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 $\hat{\beta}_{\mathrm{LSE}}$ and $\hat{\beta}_{\mathrm{BLUE}}$ 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*

$$n^{-2d} D_n \operatorname{var}(\hat{\beta}_{\mathrm{LSE}}) D_n \to V_{\mathrm{LSE}} = 2\pi R^{-1}(0) V^* R^{-1}(0). \tag{7.24}$$

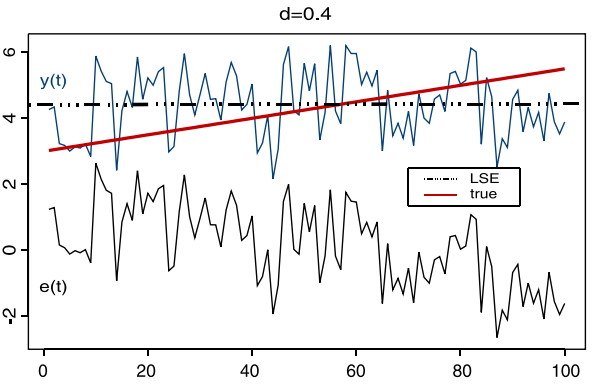

Fig. 7.1 $Y_t = 3 + 0.025t + e_t$ $(t = 1, 2, \ldots, 1000)$ where e_t is a FARIMA$(0, d, 0)$ process $e_t = (1 - B)^{-d}\varepsilon_t$ with $d = 0.4$ and $\mathrm{var}(\varepsilon_t) = 1$. The true trend function (*full line*) and the fitted least squares line (*dotted line*) are also plotted

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.

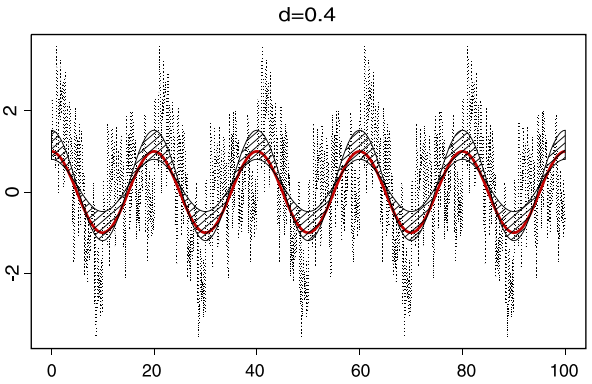

Fig. 7.2 $Y_t = \cos(2\pi t/100) + e_i$ $(t = 1, 2, \ldots, 1000)$ where e_t is a FARIMA$(0, d, 0)$ process $e_t = (1 - B)^{-d}\varepsilon_t$ with $d = 0.4$ and $\mathrm{var}(\varepsilon_t) = 1$. The true trend function (*full line*) is also plotted. The *shaded area* represents a 95 %-confidence region 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 $C_n = [c_{jl}^{(n)}]_{j,l=1,\dots,p} = D_n^{-1} X^T A_n X D_n^{-1}$,

$$c_{jl}^{(n)} = \sum_{t,s=1}^{n} \frac{x_{tj}}{\|x_{\cdot j}\|} \frac{x_{tl}}{\|x_{\cdot l}\|} \int_{-\pi}^{\pi} e^{i(j-l)\lambda} g(\lambda)\, d\lambda = \int_{-\pi}^{\pi} g(\lambda)\, dM_{jl}^{(n)}(\lambda)$$

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 $x_{tj} = t^{j-1}$ $(j = 1, \dots, p)$. 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 $W = [w_{jl}]_{j,l=1,\dots,p}$ 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

$$W = w_{11}\begin{pmatrix} 1 & \frac{\sqrt{3}(1-d)}{2-2d} \\ \frac{\sqrt{3}(1-d)}{2-2d} & \frac{3(1-d)^2}{(3-2d)(1-2d)} \end{pmatrix}.$$

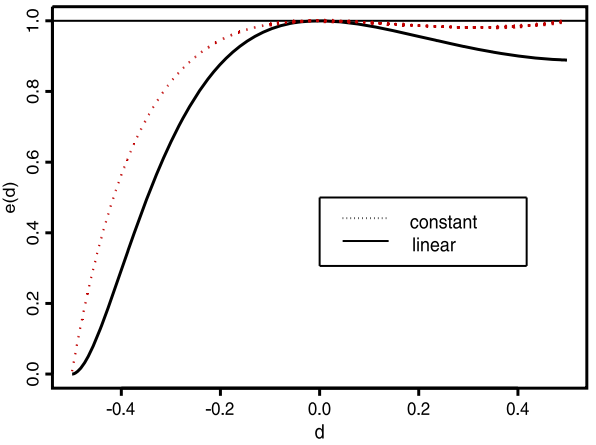

Fig. 7.3 Relative asymptotic efficiency $e(d) = \det(V_{\mathrm{BLUE}})/\det(V_{\mathrm{LSE}})$ of the least squares estimator in a linear regression model $Y_t = \beta_0 + \beta_1 t + e_t$ (*full linear*) and a regression model with $\beta_1 = 0$, i.e. $Y_t = \beta_0 + e_t$ (*dotted line*)

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,

$$B_k = \big\{(k-1)l + 1, \ldots, kl\big\}$$

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

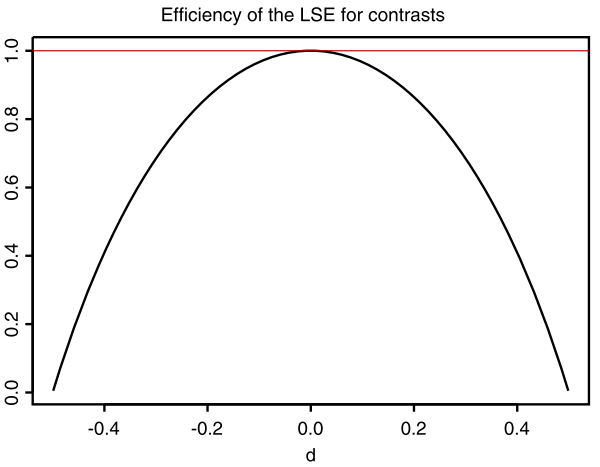

Fig. 7.4 Relative asymptotic efficiency of the LSE of a contrast $\beta_j - \beta_k$ compared to the BLUE, as a function of d. The model considered here is a FARIMA$(0, d, 0)$ process

$$E[V_{n,\mathrm{LSE}}] = \left(\frac{1}{n_j} + \frac{1}{n_k}\right)\left[\sigma_e^2 - \frac{2}{l-1}\sum_{k=1}^{l-1}\left(1 - \frac{k}{l}\right)\gamma_e(k)\right]$$
$$= \left(\frac{1}{n_j} + \frac{1}{n_k}\right)\sigma_l^2.$$

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, $s_{\beta_0}^2$ and $s_{\beta_1}^2$ 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

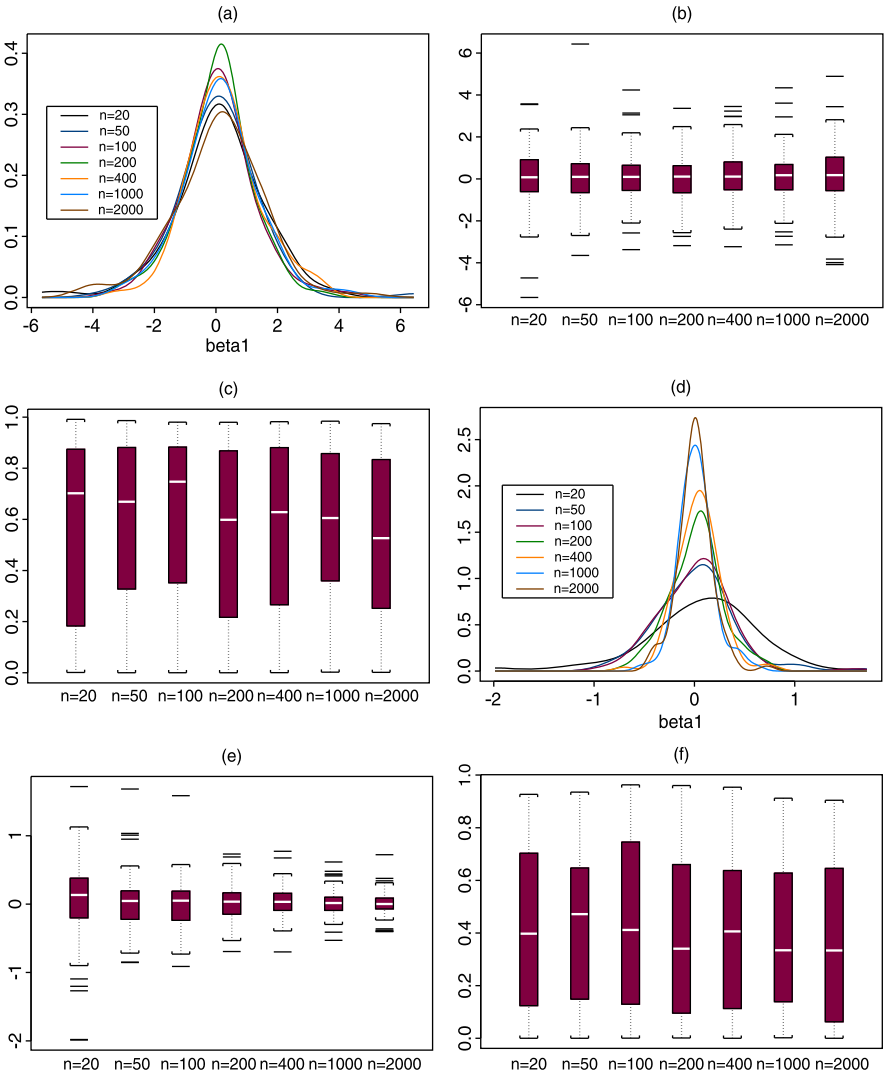

Fig. 7.5 Simulated distributions and boxplots of $\hat{\beta}_1$ in a regression of two independent integrated FARIMA$(0, d, 0)$ processes with $d_1 = d_2 = 0.4$ ((**a**) and (**b**)) and $d_1 = 0.1$, $d_2 = 0.4$ ((**d**) and (**e**)), respectively. The sample sizes are $n = 20, 50, 100, 200, 400, 1000$ 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

$$\mathbf{M}_{n+} = \big[m_{(j+)}(t/n)\big]_{t=1,\ldots,n;\, j=1,\ldots,p+1} \in \mathbb{R}^{n\times(p+1)} \tag{7.80}$$

the limit

$$\lim_{n\to\infty} n^{-1}\big(\mathbf{M}_{n+}^T \mathbf{M}_{n+}\big)_{jk} = \int_0^1 m_{(j+)}(s,\theta) m_{(k+)}(s,\theta)\, ds \tag{7.81}$$

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

$$\Lambda = \lim_n n\big(\mathbf{M}_{n+}^T \mathbf{M}_{n+}\big)^{-1}. \tag{7.82}$$

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

$$\lim_{n\to\infty} cov\big(n^{\frac{1}{2}-d} v^{-\frac{1}{2}}(d)\big(\mathbf{M}_{n+}^T \mathbf{M}_{n+}\big)^{-1} \mathbf{M}_{n+}^T e(n)\big) = \Lambda \Sigma_0 \Lambda \tag{7.83}$$

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 $n \to \infty$, $\operatorname{var}(\hat{\gamma}(k)) = o(1)$, but

$$\text{Bias} = E\big[\hat{\gamma}(k)\big] - \gamma_e(k) \sim \int \left[m(t) - \int m(s)\, ds \right]^2 dt, \tag{7.91}$$

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

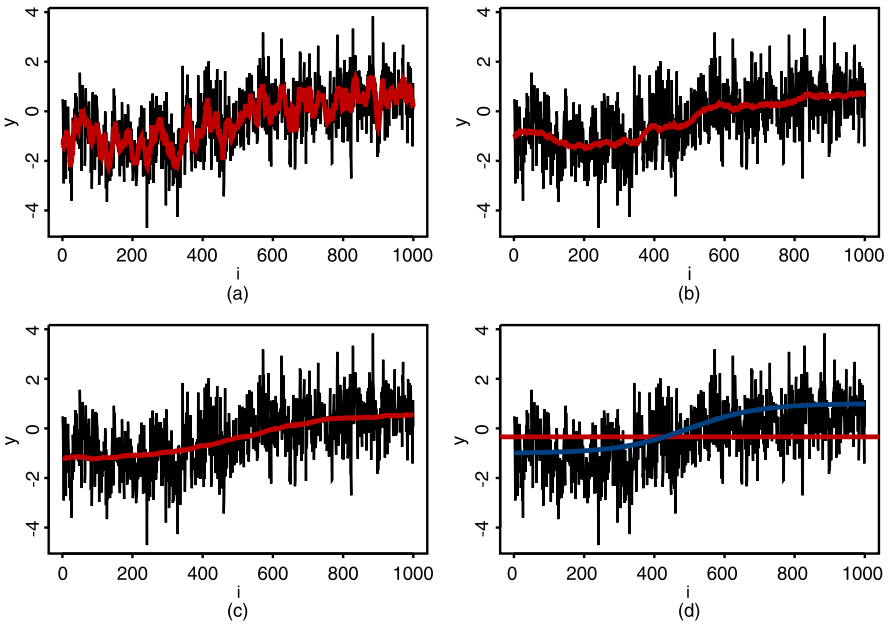

Fig. 7.6 The four pictures show the same series $Y_i = m(t_i) + e_i$ with $m(t) = \tanh(\frac{1}{2}(t - \frac{1}{2}))$ and e_i generated by a FARIMA$(0, 0.3, 0)$ process with innovation variance one. The four figures show nonparametric fits $\hat{m}(t)$ based on kernel regression with the rectangular 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

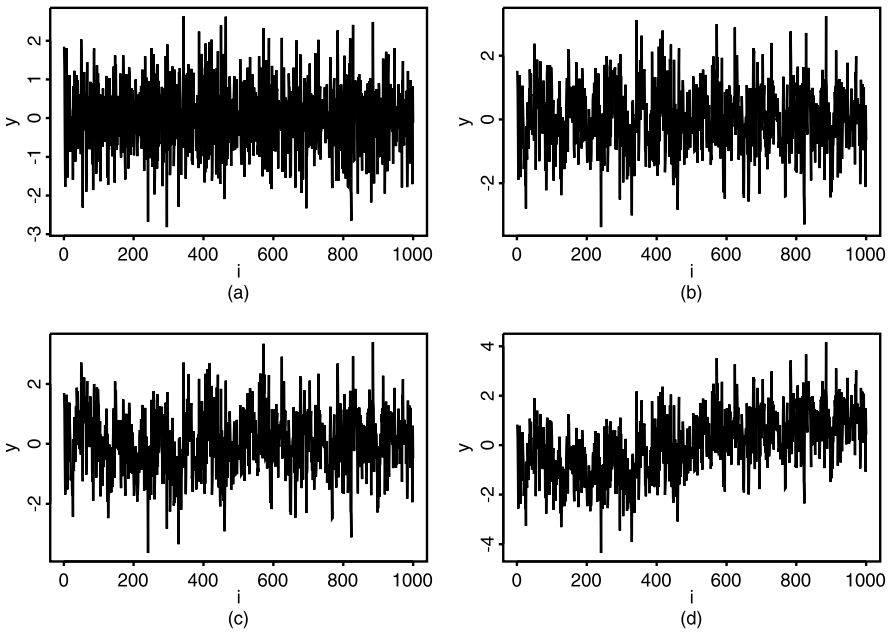

Fig. 7.7 Residuals $\hat{e}_i = Y_i - \hat{m}(t_i)$ based on the fits in Figs. 7.6(a)–(d)

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	∞	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$ $(-\infty < u < \infty)$

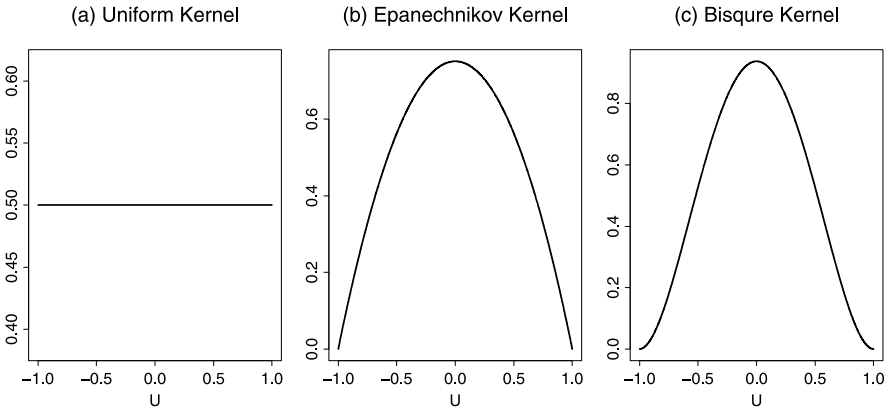

Fig. 7.8 Three commonly used second-order kernels with compact support

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 $\mathbf{N}_p^{-1} = (a_{ij})_{i,j=1,\dots,p+1}$. 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 $t \in [b, 1 - b]$,

$$\text{Bias} = E\big[\hat{m}(t)\big] - m(t) = b^2 \frac{1}{6} m^{(2)}(t) + o\big(b^2\big)$$

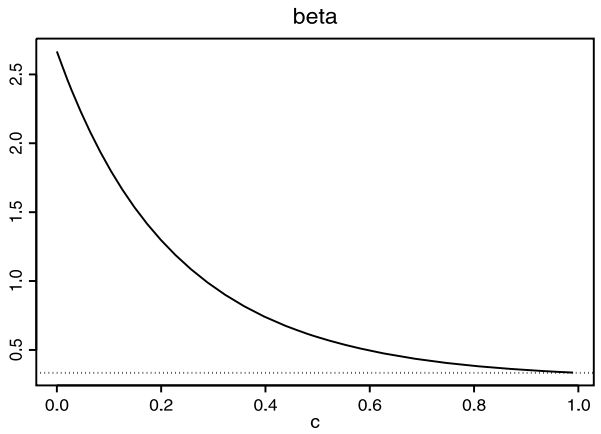

Fig. 7.9 Plot of $\beta_{(0,2,c)}$ for $0 \le c < 1$ and $K_{(0,2,c)}$ derived from the rectangular kernel

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.

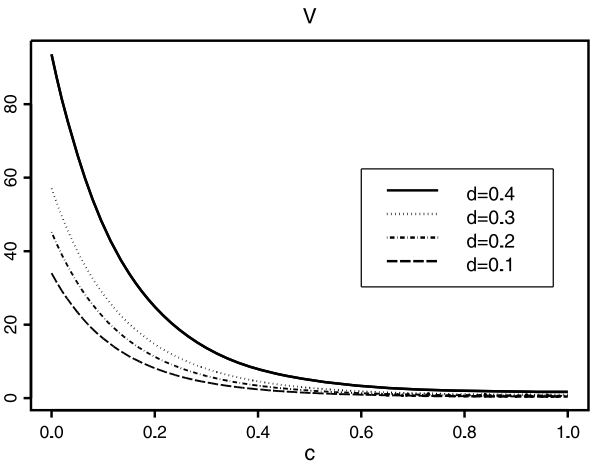

Fig. 7.10 $V_{(0,2,c)}(d)$ plotted as a function of $c \in [0, 1)$ for different values of $d \in (0, \frac{1}{2})$

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

$$IMSE = \int_0^1 MSE(t)\,dt \tag{7.131}$$

$$\sim b^{2(k-j)} \left(\frac{\beta_{(j,k,1)}}{k!} \right)^2 I_k + (nb)^{2d-1} b^{-2j} V_{(j,k,1)}(d) \tag{7.132}$$

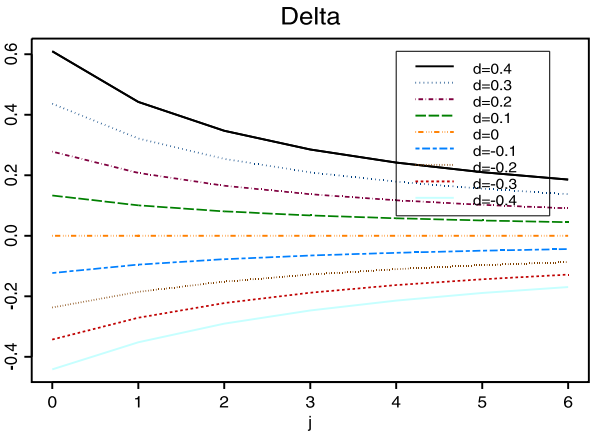

Fig. 7.11 Change Δr of the optimal exponent r_{opt} in $MSE_{\text{opt}}(\hat{m}^{(j)}) = O(n^{-r_{\text{opt}}})$ compared to the case of short memory, as a function of j, plotted for different values of d

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.*

$$b = c_b \cdot n^{-\alpha_{\mathrm{opt}}}$$

(*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,

$$(nb)^{1/2-d} V_{(0,2,1)}^{-\frac{1}{2}} \big(\hat{m}(t_1) - m(t_1), \ldots, \hat{m}(t_l) - m(t_l)\big) \underset{d}{\to} (Z_1, \ldots, Z_l) \qquad (7.144)$$

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):

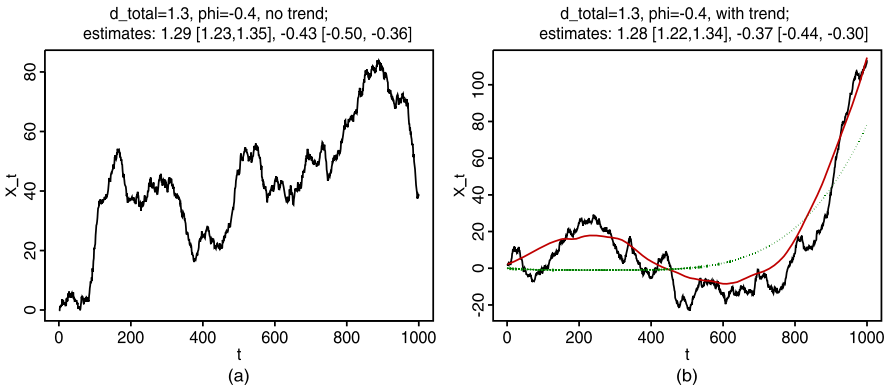

Fig. 7.12 (**a**) Simulated FARIMA$(p^0, d^0, 0)$ series with $p^0 = 1$, $d^0_{\text{total}} = 1.3$ ($r^0 = 1$, $d = 0.3$) and $\varphi_1^0 = -0.4$. This is the same as a SEMIFAR model with the same parameters and $m(t) \equiv 0$. (**b**) SEMIFAR process with the same parameters as in (**a**), but including 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 $\varphi_1^0 = -0.4$. 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.

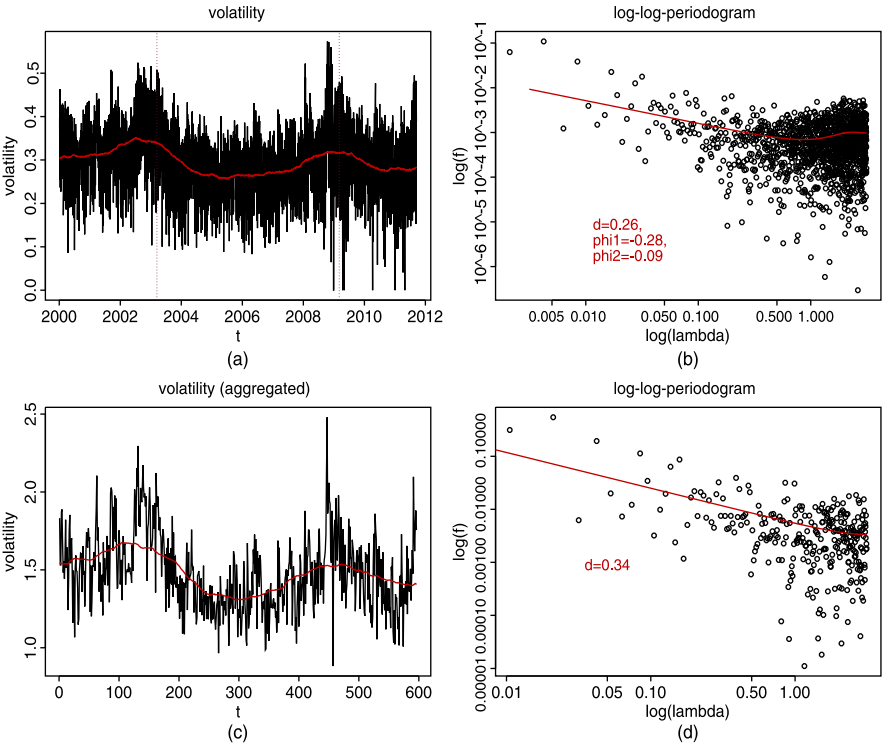

Fig. 7.13 Volatility series for the DAX between January 3, 2000 and September 12, 2011. (**a**) Shows daily data together with a nonparametric trend function fitted by Algorithm B. The corresponding log–log-plot of the periodogram together with the fitted spectral density is displayed in (**b**). (**c**) 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

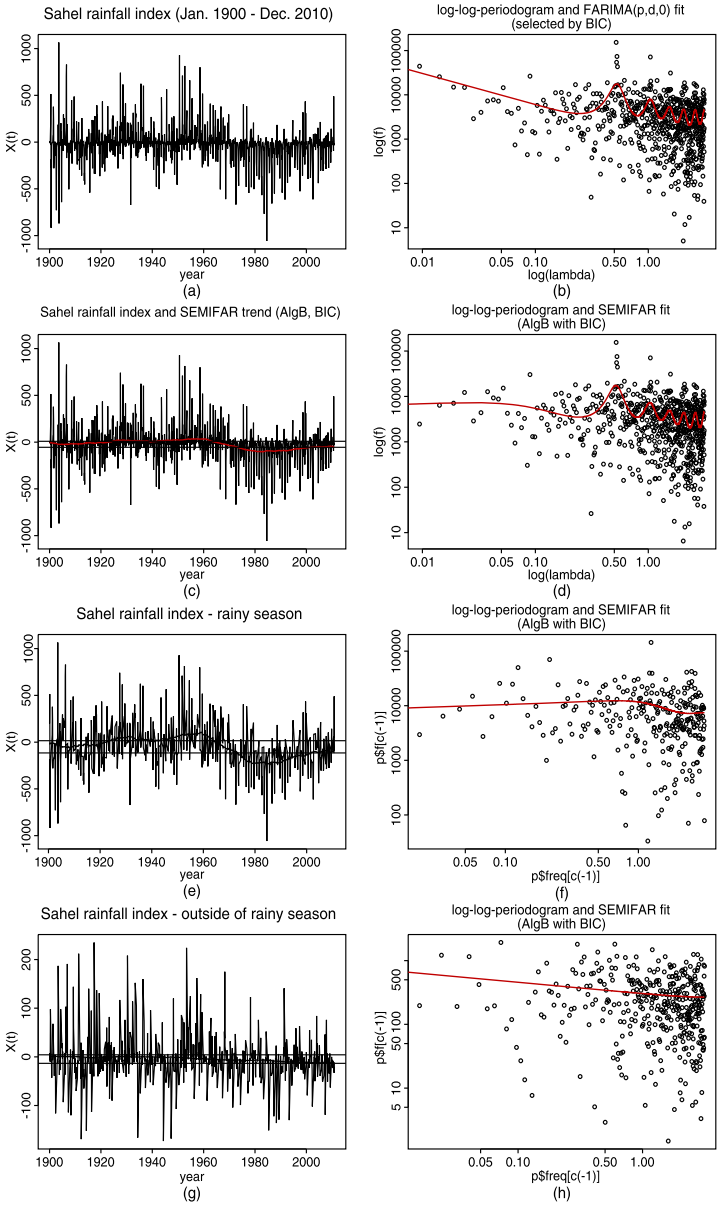

Fig. 7.14 Monthly precipitation anomalies for the Sahel region between January 1900 to December 2011 (data courtesy of Todd Mitchell, JISAO, University of Washington; http://www.ncdc.noaa.gov/temp-and-precip/ghcn-gridded-products.php): (**a**) original series; (**b**) log–log-periodogram and spectral 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 $b_{\text{opt}} = C_{\text{opt}} n^{-\frac{1}{5}}$, 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 $b_{\text{opt}} = C_{\text{opt}} n^{-\frac{1}{5}}$ 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

$$\gamma = \frac{v r \alpha}{2r + \alpha} \quad \text{if } r \geq \frac{\alpha}{2}\left(\frac{v}{\lambda} - 1\right), \tag{7.204}$$

$$r - \left(\frac{1}{\lambda} - \frac{1}{v}\right)_+ > \frac{r}{2r + \alpha}, \tag{7.205}$$

$$\gamma = \frac{\alpha v (r - \frac{1}{\lambda} + \frac{1}{v})}{2(r - \frac{1}{\lambda} + \frac{\alpha}{2})} \quad \text{if } \frac{1}{\lambda} < r < \frac{\alpha}{2}\left(\frac{v}{\lambda} - 1\right). \tag{7.206}$$

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 $(0 < \alpha < 1)$

$$\theta_t(\alpha) = \inf_y \big\{ y | F_t(y) \ge \alpha \big\},$$

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,*

$$\begin{aligned}
\operatorname{Bias}\big(\hat{\theta}_t(\alpha)\big) &= \frac{b^2}{f_t(\theta_t(\alpha))} A\big(t, \theta_t(\alpha)\big) + o\big(b^2\big), \\
\operatorname{Var}\big(\hat{\theta}_t(\alpha)\big) &= (nb)^{m(2H-2)} \frac{B(t, \theta_t(\alpha))}{f_t^2(\theta_t(\alpha))} + o\big((nb)^{m(2H-2)}\big), \\
MSE\big(\hat{\theta}_t(\alpha)\big) &= \left[\frac{A^2(t, \theta_t(\alpha))}{f_t^2(\theta_t(\alpha))} b^4 + \frac{B(t, \theta_t(\alpha))}{f_t^2(\theta_t(\alpha))} (nb)^{m(2H-2)} \right] \\
&\quad + o\big(\max\big(b^4, (nb)^{m(2H-2)}\big)\big).
\end{aligned}$$

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

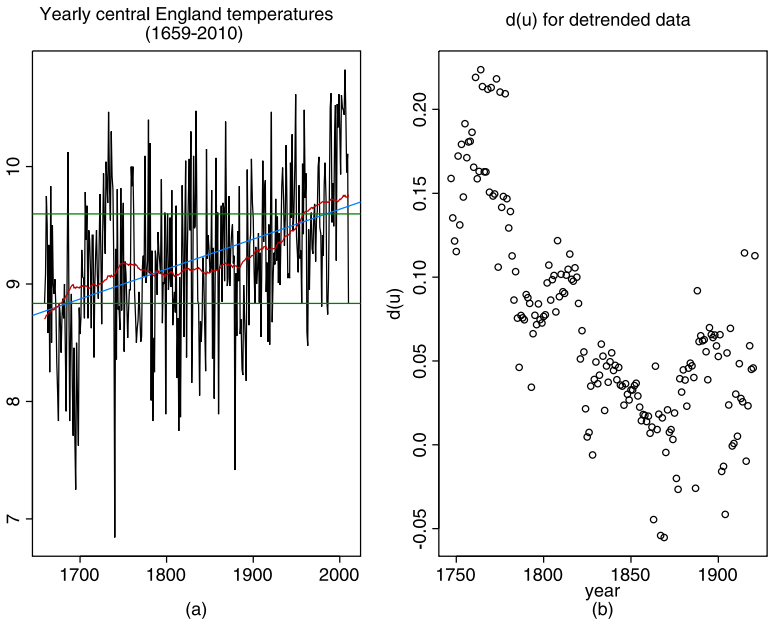

Fig. 7.15 (**a**) Central England temperature series with fitted linear and nonparametric trend function respectively; (**b**) local maximum likelihood estimates of d for detrended series, based on moving blocks of 176 years and a fractional ARIMA$(0, d, 0)$ model

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*:

$$E\big[\hat{d}(u_0)\big] - d^0(u_0) = b^2\frac{1}{2}d''(u_0)\int_{-1}^{1} K(x)x^2\,dx + o\big(b^2\big); \tag{7.234}$$

2. *Variance*:

$$\operatorname{var}\big[\hat{d}(u_0)\big] = (nb)^{-1}J^{-1}\int_{-1}^{1} K^2(x)\,dx + o\big((nb)^{-1}\big) \tag{7.235}$$

$$= (nb)^{-1}\frac{6}{\pi^2}\int_{-1}^{1} K^2(x)\,dx + o\big((nb)^{-1}\big); \tag{7.236}$$

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 $\operatorname{var}(S_n(1)) \sim c_S n$. Thus, under H_0, we have a functional limit theorem with $\tilde{Z}_n(u) = (S_n(u) - u S_n(1)) c_S^{-\frac{1}{2}} n^{-\frac{1}{2}}$ converging to a Brownian bridge $\tilde{B}(u) = B(u) - u B(1)$, 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

$$\tilde{Z}_n(u) \approx L_S^{-\frac{1}{2}}(n) n^{-H} \left(S_n(u) - u S_n(1)\right)$$

(with L_S a slowly varying function as defined in Sect. 4.2.2) converges to a fractional Brownian bridge $\tilde{B}_H(u) = B_H(u) - u B_H(1)$. 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 $S_n(u) - u S_n(1)$ 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

$$L_S^{\frac{1}{2}}(n)n^H = \sqrt{\nu(d)c_{f_X}}\,n^{d+\frac{1}{2}} = \sqrt{\nu(d)\,f_X\big(n^{-1}\big)}\,n^{\frac{1}{2}}$$

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

$$P\Big(\sup_{u\in[0,1]}\big|\tilde{B}_{\hat{H}}(u)\big| > q_{1-\alpha}\Big) = \alpha.$$

(Note that here the probability is evaluated for a fractional Brownian bridge with $\hat{H}$ being fixed.)

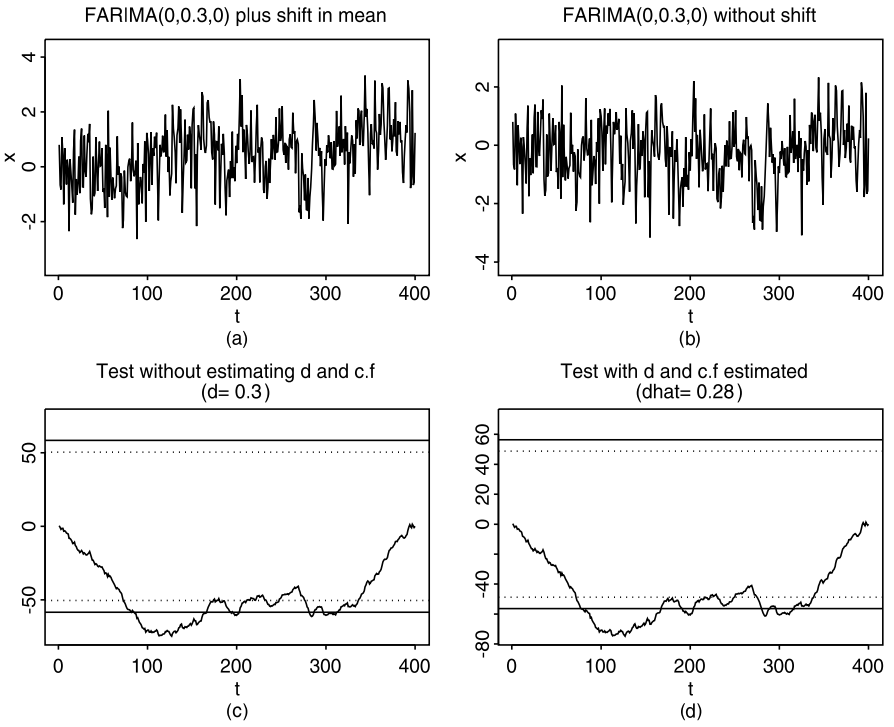

Fig. 7.16 Simulated sample paths of $Y_t = \Delta \cdot 1\{t \geq 120\} + X_t$ (**a**) and X_t (**b**) where X_t is a FARIMA$(0, 0.3, 0)$ process and $\Delta = 1$. The values of $V_i = S_{1,i} - (i/n)S_{1,n}$ are plotted against i in (**c**) and (**d**), with 5 %- and 10 %-critical values (*horizontal 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.

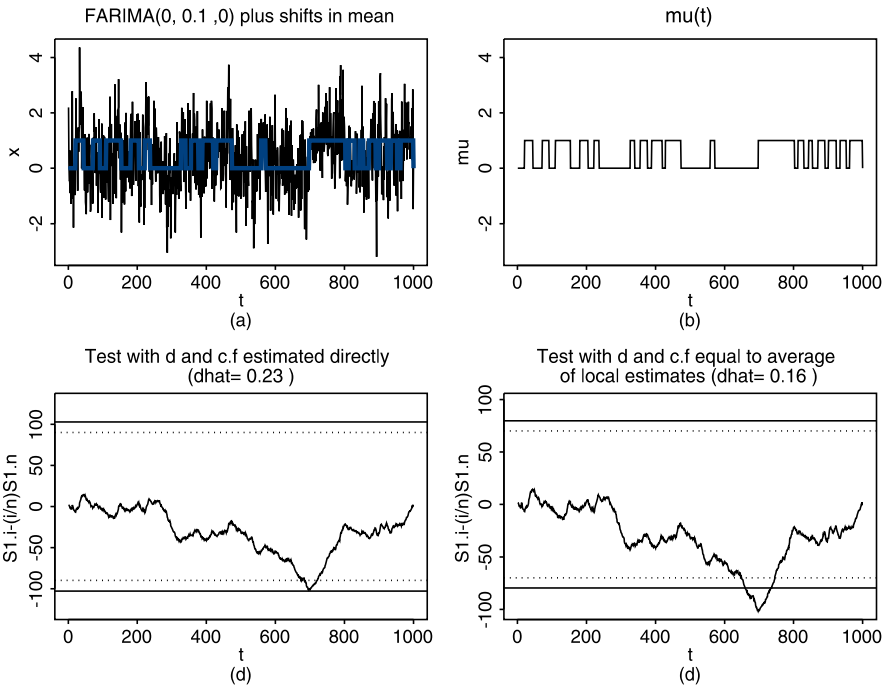

Fig. 7.17 Figure (**a**) shows a sample path of $Y_t = \mu(t) + X_t$ where X_t is a FARIMA$(0, 0.1, 0)$ process and $\mu(t)$ has multiple change points with values switching between 0 and 1 as displayed in (**b**). The values of $V_i = S_{1,i} - (i/n)S_{1,n}$ are plotted in (**c**) and (**d**). The *horizontal 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).

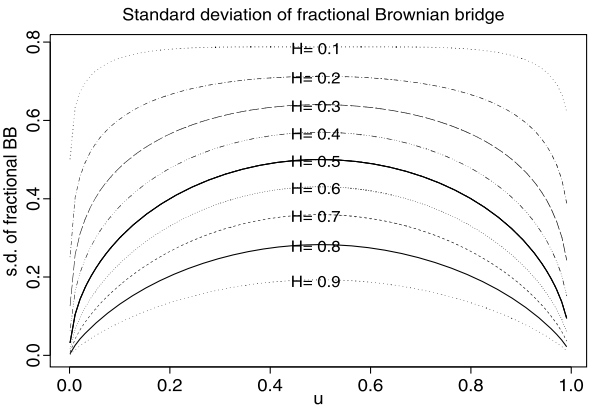

Fig. 7.18 Standard deviation of a fractional Brownian bridge $\tilde{B}_H(u)$

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,

$$L_Y(n) = \Gamma(1-2d)/\big[\Gamma(d)\Gamma(1-d)\big]$$

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

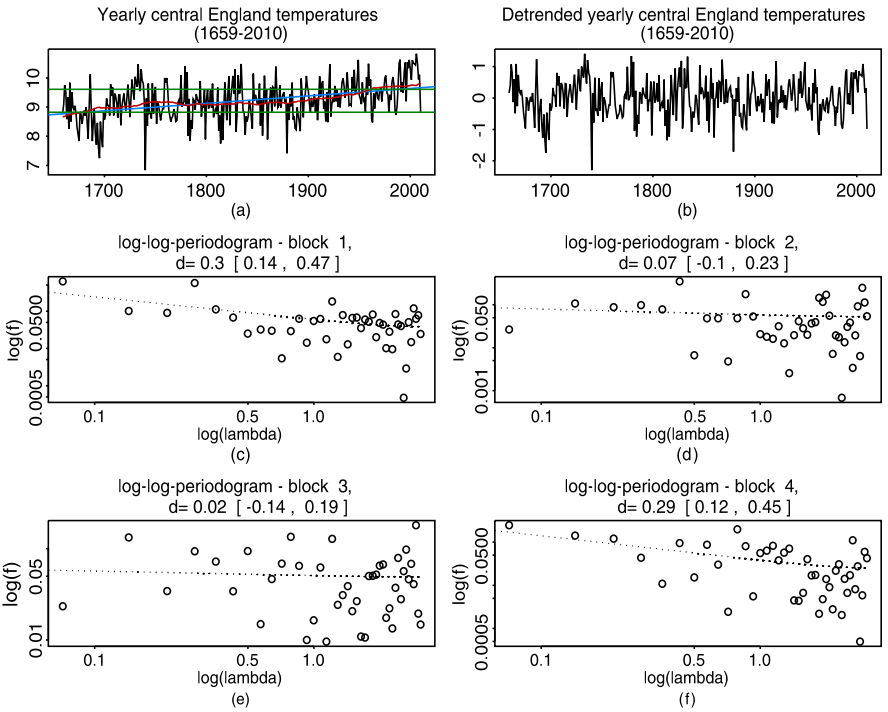

Fig. 7.19 Yearly Central England temperatures 1659–2010 (**a**) and the detrended series (**b**) after subtracting a nonparametric trend function. Also displayed are log–log-periodograms and FARIMA$(0, d, 0)$ spectral densities fitted to four disjoint blocks of length $n_j = 88$

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

$$\tilde{T} > \frac{c(x_\alpha)}{\sqrt{2\log n}},$$

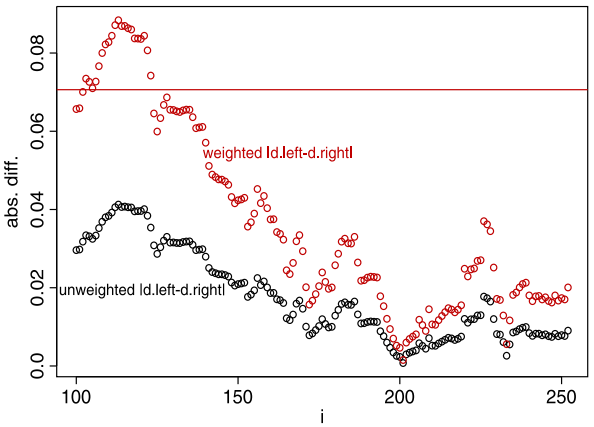

Fig. 7.20 Plot of $|\frac{i}{n}(1-\frac{i}{n})(\hat{d}_{1,i}-\hat{d}_{i+1,n})|$ and $|\sqrt{\frac{i}{n}(1-\frac{i}{n})}(\hat{d}_{1,i}-\hat{d}_{i+1,n})|$ against $i = 100, \ldots, 252$ for detrended yearly Central England temperatures. The *horizontal line* corresponds to the 5 %-critical value for the 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.

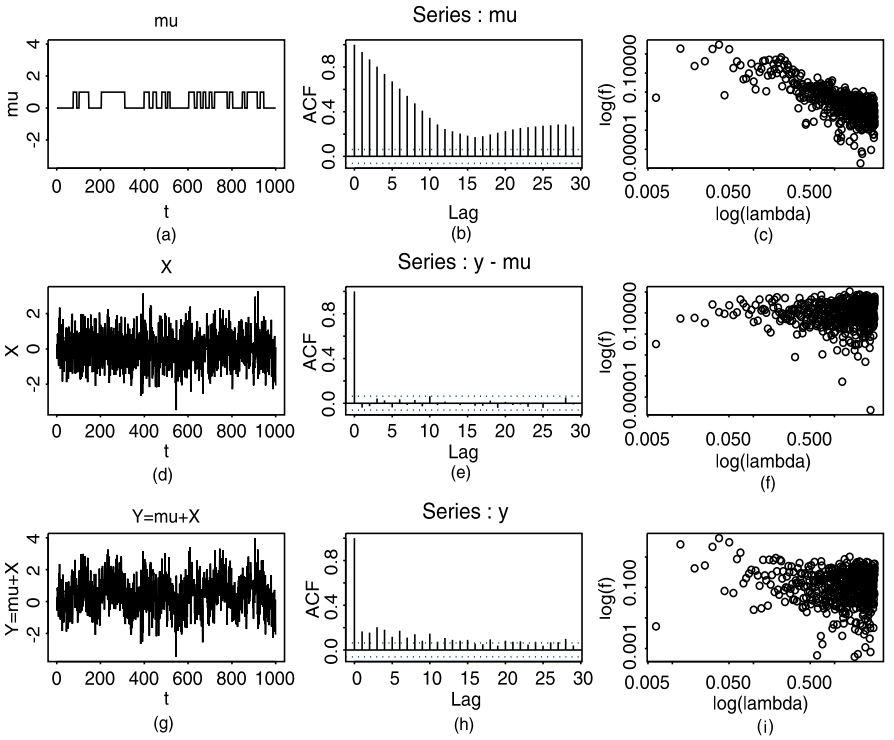

Fig. 7.21 Figure (**g**) shows a simulated sample path of $Y_t = \mu(t/n) + X_t$ where X_t are i.i.d. $N(0, 1)$-variables and $\mu(u)$ $(u \in [0, 1])$ is generated by an ON–OFF-process with long-range dependence. The ON–OFF-process is displayed in (**a**), the residual process X_t in (**d**). Also shown 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* $t \in (0,1)$:

$$\begin{aligned}
\operatorname{Bias}\big(\hat{m}^{(\nu)}(t)\big) &= E\big[\hat{m}^{(\nu)}(t)\big] - m^{(\nu)}(t) = b^{k-\nu}J_{\nu,k} + o\big(b^{k-\nu}\big),\\
\operatorname{Var}\big(\hat{m}^{(\nu)}(t)\big) &= b^{-2\nu}(T_n b)^{(2H-2)q} I_q(t) + o\big(b^{-2\nu}(T_n b)^{(2H-2)q}\big),\\
MSE\big(\hat{m}^{(\nu)}(t)\big) &= E\big[\big(\hat{m}^{(\nu)}(t) - m^{(\nu)}(t)\big)^2\big]\\
&= b^{2(k-\nu)}J_{\nu,k}^2(t) + b^{-2\nu}(T_n b)^{(2H-2)q} I_q(t)\\
&\quad + o\big(\max\big(b^{2(k-\nu)}, b^{-2\nu}(T_n b)^{(2H-2)q}\big)\big)
\end{aligned}$$

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*

$$\mu_i = \left[\frac{m^{(k)}(\tau_i)}{k!}\int_{-1}^{1} K^{(\nu)}(u)u^{k-\nu}\,du\right]\Big/ m^{(3)}(\tau_i).$$

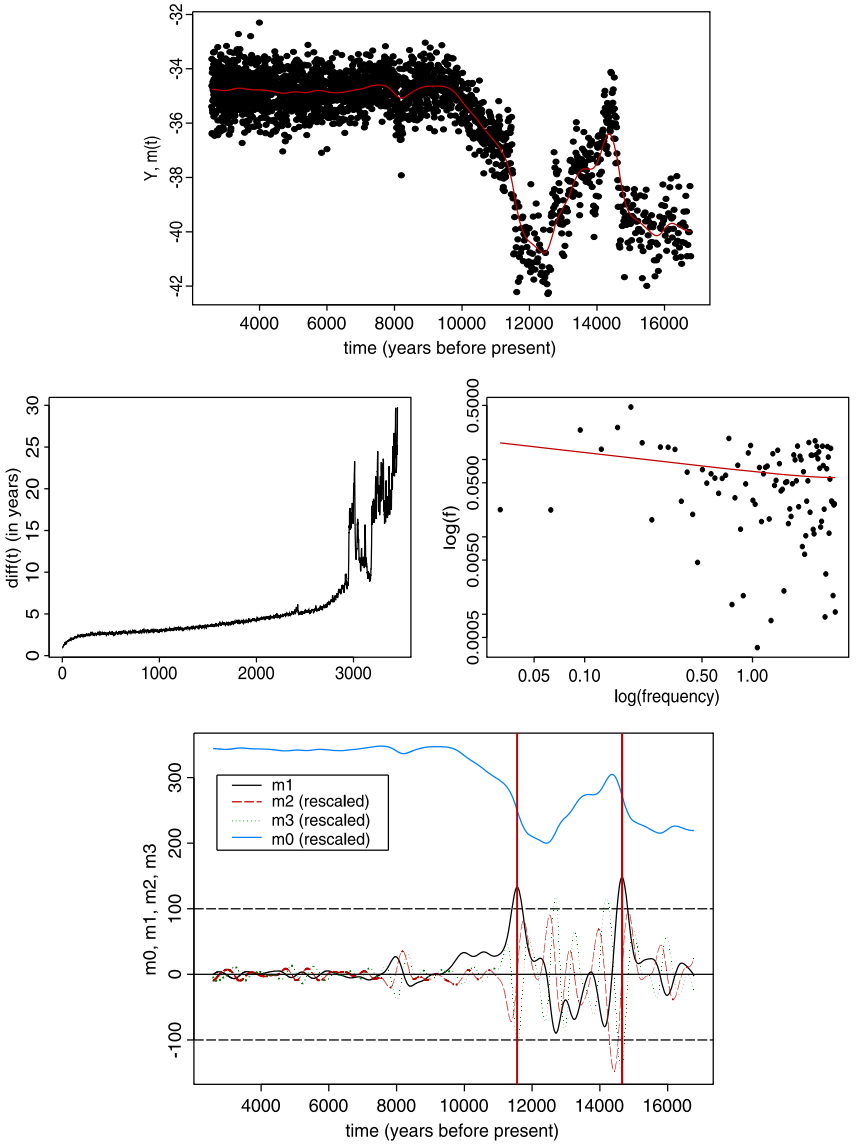

Fig. 7.22 *Top*: Oxygen isotope values plotted against age (years before present or 1989) and an estimated trend curve. *Left middle*: Distance between successive time points. *Right middle*: Periodogram of residuals and fitted spectral density in log-log coordinates. *Bottom*: Estimated trend derivatives $\widehat{m^{(\nu)}}$ ($\nu = 0, 1, 2, 3$). 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).