R., Hsieh, M.-C., Hurvich, C. M., & Soulier, P. (2006a). Long memory in nonlinear processes. In *Lecture notes in statist.: Vol. 187. Dependence in probability and statistics* (pp. 221–244). New York: Springer.

Deo, R., Hurvich, C., & Lu, Y. (2006b). Forecasting realized volatility using a long-memory stochastic volatility. *Journal of Econometrics*, *131*, 29–58.

Deo, R., Hsieh, M.-C., & Hurvich, C. M. (2007). Long memory in intertrade durations, counts and realized volatility of NYSE stocks. Preprint.

Deo, R., Hurvich, C., Soulier, P., & Wang, Y. (2009). Conditions for the propagation of memory parameter from durations to counts and realized volatility. *Econometric Theory*, *25*, 764–792.

Dette, H., & Sen, K. (2010). *Goodness-of-fit tests in long-range dependent processes under fixed alternatives* (SFB 823, Discussion Paper No. 48/2010). http://hdl.handle.net/2003/27523.

Dette, H., Leonenko, N., Pepelyshev, A., & Zhigljavsky, A. (2009). Asymptotic optimal designs under long-range dependence error structure. *Bernoulli*, *15*, 1036–1056.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, *49*, 1057–1072.

Diebold, F., & Inoue, A. (2001). Long memory and structural change. *Journal of Econometrics*, *105*, 131–159.

Diggle, P. J. (1990). *Time series: a biostatistical introduction*. Oxford: Oxford University Press.

Ding, Z., & Granger, C. W. J. (1996). Modeling volatility persistence of speculative returns: a new approach. *Journal of Econometrics*, *73*(1), 185–215.

Ding, Z., Granger, C., & Engle, R. (1993). A long-memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*, 83–106.

Dobrushin, R. L. (1965). Existence of a phase transition in two and three dimensional Ising models. *Theory of Probability and Its Applications*, *10*, 193–213.

Dobrushin, R. L. (1968a). Problem of uniqueness of a Gibbs random field and phase transitions. *Functional Analysis and Applications*, *2*(4), 44–57 (in Russian).

Dobrushin, R. L. (1968b). Gibbsian random fields for lattice systems with pairwise interactions. *Funktsionalnyi Analiz I Ego Prilozheniya*, *2*(4), 31–43. [Funct. Anal. Appl., 2, 292–301 (1968)].

Dobrushin, R. L. (1968c). The description of a random field by means of conditional probabilities and conditions of its regularity. *Teoriâ Veroâtnostej I Ee Primeneniâ*, *13*, 201–229. [Theor. Prob. Appl., 13, 197–224 (1968)].

Dobrushin, R. L. (1969). Gibbs field: the general case. *Functional Analysis and Applications*, *3*(1), 27–35 (in Russian).

Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Teoriâ Veroâtnostej I Ee Primeneniâ*, *15*, 469–497. [Theory Probab. Appl., 15, 458–486].

Dobrushin, R. L. (1980). Gaussian random fields—Gibbsian point of view. In R. L. Dobrushin & Ya. G. Sinai (Eds.), *Multicomponent random systems*. *Advances in probability and related topics* (Vol. 6, pp. 119–152). New York: Dekker.

Dobrushin, R. L., & Major, P. (1979). Non-central limit theorems for nonlinear functionals of Gaussian fields. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *50*(1), 27–52.

Dolado, J. J., Gonzalo, J., & Mayoral, L. (2003). Long range dependence in Spanish political opinion poll data. *Journal of Applied Econometrics*, *18*, 137–155.

Domb, C. (1985). Critical phenomena: a brief historical survey. *Contemporary Physics*, *26*(1), 49–72.

Domb, C. & Lebowitz, J. L. (Eds.) (2001). *Phase transitions and critical phenomena* (Vol. 18). San Diego: Academic Press.

Dombry, C., & Kaj, I. (2011). The on–off network traffic model under intermediate scaling. *Queueing Systems*, *69*(1), 29–44.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, *81*, 425–455.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224.

Donoho, D. L., & Johnstone, I. M. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, *59*(2), 319–351.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., & Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B. Methodological*, *57*, 301–369.

Doob, J. L. (1953). *Stochastic processes*. New York: Wiley.

Doornik, J. A., & Ooms, M. (2003). Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models. *Computational Statistics & Data Analysis*, *42*(3), 333–348.

Douc, R., Roueff, F., & Soulier, P. (2008). On the existence of some processes. *Stochastic Processes and Their Applications*, *118*(5), 755–761.

Doukhan, P., Oppenheim, G., & Taqqu, M. S. (2003). *Theory and application of long-range dependence*. Boston: Birkhäuser.

Draghicescu, D. (2002). *Nonparametric quantile estimation for dependent data*. Ph.D. thesis, No. 2592, EPF Lausanne.

Draghicescu, D., & Ghosh, S. (2003). Smooth nonparametric quantiles. In *Proceedings of the 2nd international colloquium of mathematics in engineering and numerical physics (MENP-2)* (pp. 45–52). Bucharest: Geometry Balkan Press.

Drees, H. (1998). Optimal rates of convergence for estimates of the extreme value index. *The Annals of Statistics*, *26*(1), 434–448.

Drees, H. (2000). Weighted approximations of tail processes for mixing random variables. *The Annals of Applied Probability*, *10*, 1274–1301.

du Bois-Reymond, P. (1880). Der Beweis des Fundamentalsatzes der Integralrechnung. *Mathematische Annalen*, *16*, 115–128.

Dueker, M., & Startz, R. (1998). Maximum-likelihood estimation of fractional cointegration with an application to U.S. and Canadian bond rates. *The Review of Economics and Statistics*, *80*(3), 420–426.

Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, *19*, 1471–1495.

Duncan, T. E., Hu, Y., & Pasik-Duncan, B. (2000). Stochastic calculus for fractional Brownian motion I. Theory. *SIAM Journal on Control and Optimization*, *38*(2), 582–612.

Durrett, R. (1984). An introduction to oriented percolation. *Annals of Probability*, *12*, 999–1040.

Durrett, R., & Levin, S. (1996). Spatial models for species-area curves. *Journal of Theoretical Biology*, *179*(2), 119–127.

Eberhard, J. W., & Horn, P. M. (1978). Excess 1/f noise in metals. *Physical Reviews B*, *18*, 6681–6693.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

El Attar, R. (2006). *Special functions and orthogonal polynomials*. Morrisville: Lulu Press.

Elliott, F. W. Jr, Horntrop, D. J., & Majda, A. J. (1997). Monte Carlo methods for turbulent tracers with long range and fractal random velocity fields. *Chaos*, *7*(1), 39–48.

Elliott, R. J., & van der Hoek, J. (2003). A general fractional white noise theory and applications to finance. *Mathematical Finance*, *13*, 301–330.

Embrechts, P., & Maejima, M. (2002). *Self-similar processes*. Princeton: Princeton University Press.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events*. New York: Springer.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, *50*, 987–1008.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: representation, estimation and testing. *Econometrica*, *55*(2), 251–276.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, *14*, 153–158.

Epps, T. W., Singleton, K. J., & Pulley, L. B. (1982). A test of separate families of distributions based on the empirical moment generating function. *Biometrika*, *69*(2), 391–399.

Ercolani, J. S. (2011). On the asymptotic properties of a feasible estimator of the continuous time long memory parameter. *Journal of Time Series Analysis*, *32*, 512–517.

Estévez, G., & Vieu, P. (2003). Nonparametric estimation under long memory dependence. *Journal of Nonparametric Statistics*, *15*(4–5), 535–551.

Eubank, R. L. (1987). *Spline smoothing and nonparametric regression*. New York: Dekker.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing* (2nd ed.). New York: Dekker.

Falconer, K. (2003). *Fractal geometry: mathematical foundations and applications* (2nd ed.). Chichester: Wiley.

Falconer, K., & Fernandez, C. (2007). Inference on fractal processes using multiresolution approximation. *Biometrika*, *94*(2), 313–334.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, *87*, 998–1004.

Fan, J., & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society B*, *57*, 371–394.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman & Hall.

Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, *85*(3), 645–660.

Fasen, V., & Samorodnitsky, G. (2009). A fluid cluster Poisson input process can look like a fractional Brownian motion even in the slow growth aggregation regime. *Advances in Applied Probability*, *41*(2), 393–427.

Faÿ, G., & Philippe, A. (2002). Goodness-of-fit test for long range dependent processes. *ESAIM: Probability and Statistics*, *6*, 239–258.

Faÿ, G., & Soulier, P. (2001). The periodogram of an i.i.d. sequence. *Stochastic Processes and Their Applications*, *92*(2), 315–343.

Faÿ, G., Moulines, E., & Soulier, P. (2004). Edgeworth expansions for linear statistics of possibly long range dependent linear processes. *Statistics & Probability Letters*, *66*(3), 275–288.

Faÿ, G., González-Arévalo, B., Mikosch, T., & Samorodnitsky, G. (2006). Modeling teletraffic arrivals by a Poisson cluster process. *Queueing Systems*, *54*(2), 121–140.

Faÿ, G., Roueff, F., & Soulier, P. (2007). Estimation of the memory parameter of the infinite source Poisson process. *Bernoulli*, *13*(2), 473–491.

Faÿ, G., Moulines, E., Roueff, F., & Taqqu, M. S. (2009). Asymptotic normality of wavelet estimators of the memory parameter for linear processes. *Journal of Time Series Analysis*, *30*(5), 534–558.

Feller, W. (1951). The asymptotic distributions of the range of sums of independent random variables. *The Annals of Mathematical Statistics*, *22*, 427–432.

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. *2*). New York: Wiley.

Feng, Y. (1999). *Kernel- and locally weighted regression—with application to time series decomposition*. Berlin: Verlag für Wissenschaft und Forschung.

Feng, Y. (2004a). Non- and semiparametric regression with fractional time series errors—theory and applications to financial data. Habilitation Work, University of Konstanz.

Feng, Y. (2004b). Simultaneously modelling conditional heteroskedasticity and scale change. *Econometric Theory*, *20*, 563–596.

Feng, Y., & Beran, J. (2012, in press). Optimal convergence rates in nonparametric regression with fractional time series errors. *Journal of Time Series Analysis*.

Feng, Y., Beran, J., & Yu, K. (2007). Modelling financial time series with SEMIFAR-GARCH models. *IMA Journal of Management Mathematics (Special Issue on Financial Mathematics)*, *18*(4), 395–412.

Ferger, D. (1994). On the rate of almost sure convergence of Dümbgen's change-point estimators. *Statistics & Probability Letters*, *19*, 27–31.

Ferger, D., & Stute, W. (1992). Convergence of changepoint estimators. *Stochastic Processes and Their Applications*, *42*, 345–35l.

Fernández-Pascual, R., Ruiz-Medina, M. D., & Angulo, J. M. (2006). Estimation of intrinsic processes affected by additive fractal noise. *Journal of Multivariate Analysis*, *97*(6), 1361–1381.

Ferrara, L., & Guégan, D. (2000). Forecasting financial time series with generalized long memory processes. In C. L. Dunis (Ed.), *Advances in quantitative asset management* (pp. 319–342). Dordrecht: Kluwer Academic.

Ferrara, L., & Guégan, D. (2001a). Forecasting with k-factor Gegenbauer processes. *Journal of Forecasting*, *20*(8), 581–601.

Ferrara, L., & Guégan, D. (2001b). Comparison of parameter estimation methods in cyclical long memory time series. In C. L. Dunis, A. Timmermann, & J. E. Moody (Eds.), *Developments in forecast combination and portfolio choice* (pp. 179–196). New York: Wiley.

Feuerverger, A. (1989). On the empirical saddlepoint approximation. *Biometrika*, *76*(3), 457–464.

Feuerverger, A., & Mureika, R. A. (1977). The empirical characteristic function and its applications. *The Annals of Statistics*, *5*, 88–97.

Feuerverger, A., Hall, P., & Wood, A. T. A. (1994). Estimation of fractal index and fractal dimension of a Gaussian process by counting the number of level crossings. *Journal of Time Series Analysis*, *15*, 587–606.

Fisher, M. E. (1964). Correlation functions and the critical region of simple fluids. *Journal of Mathematical Physics*, *5*(7), 944–962.

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Transactions on Information Theory*, *38*(2), 910–917.

Fleming, J., & Kirby, C. (2011). Long memory in volatility and trading volume. *Journal of Banking & Finance*, *35*(7), 1714–1726.

Föllmer, H. (1975). Phase transitions and Martin boundary. In *Lecture notes in mathematics: Vol. 465. Sém. prob. IX* (pp. 305–318). Berlin: Springer.

Fox, R., & Taqqu, M. S. (1985). Noncentral limit theorems for quadratic forms in random variables having long-range dependence. *Annals of Probability*, *13*(2), 428–446.

Fox, R., & Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics*, *14*, 517–532.

Fox, R., & Taqqu, M. S. (1987). Central limit theorems for quadratic forms in random variables having long-range dependence. *Probability Theory and Related Fields*, *74*(2), 213–240.

Franchi, M. (2010). A representation theory for polynomial cofractionality in vector autoregressive models. *Econometric Theory*, *26*, 1201–1217.

Francq, C., & Zakoian, J.-M. (2008). *Inconsistency of the QMLE and asymptotic normality of the weighted LSE for a class of conditionally heteroscedastic models* (Working paper). Université Lille 3.

Frederiksen, P., Nielsen, F. S., & Nielsen, M. Ø. (2012). Local polynomial Whittle estimation of perturbed fractional processes. *Journal of Econometrics*, *167*(2), 426–447.

Frisch, U. (1995). *Turbulence: the legacy of A.N. Kolmogorov*. Cambridge: Cambridge University Press.

Fröhlich, J., & Spencer, T. (1982). The phase transition in the one-dimensional Ising model with $1/r^2$ interaction energy. *Communications in Mathematical Physics*, *84*(1), 167–170.

Fuller, W. A. (1996). *Introduction to statistical time series* (2nd ed.). New York: Wiley.

Gabor, D. (1946). Theory of communication. *Proceedings of the Institution of Electrical Engineers*, *93*, 429–457.

Gaigalas, R. (2004). *A non-Gaussian limit process with long-range dependence*. Thesis Ph.D., Uppsala Universitet (Sweden).

Gaigalas, R., & Kaj, I. (2003). Convergence of scaled renewal processes and a packet arrival model. *Bernoulli*, *9*(4), 671–703 (English summary).

Gajek, L., & Mielniczuk, J. (1999). Long- and short-range dependent sequences under exponential subordination. *Statistics & Probability Letters*, *43*(2), 113–121.

Galán, R. F., Weidert, M., Menzel, R., Herz, A. V. M., & Galizia, C. G. (2006). Sensory memory for odors is encoded in spontaneous correlated activity between olfactory Glomeruli. *Neural Computation*, *18*, 10–25.

Galizia, C. G., & Menzel, R. (2001). The role of Glomeruli in the neural representation of odours: results from optical recording studies. *Journal of Insect Physiology*, *47*, 115–130.

Gao, J., & Rubin, I. (2001). Multiplicative multifractal modeling of long-range-dependent (LRD) traffic in computer communications networks. *Nonlinear Analysis*, *47*(9), 5765–5774 (English summary). Proceedings of the Third world congress of nonlinear analysts, Part 9 (Catania, 2000)

Gasser, T., & Müller, H. G. (1979). Kernel estimation of regression functions. In T. Gasser & M. Rosenblatt (Eds.), *Smoothing techniques for curve estimation* (pp. 23–68). Heidelberg: Springer.

Gasser, T., & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, *11*, 171–185.

Gasser, T., Müller, H. G., & Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society B*, *47*, 238–252.

Gasser, T., Kneip, A., & Köhler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, *86*, 643–652.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*, 320–328.

Gelfand, I. M., & Shilov, G. E. (1966–1968). *Generalized functions* (Vols. 1–5). San Diego: Academic Press.

Georgii, H. O. (1988). *Gibbs measure and phase transitions*. Berlin: De Gruyter.

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, *4*, 221–238.

Ghosh, S. (1996). A new graphical tool to detect non-normality. *Journal of the Royal Statistical Society B*, *58*, 691–702.

Ghosh, S. (2001). Nonparametric trend estimation in replicated time series. *Journal of Statistical Planning and Inference*, *97*, 263–274.

Ghosh, S. (2009). The unseen species number revisited. *Sankhya, The Indian Journal of Statistics*, *71-B(2)*, 137–150.

Ghosh, S., & Beran, J. (2000). Two sample T3-plot: a graphical comparison of two distributions. *Journal of Computational and Graphical Statistics*, *9*(1), 167–179.

Ghosh, S., & Beran, J. (2006). On estimating the cumulant generating function of linear processes. *Annals of the Institute of Statisitcal Mathematics*, *58*, 53–71.

Ghosh, S., & Draghicescu, D. (2002a). Predicting the distribution function for long-memory processes. *International Journal of Forecasting*, *18*, 283–290.

Ghosh, S., & Draghicescu, D. (2002b). An algorithm for optimal bandwidth selection for smooth nonparametric quantiles and distribution functions. In Y. Dodge (Ed.), *Statistics in industry and technology: statistical data analysis based on the L1-norm and related methods* (pp. 161–168). Basel: Birkhäuser.

Ghosh, S., & Ruymgaart, F. (1992). Applications of empirical characteristic functions in some multivariate problems. *The Canadian Journal of Statistics*, *20*, 429–440.

Ghosh, S., & Samorodnitsky, G. (2010). Long strange segments, ruin probabilities and the effect of memory on moving average processes? *Stochastic Processes and Their Applications*, *120*(12), 2302–2330.

Ghosh, S., Beran, J., & Innes, J. (1997). Nonparametric conditional quantile estimation in the presence of long memory. *Student*, *2*, 109–117.

Gil-Alana, L. A. (2004). Testing of fractional cointegration in macroeconomic time series. *Oxford Bulletin of Economics and Statistics*, *65*, 517–529.

Gil-Alana, L. A. (2008). Fractional integration and structural breaks at unknown periods of time. *Journal of Time Series Analysis*, *29*, 163–185.

Gil-Alana, L. A., & Hualde, J. (2009). Fractional integration and cointegration: an overview and an empirical application. In K. Patterson & T. C. Mills (Eds.), *Palgrave handbook of econometrics* (Vol. II, pp. 434–469). Palgrave: MacMillan.

Giraitis, L. (1985). Central limit theorem for functionals of a linear process. *Lithuanian Mathematical Journal*, *25*(1), 25–35.

Giraitis, L., & Leipus, R. (1992). Testing and estimating in the change-point problem for the spectral function. *Lithuanian Mathematical Journal*, *32*, 20–38.

Giraitis, L., & Leipus, R. (1995). A generalized fractionally differencing approach in long memory modeling. *Lithuanian Mathematical Journal*, *35*, 65–81.

Giraitis, L., & Robinson, P. M. (2001). Whittle estimation of ARCH models. *Econometric Theory*, *17*, 608–631.

Giraitis, L., & Robinson, P. M. (2003). Edgeworth expansions for semiparametric Whittle estimation of long memory. *The Annals of Statistics*, *31*(4), 1325–1375.

Giraitis, L., & Surgailis, D. (1985). CLT and other limit theorems for functionals of Gaussian processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *70*(2), 191–212.

Giraitis, L., & Surgailis, D. (1986). Multivariate Appell polynomials and the central limit theorem. In *Progr. probab. statist.: Vol. 11. Dependence in probability and statistics* (pp. 21–71). Boston: Birkhäuser.

Giraitis, L., & Surgailis, D. (1989). Limit theorem for polynomials of a linear process with long-range dependence. *Lithuanian Mathematical Journal*, *29*(2), 128–145.

Giraitis, L., & Surgailis, D. (1990). A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate. *Probability Theory and Related Fields*, *86*(1), 87–104.

Giraitis, L., & Surgailis, D. (1999). Central limit theorem for the empirical process of a linear sequence with long memory. *Journal of Statistical Planning and Inference*, *80*(1–2), 81–93.

Giraitis, L., & Surgailis, D. (2002). ARCH-type bilinear models with double long memory. *Stochastic Processes and Their Applications*, *100*, 275–300.

Giraitis, L., & Taqqu, M. S. (1997). Limit theorems for bivariate Appell polynomials I. Central limit theorems. *Probability Theory and Related Fields*, *107*(3), 359–381.

Giraitis, L., & Taqqu, M. S. (1998). Central limit theorems for quadratic forms with time-domain conditions. *Annals of Probability*, *26*(1), 377–398.

Giraitis, L., & Taqqu, M. S. (1999a). Whittle estimator for finite-variance non-Gaussian time series with long memory. *The Annals of Statistics*, *27*(1), 178–203.

Giraitis, L., & Taqqu, M. S. (1999b). Convergence of normalized quadratic forms. *Journal of Statistical Planning and Inference*, *80*(1–2), 15–35.

Giraitis, L., & Taqqu, M. S. (2001). Functional non-central and central limit theorems for bivariate Appell polynomials. *Journal of Theoretical Probability*, *14*(2), 393–426.

Giraitis, L., Koul, H. L., & Surgailis, D. (1996a). Asymptotic normality of regression estimators with long memory errors. *Statistics & Probability Letters*, *29*(4), 317–335.

Giraitis, L., Leipus, R., & Surgailis, D. (1996b). The change-point problem for dependent observations. *Journal of Statistical Planning and Inference*, *53*, 297–310.

Giraitis, L., Robinson, P. M., & Samarov, A. (1997). Rate optimal semiparametric estimation of the memory parameter of the Gaussian time series with long-range dependence. *Journal of Time Series Analysis*, *18*(1), 49–60.

Giraitis, L., Taqqu, M. S., & Terrin, N. (1998). Limit theorems for bivariate Appell polynomials II. Non-central limit theorems. *Probability Theory and Related Fields*, *110*(3), 333–367.

Giraitis, L., Kokoska, P., & Leipus, R. (2000a). Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory*, *16*, 3–22.

Giraitis, L., Kokoszka, P., Leipus, R., & Teyssière, G. (2000b). Semiparametric estimation of the intensity of long memory in conditional heteroskedasticity. *Statistical Inference for Stochastic Processes*, *3*(1–2), 113–128. 19th "Rencontres Franco-Belges de statisticiens" (Marseille, 1998).

Giraitis, L., Robinson, P. M., & Surgailis, D. (2000c). A model for long memory conditional heteroscedasticity. *The Annals of Applied Probability*, *10*(3), 1002–1024.

Giraitis, L., Hidalgo, J., & Robinson, P. M. (2001). Gaussian estimation of parametric spectral density with unknown pole. *The Annals of Statistics*, *29*(4), 987–1023.

Giraitis, L., Kokoszka, P., Leipus, R., & Teyssiere, G. (2003). Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics*, *112*(2), 265–294.

Giraitis, L., Leipus, R., Robinson, P. M., & Surgailis, D. (2004). LARCH, leverage and long memory. *Journal of Financial Econometrics*, *2*, 177–210.

Giraitis, L., Leipus, R., & Surgailis, D. (2006). Recent advances in ARCH modelling. In G. Teyssière & A. P. Kirman (Eds.), *Long memory in economics* (pp. 3–38). Berlin: Springer.

Giraitis, L., Leipus, R., & Surgailis, D. (2010). Aggregation of the random coefficient GLARCH(1, 1) process. *Econometric Theory*, *26*, 406–425.

Giraitis, L., Koul, H. L., & Surgailis, D. (2012). *Large sample inference for long memory processes*. London: Imperial College Press.

Gnedenko, B. V., & Kolmogorov, A. N. (1968). *Limit distributions for sums of independent random variables* (revised ed.). Reading: Addison-Wesley. Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. ix+293 pp.

Gneiting, T., & Schlather, M. (2004). Stochastic models which separate fractal dimension and Hurst effect. *SIAM Review*, *46*, 269–282.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. (Circulation electronic pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215); 2000 June 13. PMID: 10851218; doi:10.1161/01.CIR.101.23.e215.

Goncalves, E., & Gouriéroux, C. (1988). Aggrégation de processus autoregressifs d'ordre 1. *Annales d'Économie et de Statistique*, *12*, 127–149.

Gorodetskii, V. V. (1977). On convergence to semi-stable Gaussian process. *Theory of Probability and Its Applications*, *22*, 498–508.

Gorostiza, L. G., & Wakolbinger, A. (1991). Persistence criteria for a class of critical branching particle systems in continuous time. *Annals of Probability*, *19*, 266–288.

Gorostiza, L. G., Navarro, R., & Rodrigues, E. R. (2005). Some long-range dependence processes arising from fluctuations of particle systems. *Acta Applicandae Mathematicae*, *86*, 285–308.

Gouyet, J.-F. (1996). *Physics and fractal structures*. New York: Springer.

Gradshteyn, I. S., & Rhyzhik, I. M. (1965). *Tables of integrals, series and products*. San Diego: Academic Press.

Granger, C. W. J. (1966). The typical spectral shape of an economic variable. *Econometrica*, *34*, 150–161.

Granger, C. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, *14*, 227–238.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, *16*, 121–130.

Granger, C. W. J. (1983). *Co-integrated variables and error-correcting models* (UCSD Discussion Paper).

Granger, C. W. J. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics*, *48*, 213–228.

Granger, C. W. J. (1995). Non-linear relationships between non-stationary processes. *Econometrica*, *63*, 265–279.
Granger, C. W. J. (1998). Real and spurious long-memory properties of stock market data: comment. *Journal of Business and Economic Statistics*, *16*, 268–269.
Granger, C. W. J., & Ding, Z. (1996). Varieties of long-memory models. *Journal of Econometrics*, *73*(1), 61–77.
Granger, C. W. J., & Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance*, *11*, 399–421.
Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis*, *1*, 15–30.
Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*, 111–120.
Gray, H. L., Zhang, N.-F., & Woodward, W. A. (1989). On generalized fractional processes. *Journal of Time Series Analysis*, *10*, 233–257.
Gray, H. L., Zhang, N.-F., & Woodward, W. A. (1994). On generalized fractional processes—a correction. *Journal of Time Series Analysis*, *15*(5), 561–562.
Greiner, M., Jobmann, M., & Klüppelberg, C. (1999). Telecommunication traffic, queueing models, and subexponential distributions. Queues with heavy-tailed distributions. *Queueing Systems, Theory and Applications*, *33*(1–3), 125–152.
Grenander, U. (1954). On the estimation of regression coefficients in the case of an autocorrelated disturbance. *The Annals of Mathematical Statistics*, *25*, 252–272.
Grenander, U., & Rosenblatt, M. (1957). *Statistical analysis of stationary time series*. New York: Wiley.
Grenander, U., & Szegö, G. (1958). *Toeplitz forms and their application*. Berkeley: University of California Press.
Grimmett, G. (1999). *Percolation* (2nd ed.). Berlin: Springer.
Grossmann, A., & Morlet, J. (1985). Decomposition of functions into wavelets of constant shape and related transforms. In L. Streit (Ed.), *Mathematics and physics, lectures on recent results*, River Edge: Word Scientific.
Guasoni, P. (2006). No arbitrage under transaction costs, with fractional Brownian motion and beyond. *Mathematical Finance*, *16*, 569–582.
Guégan, D. (1999). *Note on long memory processes with cyclical behavior and heteroscedasticity* (Working paper). University of Reims, France, 99-08, 1–21.
Guégan, D. (2000). A new model: the k-factor GIGARCH process. *Journal of Signal Processing*, *4*(3), 265–271.
Guo, H., & Koul, H. L. (2008). Asymptotic inference in some heteroscedastic regression models with long memory design and errors. *The Annals of Statistics*, *36*(1), 458–487.
Guo, H., Lim, C. Y., & Meerschaert, M. (2009). Local Whittle estimator for anisotropic random fields. *Journal of Multivariate Analysis*, *100*(5), 993–1028.
Gürtler, N., & Henze, N. (2000). Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of the Institute of Statistical Mathematics*, *52*, 267–286.
Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, *69*(1), 95–105.
Guyon, X. (1995). *Random fields on a network*. New York: Springer.
Haar, A. (1910). Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, *69*, 331–371.
Haldrup, N., & Nielsen, M. O. (2007). Estimation of fractional integration in the presence of data noise. *Computational Statistics & Data Analysis*, *51*(6), 3100–3114.
Hall, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of Rural and Small Schools B*, *44*(1), 37–42.
Hall, P. (1990). Using the bootstrap to estimate the mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, *32*(2), 177–203.
Hall, P. (1992). *The bootstrap and Edgeworth expansion*. New York: Springer.

Hall, P. (1995). On the effect of measuring a self-similar process. *SIAM Journal on Applied Mathematics*, *55*(3), 800–808.

Hall, P. (1997). Defining and measuring long-range dependence. In C. D. Cutler & D. T. Kaplan (Eds.), *Nonlinear dynamics and time series (fields inst. commun. 11)* (pp. 153–160). Providence: Am. Math. Soc.

Hall, P., & Hart, J. D. (1990a). Convergence rates in density estimation for data from infinite-order moving average processes. *Probability Theory and Related Fields*, *87*(2), 253–274.

Hall, P., & Hart, J. D. (1990b). Nonparametric regression with long-range dependence. *Stochastic Processes and Their Applications*, *36*, 339–351.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its application*. New York: Academic Press.

Hall, P., & Patil, P. (1996a). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *Journal of the Royal Statistical Society, Series B*, *58*, 361–377.

Hall, P., & Patil, P. (1996b). Effect of threshold rules on performance of wavelet-based curve estimators. *Statistica Sinica*, *6*, 331–345.

Hall, P., & Roy, R. (1994). On the relationship between fractal dimension and fractal index for stationary stochastic processes. *The Annals of Applied Probability*, *4*, 241–253.

Hall, P., & Welsh, A. H. (1984). Best attainable rates of convergence for estimates of regular variation. *The Annals of Statistics*, *12*, 1079–1084.

Hall, P., & Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrika*, *71*, 285–317.

Hall, P., Lahiri, S. N., & Polzehl, J. (1995a). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *The Annals of Statistics*, *23*(6), 1921–1936.

Hall, P., Lahiri, S. N., & Truong, K. (1995b). On bandwidth choice for density estimation with dependent data. *The Annals of Statistics*, *23*(6), 2241–2263.

Hall, P., Matthews, D., & Platen, E. (1996). Algorithms for analyzing nonstationary time series with fractal noise. *Journal of Computational and Graphical Statistics*, *5*, 351–364.

Hall, P., Jing, B.-Y., & Lahiri, S. N. (1998). On the sampling window method for long-range dependent data. *Statistica Sinica*, *8*, 1189–1204.

Hallin, M. (1978). Mixed autoregressive moving-average multivariate processes with timedependent coefficients. *Journal of Multivariate Analysis*, *8*, 567–572.

Hallin, M., Taniguchi, M., Serroukh, A., & Choy, K. (1999). Local asymptotic normality for regression models with long-memory disturbance. *The Annals of Statistics*, *27*(6), 2054–2080.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. New York: Wiley.

Hannan, E. J. (1970). *Multiple time series*. New York: Wiley. xi+536 pp.

Hannan, E. J. (1973). Central limit theorems for time series regression. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *26*, 157–170.

Hannan, E. J., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: Wiley.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190–195.

Härdle, W. (1990a). *Smoothing techniques: with implementation in S*. New York: Springer.

Härdle, W. (1990b). *Applied nonparametric regression*. New York: Cambridge University Press.

Härdle, W., Hall, P., & Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum (with discussion)? *Journal of the American Statistical Association*, *83*, 86–99.

Härdle, W., Hall, P., & Marron, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association*, *87*, 227–233.

Härdle, W., Kerkyacharian, G., Picard, D., & Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*. *Lecture notes in statistics*. New York: Springer.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, *53*, 173–188.

Harte, J., Kinzig, A., & Green, J. (1999). Self-similarity in the distribution and abundance of species. *Science*, *284*, 334–336.

Harvey, A. (1998). Long memory in stochastic volatility. In J. Knight & S. Satchell (Eds.), *Forecasting volatility in the financial markets* (pp. 307–320). Oxford: Butterworth-Heineman.

Harvey, A. C. (2007). Long memory in stochastic volatility. In J. Knight & S. Satchell (Eds.), *Forecasting volatility in the financial markets* (3rd ed., pp. 351–363). Oxford: Butterworth-Heineman.

Haslett, J., & Raftery, A. E. (1989). Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. *Applied Statistics*, *38*, 1–50.

Hassler, U. (2011). Estimation of fractional integration under temporal aggregation. *Journal of Econometrics*, *162*(2), 240–247.

Hassler, U., Marmol, F., & Velasco, C. (2006). Residual log-periodogram inference for longrun relationships. *Journal of Econometrics*, *130*(1), 165–207.

Hausdorff, F. (1918). Dimension und äusseres Mass. *Mathematische Annalen*, *79*(1–2), 157–179.

Hausdorff, J. M., Mitchell, S. L., Firtion, R., Peng, C. K., Cudkowicz, M. E., Wei, J. Y., & Goldberger, A. L. (1997). Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease. *Journal of Applied Physiology*, *82*, 262–269.

Hausdorff, J. M., Lertratanakul, A., Cudkowicz, M. E., Peterson, A. L., Kaliton, D., & Goldberger, A. L. (2000). Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. *Journal of Applied Physiology*, *88*, 2045–2053.

Hauser, M. A. (1999). Maximum likelihood estimators for ARMA and ARFIMA models: a Monte Carlo study. *Journal of Statistical Planning and Inference*, *80*, 229–255.

Hauser, M. A., & Kunst, R. M. (2001). Forecasting high-frequency financial data with the ARFIMA-ARCH model. *International Journal of Forecasting*, *20*(7), 501–518.

Häusler, E., & Teugels, J. L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *The Annals of Statistics*, *13*(2), 743–756.

Heath, D., Resnick, S., & Samorodnitsky, G. (1998). Heavy tails and long range dependence in on/off processes and associated circuit models. *Mathematics of Operations Research*, *23*, 145–165.

Heck, A. & Pedang, J. M. (Eds.) (1991). *Applying fractals in astronomy*. Berlin: Springer.

Heil, C. E., & Walnut, D. F. (1989). Continuous and discrete wavelet transforms. *SIAM Review*, *31*, 628–666.

Heiler, S., & Feng, Y. (1998). A simple root n bandwidth selector for nonparametric regression. *Journal of Nonparametric Statistics*, *9*, 1–21.

Henry, M. (2007). Bandwidth choice, optimal rates and adaptivity in semiparametric estimation of long memory. In G. Teyssière & A. M. Kirman (Eds.), *Long memory in economics* (pp. 157–172). Berlin: Springer.

Henry, M., & Robinson, P. M. (1996). Bandwidth choice in Gaussian semiparametric estimation of long range dependence. In P. M. Robinson & M. Rosenblatt (Eds.), *Lecture notes in statistics: Vol. 115. Athens conference on applied probability and time series* (pp. 220–232). New York: Springer.

Henry, M., & Zaffaroni, P. (2003). The long range dependence paradigm for macroeconomics and finance. In P. Doukhan, G. Oppenheim, & M. Taqqu (Eds.), *The theory and applications of long-range dependence*, Boston: Birkhäuser.

Hernandez-Campos, F., Marron, J. S., Samorodnitsky, G., & Smith, F. D. (2002). Variable heavy tailed durations in Internet traffic: part I, understanding heavy tails. In *Proc. of the 10th IEEE intl. sympos. on modeling, analysis and simulation of computer and telecommunications systems (MASCOTS 2002), IEEE* (pp. 43–50).

Herrmann, E., Gasser, T., & Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, *79*, 783–795.

Heyde, C. C., & Dai, W. (1996). On the robustness to small trens od estimation based on the smoothed periodogram. *Journal of Time Series Analysis*, *17*, 141–150.

Heyde, C. C., & Gay, R. (1989). On asymptotic quasi-likelihood estimation. *Stochastic Processes and Their Applications*, *31*(2), 223–236.

Heyde, C. C., & Gay, R. (1993). Smoothed periodogram asymptotics and estimation for processes and fields with possible long range dependence. *Stochastic Processes and Their Applications*, *45*, 169–182.

Heyde, C. C., & Yang, Y. (1997). On defining long-range dependence. *Journal of Applied Probability*, *34*(4), 939–944.

Hidalgo, J. (1997). Non-parametric estimation with strongly dependent multivariate time series. *Journal of Time Series Analysis*, *18*(2), 95–122.

Hidalgo, J. (2005). Semiparametric estimation for stationary processes whose spectra have an unknown pole. *The Annals of Statistics*, *33*(4), 1843–1889.

Hidalgo, J., & Robinson, P. M. (1996). Testing for structural change in a long-memory environment. *Journal of Econometrics*, *70*, 159–174.

Hidalgo, J., & Soulier, P. (2004). Estimation of the location and exponent of the spectral singularity of a long memory process. *Journal of Time Series Analysis*, *25*, 55–81.

Hidalgo, J., & Yajima, Y. (2002). Prediction and signal extraction of strongly dependent processes in the frequency domain. *Econometric Theory*, *18*, 584–624.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, *3*, 1163–1174.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, *57*, 1–17.

Ho, H. C. (1996). On central and non-central limit theorems in density estimation for sequences of long-range dependence. *Stochastic Processes and Their Applications*, *63*, 153–174.

Ho, H. C. (2002). On functionals of linear processes with estimated parameters. *Statistica Sinica*, *12*(4), 1171–1190.

Ho, H. C., & Hsing, T. (1996). On the asymptotic expansion of the empirical process of long-memory moving averages. *The Annals of Statistics*, *24*(3), 992–1024.

Ho, H. C., & Hsing, T. (1997). Limit theorems for functionals of moving averages. *Annals of Probability*, *25*(4), 1636–1669.

Ho, H. C., & Sun, T. C. (1987). A central limit theorem for noninstantaneous filters of a stationary Gaussian process. *Journal of Multivariate Analysis*, *22*(1), 144–155.

Ho, H. C., & Sun, T. C. (1990). Limiting distributions of nonlinear vector functions of stationary Gaussian processes. *Annals of Probability*, *18*(3), 1159–1173.

Honda, T. (2000). Nonparametric density estimation for a long-range dependent linear process. *Annals of the Institute of Statistical Mathematics*, *52*(4), 599–611.

Horváth, L. (1993). Change in autoregressive processes. *Stochastic Processes and Their Applications*, *44*, 221–242.

Horváth, L., & Kokoszka, P. (1997). The effect of long-range dependence on change-point estimators. *Journal of Statistical Planning and Inference*, *64*(1), 57–81.

Horváth, L., & Kokoszka, P. (2008). Sample autocovariances of long-memory time series. *Bernoulli*, *14*(2), 405–418.

Horváth, L., & Shao, Q.-M. (1999). Limit theorems for quadratic forms with applications to Whittle's estimate. *The Annals of Applied Probability*, *9*(1), 146–187.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, *68*, 165–176.

Hosking, J. R. M. (1996). Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long-memory time series. *Journal of Econometrics*, *73*, 261–284.

Hosoya, Y. (1974). *Estimation problems on stationary time series models*. Ph.D. thesis, Yale.

Hosoya, Y. (1997). A limit theory for long-range dependence and statistical inference on related models. *The Annals of Statistics*, *25*(1), 105–137.

Hosoya, Y., & Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics*, *10*(1), 132–153.

Houdré, C. (1994). Wavelets, probability, and statistics: some bridges. In *Wavelets: mathematics and applications. Stud. adv. math.* (pp. 365–398). Boca Raton: CRC Press.

Houdré, C., & Kawai R, R. (2006). On fractional tempered stable motion. *Stochastic Processes and Their Applications*, *116*(8), 1161–1184.

Hristopulos, D. T. (2002). New anisotropic covariance models and estimation of anisotropic parameters based on the covariance tensor identity. *Stochastic Environmental Research and Risk Assessment*, *16*(1), 43–62.

Hsieh, M.-C., Hurvich, C. M., & Soulier, P. (2007). Asymptotics for duration-driven long range dependent processes. *Journal of Econometrics*, *141*(2), 913–949.

Hsing, T. (1991). On tail estimation using dependent data. *The Annals of Statistics*, *19*, 1547–1569.

Hsing, T. (1999). On the asymptotic distributions of partial sums of functionals of infinite-variance moving averages. *Annals of Probability*, *27*(3), 1579–1599.

Hsing, T. (2000). Linear processes, long-range dependence and asymptotic expansions. *Statistical Inference for Stochastic Processes*, *3*(1–2), 19–29 (English summary). 19th "Rencontres Franco–Belges de statisticiens" (Marseille, 1998).

Hu, Y. (2005). Integral transformations and anticipative calculus for fractional Brownian motions. *Memoirs of the American Mathematical Society*, *175*, 825.

Hu, Y., & Nualart, D. (2010). Parameter estimation for fractional Ornstein–Uhlenbeck processes. *Statistics & Probability Letters*, *80*(11–12), 1030–1038.

Hu, Y., & Øksendal, B. (2003). Fractional white noise calculus and applications to finance. *Infinite Dimensional Analysis, Quantum Probability, and Related Topics*, *6*, 1–32.

Hualde, J., & Robinson, P. M. (2007). Root-n-consistent estimation of weak fractional cointegration. *Journal of Econometrics*, *140*, 450–484.

Hualde, J., & Robinson, P. M. (2010). Semiparametric inference in multivariate fractionally cointegrated systems. *Journal of Econometrics*, *157*(2), 492–511.

Hualde, J., & Velasco, C. (2008). Distribution-free test of fractional cointegration. *Econometric Theory*, *24*, 216–255.

Huang, D., & Anh, V. V. (1992). Estimation of spatial ARMA models. *Australian Journal of Statistics*, *34*, 513–530.

Huber, P. (1981). *Robust statistics*. New York: Wiley.

Hurst, H. E. (1951). Long term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, *116*, 770–779.

Hurst, H. E., Black, R. P., & Simaika, Y. M. (1965). *Long-term storage: an experimental study*. London: Constable Press.

Hurvich, C. M. (2001). Model selection for broadband semiparametric estimation of long memory in time series. *Journal of Time Series Analysis*, *22*, 679–709.

Hurvich, C. M. (2002). Multistep forecasting of long memory series using fractional exponential models. *International Journal of Forecasting*, *18*, 167–179.

Hurvich, C. M., & Beltrao, K. I. (1993). Asymptotics for the low-frequency ordinates of the periodogram of a long-memory time series. *Journal of Time Series Analysis*, *14*, 455–472.

Hurvich, C. M., & Beltrao, K. I. (1994a). Automatic semiparametric estimation of the memory parameter of a long memory time series. *Journal of Time Series Analysis*, *15*, 285–302.

Hurvich, C. M., & Beltrao, K. I. (1994b). Acknowledgement of priority for "Asymptotics for the low-frequency ordinates of the periodogram of a long-memory time series." *Journal of Time Series Analysis*, *15*, 64.

Hurvich, C. M., & Brodsky, J. (2001). Broadband semiparametric estimation of the memory parameter of a long-memory time series using fractional exponential models. *Journal of Time Series Analysis*, *22*, 221–249.

Hurvich, C. M., & Chen, W. W. (2000). An effcient taper for potentially overdifferenced long-memory time series. *Journal of Time Series Analysis*, *21*(2), 155–180.

Hurvich, C. M., & Deo, R. (1999). Plug-in selection of the number of frequencies in regression estimates of the memory parameter of a long memory time series. *Journal of Time Series Analysis*, *20*, 331–341.

Hurvich, C. M., & Ray, B. K. (1995). Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis*, *16*, 17–42.

Hurvich, C. M., & Soulier, P. (2002). Testing for long memory in volatility. *Econometric Theory*, *18*(6), 1291–1308.

Hurvich, C. M., Deo, R., & Brodsky, J. (1998). The mean squared error of Geweke and Porter–Hudak's estimator of a long memory time series. *Journal of Time Series Analysis*, *19*, 19–46.

Hurvich, C. M., Moulines, E., & Soulier, P. (2002). The FEXP estimator for potentially non-stationary linear time series. *Stochastic Processes and Their Applications*, *97*, 307–340.

Hurvich, C. M., Lang, G., & Soulier, P. (2005a). Estimation of long memory in the presence of a smooth nonparametric trend. *Journal of the American Statistical Association*, *100*(471), 853–871.

Hurvich, C. M., Moulines, E., & Soulier, P. (2005b). Estimating long memory in volatility. *Econometrica*, *73*(4), 1283–1328.

Hwang, S. (2000). The effects of systematic sampling and temporal aggregation on discrete time long memory processes and their finite sample properties. *Econometric Theory*, *16*, 347–372.

Igloi, E., & Terdik, G. (1999). Long-range dependence through gamma-mixed Ornstein–Uhlenbeck process. *Electronic Journal of Probability*, *4*(16), 1–33.

Imbrie, J., & Newman, C. (1988). An intermediate phase with slow decay of correlations in one dimensional $1/|x-y|^2$ percolation, Ising and Potts models. *Communications in Mathematical Physics*, *118*, 303–336.

Ising, E. (1924). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, *31*, 253.

Istas, J., & Lang, G. (1997). Quadratic variations and estimation of the local Hölder index of a Gaussian process. *Annales de L'Institut Henri Poincare, Probabilites Et Statistiques*, *33*, 407–436.

Jach, A., & Kokoszka, P. (2008). Wavelet-domain test for long-range dependence in the presence of a trend. *Statistics: A Journal of Theoretical and Applied Statistics*, *42*, 101–113.

Jach, A., McElroy, T., & Politis, D. N. (2012). Subsampling inference for the mean of heavy-tailed long memory time series. *Journal of Time Series Analysis*, *33*, 96–111.

Jackson, D. (1941). *Fourier series and orthogonal polynomials*. New York: Dover.

Jackson, D. (2004). *Fourier series and orthogonal polynomials*. New York: Dover.

Járai, A. A. (2003). Invasion percolation and the incipient infinite cluster in 2D. *Journal of Communications in Mathematical Physics*, *236*(2), 311–334.

Jasiak, J. (1998). Persistence in intratrade durations. *Finance*, *19*, 166–195.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

Jeffreys, H. (1948). *Theory of probability*. Oxford: Clarendon Press.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Clarendon Press.

Jeganathan, P. (1999). On asymptotic inference in cointegrated time series with fractionally integrated errors. *Econometric Theory*, *15*, 583–621.

Jelenkovič, P. R., & Lazar, A. A. (1999). Asymptotic results for multiplexing subexponential on–off processes. *Advances in Applied Probability*, *31*(2), 394–421.

Joerges, J., Küttner, A., Galizia, C. G., & Menzel, R. (1997). Representations of odours and odour mixtures visualized in the honeybee brain. *Nature*, *387*, 285–288.

Johansen, S. (1996). Likelihood-based inference. In *Cointegrated vector autoregressive models* (2nd ed.). Oxford: Oxford University Press.

Johansen, S. (2008). Representation of cointegrated autoregressive processes with application to fractional processes. *Econometric Reviews*, *28*(1–3), 121–145.

Johansen, S. (2008). A representation theory for a class of vector autoregressive models for fractional processes. *Econometric Theory*, *24*, 651–676.

Johansen, S. (2010a). *The analysis of nonstationary time series using regression, correlation and cointegration—with an application to annual mean temperature and sea level* (Discussion Paper). Department of Economics, University of Copenhagen.

Johansen, S. (2010b). *An extension of cointegration to fractional autoregressive processes* (Discussion Paper). Department of Economics, University of Copenhagen.

Johansen, S., & Nielsen, M. O. (2010a). *Likelihood interference for a vector autoregressive model which allows for fractional and cofractional processes* (Discussion Paper). Department of Economics, University of Copenhagen.

Johansen, S., & Nielsen, M. Ø. (2010b). Likelihood inference for a nonstationary fractional autoregressive model. *Journal of Econometrics*, *158*, 51–66.

Johnsen, S. J., Clausen, H. B., Dansgaard, W., Gundestrup, N. S., Hammer, C. U., Andersen, U., Andersen, K. K., Hvidberg, C. S., Dahl-Jensen, D., Steffensen, J. P., Shoji, H., Sveinbjörnsdóttir, A. E., White, J., Jouzel, J., & Fisher, D. (1997). The $\delta^{18}o$ record along the Greenland ice core project deep ice core and the problem of possible Eemian climatic instability. *Journal of Geophysical Research*, *102*, 26397–26410.

Johnson, R. A., & Bagshaw, M. (1974). The effect of serial correlation on the performance of CUSUM tests. *Technometries*, *16*, 103–112.

Johnstone, I. M. (1999). Wavelet threshold estimators for correlated data and inverse problems: adaptivity results. *Statistica Sinica*, *9*, 51–83.

Johnstone, I. M., & Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, *59*, 319–351.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*, 401–407.

Kaj, I. (2002). *Stochastic modeling in broadband communications systems*. *SIAM monographs on mathematical modeling and computation*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM). xvi+177 pp.

Kallenberg, O. (1997). *Foundations of modern probability Probability and its applications (New York)*. New York: Springer. xii+523 pp.

Kalliorasa, A. G., Koutrouvelis, I. A., & Canavos, G. C. (2006). Testing the fit of gamma distributions using the empirical moment generating function. *Communications in Statistics, Theory and Methods*, *35*(3), 527–540.

Kanter, M., & Steiger, W. L. (1974). Regression and autoregression with infinite variance. *Advances in Applied Probability*, *6*, 768–783.

Kanwal, R. P. (2004). *Generalized functions: theory and applications* (3rd ed.). Boston: Birkhäuser.

Karagiannis, T., Molle, M., & Faloutsos, M. (2004). Long range dependence—ten years of Internet traffic modeling. *IEEE Internet Computing*, *8*, 57–64.

Karamata, J. (1930a). Sur un mode de croissance régulière des fonctions. *Mathematica (Cluj)*, *4*, 38–53.

Karamata, J. (1930b). Sur certains "Tauberian theorems" de M. M. Hardy et Littlewood. *Mathematica (Cluj)*, *3*, 33–48.

Karamata, J. (1933). Sur un mode de croissance régulière. Théorèmes fondamentaux. *Bulletin de la Société Mathématique de France*, *61*, 55–62.

Kasahara, Y., & Maejima, M. (1986). Functional limit theorems for weighted sums of i.i.d. random variables. *Probability Theory and Related Fields*, *72*(2), 161–183.

Kasahara, Y., & Maejima, M. (1988). Weighted sums of i.i.d. random variables attracted to integrals of stable processes. *Probability Theory and Related Fields*, *78*(1), 75–96.

Kasahara, Y., Maejima, M., & Vervaat, W. (1988). Logfractional stable processes. *Stochastic Processes and Their Applications*, *30*, 329–339.

Kashyap, R. L. (1984). Characterization and estimation of two-dimensional ARMA models. *IEEE Transactions on Information Theory*, *IT-30*, 736–745.

Kaufman, B., & Onsager, L. (1949). Crystal statistics III: short-range order in a binary Ising lattice. *Physical Review*, *76*, 1244–1252.

Kazakevičius, V., & Leipus, R. (2002). On stationarity in the ARCH(∞) model. *Econometric Theory*, *18*, 1–16.

Kazakevičius, V., & Leipus, R. (2003). A new theorem on the existence of invariant distributions with applications to ARCH processes. *Journal of Applied Probability*, *40*(1), 147–162.

Kazakevičius, V., Leipus, R., & Viano, M.-C. (2004). Stability of random coefficient ARCH models and aggregation schemes. *Journal of Econometrics*, *120*(1), 139–158.

Kazmin, Yu. A. (1969a). On expansions in series of Appell polynomials. *Matematičeskie Zametki*, *5*(5), 509–520.

Kazmin, Yu. A. (1969b). On Appell polynomials. *Mathematical Notes*, *6*(2), 556–562.

Kelbert, M., Leonenko, N. N., & Ruiz-Medina, M. D. (2005). Fractional random fields associated with stochastic fractional heat equations. *Advances in Applied Probability*, *37*, 108–133.

Kent, J. T., & Wood, A. T. A. (1997). Estimating the fractal dimension of a locally self-similar Gaussian process by using increments. *Journal of the Royal Statistical Society, Series B*, *59*, 679–699.

Kerkyacahrian, G., & Picard, D. (2000). Minimax or Maxisets? Prépublication PMA-556, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII.

Kesten, H. (1982). *Progress in probability and statistics: Vol. 2. Percolation theory for mathematicians*. Boston: Birkhäuser.

Kim, Y. M., & Nordman, D. J. (2011) Properties of a block bootstrap under long-range dependence. *Sankhya A*, *73*(1), 79–109.

Kim, C. S., & Phillips, P. C. B. (1999). *Log periodogram regression: the nonstationary case* (Mimeo). Cowles Foundation, Yale University.

Kim, C., & Phillips, P. (2001). *Fully modified estimation of fractional cointegration models*. Preprint, Yale University.

Kirman, A., & Teyssière, G. (2002). Microeconomic models for long memory in the volatility of financial time series. *Studies in Nonlinear Dynamics & Econometrics*, *5*(4), 40–61.

Klemes, V. (1974). The Hurst phenomenon: a puzzle? *Water Resources Research*, *10*, 675–688.

Kleptsyna, M. L., & Le Breton, A. (2002). Statistical analysis of the fractional Ornstein–Uhlenbeck type process. *Statistical Inference for Stochastic Processes*, *5*, 229–248.

Kleptsyna, M. L., Le Breton, A., & Roubaud, M.-C. (2000). Parameter estimation and optimal filtering for fractional type stochastic systems. *Statistical Inference for Stochastic Processes*, *3*, 173–182.

Klüppelberg, C., & Kühn, C. (2004). Fractional Brownian motion as a weak limit of Poisson shot noise processes—with applications to finance. *Stochastic Processes and Their Applications*, *113*(2), 333–351.

Klüppelberg, C., & Mikosch, T. (1996). Gaussian limit fields for the integrated periodogram. *The Annals of Applied Probability*, *6*(3), 969–991.

Klüppelberg, C., Mikosch, T., & Schärf, A. (2003). Regular variation in the mean and stable limits for Poisson shot noise. *Bernoulli*, *9*(3), 467–496 (English summary).

Kokoszka, P. S. (1996). Prediction of infinite variance fractional ARIMA. *Probability and Mathematical Statistics*, *16*, 65–83.

Kokoszka, P. S., & Leipus, R. (2000). Change-point estimation in ARCH models. *Bernoulli*, *6*, 1–28.

Kokoszka, P. S., & Mikosch, T. (1997). The integrated periodogram for long-memory processes with finite or infinite variance. *Stochastic Processes and Their Applications*, *66*(1), 55–78.

Kokoszka, P. S., & Taqqu, M. S. (1993). Asymptotic dependence of moving average type self-similar stable random fields. *Nagoya Mathematical Journal*, *130*, 85–100.

Kokoszka, P. S., & Taqqu, M. S. (1995a). Fractional ARIMA with stable innovations. *Stochastic Processes and Their Applications*, *60*, 19–47.

Kokoszka, P. S., & Taqqu, M. S. (1995b). Infinite variance stable moving averages with long memory. *Journal of Econometrics*, *73*, 79–99.

Kokoszka, P. S., & Taqqu, M. S. (1996). Parameter estimation for infinite variance fractional ARIMA. *The Annals of Statistics*, *24*(5), 1880–1913.

Kokoszka, P. S., & Taqqu, M. S. (1997). The asymptotic behavior of quadratic forms in heavy-tailed strongly dependent random variables. *Stochastic Processes and Their Applications*, *66*(1), 21–40.

Kokoszka, P. S., & Taqqu, M. S. (1999). Discrete time parametric models with long memory and infinite variance. *Mathematical and Computer Modelling*, *29*(10–12), 203–215.

Kokoszka, P. S., & Taqqu, M. S. (2001). Can one use the Durbin–Levinson algorithm to generate infinite variance fractional ARIMA time series? *Journal of Time Series Analysis*, *22*, 317–337.

Kolmogorov, A. N. (1937). Zur Umkehrbarkeit der statistischen Naturgesetze. *Mathematische Annalen*, *113*, 766–772.

Kolmogorov, A. N. (1940). Wienersche Spiralen und einige andere interessante Kurven in Hilbertschen Raum. *Comptes Rendus (Doklady) Academy of Sciences of the USSR (N.S.)*, *26*, 115–118.

Kolmogorov, A. N. (1941). Local structure of turbulence in fluid for very large Reynolds numbers. In S. K. Friedlander & L. Topper (Eds.), *Transl. in turbulence* (pp. 151–155). New York: Interscience. 1961.

Kôno, N. (1986). Hausdorff dimension of sample paths for self-similar processes. In E. Eberlein & M. S. Taqqu (Eds.), *Dependence in probability and statistics* (pp. 109–117). Boston: Birkhäuser.

Konstantopoulos, T., & Lin, S.-J. (1998). Macroscopic models for long-range dependent network traffic. *Queueing Systems, Theory and Applications*, *28*(1–3), 215–243.

Kosterlitz, J. M., & Thouless, D. J. (1978). *Progress in low temperature physics: Vol. VIIB. Two-dimensional physics* (p. 371). Amsterdam: North-Holland.

Koul, H. L. (1992). M-estimators in linear models with long range dependent errors. *Statistics & Probability Letters*, *14*, 153–164.

Koul, H. L., & Mukherjee, K. (1993). Asymptotics of R-, MD- and LAD-estimators in linear regression models with long range dependent errors. *Probability Theory and Related Fields*, *95*, 535–553.

Koul, H. L., & Surgailis, D. (1997). Asymptotic expansion of M-estimators with long memory errors. *The Annals of Statistics*, *25*, 818–850.

Koul, H. L., & Surgailis, D. (2000). Second-order behavior of M-estimators in linear regression with long-memory errors. *Journal of Statistical Planning and Inference*, *91*(2), 399–412.

Koul, H. L., & Surgailis, D. (2001). Asymptotics of empirical processes of long memory moving averages with infinite variance. *Stochastic Processes and Their Applications*, *91*(2), 309–336.

Koul, H. L., Baillie, R. T., & Surgailis, D. (2004). Regression model fitting with a long memory covariate process. *Econometric Theory*, *20*(3), 485–512.

Krämer, W., & Sibbertsen, P. (2000). Testing for structural change in the presence of long memory. *International Journal of Business and Economics*, *1*, 235–242.

Krämer, W., & Sibbertsen, P. (2003). Testing for structural change in the presence of long memory. *International Journal of Business and Economics*, *1*, 235–243.

Krämer, W., Sibbertsen, P., & Kleiber, C. (2002). Long memory versus structural change in financial time series. *Allgemeines Statistisches Archiv*, *86*, 83–96.

Krengel, U. (1985) *De Gruyter studies in mathematics: Vol. 6. Ergodic theorems*.

Kulik, R. (2008a). Sums of extreme values of subordinated long-range dependent sequences: moving averages with finite variance. *Electronic Journal of Probability*, *13*, 961–979.

Kulik, R. (2008b). Nonparametric deconvolution problem for dependent sequences. *Electronic Journal of Statistics*, *2*, 722–740.

Kulik, R. (2009). Empirical process of long-range dependent sequences when parameters are estimated. *Journal of Statistical Planning and Inference*, *139*(2), 287–294.

Kulik, R., & Lorek, P. (2011). Some results on random design regression with long memory errors and predictors. *Journal of Statistical Planning and Inference*, *141*(1), 508–523.

Kulik, R., & Lorek, P. (2012). Empirical process of residuals for regression models with long memory errors. Preprint.

Kulik, R., & Ould Haye, M. (2008). Trimmed sums of long range dependent moving averages. *Statistics & Probability Letters*, *78*(15), 2536–2542.

Kulik, R., & Raimondo, M. (2009a). L^p wavelet regression with correlated errors and inverse problems. *Statistica Sinica*, *19*, 1479–1489.

Kulik, R., & Raimondo, M. (2009b). Wavelet regression in random design with heteroscedastic dependent errors. *The Annals of Statistics*, *37*, 3396–3430.

Kulik, R., & Soulier, P. (2011). The tail empirical process for long memory stochastic volatility sequences. *Stochastic Processes and Their Applications*, *121*(1), 109–134.

Kulik, R., & Soulier, P. (2012). Limit theorems for long memory stochastic volatility models with infinite variance: partial sums and sample covariances. *Advances in Applied Probability*, *44*(4), 1113–1141.

Kulik, R., & Soulier, P. (2013, in press). Estimation of limiting conditional distributions for the heavy tailed long memory stochastic volatility process. *Extremes*.

Kulik, R., & Szekli, R. (2001). Sufficient conditions for long range count dependence of stationary point processes on the real line. *Journal of Applied Probability*, *38*, 570–581.

Kulik, R., & Wichelhaus, C. (2011). Nonparametric conditional variance and error density estimation in regression models with long memory. *Electronic Journal of Statistics*, *5*, 856–898.

Kulik, R., & Wichelhaus, C. (2012). Conditional variance estimation in regression with long memory. *Journal of Time Series Analysis*, *33*(3), 468–483.

Künsch, H. (1980). *Reellwertige Zufallsfelder auf einem Gitter: Interpolationsprobleme, Variationsprinzip und statistische Analyse*. Ph.D. thesis, ETH Zurich.

Künsch, H. R. (1986). Discrimination between monotonic trends and long-range dependence. *Journal of Applied Probability*, *23*, 1025–1030.

Künsch, H. R. (1987). Statistical aspects of self-similar processes. In Yu. Prohorov & V. V. Sazanov (Eds.), *Proc. 1st world congress of the Bernoulli society* (Vol. 1, pp. 67–74). Utrecht: VNU Science Press.

Künsch, H. R. (1989). The jackknife and bootstrap for general stationary observations. *The Annals of Statistics*, *17*, 1217–1261.

Künsch, H. R., Beran, J., & Hampel, F. (1993). Contrasts under long-range correlations. *The Annals of Statistics*, *21*(2), 943–964.

Kuswanto, H. (2011). A new simple test against spurious long memory using temporal aggregation. *Journal of Statistical Computation and Simulation*, *81*(10), 1297–1311.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*, 159–178.

Lahiri, S. N. (1993). On the moving block bootstrap under long range dependence. *Statistics & Probability Letters*, *18*, 405–413.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. New York: Springer.

Lamperti, J. W. (1962). Semi-stable stochastic processes. *Translations - American Mathematical Society*, *104*, 62–78.

Lamperti, J. W. (1972). Semi-stable Markov processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *22*, 205–225.

Lanford, O. E., & Ruelle, D. (1968). Observables at infinity and states with short-range correlations in statistical mechanics. *Communications in Mathematical Physics*, *13*(3), 194–215.

Lanford, O. E., & Ruelle, D. (1969). Observables at infinity and states with short range correlations in statistical mechanics. *Communications in Mathematical Physics*, *13*, 194–215.

Lang, G., & Soulier, P. (2000). Convergence de mesures spectrales aléatoires et applications à des principes d'invariance. (French) [Convergence of random spectral measures and applications to invariance principles] 19th "Rencontres Franco–Belges de statisticiens" (Marseille, 1998). *Statistical Inference for Stochastic Processes*, *3*(1–2), 41–51.

Lasak, K. (2010). Likelihood based testing for no fractional cointegration. *Journal of Econometrics*, *158*(1), 67–77.

Lavancier, F. (2006). Long memory random fields. In P. Doukhan, P. Bertail, & P. Soulier (Eds.), *Lecture notes in statistics: Vol. 187. Dependence in probability and statistics* (pp. 195–220). New York: Springer.

Lavancier, F. (2007). Invariance principles for non-isotropic long memory random fields. *Statistical Inference for Stochastic Processes*, *10*(3), 255–282.

Lavielle, M., & Ludena, C. (2000). The multiple change-points problem for the spectral distribution. *Bernoulli*, *6*(5), 845–869.

Lazarova, S. (2005). Testing for structural change in regression with long memory processes. *Journal of Econometrics*, *129*(1–2), 329–372.

Le Breton, A. (1998). Filtering and parameter estimation in a simple linear model driven by a fractional Brownian motion. *Statistics & Probability Letters*, *38*(3), 263–274.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1978). Conditions for the convergence in distribution of maxima of stationary normal processes. *Stochastic Processes and Their Applications*, *8*(2), 131–139.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. *Springer series in statistics*. New York: Springer.

Lee, S.-W., & Hansen, B. E. (1994). Asymptotic theory for the GARCH(1, 1) quasi-maximum likelihood estimator. *Econometric Theory*, *10*, 29–52.

Lee, D., & Schmidt, P. (1996). On the power of the KPSS test of stationarity against fractionally-integrated alternatives. *Journal of Econometrics*, *73*, 285–302.

Leipus, R. (1988). Weak convergence of two-parameter empirical fields in the change-point. *Lithuanian Mathematical Journal*, *28*, 348–352.

Leipus, R., & Surgailis, D. (2007). On long-range dependence in regenerative processes based on a general ON/OFF scheme. *Journal of Applied Probability*, *44*(2), 379–392.

Leipus, R., & Viano, M.-C. (2002). Aggregation in ARCH models. *Lithuanian Mathematical Journal*, *42*, 68–89.

Leipus, R., Oppenheim, G., Philippe, A., & Viano, M.-C. (2006). Orthogonal series density estimation in a disaggregation scheme. *Journal of Statistical Planning and Inference*, *136*(8), 2547–2571.

Lejeune, M. (1985). Estimation non-paramétrique par noyaux: régression polynômiale mobile. *Revue de Statistiques Appliqées*, *33*, 43–68.

Lejeune, M., & Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis*, *14*, 457–471.

Leland, W. E., Taqqu, M. S., Willinger, W., & Wilson, D. V. (1993a). Statistical analysis of high-time resolution Ethernet LAN traffic measurements. In M. E. Tarter & M. D. Lock (Eds.), *Computing science and statistics: Vol. 25. Statistical applications of expanding computer capabilities. Proceedings of the 25th symposium on the interface between statistics and computer science* (pp. 146–155).

Leland, W. E., Taqqu, M. S., Willinger, W., & Wilson, D. V. (1993b). On the self-similar nature of ethernet traffic. In *Proc. ACM SIGCOMM 1993*, San Francisco, CA (pp. 183–193).

Leland, W. E., Taqqu, M. S., Willinger, W., & Wilson, D. V. (1994). On the self-similar nature (extended version). *IEEE/ACM Transactions on Networking*, *2*(1), 1–15.

Leonenko, N., & Ruiz-Medina, M. (2006). Scaling laws for the multidimensional Burgers equation with quadratic external potential. *Journal of Statistical Physics*, *124*(1), 191–205.

Leonenko, N., & Sakhno, L. (2006). On the Whittle estimators for some classes of continuous-parameter random processes and fields. *Statistics & Probability Letters*, *76*, 781–795.

Leonenko, N., & Taufer, E. (2005). Convergence of integrated superpositions of Ornstein–Uhlenbeck processes to fractional Brownian motion. *Stochastics: An International Journal of Probability and Stochastic Processes*, *77*(6), 477–499.

Leonenko, N. N., Sakhno, L., & Taufer, E. (2001). On Kaplan–Meier estimator of long-range dependent sequences. *Statistical Inference for Stochastic Processes*, *4*(1), 17–40.

Leonenko, N. N., Sakhno, L., & Taufer, E. (2002). Product-limit estimator for long- and short-range dependent sequences under gamma type subordination. *Random Operators and Stochastic Equations*, *10*(4), 301–320 (B).

Lévy, P. (1938). *Plane or space curves and surfaces consisting of parts similar to the whole*. Reading: Addison-Wesley. Reprinted in: Classics on Fractals, G. A. Edgar (Ed.) (1993).

Lévy, P. (1953). Random functions: general theory with special reference to Laplacian random functions. In *Univ. Calif. Publ. in Statist.* (Vol. 1, pp. 331-390).

Levy, J. B., & Taqqu, M. S. (1986). Using renewal processes to generate long-range dependence and high variability. In *Progr. probab. statist.: Vol. 11. Dependence in probability and statistics*, Oberwolfach, 1985 (pp. 73–89). Boston: Birkhäuser Boston.

Levy, J. B., & Taqqu, M. S. (1987). On renewal processes having stable inter-renewal intervals and stable rewards. *Annales Des Sciences Mathématiques Du Québec*, *11*(1), 95–110.

Levy, J. B., & Taqqu, M. S. (2000). Renewal reward processes with heavy-tailed inter-renewal times and heavy-tailed rewards. *Bernoulli*, *6*(1), 23–44.

Levy, J. B., & Taqqu, M. S. (2001). Dependence structure of a renewal-reward process with infinite variance. *Fractals*, *9*(2), 185–192.

Levy, J. B., & Taqqu, M. S. (2005). The asymptotic codifference and covariation of log fractional stable noise. Preprint.

Li, L., & Xiao, Y. (2007). Mean integrated squared error of nonlinear wavelet-based estimators with long memory data. *Annals of the Institute of Statistical Mathematics*, *59*, 299–324.

Lieberman, O. (2001). Penalised maximum likelihood estimation for fractional Gaussian processes. *Biometrika*, *88*(3), 888–894.

Lieberman, O., & Phillips, P. C. B. (2004). Expansions for the distribution of the maximum likelihood estimator of the fractional difference parameter. *Econometric Theory*, *20*, 464–484.

Lieberman, O., Rousseau, J., & Zucker, D. (2000). Small-sample likelihood-based inference in the ARFIMA model. *Econometric Theory*, *16*(2), 231–248.

Lieberman, O., Rousseau, J., & Zucker, D. M. (2001). Valid Edgeworth expansion for the sample autocorrelation function under long range dependence. *Econometric Theory*, *17*, 257–275.

Lieberman, O., Rousseau, J., & Zucker, D. (2003). Valid asymptotic expansions for the maximum likelihood estimator of the parameter of a stationary, Gaussian, strongly dependent process. *The Annals of Statistics*, *31*(2), 586–612.

Liggett, T. M. (2004). *Interacting particle systems*. *Classics in mathematics*. Berlin: Springer.

Lighthill, M. J. (1958). *Cambridge monographs on mechanics*. *An introduction to fourier analysis and generalised functions*. Cambridge: Cambridge University Press.

Lighthill, M. J. (1962). *Introduction to Fourier analysis and generalised functions*. *Cambridge monographs on mechanics and applied mathematics*. Cambridge: Cambridge University Press.

Ling, S., & Li, W. K. (1997). On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *Journal of the American Statistical Association*, *92*, 1184–1194.

Lo, A. (1991). Long-term memory in stock market prices. *Econometrica*, *59*, 1279–1313.

Lobato, I. (1995). *Multivariate analysis of long memory series in the frequency domain*. Ph.D. Thesis, London School of Economics.

Lobato, I. N. (1997). Consistency of the averaged cross-periodogram in long memory series. *Journal of Time Series Analysis*, *18*, 137–155.

Lobato, I. N. (1999). A semiparametric two-step estimator in a multivariate long memory model. *Journal of Econometrics*, *90*, 129–153.

Lobato, I. N., & Robinson, P. M. (1996). Averaged periodogram estimation of long memory. *Journal of Econometrics*, *73*, 303–324.

Lobato, I. N., & Savin, N. E. (1998). Real and spurious long-memory properties of stock-market data. *Journal of Business & Economic Statistics*, *16*(3), 261–268.

Lowen, S. B., & Teich, M. C. (2005). *Fractal based point processes*. New York: Wiley.

Luceno, A. (1996). A fast likelihood approximation for vector general linear processes with long series: application to fractional differencing. *Biometrika*, *83*, 603–614.

Lumsdaine, R. (1996). Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1, 1) and covariance stationary GARCH(1, 1) models. *Econometrica*, *64*, 575–596.

Luo, L. (2011). *High quantile estimation for some stochastic volatility models*. M.Sc. thesis, University of Ottawa.

Lütkepohl, H. (2006). Structural vector autoregressive analysis for cointegrated variables. *AStA Advances in Statistical Analysis*, *90*(1), 75–88.

MacKinnon, J. G., & Nielsen, M. Ø. (2010). *Numerical distribution functions of fractional unit root and cointegration tests* (CREATES Research Paper 2010-59).

Madras, N., & Slade, G. (1996). *The self-avoiding walk*. Boston: Birkhäuser.

Maejima, M., & Yamamoto, K. (2003). Long-memory stable Ornstein–Uhlenbeck processes. *Electronic Journal of Probability*, *8*(19), 1–18.

Major, P. J. (1981). *Lecture notes in mathematics: Vol. 849*. *Multiple Wiener–Itô Integrals*. New York: Springer.

Makse, H. A., Havlin, S., Schwartz, M., & Stanley, H. E. (1996). Method for generating long-range correlations for large systems. *Physical Review E*, *53*(5), 5445–5449.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *11*, 674–693.
Malyshev, V. A., & Minlos, R. A. (1991). *Gibbs Random Fields*. Dordrecht: Kluwer Academic.
Man, K. S., & Tiao, G. C. (2006). Aggregation effect and forecasting temporal aggregates of long memory processes. *International Journal of Forecasting*, *22*(2), 267–281.
Man, K. S., & Tiao, G. C. (2009). ARFIMA approximation and forecasting of the limiting aggregate structure of long-memory process. *Journal of Forecasting*, *28*, 89–101.
Mandelbrot, B. B. (1965). Une classe de processus stochastiques homothétiques à soi; application à la loi climaologique de H. E. Hurst. *Comptes Rendus de L'Académie des Sciences de Paris*, *260*, 3274–3277.
Mandelbrot, B. B. (1967). How long is the coast of Britain? *Science*, *155*, 636.
Mandelbrot, B. B. (1969). Long-run linearity, locally Gaussian process, H-spectra and infinite variance. *International Economic Review*, *10*, 82–113.
Mandelbrot, B. B. (1971). When can price be arbitraged efficiently? A limit to the validity of the random walk and martingale models. *Reviews of Economics and Statistics*, *LIII*, 225–236.
Mandelbrot, B. B. (1975). Limit theorems on the self-normalized range for weakly and strongly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *31*(4), 271–285.
Mandelbrot, B. B. (1977). *Fractals: form, chance and dimension*. San Francisco: Freeman.
Mandelbrot, B. B. (1983). *The fractal geometry of nature*. San Francisco: Freeman.
Mandelbrot, B. B. (1997). *Fractals and scaling in finance: discontinuity, concentration, risk*. New York: Springer.
Mandelbrot, B. B. (1999). *Multifractals and* $1/f$ *noise: wild self-affinity in physics*. New York: Springer.
Mandelbrot, B. B. (2002). *Gaussian self-affinity and fractals. Globality, the earth,* $2/f$ *noise, and R/S*. New York: Springer.
Mandelbrot, B. B., & Taqqu, M. S. (1979). Robust R/S analysis of long-run serial correlation. Proc. of the 42nd session of the int. statistical institute, Manila, 1979. *Bulletin of the International Statistical Institute*, *48*(2), 69–104.
Mandelbrot, B. B., & van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, *10*(4), 422–437.
Mandelbrot, B. B., & Wallis, J. R. (1968a). Noah, Joseph and operational hydrology. *Water Resources Research*, *4*(5), 909–918.
Mandelbrot, B. B., & Wallis, J. R. (1968b). Robustness of the rescaled range R/S and the measurement of non-cyclic long-run statistical dependence. *Water Resources Research*, *5*, 967–988.
Mandelbrot, B. B., & Wallis, J. R. (1969a). Computer experiments with fractional Gaussian noises. *Water Resources Research*, *5*(1), 228–267.
Mandelbrot, B. B., & Wallis, J. R. (1969b). Some long-run properties of geophysical records. *Water Resources Research*, *5*, 321–340.
Mandelbrot, B. B., & Wallis, J. R. (1969c). Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. *Water Resources Research*, *5*, 967–988.
Manley, G. (1953). The mean temperature of Central England, 1698 to 1952. *Quarterly Journal of the Royal Meteorological Society*, *79*, 242–261.
Manley, G. (1974). Central England temperatures: monthly means 1659 to 1973. Q.J.R. *Quarterly Journal of the Royal Meteorological Society*, *100*, 389–405.
Mansfield, P., Rachev, S., & Samorodnitsky, G. (2001). Long strange segments of a stochastic process and long range dependence. *The Annals of Applied Probability*, *11*, 878–921.
Manstavičius, M. (2007). Hausdorff–Besicovitch dimension of graphs and p-variation of some Lévy processes. *Bernoulli*, *13*(1), 40–53.
Marinari, E., Parisi, G., Ruelle, D., & Widney, P. (1983). On the interpretation of $1/f$ noise. *Communications in Mathematical Physics*, *89*, 1–12.
Marinucci, D. (2000). Spectral regression for cointegrated time series with long-memory innovations. *Journal of Time Series Analysis*, *21*, 685–705.

Marinucci, D., & Robinson, P. M. (1999). Alternative forms of fractional Brownian motion. *Journal of Statistical Planning and Inference*, *80*(1–2), 111–122.
Marinucci, D., & Robinson, P. M. (2000). Weak convergence of multivariate fractional processes. *Stochastic Processes and Their Applications*, *86*, 103–120.
Marinucci, D., & Robinson, P. M. (2001). Semiparametric fractional cointegration analysis. *Journal of Econometrics*, *105*(1), 225–247.
Marmol, F. (1995). Spurious regressions between $I(d)$ processes. *Journal of Time Series Analysis*, *16*, 313–321.
Marmol, F., & Velasco, C. (2004). Consistent testing of cointegrating relationships. *Econometrica*, *72*, 1809–1844.
Marron, J. S. (1989). Automatic smoothing parameter selection. In A. Ullah (Ed.), *Semiparametric and nonparametric econometrics* (pp. 65–86). Heidelberg: Physica.
Martin, R. J. (1979). A subclass of lattice processes applied to a problem in planar sampling. *Biometrika*, *66*(2), 209–217.
Mason, D. (1982). Laws of large numbers of sums of extreme values. *The Annals of Probability*, *10*, 168–177.
Masry, E. (1993). The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion. *IEEE Transactions on Information Theory*, *39*(1), 260–264.
Masry, E. (2001). Local linear regression estimation under long-range dependence: strong consistency and rates. *IEEE Transactions on Information Theory*, *47*(7), 2863–2875.
Masry, E., & Mielniczuk, J. (1999). Local linear regression estimation for time series with long-range dependence. *Stochastic Processes and Their Applications*, *82*, 173–194.
Mathéron, G. (1962). *Traité de Géostatistique Appliquée*. *Cambridge philos. soc., Tome 1*. Paris: Editions Technip.
Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, *5*, 439–468.
Matsui, M., & Shieh, N.-R. (2009). On the exponentials of fractional Ornstein–Uhlenbeck processes. *Electronic Journal of Probability*, *14*(23), 594–611.
Maulik, K., Resnick, S., & Rootzen, H. (2002). Asymptotic independence and a network traffic model. *Journal of Applied Probability*, *39*(4), 671–699.
McCauley, J. L. (1993). *Chaos, dynamics and fractals*. Cambridge: Cambridge University Press.
McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall/CRC Press.
McElroy, T., & Politis, D. N. (2007). Self-normalization for heavy-tailed time series with long memory. *Statistica Sinica*, *17*(1), 199–220.
McKean, H. P. (1973). Geometry of differential space. *Annals of Probability*, *1*(2), 197–206.
McMullen, C. T. (1994). *Complex dynamics and renormalization*. Princeton: Princeton University Press.
Meakin, P. (1998). *Fractals, scaling and growth far from equilibrium*. Cambridge: Cambridge University Press.
Meester, R., & Steif, J. E. (1996). On the continuity of the critical value for long-range percolation in the exponential case. *Communications in Mathematical Physics*, *180*, 483–504.
Meixner, J. (1934). Orthogonale Polynomsysteme mit einer besonderen Gestalt der erzeugenden Funktion. *Journal of the London Mathematical Society*, *1*(9), 6–13.
Menéndez, P. (2009). *Statistical tools for palaeo data*. Unpublished Ph.D. thesis. Diss. ETH number 18060.
Menéndez, P., Ghosh, S., & Beran, J. (2010). On rapid change points under long memory. *Journal of Statistical Planning and Inference*, *140*(11), 3343–3354.
Menéndez, P., Ghosh, S., Künsch, H., & Tinner, W. (2012). A note on trend estimation under monotone Gaussian subordination with long memory: Application to fossil pollen series. Manuscript.
Menshikov, M., Sidoravicius, V., & Vachkovskaia, M. (2001). A note on two-dimensional truncated long-range percolation. *Advances in Applied Probability*, *33*, 912–929.

Messer, K., & Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *The Annals of Statistics*, *21*, 179–195.

Meyer, Y., Sellan, F., & Taqqu, M. S. (1999). Wavelets, generalized white noise and fractional integration: the synthesis of fractional Brownian motion. *Journal of Fourier Analysis and Applications*, *5*(5), 465–494.

Mielniczuk, J. (1997). Short-range and long-range dependence sums for infinite-order moving averages and regression estimation. *Acta Scientiarum Mathematicarum (Szeged)*, *67*, 301–316.

Mielniczuk, J., & Wojdyłło, P. (2007a). Decorrelation of wavelet coefficients for long-range dependent processes. *IEEE Transactions on Information Theory*, *53*(5), 1879–1883.

Mielniczuk, J., & Wojdyłło, P. (2007b). Estimation of Hurst exponent revisited. *Computational Statistics & Data Analysis*, *51*(9), 4510–4525.

Mielniczuk, J., & Wu, W. B. (2004). On random-design model with dependent errors. *Statistica Sinica*, *14*, 1105–1126.

Mikosch, T., & Samorodnitsky, G. (2000). Ruin probability with claims modeled by a stationary ergodic stable process. *The Annals of Probability*, *28*(4), 1814–1851.

Mikosch, T., & Samorodnitsky, G. (2007). Scaling limits for cumulative input processes. *Mathematics of Operations Research*, *32*(4), 890–918.

Mikosch, T., & Starica, C. (2000). Is it really long memory we see in financial returns? In P. Embrechts (Ed.), *Extremes and integrated risk management* (pp. 149–168). London: Risk Books.

Mikosch, T., & Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *The Review of Economics and Statistics*, *86*(1), 378–390.

Mikosch, T., Resnick, S., Rootzen, H., & Stegeman, A. (2002). Is network traffic approximated by stable Lévy motion or fractional Brownian motion? *The Annals of Applied Probability*, *12*(1), 23–68.

Milhoj, A. (1981). A test of fit in time series models. *Biometrika*, *68*, 177–188.

Mills, T. C. (2007). Time series modelling of two millennia of northern hemisphere temperatures: long memory or shifting trends? *Journal of the Royal Statistical Society. Series A. Statistics in Society*, *170*, 83–94.

Mishura, Y. (2008). *Lecture notes in mathematics: Vol. 1929*. *Stochastic calculus for fractional Brownian motion and related processes*. Berlin: Springer.

Mokkadem, A. (1997). A measure of information and its applications to test for randomness against ARMA alternatives and to goodness-of-fit test. *Stochastic Processes and Their Applications*, *72*(2), 145–159.

Morana, C. (2007). Multivariate modelling of longmemory processes with common components. *Computational Statistics & Data Analysis*, *52*(2), 919–934.

Morana, C., & Beltratti, A. (2004). Structural change and long range dependence in volatility of exchange rates: either, neither or both. *Journal of Empirical Finance*, *11*(4), 629–658.

Morlet, J., Arens, G., Fourgeau, E., & Giard, D. (1982). Wave propagation and sampling theory. *Geophysics*, *47*, 203–236.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Reading: Addison-Wesley.

Moulines, E., & Soulier, P. (1999). Broadband log-periodogram regression of time series with long-range dependence. *The Annals of Statistics*, *27*, 1415–1439.

Moulines, E., & Soulier, P. (2000). Data driven order selection for projection estimation of the spectral density of time series with long range dependence. *Journal of Time Series Analysis*, *21*, 193–218.

Moulines, E., & Soulier, P. (2003). Semiparametric spectral estimation for fractional processes. In P. Doukhan, G. Oppenheim, & M. S. Taqqu (Eds.), *Theory and applications of long-range dependence* (pp. 251–301). Boston: Birkhäuser.

Moulines, E., Roueff, F., & Taqqu, M. S. (2007a). On the spectral density of the wavelet coefficients of long-memory time series with application to the log-regression estimation of the memory parameter. *Journal of Time Series Analysis*, *28*(2), 155–187.

Moulines, E., Roueff, F., & Taqqu, M. S. (2007b). Central limit theorem for the log-regression wavelet estimation of the memory parameter in the Gaussian semi-parametric context. *Fractals*, *15*(4), 301–313.

Moulines, E., Roueff, F., & Taqqu, M. S. (2008). A wavelet Whittle estimator of the memory parameter of a non-stationary Gaussian time series. *The Annals of Statistics*, *36*(4), 1925–1956.

Mukherjee, K. (1999). The asymptotic distribution of a class of L-estimators under long range dependence. *Canadian Journal of Statistics*, *27*, 345–360.

Müller, H. G. (1984). Smooth optimum kernel estimators of regression curves, densities and modes. *Annals of Statistics*, *12*, 766–774.

Müller, H. G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statististics and Decisions, Supp. Issue, 2*, 193–206.

Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, *82*, 231–238.

Müller, H. G. (1988). *Nonparametric analysis of longitudinal data*. Berlin: Springer.

Müller, H. G. (1991). Smoothing optimal kernel estimators near the endpoints. *Biometrika*, *78*, 521–530.

Müller, H. G., & Wang, J. L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, *50*, 61–76.

Müller, U. K., & Watson, M. W. (2008). Testing models of low-frequency variability. *Econometrica*, *76*, 979–1016.

Murota, K., & Takeuchi, K. (1981). The studentized empirical characteristic function and its application to test for the shape of the distribution. *Biometrika*, *68*, 55–65.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, *9*, 141–142.

Narukawa, M., & Matsuda, Y. (2011). Broadband semi-parametric estimation of long-memory time series by fractional exponential models. *Journal of Time Series Analysis*, *32*, 175–193.

Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory*, *6*, 318–334.

Nelson, D. B., & Cao, C. Q. (1992). Inequality constraints in the univariate GARCH model. *Journal of Business & Economic Statistics*, *10*(2), 229–235.

Neumann, M. H., & von Sachs, R (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In A. Antoniadis & G. Oppenheim (Eds.), *Lecture notes in statistics: Vol. 103. Wavelets and statistics* (pp. 301–329). New York: Springer.

Newcomb, S. (1895). *Astronomical constants (the elements of the four inner planets and the fundamental constants of astronomy). Supplement to the American ephemeris and nautical almanac for 1897*. Washington D.C.: US Government Printing Office.

Newman, C. M., & Shulman, L. S. (1986). One dimensional $1/|i-j|^s$ percolation models: the existence of a transition for $s \leq 2$. *Communications in Mathematical Physics*, *104*, 547–571.

Newton, H. J., & Pagano, M. (1983). A method for determining periods in time series. *Journal of the American Statistical Association*, *78*, 152–157.

Nielsen, M. Ø. (2004a). Local empirical spectral measure of multivariate processes with long range dependence. *Stochastic Processes and Their Applications*, *109*, 145–166.

Nielsen, M. Ø. (2004b). Optimal residual-based test for fractional cointegration and exchange rate dynamics. *Journal of Business & Economic Statistics*, *22*, 331–345.

Nielsen, M. Ø. (2004c). Spectral analysis of fractionally cointegrated systems. *Economics Letters*, *83*, 225–231.

Nielsen, M. Ø. (2005a). Semiparametric estimation in time-series regression with long-range dependence. *Journal of Time Series Analysis*, *26*(2), 279–304.

Nielsen, M. Ø. (2005b). Multivariate Lagrange multiplier tests for fractional integration. *Journal of Financial Econometrics*, *3*(3), 372–398.

Nielsen, M. Ø. (2010). Nonparametric cointegration analysis of fractional systems with unknown integration orders. *Journal of Econometrics*, *155*(2), 170–187.

Nielsen, F. S. (2011, in press). Local Whittle estimation of multi-variate fractionally integrated processes. *Journal of Time Series Analysis*.

Nielsen, M. Ø., & Frederiksen, P. (2011). Fully modified narrow-band least squares estimation of weak fractional cointegration. *The Econometrics Journal*, *14*(1), 77–120.
Nielsen, M. Ø., & Shimotsu, K. (2007). Determining the cointegrating rank in nonstationary fractional systems by the exact local Whittle approach. *Journal of Econometrics*, *141*, 574–596.
Nolan, J. P. (1988). Path properties of index-β stable fields. *The Annals of Probability*, *16*(4), 1596–1607.
Nolan, J. P. (2011). *Distributions—models for heavy tailed data*. Boston: Birkhäuser. (In progress, Chap. 1 online at academic2.american.edu/~jpnolan.)
Nordman, D. J., & Lahiri, S. N. (2005). Validity of the sampling window method for long-range dependent linear processes. *Econometric Theory*, *21*, 1087–1111.
Nordman, D. J., & Lahiri, S. N. (2006). A frequency domain empirical likelihood for short- and long-range dependence. *The Annals of Statistics*, *34*(6), 3019–3050.
Nordman, D. J., Sibbertsen, P., & Lahiri, S. N. (2006). Empirical likelihood confidence intervals for the mean of a long-range dependence process. *Journal of Time Series Analysis*, *28*(4), 576–599.
Norros, I., Valkeila, E., & Virtamo, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions. *Bernoulli*, *5*(4), 571–587.
Nourdin, I., & Rosinski, J. (2012). Asymptotic independence of multiple Wiener–Ito integrals and the resulting laws. Preprint.
Ohanissian, A., Russell, J. R., & Tsay, R. S. (2008). True or spurious long memory? A new test. *Journal of Business & Economic Statistics*, *26*, 161–175.
Olhede, S. C., McCoy, E. J., & Stephens, D. A. (2004). Large-sample properties of the periodogram estimator of seasonally persistent processes. *Biometrika*, *91*(3), 613–628.
Onsager, L. (1944). Crystal statistics I: a two dimensional model with order-disorder transition. *Physical Review*, *65*, 117–149.
Oppenheim, G., & Viano, M.-C. (2004). Aggregation of random parameters Ornstein–Uhlenbeck or AR processes: some convergence results. *Journal of Time Series Analysis*, *25*, 335–350.
Opsomer, J. D., Wang, Y., & Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, *16*, 134–153.
Ozhegov, V. B. (1965). On generalized Appell polynomials. In *Investigations in modern problems of the constructive theory of functions*, Baku (pp. 595–601). (In Russian).
Ozhegov, V. B. (1967). On certain extremal properties of generalized Appell polynomials. *Doklady Akademii Nauk SSSR*, *159*(5), 985–987.
Page, E. S. (1954). Continuous inspection scheme. *Biometrika*, *41*(1–2), 100–115.
Palma, W. (2007). *Long-memory time series—theory and methods*. New York: Wiley.
Palma, W., & Chan, N. H. (1997). Estimation and forecasting of long-memory processes with missing values. *International Journal of Forecasting*, *16*(6), 395–410.
Palma, W., & Olea, R. (2010). An efficient estimator for locally stationary Gaussian long-memory processes. *The Annals of Statistics*, *38*(5), 2958–2997.
Park, K. & Willinger, W. (Eds.) (2000). *Self-similar network traffic and performance evaluation*. New York: Wiley-Interscience.
Parke, W. R. (1999). What is fractional integration? *Review of Economics and Statistics*, *81*, 632–638.
Parker, D. E., & Horton, E. B. (2005). Uncertainties in the Central England temperature series since 1878 and some changes to the maximum and minimum series. *International Journal of Climatology*, *25*, 1173–1188.
Parker, D. E., Legg, T. P., & Folland, C. K. (1992). A new daily Central England temperature series, 1772–1991. *International Journal of Climatology*, *12*, 317–342.
Parzen, E. (1968). *Statistical spectral analysis* (*single channel case*) (Technical Report No. 11). Stanford University, Statistics Department.
Parzen, E. (1969). Mulitple time series modelling. In P. R. Krishnaiah (Ed.), *Multivariate analysis II* (pp. 389–409). New York: Academic Press.
Parzen, E. (1974). Some recent advances in time series modeling. *IEEE Transactions on Automatic Control*, *AC-19*, 723–730.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, *74*, 105–121.

Pawlak, M., & Stadtmüller, U. (2007). Signal sampling and recovery under dependent errors. *IEEE Transactions on Information Theory*, *53*(7), 2526–2541.

Paxson, V., & Floyd, S. (1995). Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, *3*(3), 226–244.

Paya, I., Duarte, A., & Holden, K. (2007). On the relationship between inflation persistence and temporal aggregation. *Journal of Money, Credit, and Banking*, *39*, 1521–1531.

Pearson, K. (1902). On the mathematical theory of errors of judgement, with special reference to the personal equation. In *Philosophical transactions of the royal society of London* (pp. 235–299).

Peirce, C. S. (1873). Theory of errors of observations. Appendix No. 21 (pp. 200–224 and plate 28) of report of the superintendent of the US coast survey for the year ending November 1870). G.P.O., Washington. Reprinted in the new elements of mathematics by C.S. Peirce, ed. by C. Eisele, Humanities Press, Atlantic Highlands, 1976, Vol. 3, pt. 1, pp. 639–676.

Peiris, M. S., & Pereira, B. J. C. (1988). On prediction with fractionally differenced ARIMA processes. *Journal of Time Series Analysis*, *9*, 215–220.

Peitgen, H. O., & Richter, P. H. (1986). *The beauty of fractals*. Berlin: Springer.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., & Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, *49*(2), 1685–1689.

Percival, D. B. (1983). *The statistics of long-memory processes*. Ph.D. thesis, Dept. of Statistics, University of Washington, Seattle.

Percival, D. B., & Guttorp, P. (1994). Long-memory processes, the Allan variance and wavelets. In E. Foufoula-Georgiu & P. Kumar (Eds.), *Wavelets in geophysics*. New York: Academic Press.

Percival, D. P., & Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge: Cambridge University Press.

Petersen, K. (1989). *Cambridge studies in advanced mathematics: Vol. 2. Ergodic theory*. Cambridge: Cambridge University Press.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics*, *33*, 311–340.

Phillips, P. C. B. (1995). Nonstationary time series and cointegration. *Journal of Applied Econometrics*, *10*, 87–94.

Phillips, P. C. B. (2007). Unit root log periodogram regression. *Journal of Econometrics*, *138*(1), 104–124.

Phillips, P. C. B., & Loretan, M. (1991). Estimating long-run economic equilibria. *The Review of Economic Studies*, *58*(3), 407–436. Special Issue: The Econometrics of Financial Markets.

Phillips, P. C. B., & Shimotsu, K. (2004). Local Whittle estimation in nonstationary and unit root cases. *The Annals of Statistics*, *32*(2), 656–692.

Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, *17*, 841–867.

Pietronero, L. & Tosatti, E. (Eds.) (1986). *Fractals in physics*. Amsterdam: North-Holland.

Pinsky, M. A. (2002). *Introduction to Fourier analysis and wavelets. The Brooks/Cole series in advanced mathematics*. Pacific Grove: Brooks/Cole.

Pipiras, V., & Taqqu, M. S. (2000a). Integration questions related to fractional Brownian motion. *Probability Theory and Related Fields*, *118*(2), 251–291.

Pipiras, V., & Taqqu, M. S. (2000b). The limit of a renewal reward process with heavy-tailed rewards is not a linear fractional stable motion. *Bernoulli*, *6*(4), 607–614.

Pipiras, V., & Taqqu, M. S. (2000c). Convergence of weighted sums of random variables with long-range dependence. *Stochastic Processes and Their Applications*, *90*, 157–174.

Pipiras, V., & Taqqu, M. S. (2003). Fractional calculus and its connect on to fractional Brownian motion. In *Long range dependence* (pp. 166–201). Basel: Birkhäuser.

Pipiras, V., & Taqqu, M. S. (2012, to appear). Long-range dependence of the two-dimensional Ising model at critical temperature. *Fractals*.

Pipiras, V., Taqqu, M. S., & Levy, J. B. (2004). Slow, fast and arbitrary growth conditions for renewal-reward processes when both the renewals and the rewards are heavy-tailed. *Bernoulli*, *10*(1), 121–163.

Politis, D. N., & Romano, J. P. (1993). Nonparametric resampling for homogeneous strong mixing random fields. *Journal of Multivariate Analysis*, *47*(2), 301–328.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. New York: Springer.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Ponson, L., Bonamy, D., Auradou, H., Mourot, G., Morel, S., Bouchaud, E., Guillot, C., & Hulin, J. P. (2005). Anisotropic self-affine properties of experimental fracture surfaces. *International Journal of Fracture*, *140*(1–4), 27–37.

Porter-Hudak, S. (1990). An application to the seasonal fractionally differenced model to the monetary aggregates. *Journal of the American Statistical Association*, *85*(410), 338–344.

Poskitt, D. S. (2007a). Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases. *Annals of the Institute of Statistical Mathematics*, *59*(4), 697–725.

Poskitt, D. S. (2007b). Properties of the sieve bootstrap for fractionally integrated and non-invertible processes. *Journal of Time Series Analysis*, *29*(2), 224–250.

Priestley, M. B. (1981). *Spectral analysis and time series*. San Diego: Academic Press.

Priestley, M. E., & Chao, M. T. (1972). Nonparametric function fitting. *Journal of The Royal Statistical Society, Series B*, *34*, 385–392.

Qu, Z. (2010). A test against spurious long memory. *Journal of Business & Economic Statistics*, *26*(2), 161–175.

Rachev, S. T., & Samorodnitsky, G. (2001). Long strange segments in a long-range dependent moving average. *Stochastic Processes and Their Applications*, *93*(1), 119–148.

Racheva-Iotova, B., & Samorodnitsky, G. (2003). Long range dependence and heavy tails. In S. T. Rachev (Ed.), *Handbook of heavy tailed distributions in finance* (pp. 641–662). Amsterdam: Elsevier. Ch. 16.

Ramjee, R., Crato, N., & Ray, B. K. (2002). A note on moving average forecasts of long memory processes with an application to quality control. *International Journal of Forecasting*, *18*(2), 291–297.

Ramm, A. G. (1980). *Theory and applications of some new classes of integral equations*. New York: Springer.

Rangarajan, G. & Ding, M. (Eds.) (2003). *Lecture notes in physics: Vol. 621. Processes with long-range correlations—theory and applications*.

Rao, C. R. (1965). *Linear statistical inference and its applications*. New York: Wiley.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Ravishanker, N., & Ray, B. K. (1997). Bayesian analysis of vector ARFIMA processes. *Australian Journal of Statistics*, *39*, 295–312.

Ravishanker, N., & Ray, B. K. (2002). Bayesian prediction for vector ARFIMA processes. *International Journal of Forecasting*, *18*(2), 207–214.

Ray, B. K., & Tsay, R. S. (1997). Bandwidth selection for kernel regression with long-range dependent errors. *Biometrika*, *84*(4), 791–802.

Ray, B. K., & Tsay, R. S. (2000). Long-range dependence in daily stock volatilities. *Journal of Business & Economic Statistics*, *18*(2), 254–262.

Ray, B. K., & Tsay, R. S. (2002). Bayesian methods for change-point detection in long-range dependent processes. *Journal of Time Series Analysis*, *23*(6), 687–705.

Reinsel, G. C., & Lewis, R. A. (1987). Prediction mean square error for nonstationary multivariate time series using estimated parameters. *Economics Letters*, *24*, 57–61.

Resnick, S. I. (1992). *Adventures in stochastic processes*. Boston: Birkhäuser.

Resnick, S. I. (1997). Heavy tail modelling and teletraffic data: special invited paper. *The Annals of Statistics*, *25*(5), 1805–1869.

Resnick, S. I. (2007). *Heavy-tail phenomena*. New York: Springer.

Resnick, S. I., & Samorodnitsky, G. (2004). Point processes associated with stationary stable processes. *Stochastic Processes and Their Applications*, *114*(2), 191–209.

Resnick, S., & Starica, C. (1995). Consistency of Hill's estimator for dependent data. *Journal of Applied Probability*, *32*, 139–167.

Resnick, S., & Starica, C. (1998). Tail index estimation for dependent data. *The Annals of Applied Probability*, *8*(4), 1156–1183.

Resnick, S. I., & van den Berg, E. (2000). Weak convergence of high-speed network traffic models. *Journal of Applied Probability*, *37*(2), 575–597.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, *12*, 1215–1230.

Rice, J. (1986). Bandwidth choice for differentiation. *Journal of Multivariate Analysis*, *19*, 251–264.

Robinson, P. M. (1978). Statistical inference for a random coefficient autoregressive model. *Scandinavian Journal of Statistics*, *5*, 163–168.

Robinson, P. M. (1991). Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics*, *47*, 67–84.

Robinson, P. M. (1994a). Time series with strong dependence. In C. A. Sims (Ed.), *Advances in econometrics: sixth world congress* (Vol. 1, pp. 47–95). Cambridge: Cambridge University Press.

Robinson, P. M. (1994b). Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association*, *89*, 1420–1437.

Robinson, P. M. (1994c). Semiparametric analysis of long-memory time series. *The Annals of Statistics*, *22*, 515–539.

Robinson, P. M. (1995a). Log-periodogram regression of time series with long range dependence. *The Annals of Statistics*, *23*(3), 1048–1072.

Robinson, P. M. (1995b). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, *23*(5), 1630–1661.

Robinson, P. M. (1997). Largesample inference for nonparametric regression with dependent errors. *Annals of Statistics*, *25*(5), 2054–2083.

Robinson, P. M. (Ed.) (2002). *Time series with long memory*. Oxford: Oxford University Press.

Robinson, P. M. (2005). Efficiency improvements in inference on stationary and nonstationary fractional time series. *The Annals of Statistics*, *33*(4), 1800–1842.

Robinson, P. M. (2007). Nonparametric spectrum estimation for spatial data. *Journal of Statistical Planning and Inference*, *137*(3), 1024–1034.

Robinson, P. M. (2008). Multiple local Whittle estimation in stationary systems. *The Annals of Statistics*, *36*(5), 2508–2530.

Robinson, P., & Henry, M. (1996). Bandwidth choice in Gaussian semiparametric estimation of long range dependence. In P. Robinson & M. Rosenblatt (Eds.), *Time series analysis in memory of E.J. Hannan: Vol. II. Athens conference on applied probability and time series analysis*. Berlin: Springer.

Robinson, P. M., & Henry, M. (2003). Higher-order kernel semiparametric M-estimation of long memory. *Journal of Econometrics*, *114*, 1–27.

Robinson, P. M., & Hidalgo, F. J. (1997). Time series regression with long-range dependence. *The Annals of Statistics*, *25*(1), 77–104.

Robinson, P. M., & Hualde, J. (2003). Cointegration in fractional systems with unknown integration orders. *Econometrica*, *71*, 1727–1766.

Robinson, P. M., & Iacone, F. (2005). Cointegration in fractional systems with deterministic trends. *Journal of Econometrics*, *129*, 263–298.

Robinson, P. M., & Marinucci, D. (2001). Narrow-band analysis of nonstationary processes. *The Annals of Statistics*, *29*, 947–986.

Robinson, P. M., & Marinucci, D. (2003). *Semiparametric frequency-domain analysis of fractional cointegration*. *Time series with long memory* (pp. 334–373). Oxford: Oxford University Press.

Robinson, P. M., & Yajima, Y. (2002). Determination of cointegrating rank in fractional systems. *Journal of Econometrics*, *106*, 217–241.

Robinson, P. M., & Zaffaroni, P. (1997). Modelling nonlinearity and long memory in time series. *Fields Institute Communications*, *11*, 161–170.

Robinson, P. M., & Zaffaroni, P. (1998). Nonlinear time series with long memory: a model for stochastic volatility. *Journal of Statistical Planning and Inference*, *68*, 359–371.

Robinson, P. M., & Zaffaroni, P. (2006). Pseudo-maximum likelihood estimation of ARCH(∞) models. *The Annals of Statistics*, *34*(3), 1049–1074.

Rodrigues, O. (1816). De l'attraction des sphéroïdes. *Correspondence sur l'Ecole Impériale Polytechnique*, *3*(3), 361–385.

Rodriguez-Iturbe, I., & Rinaldo, A. (1997). *Fractal river basins*. Cambridge: Cambridge University Press.

Rogers, L. C. G. (1997). Arbitrage with fractional Brownian motion. *Mathematical Finance*, *7*, 95–105.

Rolls, D. A. (2010). Reduced long-range dependence combining Poisson bursts with on–off sources. *Brazilian Journal of Probability and Statistics*, *24*(3), 479–501.

Rootzén, H. (1986). Extreme value theory for moving average processes. *Annals of Probability*, *14*(2), 612–652.

Rootzén, H. (2009). Weak convergence of the tail empirical process for dependent sequences. *Stochastic Processes and Their Applications*, *119*(2), 468–490.

Rootzén, H., Leadbetter, M. R., & de Haan, L. (1998). On the distribution of tail array sums for strongly mixing stationary sequences. *The Annals of Applied Probability*, *8*(3), 868–885.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*, 832–837.

Rosenblatt, M. (1961). Independence and dependence. In *Proc. 4th Berkeley sympos. math. statist. and prob.* (Vol. II, pp. 431–443). Berkeley: University of California Press.

Rosenblatt, M. (1971). Curve estimates. *The Annals of Mathematical Statistics*, *42*, 1815–1842.

Rosenblatt, M. (1979). Some limit theorems for partial sums of quadratic forms in stationary Gaussian variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *49*(2), 125–132.

Rosenblatt, M. (1985). *Stationary sequences and random fields*. Boston: Birkhäuser.

Rosiński, J. (1995). On the structure of stationary stable processes. *Annals of Probability*, *23*(3), 1163–1187.

Roueff, F., & von Sachs, R. (2011). Locally stationary long memory estimation. *Stochastic Processes and Their Applications*, *121*(4), 813–844.

Roughan, M., & Veitch, D. (1999). Measuring long-range dependence under changing traffic conditions. In *Proceedings of IEEE INFOCOM* (Vol. 3, pp. 1513–1521).

Ruelle, D. (1968). Statistical mechanics of one-dimensional lattice gas. *Communications in Mathematical Physics*, *9*(4), 267–278.

Ruelle, D. (1970). Superstable interactions in classical statistical mechanics. *Communications in Mathematical Physics*, *18*, 127–159.

Ruiz-Medina, M. D., Angulo, J. M., & Anh, V. V. (2003). Fractional-order regularization and wavelet approximation to the inverse estimation problem for random fields. *Journal of Multivariate Analysis*, *85*, 192–216.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, *22*, 1346–1370.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, *90*, 1257–1270.

Sain, R. S., & Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics*, *140*(1), 226–259.

Samarov, A., & Taqqu, M. S. (1988). On the efficiency of the sample mean in long-memory noise. *Journal of Time Series Analysis*, *9*(2), 191–200.

Samko, S. G., Kilbas, A. A., & Marichev, O. I. (1987). *Integrals and derivatives of fractional order and some its applications*. Minsk: Nauka i Tekhnika or fractional integrals and derivatives theory and applications. New York: Gordon and Breach (1993).

Samorodnitsky, G. (2002). Long range dependence, heavy tails and rare events. *MaPhySto, Centre for Mathematical Physics and Stochastics, Aarhus. Lecture Notes.*

Samorodnitsky, G. (2004). Extreme value theory, ergodic theory, and the boundary between short memory and long memory for stationary stable processes. *Annals of Probability*, *32*, 1438–1468.

Samorodnitsky, G. (2005). Null flows, positive flows and the structure of stationary symmetric stable processes. *The Annals of Probability*, *33*(5), 1782–1803.

Samorodnitsky, G. (2006). Long range dependence. *Foundations and Trends in Stochastic Systems*, *1*(3), 163–257.

Samorodnitsky, G., & Taqqu, M. S. (1994). *Stable non-Gaussian random processes: stochastic models with infinite variance*. New York: Chapman & Hall/CRC Press.

Scharth, M., & Medeiros, M. C. (2009). Asymmetric effects and long memory in the volatility of Dow Jones stocks. *International Journal of Forecasting*, *25*, 304–327.

Scheuring, I. (1991). The fractal nature of vegetation and the species-area relation. *Theoretical Population Biology*, *39*, 170–177.

Schützner, M. (2006). *Appell-Polynome und Grenzwertsätze*. Diploma Thesis, University of Konstanz.

Schützner, M. (2009). *Asymptotic statistical theory for long memory volatility models*. Ph.D. thesis, University of Konstanz.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *2*(6), 461–464.

Sedletskii, A. M. (2000). *Fourier transforms and approximations*. Boca Raton: CRC press.

Sela, R. J., & Hurvich, C. M. (2009). Computationally efficient methods for two multivariate fractionally integrated models. *Journal of Time Series Analysis*, *30*, 631–651.

Seneta, E. (1976). *Lecture notes in mathematics: Vol. 508. Regularly varying functions*. New York: Springer.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Sethuraman, S., & Basawa, I. V. (1995). Maximum likelihood estimation for a fractionally differenced autoregressive model on a two-dimensional lattice. *Journal of Statistical Planning and Inference*, *44*, 219–235.

Shao, X., & Wu, W. B. (2007). Local Whittle estimation of fractional integration for nonlinear processes. *Econometric Theory*, *23*(5), 899–929.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*(1), 117–126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, *8*, 147–164.

Shibata, R. (1981). An optimal autoregressive spectral estimate. *The Annals of Statistics*, *9*, 300–306.

Shimotsu, K. (2006). *Simple (but effective) tests of long memory versus structural breaks*. Queen's Economics Dept. Working Paper.

Shimotsu, K. (2012). Exact local Whittle estimation of fractionally cointegrated systems? *Journal of Econometrics*, *169*(2), 266–278.

Shimotsu, K., & Phillips, P. C. B. (2006). Local Whittle estimation of fractional integration and some of its variants. *Journal of Econometrics*, *130*(2), 209–233. 2006.

Shorack, G. R., & Wellner, J. A. (1986). *Empirical processes with applications to statistics*. New York: Wiley.

Sibbertsen, P. (2001). *S*-estimation in the linear regression model with long-memory error terms under trend. *Journal of Time Series Analysis*, *22*, 353–363.

Sibbertsen, P. (2004). Long memory versus structural breaks: an overview. *Statistical Papers*, *45*(4), 465–515.

Sierpinksi, M. (1915). Sur une courbe dont tout point est un point de ramification. *Comptes Rendus de L'Académie des Sciences de Paris*, *160*, 302–305.

Silveira, G. (1991). *Contributions to strong approximations in time series with applications in nonparametric statistics and functional central limit theorems*. Ph.D. Thesis, University of London.

Silverman, B. W. (1986). *Density estimation*. London: Chapman & Hall.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.

Simos, T. (2008). The exact discrete model of a system of linear stochastic differential equations driven by fractional noise. *Journal of Time Series Analysis*, *29*, 1019–1031.

Skorokhod, A. V. (1957). Limit theorems for processes with independent increments. *Teoriâ Veroâtnostej I Ee Primeneniâ*, *2*, 145–177. (English translation in *Theory Probability and Its Applications, 2*, 138–171).

Sly, A., & Heyde, C. (2008). Nonstandard limit theorem for infinite variance functionals. *Annals of Probability*, *36*(2), 796–805.

Smith, H. J. S. (1875). On the integration of discontinuous functions. *Proceedings of the London Mathematical Society, Series 1*, *6*, 140–153.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, *28*, 1–23.

Smith, R. L. (1992). Estimating dimension in noisy chaotic time series. *Journal of the Royal Statistical Society*, *54*(2), 329–351.

Smith, J., & Yadav, S. (1994). Forecasting costs incurred from unit differencing fractionally integrated processes. *International Journal of Forecasting*, *10*, 507–514.

Smith, A. A. J., Sowell, F. B., & Zin, S. E. (1997). Fractional integration with drift: estimation in small samples. *Empirical Economics*, *22*, 103–116.

Sokal, A. D. (1981). Existence of compatible families of proper regular conditional probabilities. *Probability Theory and Related Fields*, *56*(4), 537–548.

Solo, V. (1992). Intrinsic random functions and the paradox of $1/f$ noise. *SIAM Journal on Applied Mathematics*, *52*(1), 270–291.

Soulier, P. (2010). Best attainable rates of convergence for the estimation of the memory parameter. In P. Doukhan, G. Lang, D. Surgailis, & G. Teyssiere (Eds.), *Lecture notes in statistics: Vol. 200. Dependence in probability and statistics*. New York: Springer.

Souza, L. R. (2005). A note on Chambers's long memory and aggregation in macroeconomic time series. *International Economic Review*, *46*, 1059–1062.

Souza, L. R. (2007). Temporal aggregation and bandwidth selection in estimating long memory. *Journal of Time Series Analysis*, *28*(5), 701–722.

Souza, L. R. (2008). Why aggregate long memory time series? *Econometric Reviews*, *27*, 298–316.

Souza, L. R., & Smith, J. (2004). Effects of temporal aggregation on estimates and forecasts of fractionally integrated processes: a Monte-Carlo study. *International Journal of Forecasting*, *20*, 487–502.

Sowell, F. B. (1990). The fractional unit root distribution. *Econometrica*, *58*(2), 495–505.

Sowell, F. B. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics*, *53*, 165–188.

Speckman, P. (1988). *Kernel smoothing in partial linear models*.

Stanley, H. E. (1971). *Introduction to phase transitions and critical phenomena*. Oxford: Oxford University Press.

Stanley, H. E. (1987). *Introduction to phase transitions and critical phenomena*. Oxford: Oxford University Press.

Stauffer, D., & Aharony, A. (1994). *Introduction to percolation theory*. Boca Raton: CRC Press.

Steeb, W.-H. (1998). *Hilbert spaces, wavelets, generalised functions and modern quantum mechanics*. Dordrecht: Kluwer Academic.

Stoev, S., & Taqqu, M. S. (2005a). Path properties of the linear multifractional stable motion. *Fractals*, *13*(2), 157–178.

Stoev, S., & Taqqu, M. S. (2005b). Asymptotic self-similarity and wavelet estimation for long-range dependent fractional autoregressive integrated moving average time series with stable innovations. *Journal of Time Series Analysis*, *26*(2), 211–249.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, *36*, 111–147.

Stout, W. F. (1974). *Almost sure convergence*. New York: Academic Press.

Strang, G. (1989). Wavelets and dilation equations: a brief introduction. *SIAM Review*, *31*(4), 614–627.

Straumann, D. (2004). *Lecture notes in statistics: Vol. 181. Estimation in conditionally heteroskedastic time series models*. New York: Springer.

Strichartz, R. (1994). *A guide to distribution theory and Fourier transforms*. Boca Raton: CRC Press.

Student (1927). Errors of routine analysis. *Biometrika*, *19*, 151–164.

Subba Rao, T. (1970). The fitting of non-stationary time-series models with time-dependent parameters. *Journal of the Royal Statistical Society. Series B*, *32*(2), 312–322.

Suquet, C. (1996). Tightness in Schauder decomposable Banach spaces. *Translations—American Mathematical Society*, *193*(2), 201–224.

Surgailis, D. (1981). Convergence of sums of non-linear functions of moving averages to self-similar processes. *Soviet Mathematics, Doklady*, *23*, 247–250.

Surgailis, D. (1982). Zones of attraction of self-similar multiple integrals. *Lithuanian Mathematical Journal*, *22*, 327–340.

Surgailis, D. (2000). Long-range dependence and Appell rank. *Annals of Probability*, *28*(1), 478–497.

Surgailis, D. (2002). Stable limits of empirical processes of moving averages with infinite variance. *Stochastic Processes and Their Applications*, *100*, 255–274.

Surgailis, D. (2003). CLTs for polynomials of linear sequences: diagram formula with illustrations. In P. Doukhan, G. Oppenheim, & M. S. Taqqu (Eds.), *Theory and applications of long-range dependence* (pp. 111–127). Boston: Birkhäuser Boston.

Surgailis, D. (2004). Stable limits of sums of bounded functions of long-memory moving averages with finite variance. *Bernoulli*, *10*(2), 327–355.

Surgailis, D. (2008). A quadratic ARCH(∞) model with long memory and Lévy stable behavior of squares. *Advances in Applied Probability*, *40*(4), 1198–1222.

Surgailis, D., & Vaičiulis, M. (1999). Convergence of Appell polynomials of long range dependent moving averages in martingale differences. *Acta Applicandae Mathematicae*, *58*(1–3), 343–357.

Surgailis, D., & Viano, M.-C. (2002). Long memory properties and covariance structure of the EGARCH model. *ESAIM: Probability and Statistics*, *6*, 311–329.

Szegö, G. (1939). *Orthogonal polynomials. Colloquium publications*. Providence: Am. Math. Soc.

Szegö, G. (1974). *Orthogonal polynomials* (3rd ed.). Providence: Am. Math. Soc.

Sznitman, A. S. (2010). Vacant set of random interlacements and percolation. *Annals of Mathematics*, 2039–2087.

Talagrand, M. (1995). Hausdorff measure of trajectories of multiparameter fractional Brownian motion. *Annals of Probability*, *23*, 767–775.

Tang, S. M., & MacNeill, I. B. (1993). The effect of serial correlation on tests for parameter change at unknown time. *The Annals of Statistics*, *21*, 552–575.

Taqqu, M. S. (1975). Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *31*, 287–302.

Taqqu, M. S. (1977). Law of the iterated logarithm for sums of non-linear functions of Gaussian variables that exhibit a long range dependence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *40*(3), 203–238.

Taqqu, M. S. (1978). A representation for self-similar processes. *Stochastic Processes and Their Applications*, *7*(1), 55–64.

Taqqu, M. S. (1979). Convergence of integrated processes of arbitrary Hermite rank. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *50*, 53–83.

Taqqu, M. S. (2002). The modelling of ethernet data and of signals that are heavy-tailed with infinite variance. Large structured models in applied sciences; challenges for statistics (Grimstad, 2000). *Scandinavian Journal of Statistics*, *29*(2), 273–295.

Taqqu, M. S. (2003). Fractional Brownian motion and long-range dependence. In *Theory and applications of long-range dependence* (pp. 5–38). Boston: Birkhäuser.

Taqqu, M. S., & Levy, J. B. (1986). Using renewal processes to generate long-range dependence and high variability. In *Progr. Probab. Statist.: Vol. 11. Dependence in probability and statistics*, Oberwolfach, 1985 (pp. 73–89). Boston: Birkhäuser Boston.

Taqqu, M. S., & Wolpert, R. L. (1983). Infinite variance self-similar processes subordinate to a Poisson measure. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *62*(1), 53–72.

Taqqu, M. S., Teverovsky, V., & Willinger, W. (1995). Estimators for long-range dependence: an empirical study. *Fractals*, *3*(4), 785–798.

Taqqu, M. S., Willinger, W., & Sherman, R. (1997). Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review*, *27*, 5–23.

Taylor, C. C., & Taylor, S. J. (1991). Estimating the dimension of a fractal. *Journal of the Royal Statistical Society*, *53*(2), 353–364.

Teles, P., Wei, W. W. S., & Crato, N. (1999). The use of aggregate time series in testing for long memory. In *Bulletin of the international statistical institute, 52nd session* (pp. 341–342).

Terrin, N., & Taqqu, M. S. (1990). A noncentral limit theorem for quadratic forms of Gaussian stationary sequences. *Journal of Theoretical Probability*, *3*(3), 449–475.

Terrin, N., & Taqqu, M. S. (1991). Convergence in distribution of sums of bivariate Appell polynomials with long-range dependence. *Probability Theory and Related Fields*, *90*(1), 57–81.

Teugels, J. L. (1968). Renewal theorems when the first or the second moment is infinite. *The Annals of Mathematical Statistics*, *39*, 1210–1219.

Teverovsky, V., & Taqqu, M. S. (1997). Testing for long-range dependence in the presence of shifting means or a slowly declining trend, using a variance-type estimator. *Journal of Time Series Analysis*, *18*(3), 279–304.

Teverovsky, V., Taqqu, M. S., & Willinger, W. (1999). A critical look at Lo's modified R/S statistic. *Journal of Statistical Planning and Inference*, *80*, 211–227.

Tewfik, A. H., & Kim, M. (1992). Correlations structure of the discrete wavelet coefficients of fractional Brownian motion. *IEEE Transactions on Information Theory*, *38*(2), 904–909.

Teyssière, G., & Abry, P. (2006). Wavelet analysis of nonlinear long–range dependent processes. Applications to financial time series. In G. Teyssiere & A. Kirman (Eds.), *Long memory in economics* (pp. 173–238). Berlin: Springer.

Teyssière, G., & Kirman, A. (2005). *Long memory in economics*. New York: Springer.

Thavaneswaran, A., & Peiris, M. S. (2001). Recursive estimation for regression with infinite variance fractional ARIMA noise. *Mathematical and Computer Modelling*, *34*(9–11), 1133–1137.

Tjostheim, D. (1978). Statistical spatial series modelling. *Advances in Applied Probability*, *10*(1), 130–154.

Tong, H. (1993). *Non-linear time series: a dynamical system approach*. Oxford: Oxford University Press.

Trapman, P. (2010). The growth of the infinite long-range percolation cluster. *Annals of Probability*, *38*, 1583–1608.

Truquet, L. (2008). *A new smoothed QMLE for AR processes with LARCH errors* (Working paper). Université Panthéon-Sorbonne Paris I.

Tsai, H. (2006). Quasi-maximum likelihood estimation of long-memory limiting aggregate processes. *Statistica Sinica*, *16*, 213–226.

Tsai, H. (2009). On continuous-time autoregressive fractionally integrated moving average processes. *Bernoulli*, *15*(1), 178–194.

Tsai, H., & Chan, K. S. (2005a). Temporal aggregation of stationary and nonstationary discrete-time processes. *Journal of Time Series Analysis*, *26*(4), 613–624.

Tsai, H., & Chan, K. S. (2005b). Temporal aggregation of stationary and non-stationary continuous-time processes. *Scandinavian Journal of Statistics*, *32*, 583–597.

Tsai, H., & Chan, K. S. (2005c). Quasi-maximum likelihood estimation for a class of continuous-time long-memory processes. *Journal of Time Series Analysis*, *26*, 691–713.

Tsai, H., & Chan, K. S. (2005d). Maximum likelihood estimation of linear continuous-time long memory processes with discrete time data. *Journal of the Royal Statistical Society, Series B*, *67*, 703–716.

Tsay, W. J. (2000). Maximum likelihood estimation of stationary multivariate ARFIMA processes. *Journal of Statistical Computation and Simulation*, *80*(7), 729–745.

Tsay, W. J., & Chung, C.-F. (2000). The spurious regression of fractionally integrated processes. *Journal of Econometrics*, *96*(1), 155–182.
Tsybakov, A. B. (2010). *Introduction to nonparametric estimation*. New York: Springer.
Tukey, J. W. (1967). An introduction to the calculations of numerical spectrum analysis. In B. Harris (Ed.), *Advanced seminar on spectral analysis of time series* (pp. 25–46). New York: Wiley.
Turcotte, D. L. (1997). *Fractals and chaos in geology and geophysics*. Cambridge: Cambridge University Press.
Ullah, A. (1988). Nonparametric estimation of econometric functionals. *Canadian Journal of Economics*, *21*, 625–658.
Ullah, A. (1989). Nonparametric estimation and hypothesis testing in econometric models. In A. Ullah (Ed.), *Semiparametric and nonparametric econometrics* (pp. 101–129). Heidelberg: Physica.
Vaičiulis, M. (2003). Convergence of sums of Appell polynomials with infinite variance. *Lithuanian Mathematical Journal*, *43*(1), 67–82.
Van Bellegen, S., & Dahlhaus, R. (2006). Semiparametric estimation by model selection for locally stationary processes. *Journal of the Royal Statistical Society B*, *68*, 721–764.
Vanderzande, C. (1998). *Lattice models of polymers*. *Cambridge lecture notes in physics*. Cambridge: Cambridge University Press.
Vasilkov, A. P., Joiner, J., Spurr, R. J. D., Bhartia, P. K., Levelt, P., & Stephens, G. (2008). Evaluation of the OMI cloud pressures derived from rotational Raman scattering by comparisons with other satellite data and radiative transfer simulations. *Geophysical Research*, *113*, D15S19. doi:10.1029/2007JD008689.
Veitch, D., & Abry, P. (1999). A wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Transactions on Information Theory*, *45*, 878–897.
Veitsch, D. N., Taqqu, M. S., & Abry, P. (2000). Meaningful MRA initialisation for discrete time series. *Signal Processing*, *80*(9), 1971–1983.
Velasco, C. (1999a). Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis*, *20*, 87–127.
Velasco, C. (1999b). Non-stationary log-periodogram regression. *Journal of Econometrics*, *91*, 325–371.
Velasco, C. (2000). Non-Gaussian log-periodogram regression. *Econometric Theory*, *16*, 44–79.
Velasco, C. (2003). Gaussian semi-parametric estimation of fractional cointegration. *Journal of Time Series Analysis*, *24*, 345–378.
Velasco, C. (2007). The periodogram of fractional processes. *Journal of Time Series Analysis*, *28*(4), 600–627.
Velasco, C., & Robinson, P. M. (2000). Whittle pseudo-maximum likelihood estimates for nonstationary time series. *Journal of the American Statistical Association*, *95*, 1229–1243.
Veres, S., & Boda, M. (2000). The chaotic nature of TCP congestion control. In *Proceedings of IEEE INFOCOM* (Vol. 3, pp. 1715–1723).
Vervaat, W. (1972). Functional central limit theorems for processes with positive drift and their inverses. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *23*, 245–253.
Vesilo, R. A., & Chan, A. (1996). Detecting change points in long range dependency traffic. In *Proceedings of the Australian telecommunication networks & applications conference*, Melbourne, 3–6 December 1996 (pp. 567–572).
Viano, M. C., Deniau, C., & Oppenheim, G. (1994). Continuous-time fractional ARMA processes. *Statistics & Probability Letters*, *21*, 323–336.
Vicsek, T. (1992). *Fractal growth phenomena* (2nd ed.). River Edge: World Scientific.
Vidakovic, B. (1999). *Statistical modeling by wavelets*. *Wiley series in probability and statistics*. New York: Wiley.
Vladimirov, V. S. (2002). *Methods of the theory of generalized functions*. London: Taylor & Francis.
Volterra, V. (1881). Alcune osservazioni sulle funzioni punteggiate discontinue. *Giornale di Matematiche*, *19*, 76–86.

von Koch, H. (1904). Sur une courbe continue sans tangente obtenus par une construction géométrique élémentaire. *Arkiv for Mathematik, Astronomi och Fysich*, *1*, 681–704.
Wackernagel, H. (1998). *Multivariate geostatistics* (2nd ed.). Berlin: Springer.
Walters, P. (1989). *Graduate texts in mathematics: Vol. 79. An introduction to ergodic theory*. New York: Springer.
Walters, P. (2000). *An introduction to ergodic theory*. New York: Springer.
Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. London: Chapman & Hall/CRC Press.
Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics*, *24*(2), 466–484.
Wang, Y. (1997). Mimax estimation via wavelets for indirect long-memory data. *Journal of Statistical Planning and Inference*, *64*(1), 45–55.
Wang, L. (2008). Change-point detection with rank statistics in long-memory time-series models. *Australian & New Zealand Journal of Statistics*, *50*, 241–256.
Wang, Q., Lin, Y.-X., & Gulati, C. M. (2003). Strong approximation for long memory processes with applications. *Journal of Theoretical Probability*, *16*(2), 377–389.
Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Series A*, *26*, 359–372.
Weinrib, A. (1984). Long-range correlated percolation. *Physical Review, B*, *29*, 387–395.
Weiss, A. A. (1986). Asymptotic theory for ARCH models: estimation and testing. *Econometric Theory*, *2*, 107–131.
Whistler, D. E. N. (1990). *Semiparametric models of daily and intra-daily exchange rate volatility*. Ph.D. dissertation, Univ, London.
Whitcher, B., & Jensen, M. J. (2000). Wavelet estimation of a local long memory parameter. *Exploration Geophysics*, *31*, 94–103.
Whitcher, B., Guttorp, P., & Percival, D. B. (2000). Multiscale detection and location of multiple variance changes in the presence of long memory. *Journal of Statistical Computation and Simulation*, *68*(1), 65–87.
Whitcher, B., Byers, S. D., Guttorp, P., & Percival, D. B. (2002). Testing for homogeneity of variance in time series: long memory, wavelets and the Nile river. *Water Resources Research*, *38*(5), 1000–1029.
Whitt, W. (2002). *Stochastic-process limits. An introduction to stochastic-process limits and their application to queues*. *Springer series in operations research*. New York: Springer.
Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för Matematik*, *2*, 423–434.
Whittle, P. (1956). On the variation of yield variance with plot size. *Biometrika*, *43*, 337–343.
Whittle, P. (1962). Gaussian estimation in stationary time series. *Bulletin de L'Institut International de Statistique*, *39*, 105–129.
Willinger, W., Paxson, V., Riedi, R. H., & Taqqu, M. S. (2003). Long-range dependence and data network traffic. In *Theory and applications of long-range dependence* (pp. 373–407). Boston: Birkhäuser Boston. 2003.
Wilson, E. B., & Hilferty, M. M. (1929). Note on C.S. Peirce's experimental discussion of the law of errors. *Proceedings of the National Academy of Sciences of the United States of America*, *15*(2), 120–125.
Withers, C. S. (2000). A simple expression for the multivariate Hermite polynomials. *Statistics & Probability Letters*, *47*(2), 165–169.
Woodward, W. A., Cheng, Q. C., & Gray, H. L. (1998). A k-factor GARMA long-memory model. *Journal of Time Series Analysis*, *19*(5), 485–504.
Wornell, W., & Oppenheim, A. V. (1992). Estimation of fractal signals from noisy measurements using wavelets. *IEEE Transactions on Signal Processing*, *40*(3), 611–623.
Wright, J. H. (1998). Testing for a structural break at unknown date with long-memory disturbances. *Journal of Time Series Analysis*, *19*(3), 369–376.
Wright, J. H. (2002). Log-periodogram estimation of long memory volatility dependencies with conditionally heavy tailed returns. *Econometric Reviews*, *21*(4), 397–417.
Wu, W. B. (2003). Empirical processes of long-memory sequences. *Bernoulli*, *9*(5), 809–831.

Wu, W. B. (2005). On the Bahadur representation of sample quantiles for dependent sequences. *The Annals of Statistics*, *33*(4), 1934–1963.

Wu, W. B., & Mielniczuk, J. (2002). Kernel density estimation for linear processes. *The Annals of Statistics*, *30*, 1441–1459.

Xiao, Y. (1997a). Hausdorff measure of the graph of fractional Brownian motion. *Mathematical Proceedings of the Cambridge Philosophical Society*, *122*, 565–576.

Xiao, Y. (1997b). Hausdorff-type measures of the sample paths of fractional Brownian motion. *Stochastic Processes and Their Applications*, *74*, 251–272.

Xie, H. (1993). *Fractals in rock mechanics*. Rotterdam: Balkema.

Yajima, Y. (1985). On estimation of long-memory time series models. *Australian Journal of Statistics*, *27*(3), 303–320.

Yajima, Y. (1988). On estimation of a regression model with long term errors. *The Annals of Statistics*, *16*(2), 791–807.

Yajima, Y. (1991). Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *The Annals of Statistics*, *19*, 158–177.

Yajima, Y. (1993). Asymptotic properties of estimates in incorrect ARMA models for long-memory time series. In *New directions in time series analysis, Part II* (pp. 375–382). New York: Springer.

Yajima, Y. (1996). *Estimation of the frequency of unbounded spectral densities* (Discussion Paper). Faculty of Economics, University of Tokyo.

Yao, Y. C. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *The Annals of Statistics*, *15*, 1231–1238.

Youndjé, E., & Vieu, P. (2006). A note on quantile estimation for long-range dependent stochastic processes. *Statistics & Probability Letters*, *76*(2), 109–116.

Zaffaroni, P. (2004). Contemporaneous aggregation of linear dynamic models in large economies. *Journal of Econometrics*, *120*, 75–102.

Zaffaroni, P. (2007a). Aggregation and memory of models of changing volatility. *Journal of Econometrics*, *136*, 237–249.

Zaffaroni, P. (2007b). Contemporaneous aggregation of GARCH processes. *Journal of Time Series Analysis*, *28*, 521–544.

Zaffaroni, P. (2009). Whittle estimation of EGARCH and other exponential volatility models. *Journal of Econometrics*, *151*, 190–200.

Zähle, M. (1998). Integration with respect to fractal functions and stochastic calculus. *Probability Theory and Related Fields*, *111*(3), 333–374.

Zemanian, A. H. (2010). *Distribution theory and transform analysis: an introduction to generalized functions, with applications*. New York: Dover.

Zhao, Z., & Wu, W. B. (2008). Confidence bands in nonparametric time series regression. *The Annals of Statistics*, *36*(4), 1854–1878.

Zhou, Z., & Wu, W. B. (2010). On linear models with long memory and heavy-tailed errors. *Journal of Multivariate Analysis*, *102*(2), 349–362.

Zivot, E., & Wang, J. (2003). *Modeling financial time series with S-PLUS*. New York: Springer.

Zolotarev, V. M. (1986). *Translations of mathematical monographs: Vol. 65. One dimensional stable distributions*. Providence: Am. Math. Soc. Translation from the original 1983 Russian edition.

Zygmund, A. (1968). *Trigonometric series* (Vol. 1). Cambridge: Cambridge University Press.

Author Index

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013

Subject Index

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7, © Springer-Verlag Berlin Heidelberg 2013