Dietmar Ferger · Wenceslao González Manteiga
Thorsten Schmidt · Jane-Ling Wang *Editors*

From Statistics to Mathematical Finance

Festschrift in Honour of Winfried Stute

Springer

From Statistics to Mathematical Finance

Dietmar Ferger
Wenceslao González Manteiga
Thorsten Schmidt · Jane-Ling Wang
Editors

From Statistics to Mathematical Finance

Festschrift in Honour of Winfried Stute

Springer

Editors
Dietmar Ferger
Institute of Mathematical Stochastics
Technische Universität Dresden
Dresden
Germany

Thorsten Schmidt
Institute of Mathematics
University of Freiburg
Freiburg
Germany

Wenceslao González Manteiga
Faculty of Mathematics
University of Santiago de Compostela
Santiago de Compostela
Spain

Jane-Ling Wang
Department of Statistics
University of California
Davis, CA
USA

ISBN 978-3-319-50985-3 ISBN 978-3-319-50986-0 (eBook)
DOI 10.1007/978-3-319-50986-0

Library of Congress Control Number: 2017945698

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Scenes from my Life—Winfried Stute

I was born on 20 November 1946, in Bochum (West Germany), a city which is located in the Ruhr Valley, a heavily industrialized region comprising around 40 cities, whose economic basis at that time was the coal and steel industry. People from the area take pride in their working-class background, though they do not usually talk much about it. What they do make known is their passion for football, and wearing a shirt with the right team colours is almost a matter of religion.

In the early 1950s, family apartments were still small, so to meet friends one had to play in the streets or the ruins left over from the war. I had no brothers and sisters, but my cousins all lived in the same neighbourhood. What I remember clearly was the extreme optimism of the people at the time and their planning for the future. Only when I became older did I begin to understand that, having lived through the war, they were all survivors of a great disaster and that their attitude really reflected a celebration of their survival.

When I was four or so I developed some "emotional distance" to people dressed in white, e.g. medical doctors, pharmacists or the nuns running the kindergarten. I decided to escape on the very first day and not to go back again. School was later something which I took more seriously. Unfortunately, when I was seven, my father, who was a manager in the furniture business, started what would become a random walk-through the cluster of cities in the region. While changing jobs was good for his career, I always had to make new friends and adapt to levels of teaching, which were potentially different at each new school. The nice consequence of this was that because I was able to teach myself and no tutoring was necessary, my mathematical level in particular was often much higher than that of the rest of the class. So at school mathematics became my success story. As to football, all of the teams I supported were dressed in blue. So blue of course became my favourite colour.

As we became older, teachers expected us to learn a lot by heart, for example in history and geography. At the end of my time at school, I felt fairly exhausted and was ready for a big change. I joined the army for two years and became a truck driver. Unlike other boys of my generation, I never complained about this time, but considered it a pleasant interlude between school and what I was planning to do,

becoming a student of mathematics at the university in my hometown Bochum. The change when I began my studies was dramatic. In the fall of 1968, as in many other western countries, there were a lot of violent protests at German universities. At the mathematical institute in Bochum, it was relatively peaceful, and our provocations were mild. Many of our professors were young, only 10–15 years older than we were ourselves. Some were members of the Bourbaki school and presented the material in the most abstract and concise way. To digest Zorn's lemma was not that easy for somebody who had been driving trucks just weeks before. But overall my new life went on smoothly.

Significant changes then came in 1971. In 1969, I had met my girlfriend Gerti, who came from a city in the region whose football team was considered the most important local rival of my "blue team". Though my friends all warned me about this new "black and yellow" influence, I—luckily—ignored them and we married in 1971. At the same time, a new professor came to our faculty, Peter Gaenssler, who had just completed his habilitation in Cologne and accepted the offer of a professorship from Bochum. He was very different from the "Bourbakis" at the institute, whom we never tried to meet in their office hours. Gaenssler was open-minded and friendly. In his lectures, he focused on topological measure theory and probability. Deep in his heart he was a measure theorist with strong connections to functional analysis, and Donsker's invariance principles, which had become popular through Billingsley's monograph on *Convergence of Probability Measures*, were major applications in these areas, so that students of Gaenssler were quite familiar with concepts such as tightness, convergence of distributions and the continuous mapping theorem.

In 1972 I started to work on my Diploma (master's) thesis, of course on a topic related to convergence of empirical measures. Our daughter Petra was born in the same year, and everything went on very nicely. My thesis was completed in January 1973, and Prof. Gaenssler offered me a position as a teaching assistant. We had several outstanding young students at the time. Two of them were Erich Haeusler who later became my colleague in Giessen, and Michael Falk, who became a professor in Wuerzburg. In 1974, our son Volker was born, and our small family was complete. Regarding work, after my Ph.D. in 1975, Prof. Gaenssler proposed a joint book project, a textbook on advanced probability theory in German, which was completed in 1977 and became very successful.

The year 1978 brought several changes. My advisor had accepted the offer of a position from the University of Munich. So, in May 1978, we all moved to Bavaria. In the following years, we had several visitors from the USA, who would all become important members of the school of empirical processes: D. Mason, D. Pollard and J. Wellner, among others. At that time, Gaenssler became one of the leading researchers in the Vapnik–Chervonenkis group. There were other subgroups of the "empiricals", like the Hungarians with their strong approximations or the counting process group who brought stochastic calculus to statistics. My situation was slightly different. In the German system, the next step in my career was habilitation, an advanced postdoctoral thesis needed to qualify for the offer of a

professorship. Since my knowledge of statistics at the time was still poor, I continued with something which I knew best: tightness of stochastic processes. In the case of empirical processes, this led me to the study of their oscillation modulus. What I did not know in the beginning was that this would immediately lead me to the estimation of local quantities such as densities, regression functions or hazards. In the multivariate case, I came to conditional empirical processes and their inverses, which are now referred to in the research as "quantile regression". My personal hero at that time was Prof. Jack Kiefer (Cornell and Berkeley) whose work and ideas on empiricals gave me a lot. Actually, what I liked very much during the 1980s was to develop applications combining my local results with the global bound in the Dvoretzky–Kiefer–Wolfowitz inequality.

In 1981, one year after my habilitation in Munich, I became an associate professor in Siegen and a colleague of Rolf-Dieter Reiss, an internationally known expert in extreme value theory. Two years later I moved to Giessen, a mid-sized city of 80,000 people 70 km north of Frankfurt/Main. Its university was founded in 1607 and named after the famous chemist Justus von Liebig. The atmosphere at the institute was friendly and open-minded so that I immediately felt at home. My colleague in stochastics at the time was Georg Pflug, who a few years later went back to his hometown of Vienna. Geographically speaking, the move to Giessen was the last big change I would make. As far as my research network and cooperative projects, however, things began to change rapidly. In August 1985, I attended a conference in Bilbao, in the Basque country of Spain. After my talk a young Spanish scientist approached me and introduced himself as Wenceslao Manteiga from Santiago de Compostela, Galicia. It was the starting point of a long and fruitful friendship and cooperation. The next year he invited me to Santiago to give a workshop on empiricals. Many of his Ph.D. students later visited me in Giessen, and my cooperation with researchers across Spain expanded to include many universities in the north and west of the country, and Carlos III from Madrid. The contributions to this volume clearly reflect this part of my activities.

At the end of the 1980s, my interest in local studies, i.e. smoothing, had declined. At the same time, I had intensified my contact with researchers from UC Davis, and in September 1990, Jane Ling Wang came to Giessen for four months. Our plan was to work on the extension of the strong law of large numbers to Kaplan–Meier integrals. To prove it, I proposed a technique which I had learned from Gaenssler, in the simpler classical situation, which used the convergence theorem for reverse martingales. September and October were terrible and full of frustration. As it would turn out later, we were on an impossible mission and that actually, due to censorship, the Kaplan–Meier estimator was biased so that attempting to prove a martingale property was hopeless. As all of our attempts failed Jane Ling made a suggestion which brought my hero from the 1970s back into play. Since she was a former Ph.D. student of Prof. Kiefer, she could tell me what he would have done when things became complicated. In our situation, her simple recipe was to study the martingale property first by comparing the simplest cases, namely sample size $n = 1$ and $n = 2$. The results were striking: the martingale property could be disproved, but the sub- or super-martingale property, depending

on the sign of the integrand, was still valid. After we knew what to look for, we rapidly—at least in my memory—completed the proof. In the 1990s, I continued to work on various open questions in survival analysis, for censored and truncated data and combinations of both. After this, it was time for another change, and my interests began to move towards statistical model checking. Besides my work with my Spanish friends, the list of partners now also included Li Xing Zhu from Hong Kong and Hira Koul from Michigan State. Both visited me on a Humboldt-Award grant, and in our joint work, I enjoyed learning from their ideas and experience.

These developments came to an abrupt end on a Monday afternoon in October 1998. For the opening lecture of the new winter semester, we had invited a former student of ours who had a successful career in the financial sector in Frankfurt. The success of this talk was overwhelming. The following day five students came to my office and told me that they were quite impressed by yesterday's speaker. They also asked me to start a new programme on financial mathematics. I must say that I always liked students who were committed to their subject. So I agreed and began teaching a course the following spring semester. What I found out was that talking about the Black–Scholes formula required not only Ito-calculus, but also a thorough knowledge of the economic theory. To improve my own understanding, I invested some money in highly risky financial derivatives—and survived. I also realized that facing financial risks was very different from erroneously rejecting a statistical hypothesis at the 5% level. My interest in the behaviour of the market grew further, so when I was ready to begin research I looked for models which could incorporate aspects such as shocks, profit-taking or rallies. Finding the Girsanov martingale measure and seeing how things worked in simulations was a new experience for me, which was quite different from what I had done before. And I began to like it. It eventually led me to some cooperative projects in the marketing industry which was looking for new dynamic models describing the impact of promotion events and television advertisements on consumption. The result was an investigation of so-called self-exciting processes.

Looking back, my interest in mathematics really began at school. I liked its hierarchical structure and the fact that learning by heart was kept to a minimum. Later, at university, learning so much about Lindelöf spaces and tight measures did not dampen my interest, and we students shared our teacher's enthusiasm for Donsker and Billingsley. Probability theory was also well structured, and concepts like martingales became very familiar to us. When I came to statistics, it was just the opposite, a big building with many dark chambers, so that I was looking for someone to guide me. What I realized only later was that this friend was already there when I worked on my diploma thesis: the empirical distribution. Of course, as time went by, one had to be prepared to adapt and become quite flexible and to study new techniques when approaching a new problem.

Of all the areas of mathematics I have worked in I found statistics the most demanding. Maybe this is so because real life is so colourful and data can often contain hidden features, which can be detected if one is prepared to take the time to look. Therefore, when young students ask me to describe, in a non-technical way,

the nature of statistics, I usually point out that statistics is the *art* of how to properly weight the available information in the data.

I'd like to thank all who made this volume feasible. It is good to know that when walking down the road one was not alone, but could share ideas and time with good friends.

March 2017

Winfried Stute
Mathematical Institute
University of Giessen, Giessen, Germany

Contents

Part I
Survival Analysis

An Odyssey to Incomplete Data: Winfried Stute's Contribution to Survival Analysis

1

Jane-Ling Wang

1.1 Introduction

Winfried Stute is one of the pioneers and key contributors to empirical process theory. His interest in empirical processes dates back to his students days with a Diploma thesis on this topic under the guidance of Peter Gaenessler, which was later published in Z. Wahrscheinlichkeitstheorie und verw. Gebiete in 1976, a premier journal in probability. This was a feat for a diploma (comparable to M.Sc.) student. Winfried continued to work on problems in empirical processes for the next ten years and gained international acclaim for this work. His 1982 paper (Stute 1982) on the oscillation behavior of empirical processes remains a classic and became a foundation for research in density estimation, nonparametric regression, and beyond. His odyssey into the terrain of survival analysis was not accidental and was a consequence of his interest in expanding the horizon of empirical processes from the i.i.d. setting to incomplete data. Here we define survival analysis in the narrow sense that it involves incomplete data, such as randomly right censored or truncated data, or doubly censored data etc. With this narrow interpretation, Winfried's first publications in survival analysis appeared in Diehl and Stute (1988) and Dikta et al. (1989). They involved density and hazard function estimation as well as sequential confidence bands for distribution functions, all for randomly right censored data. Applying empirical process theory to censored data is a natural path that many theoreticians have partaken in, thereafter his interest in survival analysis intensified. From 1992 to 1997, he had more than 20 papers in survival analysis and continued to plow the field till his retirement in 2012 and beyond. In fact, his most recent papers including Sen and Stute (2014) and Azarang et al. (2013) are on this topic. To date he has produced

J.-L. Wang (✉)
University of California, Davis, CA 95616, USA
e-mail: janelwang@ucdavis.edu

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_1

nearly 40 papers in survival analysis, which accounts for roughly one-third of his publications. It is fitting and a pleasure for me to comment on his contributions in survival analysis, especially as I have worked with him on several projects in this area. Instead of an exhaustive review of his work in survival analysis, I will focus on the projects that I co-authored or am most familiar with, complemented by a few anecdotes.

We began to collaborate in the Fall of 1990 when I spent a sabbatical leave at the University of Giessen, where Winfried was a faculty member in the Department of Mathematics until his retirement. We were looking for a topic of common interest and survival analysis was the obvious choice, as he had just entered the field and I was in the midst of several projects on incomplete data. It took little time for us to settle on a topic, the strong law of large number (SLLN) for censored data, which seemed of great interest given that the SLLN is one of the most fundamental theoretical concepts in statistics. There were many results on the strong consistency of the Kaplan-Meier (hereafter abbreviated as K-M) estimator at that time but little was known for the general setting that involves the K-M integral defined as $\int \phi(x)\, d\hat{F}_n(x)$, where $\hat{F}_n$ is the K-M estimator (defined in (1.5) of section "Strong Law of Large Numbers: Random Right Censoring") of the true lifetime distribution function F and ϕ is a Borel measurable function on the real line such that $\int |\phi(x)|\, dF(x) < \infty$. The open problem we addressed was under what conditions $\int \phi(x)\, d\hat{F}_n(x)$ would converge to $\int \phi(x)\, dF(x)$ with probability one; the answer was provided in Stute and Wang (1993b).

This was a most memorable experience for me, especially as we were able to solve the problem with the minimum requirement that ϕ is F-integrable, which is the same condition that is needed for the classical SLLN to hold for i.i.d. data. The proof is fairly elaborate and involves several cases where both the lifetime and censoring distributions could be continuous, discrete, or neither, as long as they do not have common jumps (see Sect. 1.2 for details). This collaboration led to three subsequent joint papers (Gürler et al. 1993; Stute and Wang 1993a, 1994) and a series of papers by Winfried and his other collaborators (Stute 1993a, 1994a, b, c, d, 1995a) within the next two years.

Shortly after solving the SLLN for censored survival data, Winfried tackled the next most fundamental result, the central limit theorem (CLT) for the K-M integral (Stute 1995b), to be discussed further in Sect. 1.3. Besides randomly right censored data, Winfried also made landmark contributions to truncated data, another type of incomplete data that are challenging for two reasons: the sample is biased and there are technical difficulties at both the left and right tails of the lifetime distribution F, in contrast to the right censoring case where the left tail pose no difficulties. For truncated data the counterpart of the K-M estimator is the Lynden-Bell estimator (Lynden-Bell 1971) $\hat{F}_n^T$ defined in (1.27), which will be further discussed in Sect. 1.2. In Stute (1993a) Winfried established an i.i.d. representation for the Lynden-Bell estimator, which facilitated further asymptotic analysis for truncated data. This work sparked further research interest for truncated data, however a fundamental result regarding a CLT for the Lynden-Bell integral $\int \phi(x)\, d\hat{F}_n^T$ had proved elusive for a long time and remained an open problem. By 2000 we both drifted away from

survival analysis but were acutely aware of the need to fill this major theoretical gap. Finally, in the fall of 2005 (or around that period) I returned to Giessen to work on this project with Winfried and the results were published in Stute and Wang (2008). This was my last paper with Winfried although we both maintained interest in survival analysis and continued to dash into the field occasionally.

In the remainder of this paper, we discuss Winfried's four papers with focus on the SLLN and CLT for survival data.

1.2 Strong Law of Large Numbers: Random Right Censoring

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables from a distribution function F, and let F_n be their empirical distribution function. The classical SLLN implies that, with probability one, $\int \phi(x)\, dF_n(x) = \frac{1}{n}\sum_{i=1}^n \phi(X_i) \to E(\phi(X_1)) = \int \phi(x)\, dF(x)$, as long as $E(|\phi(X_1)|) < \infty$. Here the empirical distribution is a discrete probability measure that assigns equal point mass 1/n to each observation X_i, hence the classical SLLN and CLT hold automatically for $\int \phi(x)\, dF_n(x)$. When F is an event-time or lifetime distribution, a longitudinal follow-up study is needed to track the event-time X_i and as in many studies patients/subjects may be lost during the follow-up period or the study has to end before the event, which could be death. Therefore the event time cannot be observed for all patients. This triggers right censoring for which Winfried has made major contributions.

In the setting of random right censoring, X_i are no longer observed directly as they are subject to potential censoring by an independent variable Y_i. Instead, one can only observe $Z_i = \min(X_i, Y_i)$ along with the censoring indicator, $\delta_i = 1_{\{X_i \le Y_i\}}$. Unless otherwise mentioned, we make the standard assumption that $Y_1, \ldots, Y_n$ is an independent sequence of i.i.d. censoring variables with distribution function G that is independent of the sequence X_i. The counterpart of the empirical distribution in the presence of right censoring is the Kaplan-Meier estimator $\hat{F}_n$, which has been defined in several different but equivalent ways. We will use the form that has the most intuitive appeal for the purpose we want to serve.

One of the most intuitive ways to understand the principle of estimation for incomplete or biasedly sampled data is to first identify what parametric or nonparametric components could be estimated empirically from the observed data and then relate these components to the main target. In the random right censoring setting the main target is the lifetime distribution F, or equivalently its cumulative hazard function, which is defined as

$$\Lambda(x) = \int_0^x \frac{dF(t)}{1 - F(t-)}, \tag{1.1}$$

where the notation $F(t-)$, for any distribution F, stands for $F(t-) = \lim_{y \uparrow t} F(y)$, the limit of $F(y)$ as y approaches t from below.

It is obvious that Z_i can always be observed, so its empirical distribution function, $H_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{Z_i \le x}$, is the natural estimate for $H(x) = \Pr(Z_1 \le x)$. Likewise,

$H_1(x) = Pr(Z_1 \le x, \delta = 1)$, a subdistribution of the distribution of the Z_i, can be estimated empirically by $H_{1n}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \le x,\ \delta_i = 1\}}$. It is not difficult to show that

$$\Lambda(x) = \int_0^x \frac{dH_1(t)}{1 - H(t-)}. \tag{1.2}$$

So Λ can be estimated by replacing H and H_1 in (1.2) with their respective empirical estimates, H_n and H_{1n}.

To include the case with tied observations, let $Z_{(1)} < Z_{(2)} < \ldots < Z_{(K)}$ denote the K distinct and ordered observed lifetimes among $\{Z_1, \ldots, Z_n\}$, i.e. $Z_{(i)} = Z_j$ for some j for which $\delta_j = 1$. Then the resulting estimate of Λ is

$$\hat{\Lambda}_n(x) = \int_0^x \frac{dH_{1n}(t)}{1 - H_n(t-)} = \sum_{i=1}^{K} \left[\frac{d_i}{n_i}\right]^{1_{\{Z(i) \le x\}}}, \tag{1.3}$$

where $d_i = \sum_{j=1}^{n} 1_{\{Z_j = Z_{(i)}, \delta_j = 1\}}$ is the number of deaths observed at time $Z_{(i)}$ and $n_i = n[1 - H_n(Z_{(i)}-)] = \sum_{j=1}^{n} 1_{\{Z_j \ge Z_{(i)}\}}$ is the number of subjects still at risk at $Z_{(i)}$.

The estimator $\hat{\Lambda}_n(x)$ has an intuitive interpretation as the cumulative risk up to time x and is referred to in the literature as the Nelson-Aalen estimator. It is also the cumulative hazard function of the Kaplan-Meier estimate if one adopts a general result that provides a one-to-one correspondence between a cumulative distribution function (F) and its cumulative hazard function (Λ) for any random variable, be it continuous, discrete, or neither. Specifically, for any F let $F\{x\} = F(x) - F(x-)$ denote the point mass at x and similarly for Λ, then $F\{x\} = \Lambda\{x\} = 0$, except for $x \in A_F$, where A_F is the set of atoms of F, i.e. A_F is the set of all x at which $F(x)$ is discontinuous. Decomposing Λ into $\Lambda = \Lambda_c + \Lambda_d$, where Λ_c is a continuous function and Λ_d is a step function with jumps at A_F and jump sizes $\Lambda\{x\}$, the following relation holds for any distribution function F:

$$1 - F(x) = e^{-\Lambda_c(x)} \prod_{a_j \in A_F, a_j \le x} [1 - \Lambda\{a_j\}]. \tag{1.4}$$

Since the Nelson-Aalan estimator in (1.3) is a step function with atoms in $A_H = \{Z_{(1)}, \ldots, Z_{(K)}\}$, following (1.4) its corresponding survival function is:

$$1 - \hat{F}_n(x) = \prod_{i=1}^{K} \left[1 - \frac{d_i}{n_i}\right]^{1_{\{Z(i) \le x\}}}, \tag{1.5}$$

which is the Kaplan-Meier estimator.

It follows from (1.5) that $\hat{F}_n$ is a discrete distribution function with atoms at $Z_{(i)}$ and jump sizes

$$w_i = \frac{d_i}{n_i} \prod_{j=1}^{i-1} \left[1 - \frac{d_j}{n_j}\right]. \tag{1.6}$$

With this interpretation it is easy to see that the Kaplan-Meier estimate $\hat{F}_n$ collapses to the empirical distribution F_n in the absence of censoring, i.e. when all $\delta_i = 1$, and the SLLN implies $\int \phi(x)\, dF_n(x) = \frac{1}{n}\sum_{i=1}^n \phi(X_i) \to E(\phi(X_1)) = \int \phi(x)\, dF(x)$ as long as $\int |\phi(x)|\, dF(x) < \infty$. When censoring is present, the Kaplan-Meier integral, defined as

$$S_n = \int \phi(x)\, d\hat{F}_n(x) = \sum_{i=1}^{K} w_i\, \phi(Z_{(i)}), \tag{1.7}$$

now has random weights w_i (1.6) at $Z_{(i)}$, and this poses technical challenges for the SLLN in Theorem 1. Moreover, S_n cannot converge to $\int \phi(x) dF(x)$, if $\tau_H = inf\{x : H(x) = 1\} < \tau_F = inf\{x : F(x) = 1\}$.

The correct limit for S_n turns out to be

$$\begin{aligned} S &= \int_{x<\tau_H} \phi(x)\, dF(x) + 1_{\{\tau_H \in A_H\}}\, \phi(\tau_H)\, F\{\tau_H\} \\ &= \int_{x\le\tau_H} \phi(x)\, d\tilde{F}(x), \end{aligned} \tag{1.8}$$

where A_H is the set of atoms of H and

$$\begin{aligned} \tilde{F}(x) &= F(x), && \text{if } x < \tau_H, \\ &= F(\tau_H{}^-) + 1_{\{\tau_H \in A_H\}}\, F\{\tau_H\}, && \text{if } x \ge \tau_H. \end{aligned} \tag{1.9}$$

We now state the SLLN for censored data in Stute and Wang (1993b), hereafter abbreviated as SW93, which had more than 300 citations according to Google Scholar in August, 2017.

Theorem 1 (Stute and Wang 1993b)
Assume that $\int |\phi|\, dF < \infty$, *then*

$$S_n = \int \phi(x)\, d\hat{F}_n(x) \to S = \int \phi(x)\, d\tilde{F}(x),$$

with probability one and in the mean.

Remark 1 Obviously, the limit in the r.h.s. is $\int_{\{x\le\tau_H\}} \phi(x)\, dF(x)$ unless F is discontinuous at τ_H and $G(\tau_H-) = 1$. Otherwise, the limit is $\int_{\{x<\tau_H\}} \phi(x)\, dF(x)$. Also, the limit would be $\int \phi(x)\, dF(x)$ if $\tau_H = \tau_F$ and F is either continuous at τ_H or F has a jump at τ_H but $G(\tau_H-) < 1$.

Remark 2 The original SLLN in SW93 involved an extra condition that F and G have no jumps in common but this condition can be removed using a new time scale. This extension was briefly mentioned in a review paper (Stute 1995b), which

is a highly recommend reading for anyone interested in studying Winfried's striking results on K-M integrals. With this extension, the only assupmtion for the SLLN to hold under the random censoring scheme is exactly the same as its empirical counterpart with no censoring. That is, censoring does not cost any theoretical compromise for the SLLN but this is not the case for the central limit theorem which will be explored in Sect. 1.3.

Remark 3 Applications of Theorem 1 are plentiful. For example the choice $\phi(x) = 1_{(-\infty,t]}(x)$ leads to the strong consistency of the K-M estimator, and the choice $\phi(x) = x^k$ leads to the convergence of the K-M moment estimators, among others results. Needless to say, it is useful to establish the SLLN for U-statistics (Stute and Wang 1993a) and M-estimators (Wang 1995). The results in Stute (1976) can be further used to derive the strong uniform consistency of the K-M estimator. We refer the readers to the corollaries on Page 1595 there and the additional discussions in SW93.

Last but not least, we mention three additional applications of the SLLN that are provided in Stute (1993b, 1994a) and Stute and Wang (1994). An extension to the multivariate case was studied in Stute (1993b) in the presence of a p-dimensional covariate when these covariates are not subject to censoring. The SLLN for the multivariate joint distribution of the censored response and its covariates is presented there with a very neat application to the linear censored regression model and a proposal for a new and simple estimator for the slope regression parameter. This estimator was shown to perform favorably against its competitors, such as the Buckley-James estimator (Buckley and James 1979). In Stute (1994a) an explicit expression for the bias of a Kaplan-Meier integral was established, while Stute and Wang (1994) provides an explicit formula for the jackknife estimate of a Kaplan-Meier integral.

1.2.1 Key Ideas of the SLLN

The most studied case in the literature is for $\phi(x) = 1_{(-\infty,t]}$, which amounts to showing $\hat{F}_n(t) \to F(t)$ almost surely (a.s.). Because of the nice empirical expression (1.3), most approaches in the literature by 1990 took a two step approach by first showing that $\hat{\Lambda}_n(t) \to \Lambda(t)$ a.s. and then showing that $\log(1 - \hat{F}_n(t)) + \hat{\Lambda}_n(t) \to 0$ a.s. This completes the proof for continuous F, since $\log(1 - F(t)) = -\Lambda(t)$. One drawback with such a two-step approach is that the convergence can only be established for t such that $F(t) < 1$ or along a sequence of t_n such that $F(t_n) \to 1$ slowly, since $\Lambda(t)$ needs to stay away from ∞ at a certain rate. It turns out that this problem on the right tail can be avoided if one bypasses the cumulative hazard function and works instead with the distribution function directly. This is the approach taken in SW93.

Since the idea is to stick to the target (1.7), the goal is to explore what kind of structure it possesses. There are three classical techniques to prove SLLNs for the case where there is no censoring, (i) Kolmogorov's original proof, (ii) the ergodic

theorem for strictly stationary and ergodic sequences, and (iii) the reverse-time martingale approach (Neveu 1975) for a proper sequence of $\sigma-$fields so the martingale convergence theorem can be applied. When we first looked into this problem in 1990, we knew immediately that the first two approaches could not be extended to censored data easily and the reverse-time martingale structure does not hold for S_n since $E(S_n)$ varies with n. But Winfried had a hunch that it might still work if we can show that S_n endowed with a proper sequence of decreasing $\sigma-$field is a reverse-time supermartingale for positive ϕ functions (we knew that S_n could not be a reverse-time submartingale because the K-M estimator is biased downward). Although it was not hard to construct the proper $\sigma-$fields (cf. Step 1 of Sect. 1.2.2) needed for this martingale structure it was not easy to pin down the supermartingale structure.

This went on for some time and we still had no clue about the martingale structure of S_n. After another uneventful day (Wish I remembered the date !) I left the institute frustrated. After dinner and some soothing German dinner that night, my spirit was lifted and I decided to give a final shot to see if we should continue to invest our time on this problem. My plan was simple, just check the simple cases of $n = 1, 2$ and 3, and the truth would be revealed. I settled down to do the calculation: The case of $n = 1$ was trivial, and YES, S_n is a reverse-time supermartingale when $n = 2$. When it was revealed that S_n is also a reverse-time supermartingale for $n = 3$, I thought that I had hit the jackpot—it had to be (ok, just might be) true for general n. I learned this naive strategy to do research on difficult problems from the late Jack Kiefer who taught me that if something is true for $n = 1$ to 3, it is probably true for all n!

A side note about my two mentors in research, Professors Jack Kiefer and Lucien LeCam, two geniuses who approached open problems from opposite ends. Both were extremely kind and generous. Kiefer was my thesis advisor until his sudden death at the age of 57, less than a year before my graduation, and Le Cam graciously took me under his wings after Kiefer's death and remained a mentor and friend until his death in 2000. I am forever indebted to their inspiration and guidance. As mentioned, Kiefer taught me how to approach a problem from the simplest scenario, e.g., try the one-dimensional case first, then general Euclidean space, before launching to the infinite-dimensional abstract space. LeCam favored the opposite, top-down, approach, as he could see things high up in the abstract space that few others could, so he typically approached a problem in its most general and abstract setting. I was extremely fortunate to witness their differences and took advantage of both approaches. Often, I would start to work on a problem at the ground level and work my way up as Kiefer had taught me to do, but once reaching the higher ground, I would look for powerful tools that could simplify the proofs or expand the results with less stringent assumptions. These two reversed approaches are effective in their own individual way but together they form a powerful team.

Going back to the story of the paper SW93 on the SLLN, the next morning I arrived early at the institute to eagerly share the discovery from the night before. I ran straight to Winfried's office to announce the big news—the supermartingale structure must be true for S_n because it holds for $n = 1, 2$ and 3. He did not laugh at the obviously flawed and naive statement and instead immediately realized that

we had work to do. We went downstairs to a classroom which had huge blackboards and began to explore our options one by one. After we understood what was going on and believe by then that the statement must be true we still could not come up with a one-shot proof that S_n is a reverse-time supermartingale. So we took the last resort—to prove it by mathematical induction! I don't know about Winfried but I never would have thought that one day I would use induction to prove the theorem of my life. It is perhaps not the most elegant way to prove the key result, Lemma 2.2 in SW93, and I still wonder whether there is a more direct way to prove this lemma. But even with the induction the proof of Lemma 2.2 is non-trivial and involves elegant use of order statistics, concomitants and ranks.

We made significant progress over the next few days but there were still multiple hurdles ahead. It must have taken more than a week before we had a proof for the special case when H is a continuous distribution function. I was exhilarated and ready to call it quits when we had the proof for continuous H as that was already far better than any existing result. But Winfried was not content, he was keen to get rid of the continuity assumption. So upon his insistence we eliminated this assumption and showed that the SLLN holds as long as F and G do not have common jumps. This was the version published in our 1993 paper but Winfried learned of a trick later to get rid of this assumption and included that extension in Stute (1995b). In the end he fulfilled his dream to show that the SLLN for K-M integral holds under no assumptions other than the trivial one that $\int |\phi| \, dF < \infty$. Overall, this project involved the most unusual route towards a proof that I have ever encountered in my career. Still until today, Theorem 1.1 in SW93 remains my favorite theorem of all time.

1.2.2 Outline of the Proof

Step 1. The first step is to identify the $\sigma-$fields for the reverse time martingale. Towards this goal, it is easier to consider an alternative form of the K-M estimator, which aims at breaking tied observations so that any lifetime precedes a tied censoring time but the ordering within tied lifetimes or tied censoring times can be arbitrary. With this rule, the K-M estimator in (1.5) is equivalent to the one in formula (1.2) of SW93 and the associated K-M integral (1.7) can be expressed as

$$S_n = \sum_{i=1}^{n} W_{in} \, \phi(Z_{i:n}), \text{ where } Z_{i:n} \text{ is the } i\text{th ordered-statistics among } \{Z_1, \ldots, Z_n\}, \quad (1.10)$$

$$W_{in} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1} \right]^{\delta_{[i:n]}} \text{ with } \delta_{[i:n]} = \delta_j, \text{ if } Z_j = Z_{i:n}.$$

The $\delta_{[i:n]}$ above are often called the concomitants of the $Z_{i:n}$.

With these notations define $\mathcal{F}_n$ to be the σ field generated by $\{Z_{i:n}, \delta_{[i:n]}, 1 \le i \le n, Z_{n+1}, \delta_{n+1} \ldots\}$. Then S_n is adapted to $\mathcal{F}_n$ with $\mathcal{F}_n \downarrow \mathcal{F}_\infty = \cap_{n \ge 1} \mathcal{F}_n$ and $\mathcal{F}_\infty$ is

trivial by the Hewitt-Savage zero-one law.

Step 2. Next, we show that, for every $\phi \geq 0$ and continuous H, $E(S_n | \mathcal{F}_{n+1}) \leq S_{n+1}$. Hence $\{S_n, \mathcal{F}_n\}_{n \geq 1}$ is a reverse-time supermartingale. The proof uses mathematical induction and is included in Lemma 2.2 of SW93. This is the key step towards the final SLLN.

Step 3. Proposition 5-3-11 in Neveu (1975) then implies that, for every $\phi \geq 0$, S_n converges a.s. and in the mean to some random variable S_∞, which must be a constant S by the Hewitt-Savage zero-one law. Hence $S_n \to S$ a.s. and $E|S_n - S| \to 0$.

This result can be extended to general $\phi = \phi^+ + \phi^-$ by decomposing into positive (ϕ^+) and negative (ϕ^-) parts.

Step 4. It now remains to identify the constant S and this was achieved in Lemma 2.7 of SW93 for continuous H, which implies that

$$S = \int \phi(x)\, m(x)\, \gamma_0(x)\, dH(x), \tag{1.11}$$

where $m(x) = P(\delta = 1 | Z = x)$, and $\gamma_0(x) = \int_0^{x-} \frac{1-m(y)}{1-H(y)}\, dH(y)$.

Under independence of lifetime T and censoring variable C, the limit S then takes the form in (1.8).

Step 5. To show the result for a general H, we first look at the case where F and G have no common jumps, hence there are no tied observations between the censored and uncensored observations. Under this assumption, apply a quantile transformation, $H^{-1}(U_i)$, to a specially constructed sequence U_i of uniform [0, 1] random variables as in Lemma 2.8 in SW93, so that $Z_i = H^{-1}(U_i)$. Then

$$S_n = \sum_{i=1}^{n} W_{in}\, \phi(H^{-1}(U_{i:n})). \tag{1.12}$$

The SLLN now follows by replacing ϕ with $\phi \circ H$. This is what was obtained in SW93, where the only assumption needed for the SLLN of K-M integrals is that F and G have no common jump points. It turns out that this restriction can be removed because the K-M estimator treats an uncensored observation as if it precedes a censored one slightly if there is a tie between them. A trick in Gill (1980) to shift the time scale of G slightly to the right of those common jump points then implies that F and the transformed G on this new time scale no longer have common jumps and hence the SLLN holds. This trick was discussed in detail on page 437 of Stute (1995a) where he dealt with the CLT for the K-M integral. In conclusion, the SLLN for a K-M integral holds under the minimal assumption $\int |\phi(x)|\, dF(x) < \infty$, with no restriction on F and G, just like the classical SLLN when there is no censoring.

1.3 Central Limit Theorem: Random Right Censoring

Once the limit S in (1.8) of $S_n = \int \phi(x)\, d\hat{F}_n(x)$ has been identified, this facilitates to explore the limiting distribution of $S_n - S$, i.e. the CLT for a K-M integral. The special case of $\phi(x) = 1_{(\infty,x]}$ for $x < \tau_H$, was studied, for instance, by Breslow and Crowley (1974); Lo and Singh (1986); Major and Rejto (1988). The unrestricted case for all x was established in Gill (1983) and Ying (1989) by using the martingale convergence theorem, a powerful tool to handle asymptotic theory for censored data that was popular in the 1980s. However, some technical assumptions were still needed to control the censoring effect in the right tail of the lifetime distributions. Under these assumptions, the case of a ϕ-function that is of bounded variation on an interval $[0, T]$ for which $T < \tau_H$ can be handled without much difficulties by invoking integration by parts. But this is a restrictive class of functions and specifically, it excludes the estimation of the K-M mean which corresponds to $\phi(x) = x$. Susarla and Van Ryzin (1980) were able to extend the K-M mean estimate to an interval $[0, M_n]$ for which $M_n \to \infty$ at a suitable rate but the results on $[0, \infty)$ remains unresolved.

Subsequently, Schick et al. (1988) established the CLT for ϕ-functions that are nonnegative, nonincreasing and continuous. Under some regularity conditions on F, Yang (1994) extended the results to general functions ϕ that satisfy

$$\int \frac{\phi^2}{1-G}\, dF < \infty. \tag{1.13}$$

Other than the restrictions on F, the result of Yang (1994) is optimal as assumption (1.13) is needed to ensure that the limiting variance is finite. Other restrictions in Yang (1994) were removed by Winfried in his1995 paper (Stute 1995a), where he established the CLT for K-M integral for any F and G under minimal conditions. This paper had 173 citations based on Google scholar in August 2017.

How did Winfried do it? There are several ways to derive the CLT and those familiar with Winfried's technical style probably know his affinity to derive everything from scratch using basic tools. Thus, instead of employing the martingale CLT as was done in Gill (1983), he took the classical approach of expanding $\sqrt{n}\ (S_n - S)$ as a sum of i.i.d. random variables plus a small and negligible remainder term. While this i.i.d. representation approach has been explored by many before him, the key to success rests upon the conditions that are invoked to handle the remainder term. Through a clever expression of S_n as a U-statistic of degree three plus a negligible remainder term, he realized that the Hajek projection for U-statistics would provide the right platform for the desired i.i.d. decomposition. What remains is hard analysis and the tenacity to get things right.

In professional life, Winfried is a minimalist. Any extra condition for the sake of convenience would be an eyesore for him. In my experience working with him, I have witnessed repeatedly his persistence to get rid of anything that is not elegant.

This working style has served him well and attributed to his ability to produce the most elegant results time and again.

To state the CLT, we first define several quantities:
$\tilde{H}^0(z) = P(Z \le z, \delta = 0) = \int_{-\infty}^{z} (1 - F(y))\, dG(y)$,

$\tilde{H}^1(z) = P(Z \le z, \delta = 1) = \int_{-\infty}^{z} (1 - G(y-))\, dF(y)$,

$\gamma_0(x) = \exp\{\int_{-\infty}^{x-} \frac{d\tilde{H}^0(y)}{1-H(y)}\}$,

$\gamma_1(x) = \frac{1}{1-H(x)} \int 1_{\{x<w\}}\, \phi(w)\, \gamma_0(w)\, d\tilde{H}^1(w)$,

and

$\gamma_2(x) = \int\int \frac{1_{v<x, v<w}\, \phi(w)\, \gamma_0(w)}{[1-H(v)]^2}\, d\tilde{H}^0(v)\, d\tilde{H}^1(w)$.

The following two assumptions are needed for the CLT in Theorem 2:

$$\int \phi^2(x)\, \gamma_0^2(x)\, d\tilde{H}^1(x) < \infty \tag{1.14}$$

and

$$\int |\phi(x)|\, C^{1/2}(x)\, d\tilde{F}(x) < \infty, \tag{1.15}$$

where $C(x) = \int_{-\infty}^{x-} \frac{dG(y)}{[1-H(y)]\,[1-G(y)]}$ and $\tilde{F}$ is defined in (1.9).

Theorem 2 (Corollary 1.2 of Stute, 1995a) *Under assumptions (1.14) and (1.15),* $\sqrt{n}(S_n - S) = \sqrt{n} \int \phi(x)\, d(\hat{F}_n - F)(x) \to N(0, \sigma^2)$ *in distribution, where* $\sigma^2 =$ *Var* $[\phi(Z)\, \gamma_0(Z)\, \delta + \gamma_1(Z)\, (1 - \delta) - \gamma_2(Z)]$.

Remark 4 For continuous F the asymptotic variance becomes

$$\sigma^2 = \int_{-\infty}^{\tau_H} \frac{\phi^2(x)}{1 - G(x)}\, dF(x) - \left[\int_{-\infty}^{\tau_H} \phi(x)\, dF(x)\right]^2 - \int \left[\int_{x}^{\tau_H} \phi(y)\, dF(y)\right]^2 \frac{1 - F(x)}{[1 - H(x)]^2}\, dG(x), \tag{1.16}$$

which further simplifies to $\sigma^2 = \int \phi^2\, dF - [\int \phi\, dF]^2$ when there is no censoring as G then always equals zero.

Remark 5 Condition (1.14) is equivalent to condition (1.13) when F is a continuous function. Both are properly modified "second moment" conditions in the CLT for censored data. Condition (1.15), on the other hand, is used to control the bias of the K-M integral so the $\sqrt{n}$ rate can be achieved. This is the price paid by the K-M estimator and is needed for the CLT of general K-M integrals. Examples provided in Stute (1995a) imply that Theorem 2 may not hold if condition (1.15) is not satisfied.

Remark 6 As with the SLLN, applications of the CLT are plentiful. Remark 3 above listed a few such applications. In particular, the CLT was extended to the case when covariates are present in Stute (1996a). Another application is provided in Stute (1996b), where Winfried established an explicit expression for the variance of the jackknife estimator of the K-M integral and investigated the convergence of this variance estimator. Surprisingly the variance of the jackknife estimator converges to the variance of the K-M integral only when $\phi(x) \to 0$ as $x \to \tau_H$. As this is quite restrictive and in view of Winfried's low tolerance for wrinkles, he proposed a modified variance estimate, $\widehat{\text{var}}^*_{JK}$, that satisfies $n\,\widehat{\text{var}}^*_{JK} \to \sigma^2$.

1.3.1 Outline of the Proof

The proof of the CLT is based on an i.i.d. representation of the K-M integral (cf. Theorem 1.1 of Stute (1995a)), which leads to

$$\int \phi\, d(\hat{F}_n - \tilde{F}) = \frac{1}{n}\sum_{i=1}^{n} U_i + R_n, \tag{1.17}$$

where the U_i are i.i.d. with mean zero and variance σ^2 and $R_n = o_P(n^{-1/2})$.

The derivation of (1.17) follows the following steps.

Step 1. First assume that H is continuous. Then (1.10) and Lemma 2.1 in Stute (1995a) imply that $\int \phi\, d\hat{F}_n$ can be expressed as

$$\sum_{i=1}^{n} W_{in}\,\phi(Z_{i:n}) = \int \phi(x) \exp\left\{ n \int_{-\infty}^{x-} \ln\left[1 + \frac{1}{n(1 - H_n(y))}\right] d\tilde{H}^0_n(y) \right\} d\tilde{H}^1_n(x). \tag{1.18}$$

Step 2. Replace the logarithm term $\ln(1 + x)$ by x and neglecting the error terms, then the exponential term in (1.18) becomes $\exp\left\{ \int_{-\infty}^{x-} \frac{d\tilde{H}^0(y)}{1 - H_n(y)} \right\}$. Integrating this term w.r.t. $\tilde{H}^1_n$ and further expanding this exponential term leads to a U-statistic of order 3.

Step 3. The Hájek projection of this U-statistic leads to the desired i.i.d. expansion in (1.17). Details are provided in Lemma 2.2–2.7 of Stute (1995a) under the additional assumption that

$$\phi(x) = 0 \text{ for all } x > T \text{ and some } T < \tau_H, \tag{1.19}$$

so all terms appearing in the denominators of the proof are bounded away from zero, hence the denominator will cause no problem.

Step 4. Under the two assumptions (1.14) and (1.15), and for continuous F, the denominators in the proof can be controlled without assumption (1.19). The proof is thus extended to the case without assumption (1.19) but with the assumption that H is continuous.

Step 5. Finally, the assumption of continuous H can be removed just as in the case for the SLLN, as discussed in Step 5 of Sect. 1.2.2.

1.4 Random Truncation

While I was visiting Giessen in 1990, Winfried and I worked on another type of incomplete data, random truncated data (Gürler et al. 1993), which often occur in astronomy (Woodroofe 1985) or in studies with delayed entry of patients into a study. Let (X_i, Y_i), $i = 1, \ldots, N$, be a sequence of i.i.d. random vectors for which $X_i \sim F$ is also independent of $Y_i \sim G$. Random truncation occurs when the pair (X_i, Y_i) can be observed only when $X_i \geq Y_i$. That is, neither X_i nor Y_i can be observed when $X_i < Y_i$ but both are observed when $X_i \geq Y_i$. This sampling structure is quite different from the one for censored data where one, and only one (the minimum), of the lifetime and censoring variable can be observed. Consequently, the observed sample size, $n = \sum_{i=1}^{N} 1_{\{X_i \geq Y_i\}}$, is a random quantity while the latent sample size N is unknown. We denote the observed data by (X^*, Y^*) to distinguish them from the original (X, Y). Luckily (X_i^*, Y_i^*) are still i.i.d. with joint distribution

$$H^*(x, y) = P(X \leq x, Y \leq y | Y \leq X) = \frac{1}{\alpha} \int_{-\infty}^{x} G(y \wedge z)\, dF(z); \tag{1.20}$$

and marginal distributions

$$F^*(x) = H^*(x, \infty) = \frac{1}{\alpha} \int_{-\infty}^{x} G(z)\, dF(z), \tag{1.21}$$

$$G^*(y) = H^*(\infty, y) = \frac{1}{\alpha} \int_{-\infty}^{\infty} G(y \wedge z)\, dF(z), \tag{1.22}$$

where $\alpha = P(Y \leq X)$ and $y \wedge z$ denotes the minimum of y and z.

Note here that F^*, G^* and H^* can all easily be estimated empirically but the goal is to estimate F and G. We say that left truncation occurs when the primary interest is F, in which case G is called the truncation distribution. Likewise, the data are right truncated when G is the primary interest and F is the right truncation distribution. For illustration purposes we focus on the left truncation case for which F is the primary target with cumulative hazard function Λ_F defined as in (1.1).

Let $a_F = \inf\{x : F(x) > 0\}$ be the left support point of F, and $b_F = \sup\{x : F(x) < 1\}$ be its right support point. It is not surprising that F can be estimated only when

$$\begin{aligned} &(i)\ a_G < a_F, \quad \text{or} \\ &(ii)\ a_G = a_F \text{ and } F\{a_F\} = 1. \end{aligned} \tag{1.23}$$

Case (i) is much easier to deal with than case (ii). Throughout this section we make the assumption (1.23) and write

$$\begin{aligned} C(z) &= P(Y^* \le z \le X^*) = G^*(z) - F^*(z-) \\ &= \frac{1}{\alpha}\, G(z)\,[1 - F(z-)], \text{ for } a_F \le z < \infty. \end{aligned} \tag{1.24}$$

It can be easily shown that $\Lambda_F(x) = \int_{-\infty}^{x} \frac{dF^*(y)}{C(z)}$. Hence the empirical estimate of $\Lambda_F(x)$ is

$$\hat{\Lambda}_n^T(x) = \int_{-\infty}^{x} \frac{dF_n^*(z)}{C_n(z)} = \sum_{\text{distinct } X_k^* \le x} \frac{F_n^*\{X_k^*\}}{C_n(X_k^*)}, \tag{1.25}$$

where F_n^* is the empirical estimates based on $\{X_1^*, \ldots, X_n^*\}$ and

$$C_n(z) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i^* \le z \le X_i^*\}} \tag{1.26}$$

is the empirical estimate of C. The superscript T in (1.25) reminds us that this is for truncated data.

Based on (1.4) the distribution function $\hat{F}_n^T$ that corresponds to $\hat{\Lambda}_n^T$ is

$$1 - \hat{F}_n^T(x) = \prod_{k:\ X_k^* \le x} \left[1 - \frac{F_n\{X_k^*\}}{C_n(X_i^*)}\right], \tag{1.27}$$

which, when there are no tied observations among X_i^*, becomes

$$1 - \hat{F}_n^T(x) = \prod_{i:\ X_i^* \le x} \left[1 - \frac{1}{n\, C_n(X_i^*)}\right]. \tag{1.28}$$

This is the original Lynden-Bell estimate (Lynden-Bell 1971) which was shown to be the nonparametric maximum likelihood estimator of F by Woodroofe (1985) and Wang et al. (1986). One undesirable feature of the estimator $\hat{F}_n^T(x)$ is that it may jumps to 1 before x reaches the largest order statistic. To see this, consider the simpler case when there is no tied observation so (1.28) holds. Under this scenario $\hat{F}_n^T(x) = 1$ as soon as $nC_n(X_j^*) = 1$ for some $X_j^* \le x$, and all observations larger

than X_j^* will have no influence on the estimation of F. This is a soft spot of the Lynden-Bell estimator which triggers technical difficulties as we will elaborate later. Luckily, the probably that this happens is small so at the end of the day the Lynden-Bell estimator still enjoys nice properties. However, in order to establish the CLT for Lynden-Bell integrals, a revised estimator, which is asymptotically equivalent to the Lynden-Bell estimate, was constructed in Stute and Wang (2008) to facilitate the proof.

To understand when the above undesirable feature might occur, observe that $n\ C_n(X_j^*) = 1$, if X_j^* is not covered by any other interval $[Y_i^*, X_i^*]$, $i \neq j$ (note that $n\ C_n(X_j^*) \geq 1$ because $[Y_j^*, X_j^*]$ always covers X_j^*). This phenomenon occurs when there are gaps in the unions of all intervals $[Y_i^*, X_i^*]$, $1 \leq i \leq n$, and these gaps, which are intervals that are not covered by any $[Y_i^*, X_i^*]$, are referred to as the "holes" for truncated data (Strzalkowska-Kominiak and Stute 2010) or "empty inner risk sets" (Keiding and Gill 1990). On those holes, $C_n(x)$ may be zero so the probability for those holes needs to be small and tend to zero sufficiently fast as the sample size tends to infinity. Sharp probability bounds were developed in Strzalkowska-Kominiak and Stute (2010) and they have ramifications on the estimation of α, a topic of practical interest further studied in He and Yang (1998a). Below we focus on two of the fundamental results that Winfried established (Stute 1993a; Stute and Wang 2008).

1.4.1 I.I.D. Representation for Truncated Data

So far, the left truncation setting resembles the random censoring case by replacing H_1 and H in (1.2), respectively, by F^* and C and by replacing the empirical estimates H_{1n} and H_n in (1.3), respectively, by the empirical estimates F_n^* and C_n. However, there are distinctive features between the two settings in the handling of theory because C_n is not a monotone function while H_n is and also because of the problems created by the "holes" in truncated data when $nC_n(X_j^*) = 1$, for some j.

Winfried's first solo act for truncated data was Stute (1993a), which appeared around the same time as SW93. However, instead of tackling the Lynden-Bell integrals, $\int \phi(x) d\hat{F}_n^T(x)$, he focused on the Lynden-Bell estimator itself and on providing an i.i.d. representation for the Lynden-Bell estimator. Along this path, he realized that he needed stronger results on the processes of U-statistics which he developed alongside with Stute (1993a) and which subsequently appeared in Stute (1994e). Below we summarize the main result of Stute (1993a), which improved the results in Chao and Lo (1988) and had 109 citations according to Google Scholar in August, 2017.

Theorem 3 (Theorems 1 and 2 of Stute Stute 1993a) *Assume* $a_G \le a_F$ *and* $\int_{a_F}^{\infty} G^{-2}(x)\, dF(x) < \infty$, *then uniformly in* $a_F \le x \le b < b_F$ *we have*

$$(i) \quad \hat{\Lambda}_n^T(x) - \Lambda_F(x) = L_n + R_n, \text{ and}$$

$$(ii) \quad \hat{F}_n^T(x) - F(x) = (1 - F(x))\, L_n(x) + R_n^0(x),$$

$$\text{where } L_n = \int_{a_F}^{x} \frac{1}{C(z)}\, d(F_n^* - F^*)(z) - \int_{a_F}^{x} \frac{C_n(z) - C(z)}{C^2(z)}\, dF^*(z),$$

$$\sup_{a_F \le x \le b} |R_n(x)| = o(n^{-1}(\ln n)^{\delta}), \quad \text{with probability one and for any } \delta > 1.5,$$

$$\sup_{a_F \le x \le b} |R_n^0(x)| = O(n^{-1}(\ln n)^3) \text{ with probability one.}$$

Remark 7 It is clear from the theorem that L_n is a sum of i.i.d. processes which then leads to the CLT and LIL (Law of Iterated logarithm) for the Lynden-Bell estimator.

Remark 8 The order of the remainder terms (other than the log part) in Theorem 3 is $O(n^{-1})$, which is much sharper than the order $o(n^{-1/2})$ in standard i.i.d. representations. Winfried stressed the need to have such a higher order remainder term, e.g. for density and quantile estimation.

Remark 9 A version of the SLLN for truncated data was later studied in He and Yang (1998b) but an optimal solution under random truncation remains elusive at this time. Maybe Winfried will fill this gap when he has more time in his hand (he is still carrying a full teaching load at Giessen).

1.4.2 CLT for Truncated Data

For right censored data, only the right tail poses technical challenges, but for left truncated data both the left and right tails present challenges. This can be seen from the function C and its estimator C_n, as both approach zero on the left and right tail and as in particular C_n appears in the denominator. Additional challenges are due to the aforementioned "holes" in the data and to the fact neither C nor C_n is monotone. Consequently, the proof of the CLT for right censored data in Stute (1995a) does not apply directly for truncated data. To circumvent the tail problem we want to prevent $\hat{F}_n^T$ to reach its full mass (one) prematurely, which means that we need to construct a modified estimator $\tilde{F}_n^T$ which avoids this problem but satisfies $\sqrt{n}\{\int \phi\, d\hat{F}_n^T - \int \phi\, d\tilde{F}_n^T\} = o_P(n^{-1/2})$. This is the key idea in Stute and Wang (2008), which will be discussed further after we present the CLT for Lynden-Bell estimator.

The following assumptions are needed for the CLT,

$$(i) \int \frac{dF}{G} < \infty, \tag{1.29}$$

$$(ii) \int \frac{\phi^2}{G}\, dF < \infty. \tag{1.30}$$

Theorem 4 (Theorem 1.1 and Corollary 1.1 of Stute and Wang, 2008)
Under assumptions (1.23), (1.29), and (1.30) we have

$$\int \phi d\hat{F}_n^T - \int \phi dF = \int \frac{\psi(y)}{C(y)} d(F_n^* - F^*)(y) - \int \frac{C_n(y) - C(y)}{C^2(y)} \psi(y)\, dF^*(y) + o_P(n^{-1/2}), \tag{1.31}$$

where

$$\psi(y) = \phi(y)\,[1 - F(y)] - \int_{[y<x]} \frac{\phi(x)\,[1 - F(x)]}{C(x)}\, dF^*(x) = \int_{[y<x]} [\phi(y) - \phi(x)]\, dF(x).$$

Hence

$$\sqrt{n} \int \phi\, d(F_n^* - F) \;\to N(0, \sigma^2) \text{ in distribution,} \tag{1.32}$$

$$\text{with} \quad \sigma^2 = Var\left\{ \frac{\psi(X)}{C(X)} - \int_Y^X \frac{\psi(y)}{C^2(y)}\, dF^*(y) \right\}.$$

Remark 10 Assumption (1.23) as mentioned before ensures that F can be properly estimated under the left truncation setting and assumption (1.29) further ensures that there is enough information in the left tail so F can be estimated at the $\sqrt{n}$ rate. Both assumptions are standard for truncated data and were already stated in Woodroofe (1985). Assumption (1.30) is needed to ensure that the leading terms in the i.i.d. representation (1.31) have finite second moments so the asymptotic normality in (1.32) holds. Thus, the assumptions listed in Theorem 4 are mild and they are much weaker than any other existing assumptions for truncated data.

The fact that $G \leq 1$ implies $\int \phi^2 dF < \infty$, which is the second moment assumption for standard CLTs when there is no truncation. It will be implied by assumption (1.29) when $\int \phi^2 dF < \infty$ and ϕ is locally bounded in a neighborhood of a_G. Both assumptions in (1.29) and (1.30) will be satisfied when $a_G < a_F$ and $\int \phi^2 dF < \infty$.

Remark 11 Theorem 4 actually has broader implications to a class of ϕ functions if one traces its proof carefully. For instance, if we take the class of all indicators $\phi_x = 1_{(-\infty,x]}$ then it can be shown that the i.i.d representation in (1.31) holds uniformly for all ϕ_x under condition (1.23) and (1.29) (here condition (1.30) is implied by condition (i)) because the remainder term can be bounded uniformly.

1.4.3 Outline of the Proof

Step 1. The proof begins with the case that F is continuous, where the Lynden-Bell estimator takes the special form in (1.28). As mentioned, $\hat{F}_n^T$ has the undesirable property that if "holes" exist then $\hat{F}_n^T$ jumps to one as soon as $nC_n(X_j^*) = 1$ for some X_j^*. When this happens before the largest order statistic the exponential representation in (1.18) for the K-M integral cannot hold for the Lynden-Bell integral, so the method of proof for Theorem 2 is not applicable. To circumvent this problem, a modified estimator was proposed in Stute and Wang (2008). This estimator was constructed by modifying the weights of the Lynden-Bell estimator from

$$\hat{F}_n^T\{X_{i:n}^*\} = \frac{1}{C_n(X_{i:n}^*)} \prod_{j=1}^{i-1} \left[1 - \frac{1}{nC_n(X_{i:n}^*)}\right]. \tag{1.33}$$

to

$$\tilde{F}_n^T\{X_{i:n}^*\} = \frac{1}{C_n(X_{i:n}^*)} \prod_{j=1}^{i-1} \left[1 - \frac{1}{nC_n(X_{i:n}^*) + 1}\right], \tag{1.34}$$

where $X_{i:n}^*$ is the ith order statistics of $\{X_1^*, \ldots, X_n^*\}$.

This small modification in the denominator of the products now avoids the problem of "holes" so all observed data X_i^* receive positive weights and hence are properly accounted for.

Step 2. A similar proof to the CLT for censored data can then be applied to $\int \phi \, d\tilde{F}_n^T$, albeit extra care is still needed as the truncation case is challenging both in the left and right tail, whereas the censored case only faces challenges in the right tail. For instance, a new bound for the function C/C_n is needed and established in Lemma 3.1 of Stute and Wang (2008). In addition, many more bounds need to be established for quantities that involve C_n in the denominator. Since C_n is not monotone, many of the nice properties that are readily available for its censoring counter part H_n are not afforded to C_n.

After eight lemmas and three corollaries, Theorem 4 was established for continuous F.

Step 3. The extension to an arbitrary F is less treacherous and follows a similar path as the analogous extension for censored data, as described in step 5 of Sect. 1.2.2, by invoking a quantile transformation.

1.5 Conclusion

It has been 40 years since Winfried's first publication and since he obtained his Ph.D. degree (both events occurred in the same year 1976). During these 40 years, he had a very productive career with many landmark papers. It appears from his CV that the four years from 1993 to 1996 were Winfried's most productive period, during which he had a total of 27 publications, 20 of which were either in survival analysis or inspired by his interest in survival analysis. In this review, we focused on four of his papers and some of their applications in survival analysis as examples for the scope and impact of his research. Hopefully, the review gives the reader a sense of the transformative nature of his contributions to the theory of survival analysis. It is also my hope that after a brief tranquil period after his retirement Winfried gets fired up again to crack another code, perhaps for doubly or interval censored data this time. There are still lots of interesting open problems for the theory of incomplete data—the world of incomplete data is not complete yet. I look forward to another opportunity to hack the incomplete filed together. Meanwhile, I wish him the very best for his 70th birthday—and many more years to look forward to!

Acknowledgements The author is grateful for the suggestions and thorough review of two referees and a co-editor.

References

Azarang, L., J. de Uña-Álvarez, and W. Stute (2013). The jackknife estimate of covariance under censorship when covariables are present. Discussion Papers in Statistics and Operations Research, Report 13/04, Universidad de Vigo.

Breslow, N. and J. Crowley (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist. 2*(3), 437–453.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*(3), 429–436.

Chao, M.-T. and S.-H. Lo (1988). Some representations of the nonparametric maximum likelihood estimators with truncated data. *Ann. Statist. 16*(2), 661–668.

Diehl, S. and W. Stute (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal. 25*(2), 299–310.

Dikta, G., B. Kurtz, and W. Stute (1989). Sequential fixed-width confidence bands for distribution functions under random censoring. *Metrika 36*(1), 167–176.

Gill, R. (1980). Censoring and stochastic integrals. *Stat. Neerl. 34*(2), 124–124. Censoring and stochastic integrals. Math. Centre Tract 124, Math. Centrum, Amsterdam.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist. 11*(1), 49–58.

Gürler, Ü., W. Stute, and J.-L. Wang (1993). Weak and strong quantile representations for randomly truncated data with applications. *Statist. Probab. Lett. 17*(2), 139–148.

He, S. and G. L. Yang (1998a). Estimation of the truncation probability in the random truncation model. *Ann. Statist. 26*(3), 1011–1027.

He, S. and G. L. Yang (1998b). The strong law under random truncation. *Ann. Statist. 26*(3), 992–1010.

Keiding, N. and R. D. Gill (1990). Random truncation models and markov processes. *Ann. Statist. 18*(2), 582–602.

Lo, S.-H. and K. Singh (1986). The product-limit estimator and the bootstrap: Some asymptotic representations. *Probab. Theory Related Fields 71*(3), 455–465.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3cr quasars. *Mon. Not. R. Astron. Soc. 155*(1), 95–118.

Major, P. and L. Rejto (1988). Strong embedding of the estimator of the distribution function under random censorship. *Ann. Statist. 16*(3), 1113–1132.

Neveu, J. (1975). *Discrete-parameter martingales*. North-Holland, Amsterdam.

Schick, A., V. Susarla, and H. Koul (1988). Efficient estimation of functionals with censored data. *Stat. Risk Model. 6*(4), 349–360.

Sen, A. and W. Stute (2014). Identification of survival functions through hazard functions in the clayton-family. *Statist. Probab. Lett. 87*, 94–97.

Strzalkowska-Kominiak, E. and W. Stute (2010). On the probability of holes in truncated samples. *J. Statist. Plann. Inference 140*(6), 1519–1528.

Stute, W. (1976). On a generalization of the glivenko-cantelli theorem. *Z. Wahrscheinlichkeit. 35*(2), 167–175.

Stute, W. (1982). The oscillation behavior of empirical processes. *Ann. Probab. 10*(1), 86–107.

Stute, W. (1993a). Almost sure representations of the product-limit estimator for truncated data. *Ann. Statist. 21*(1), 146–156.

Stute, W. (1993b). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal. 45*(1), 89–103.

Stute, W. (1994a). The bias of kaplan-meier integrals. *Scand. J. Stat. 21*(4), 475–484.

Stute, W. (1994b). Convergence of the kaplan-meier estimator in weighted sup-norms. *Statist. Probab. Lett. 20*(3), 219–223.

Stute, W. (1994c). Improved estimation under random censorship. *Commun. Stat. Theor. M. 23*(9), 2671–2682.

Stute, W. (1994d). Strong and weak representations of cumulative hazard function and kaplan-meier estimators on increasing sets. *J. Statist. Plann. Inference 42*(3), 315–329.

Stute, W. (1994e). U-statistic processes: A martingale approach. *Ann. Probab. 22*(4), 1725–1744.

Stute, W. (1995a). The central limit theorem under random censorship. *Ann. Statist. 23*(2), 422–439.

Stute, W. (1995b). The statistical analysis of kaplan-meier integrals. *Lecture Notes-Monograph Series 27*, 231–254.

Stute, W. (1996a). Distributional convergence under random censorship when covariables are present. *Scand. J. Stat. 23*(4), 461–471.

Stute, W. (1996b). The jackknife estimate of variance of a kaplan-meier integral. *Ann. Statist. 24*(6), 2679–2704.

Stute, W. and J.-L. Wang (1993a). Multi-sample u-statistics for censored data. *Scand. J. Stat. 20*(4), 369–374.

Stute, W. and J.-L. Wang (1993b). The strong law under random censorship. *Ann. Statist. 21*(3), 1591–1607.

Stute, W. and J.-L. Wang (1994). The jackknife estimate of a kaplan-meier integral. *Biometrika 81*(3), 602–606.

Stute, W. and J.-L. Wang (2008). The central limit theorem under random truncation. *Bernoulli 14*(3), 604–622.

Susarla, V. and J. Van Ryzin (1980). Large sample theory for an estimator of the mean survival time from censored samples. *Ann. Statist. 8*(5), 1002–1016.

Wang, J.-L. (1995). M-estimators for censored data: strong consistency. *Scand. J. Stat. 22*(2), 197–205.

Wang, M.-C., N. P. Jewell, and W.-Y. Tsai (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist. 14*(4), 1597–1605.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist. 13*(1), 163–177.

Yang, S. (1994). A central limit theorem for functionals of the kaplanmeier estimator. *Statist. Probab. Lett. 21*(5), 337 – 345.

Ying, Z. (1989). A note on the asymptotic properties of the product-limit estimator on the whole line. *Statist. Probab. Lett. 7*(4), 311–314.

The Kaplan-Meier Integral in the Presence of Covariates: A Review

2

Thomas A. Gerds, Jan Beyersmann, Liis Starkopf, Sandra Frank, Mark J. van der Laan and Martin Schumacher

2.1 Introduction

In survival analysis with covariates, many parameters of interest are special cases of the integral:

$$\theta(\varphi) = \int_{\mathbb{R}^p} \int_0^\infty \varphi(t, z) F(\mathrm{d}t \mid z)\, H(\mathrm{d}z). \tag{2.1}$$

Here, T is the time of an event and Z a p-dimensional vector of covariates, φ a square integrable function, and $F(t \mid z) = \mathrm{P}(T \leq t \mid Z = z)$ and $H(\mathrm{d}z) = \mathrm{P}(Z \in \mathrm{d}z)$ denote the conditional survival distribution and the marginal law of Z, respectively. For example, $\theta(I\{t > t^*\})$ is the marginal survival probability at time t^*, $\theta(I\{t > t^*, z_1 > z_1^*\})$ the bivariate distribution at (t^*, z_1^*) (Akritas 1994), and $\theta([I\{t > t^*\} - m(t^*|z)]^2)$ the expected Brier score of a regression model m which predicts survival at time t^* conditional on the covariates (Graf et al. 1999). In the absence of covariates, using the integrand $\varphi_s(t) = \exp(st)$ in (2.1) has been used for expressing the moment generating function of multi-state survival times (Hudson et al. 2014).

T.A. Gerds (✉) · L. Starkopf
Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark
e-mail: tag@biostat.ku.dk

J. Beyersmann · S. Frank
Institute of Statistics, University of Ulm, Ulm, Germany

M.J. van der Laan
Division of Biostatistics, School of Public Health, University of California, Berkeley, USA

M. Schumacher
Institute for Medical Biometry and Statistics, University of Freiburg, Freiburg, Germany

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_2

In a remarkable series, Stute (1993, 1996, 1999) analyzed an estimator of (2.1) for right censored observations of the survival time. The estimator is called the Kaplan-Meier integral. In this paper we first show that Stute's estimator can be written as an inverse of the probability of censoring weighted (IPCW) estimator (Van der Laan and Robins 2003) and then review the structural assumptions of the estimation problem and the asymptotic properties of the estimator.

In biostatistics, Stute's method has recently been put to prominent use for estimating transition probabilities in non-Markov illness-death models (e.g., Meira-Machado et al. 2006; Andersen and Perme 2008; Allignol et al. 2014; de Uña-Álvarez and Meira-Machado 2015). For instance in oncology, illness-death models are used to jointly model progression-free survival and overall survival, and Kaplan-Meier integrals apply interpreting progression-free survival as the covariate and overall survival as time-to-event. We illustrate the general program of the present paper in this example. Using IPCW representations, we obtain simplified estimators that even allow for left-truncated data. Left-truncation is another common phenomenon in survival analysis describing a situation of delayed study entry where individuals are included in prospective cohorts after time origin, conditional on still being alive (Keiding 1992).

2.2 The Kaplan-Meier Integral

Let C be a positive random variable (the censoring time) and suppose that instead of (T, Z) one observes $X = (\tilde{T}, \Delta, Z)$ where $\tilde{T} = \min(T, C)$ and $\Delta = I\{T \leq C\}$. Stute's estimate of (2.1) is defined on a set of n iid right censored observations $X_1, \ldots, X_n$. Let $\tilde{T}_{1:n} \leq \cdots \leq \tilde{T}_{n:n}$ denote the ordered values of $\tilde{T}_1, \ldots, \tilde{T}_n$, and $(\delta_{i:n}, Z_{i:n})$ the concomitant status and covariate values. Stute (1993) introduced the estimate

$$\hat{\theta}(\varphi) = \sum_{i=1}^{n} W_{in}\, \varphi(\tilde{T}_{i:n}, Z_{i:n}) \tag{2.2}$$

where

$$W_{in} = \frac{\delta_{i:n}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{j:n}}.$$

The weights W_{in} do not only match the initials of their inventor's first name, they are also equal to the jump sizes of the Kaplan-Meier estimator for the marginal survival function of T_i and thereby justify the name "Kaplan-Meier integral".

Lemma 1 *Assume that there are no tied event times, i.e.,* $\tilde{T}_{i:n} < \tilde{T}_{(i+1):n}, i = 1, \ldots, n-1$. *The product limit forms of the Kaplan-Meier estimators of the marginal survival time distribution* $S(t) = P(T > t)$ *and the marginal censoring time distribution* $G(t) = P(C > t)$ *are given by*

$$\widehat{S}_0(t) = \prod_{i:\tilde{T}_{i:n} \le t} \left\{1 - \frac{\delta_{i:n}}{n-i+1}\right\} \qquad \widehat{G}_0(t) = \prod_{i:\tilde{T}_{i:n} \le t} \left\{1 - \frac{(1-\delta_{i:n})}{n-i+1}\right\}.$$

The corresponding IPCW *sum forms are:*

$$\widehat{S}_0(t)\widehat{G}_0(t) = \frac{1}{n}\sum_{i=1}^{n} I\{\tilde{T}_{i:n} > t\}$$

$$\widehat{S}_0(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{I\{\tilde{T}_{i:n} \le t\}\delta_{i:n}}{\widehat{G}_0(T_{i:n})},$$

and

$$\widehat{G}_0(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{I\{\tilde{T}_{i:n} \le t\}(1-\delta_{i:n})}{\widehat{S}_0(T_{i:n})}.$$

Proof These relations were readily noted by Gill (1980, page 36) in slightly more general form, that is allowing for tied times.

Lemma 2 *Under the assumption of Lemma 1 the weights of the Kaplan-Meier integral equal the jump size of the Kaplan-Meier estimator:*

$$W_{i:n} = \widehat{S}_0(T_{(i-1):n}) - \widehat{S}_0(T_{i:n})$$

The Kaplan-Meier integral has the following IPCW *representation:*

$$\hat{\theta}(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi(T_{i:n}, Z_{i:n})\delta_{i:n}}{\widehat{G}_0(T_{i:n})}.$$

Proof It follows from Lemma 1 that

$$\widehat{S}_0(T_{(i-1):n}) - \widehat{S}_0(T_{i:n}) = -\frac{1}{n}\sum_{j=1}^{i-1} \frac{\delta_{j:n}}{\widehat{G}_0(T_{j:n})} + \frac{1}{n}\sum_{j=1}^{i} \frac{\delta_{j:n}}{\widehat{G}_0(T_{j:n})} = \frac{1}{n}\frac{\delta_{i:n}}{\widehat{G}_0(\tilde{T}_{i:n})}.$$

The claim follows since

$$nW_{in} = \frac{n\,\delta_{i:n}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{j:n}} = \frac{n\,\delta_{i:n}}{n-i+1}\prod_{j=1}^{i-1}\left(1 - \frac{\delta_{j:n}}{n-j+1}\right)$$

$$= \delta_{i:n}\frac{n}{n-i+1}\widehat{S}_0(\tilde{T}_{(i-1):n}) = \frac{\delta_{i:n}}{\widehat{G}_0(\tilde{T}_{i:n})}.$$

Interestingly, Lemma 2 shows that the IPCW sum form of the Kaplan-Meier estimator (Satten and Datta 2001) is the special case of the Kaplan-Meier integral where $\varphi(t, z) = \hat{\theta}(I\{t > t^*\})$ Akritas (2000).

2.3 Identifiability and Structural Assumptions

2.3.1 Support

In biomedical applications of survival analysis, due to limited follow up times, the support of the censoring times is usually strictly smaller than the support of the survival times. This means that inference on the tail of the survival distribution is not feasible and to identify the parameter in (2.1) based on the right censored observations we have to truncate the parameter at some point in time. To formalize all this let $\tau_0 = \inf_s P(C > s) = 0$ and $\tau_1 = \inf_s P(T > s) = 0$ denote the limits of the supports of C and T, respectively. To meet the setting of typical biomedical applications of survival analysis, we assume $\tau_0 < \tau_1$, and to achieve identifiability we assume that φ satisfies the following condition for some function φ^* of the covariates only and $\epsilon > 0$:

$$\varphi(t, z) = \varphi(t, z) I\{t \leq \tau_0 - \epsilon\} + \varphi^*(z) I\{t > \tau_0 - \epsilon\}. \tag{A1}$$

For example, the mean restricted lifetime is defined as $\theta(t I\{t > t^*\})$ for a suitably chosen truncation time t^*. We refer to Stute (1993, 1996) for a rigorous discussion of the borderline cases where $\epsilon \to 0$.

2.3.2 Independence

Assumption (A1) is not sufficient to achieve identifiability and a further assumption is needed regarding the independence of the censoring mechanism (Tsiatis 1975; Grüger et al. 1991; Gill et al. 1995). To discuss the different assumptions that lead to identifiability we introduce the function G whose values are the conditional probabilities that an observation is uncensored given the event time and the covariates:

$$\mathrm{P}(\Delta = 1 \mid Z = z, T = t) = \mathrm{P}(C > t \mid Z = z, T = t) = G(t, z). \tag{2.3}$$

Even without further independence assumptions, the density of a right censored observation X (with respect to an appropriately chosen dominating measure) can be decomposed as

$$\begin{aligned}\mathrm{P}(\tilde{T} \in \mathrm{d}t, \Delta = \delta, Z \in \mathrm{d}z) = &\{\mathrm{P}(\Delta = 1 \mid Z = z, T = t)\,\mathrm{P}(T \in \mathrm{d}t, Z \in \mathrm{d}z)\}^{\delta} \\ &+ \{\mathrm{P}(\Delta = 0 \mid Z = z, C = t)\,\mathrm{P}(C \in \mathrm{d}t, Z \in \mathrm{d}z)\}^{(1-\delta)}.\end{aligned}$$

The first term can be expressed as

$$\begin{aligned}\mathrm{P}\,(\tilde{T} \in \mathrm{d}t, \Delta = 1, Z \in \mathrm{d}z) &= \mathrm{P}\,(\Delta = 1 \mid Z = z, T = t)\,\mathrm{P}\,(T \in \mathrm{d}t, Z \in \mathrm{d}z) \\ &= G(t, z)\,F(\mathrm{d}t \mid z)\,H(\mathrm{d}z) = \mathrm{P}^{(1)}(\mathrm{d}t,\ \mathrm{d}z)\end{aligned}$$

and this relation motivates the general form IPCW estimation equations for θ:

$$\theta_\varphi(F, H) = \int \varphi(t, z)\,F(\mathrm{d}t \mid z)\,H(\mathrm{d}z) = \int \varphi(t, z)\frac{\mathrm{P}^{(1)}(\mathrm{d}t,\ \mathrm{d}z)}{G(t, z)} = \nu_\varphi(\mathrm{P}^{(1)}, G). \tag{2.4}$$

Since $\mathrm{P}^{(1)}$ only depends on the right censored observations it can be estimated non-parametrically, i.e., by the empirical law of the uncensored observations $\widehat{\mathrm{P}}_n^{(1)}(A, B) = \frac{1}{n}\sum_{i=1}^n \mathcal{I}\{\tilde{T}_i \in A, \Delta_i = 1, Z_i \in B\}$. The general form of the IPCW estimate of θ is then obtained by also substituting an estimate $\widehat{G}$ for G:

$$\hat{\theta}_n(\varphi) = \hat{\nu}_n(\varphi; \widehat{\mathrm{P}}_n^{(1)}, \widehat{G}) = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i \varphi(T_i, Z_i)}{\widehat{G}(T_i, Z_i)}.$$

To justify the IPCW estimate defined in Lemma 2 above, Stute (1993, 1996) restricted the model for G by assuming

$$T \text{ and } C \text{ are independent,} \tag{A2}$$

$$\mathrm{P}\,(T \le C \mid T, Z) = \mathrm{P}\,(T \le C \mid T). \tag{A3}$$

These two conditions together imply

$$G(t, z) = \mathrm{P}\,(C > t | T = t, Z = z) \overset{A3}{=} \mathrm{P}\,(C > t | T = t) \overset{A2}{=} \mathrm{P}\,(C > t). \tag{2.5}$$

Alternatively, we may assume

$$T \text{ and } C \text{ are conditionally independent given } Z \tag{A4}$$

which is familiar from the Cox regression model (compare Begun et al. 1983, page 448). Under (A4) we have

$$G(t, z) = \mathrm{P}\,(C > t \mid Z = z). \tag{2.6}$$

Comparing (2.5) and (2.6) shows that under (A2) and (A3) the function G is a simpler parameter, because it does not depend on the covariates. Note also that neither (A2) implies (A4) nor (A4) implies (A2), and that in generality both assumptions permit that the censoring times depend on the covariates. However, we emphasize that under (A2) and (A3) the function G does not depend on the covariates and hence the conditional censoring distribution may depend on the covariates only in regions of the underlying probability space that are irrelevant for estimation of θ_φ.

Under (A2) and (A3) the function $G(t) = \mathrm{P}(C > t)$ equals the marginal survival function of the censoring times and can be estimated consistently by the marginal reverse Kaplan-Meier estimator for the survival function of the censoring times as defined in Lemma 1. Under (A4) we need to estimate the conditional censoring distribution. Only when all covariates are discrete variables this can be done without further modelling assumptions.

2.4 Large Sample Properties of the Kaplan-Meier Integral

Lemma 2 shows that the plug-in IPCW estimator $\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$ equals Stute's Kaplan-Meier integral (2.2). Stute (1993, 1996) proves strong consistency and weak convergence of $\hat{\theta}(\varphi)=\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)},\widehat{G}_0)$ and obtains the following iid representation (translated to our notation)

$$\sqrt{n}(\hat{\theta}(\varphi) - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IC}_{\hat{\theta}(\varphi)}(\tilde{T}_i, \Delta_i, Z_i) + o_P(1)$$

where the influence function $\mathrm{IC}_{\hat{\theta}(\varphi)}$ of the Kaplan-Meier integral is given in the following theorem.

Theorem 1 *Under* (A1), (A2) *and* (A3) *the Kaplan-Meier integral*

$$\hat{\theta}(\varphi) = \hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$$

is consistent and regular, asymptotically Gaussian linear with influence function

$$\begin{aligned}\mathrm{IC}_{\hat{\theta}(\varphi)}(\tilde{T}_i, \Delta_i, Z_i) = \Delta_i \frac{\varphi(\tilde{T}_i, Z_i)}{G(\tilde{T}_i)} + \frac{(1-\Delta_i)}{W(\tilde{T}_i)} \int_{\tilde{T}_i}^{\infty} \varphi(s,z)\, F(\mathrm{d}s \mid z)\, H(\mathrm{d}z) \\ - \int \left\{ \int_0^{\tilde{T}_i \wedge s} \frac{G(\mathrm{d}u)}{W(u)\, G(u-)} \right\} \varphi(s,z)\, F(\mathrm{d}s \mid z)\, H(\mathrm{d}z) - \theta(\varphi)\end{aligned} \tag{2.7}$$

where $W(t) = \mathrm{P}(\tilde{T}_i > t)$.

Proof See Stute (1993, 1996). An alternative proof can be obtained by applying the functional delta method (e.g. Van der Vaart 1998, Theorem 20.8) to the Hadamard differentiable functional $\nu(G, \mathrm{P}^{(1)})$ using that both the Kaplan-Meier estimator for the censored times $\widehat{G}_0$ and the empirical distribution function $\widehat{\mathrm{P}}_n^{(1)}$ are $\sqrt{n}$-consistent in appropriately normed spaces of distributions (Reeds 1976; Van der Vaart 1991).

2.5 Bias and Efficiency

The Kaplan-Meier integral can have a large sample bias and it is not efficient even not when assumptions (A2) and (A3) are satisfied. The bias can be seen when the conditional survival distribution of the censoring times depend on the covariates $P(C > t|Z = z) \neq P(C > t)$. In this case the marginal Kaplan-Meier estimator for the censored times $\widehat{G}_0(t)$ converges in probability to $\widetilde{G}(t) = \int_{\mathbb{R}^p} G(t \mid z)\, H(\mathrm{d}z)$ and the large sample bias of $\hat{\theta}(\varphi)$ is given by the following limit as $n \to \infty$:

$$\left|\hat{\theta}(\varphi) - \theta(\varphi)\right| \to \left| \int \varphi(t, z) \left\{ \frac{1}{\widetilde{G}(t)} - \frac{1}{G(t, z)} \right\} \mathrm{P}^{(1)}(\mathrm{d}t,\ \mathrm{d}z) \right|.$$

Rotnitzky and Robins (1995) were the first to observe that the Kaplan-Meier integral is not efficient even not when it is consistent and the survival distribution of the censored times does not depend on the covariates.

The following is a special case of Van der Laan and Robins (2003, Theorem 1.1 and Example 1.12), see also Gerds (2002).

Proposition 1 *The efficient influence function for estimation of θ based on the right censored data $(\tilde{T}_i, \Delta_i, Z_i)$ is given by*

$$\begin{aligned} \mathrm{IC}^{\mathit{eff}}(\tilde{T}_i, \Delta_i, Z_i) = \Delta_i \frac{\varphi(\tilde{T}_i, Z_i)}{G(\tilde{T}_i \mid Z_i)} + \frac{(1 - \Delta_i)}{\widetilde{W}(\tilde{T}_i \mid Z_i)} \int_{\tilde{T}_i}^{\infty} \varphi(s, Z_i)\, F(\mathrm{d}s \mid Z_i) \\ - \int \int_0^{\tilde{T}_i \wedge s} \frac{G(\mathrm{d}s \mid z)}{\widetilde{W}(s \mid z)\, G(s- \mid z)} \varphi(s, Z_i)\, F(\mathrm{d}s \mid Z_i) - \theta(\varphi) \end{aligned} \quad (2.8)$$

where $\widetilde{W}(t \mid z) = \mathrm{P}(\tilde{T} > t \mid Z = z)$.

A regular, asymptotically linear estimator is asymptotically efficient if and only if the influence function of the estimator equals the efficient influence function for the estimation problem. Hence, comparing (2.8) with (2.7) shows that $\hat{\theta}(\varphi)$ is inefficient except for the case where $G(t, z) = G(t, z')$ and $F(t, z) = F(t, z')$ for all z, z', i.e. where the covariates are independent of both survival and censoring times (Malani 1995). At first glance, the inefficiency of the Kaplan-Meier integral may appear counter-intuitive as it is not so obvious where the information is lost. A closer look however reveals that the covariate values corresponding to the right censored observations do not enter the statistic (2.2). But, there is information in the fact that no event happened until the end of followup (right censored). This information can be recovered by a model for the conditional survival function of the censored times given the covariates. For example, a standard Cox regression model fitted to the censored times yields

$$\widehat{G}_1(t, z) = \exp\left\{ - \int_0^t \exp(\widehat{\beta}\, z)\, \widehat{\Lambda}_0(\mathrm{d}s) \right\}$$

where $\widehat{\beta}$ and $\widehat{\Lambda}_0$ are the partial likelihood estimates of the regression coefficients and the Breslow estimate of the cumulative baseline hazard function, respectively. The corresponding plug-in IPWC estimator $\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_1)$ is more efficient than the IPCW estimator using Kaplan-Meier for the censoring, but it is still inefficient. The influence curve for this estimator equals $(\Delta_i \varphi)/G - \theta(\varphi)$ minus its projection on the tangent space of the scores of the censoring model, as shown in Van der Laan and Robins (2003, Sect. 2.3.7). The principle of adaptive estimation (Bickel et al. 1993) in this situation can be expressed as follows: The bigger the censoring model the more efficient the IPCW estimator. In particular, if one has available a consistent estimator in a saturated model for G, then the correspondingly defined IPCW estimator is fully efficient. Similarly, it is known that in general the traditional survival rank test needs the whole nonparametric model for its efficiency (Neuhaus 2000). But if the covariates are continuous or high dimensional such estimators perform not very nicely in small samples due to the curse of dimensionality. A practical solution is given by doubly robust estimators which rely on models for both G and F and are locally efficient if both models are correctly specified. If either the model for G or the model for F is correctly specified then the estimator is consistent and asymptotically linear.

2.6 Empirical Results

This section illustrates the magnitude of the potential bias and efficiency loss in the special case $\theta(I\{t > t^*\})$, i.e., where the parameter is the marginal survival function at t^*. Note that in this case the Kaplan-Meier integral (with $\widehat{G}_0$) equals the ordinary Kaplan-Meier estimate. See (e.g. Gerds and Schumacher 2006) for a similar simulation study of IPCW estimators of a more complex parameter. We consider two simulation scenarios. For both settings, a binary covariate is drawn from the binomial distribution with $P(X = 1) = 0.5$. The survival and censoring times were generated using parametric Cox proportional hazard models $\lambda_0^T \exp(1.5Z)$ and $\lambda_0^C \exp(\gamma Z)$, respectively, as described in Bender et al. (2005). In the first setting we set $\gamma = 1.2$ so that the censoring time distribution depends on the covariate. In the second setting we set $\gamma = 0$ so that only the survival times depend on the covariate. In both settings the baseline hazards λ_0^T and λ_0^C were chosen so that $S(t = 70) = 62\%$ and $P(C \leq 70,\ T > C) = 60\%$. We contrast estimates of the parameter $\theta(I\{t > 70\})$ obtained with the Kaplan-Meier estimate $\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$ and with the IPCW estimate $\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_2)$ where

$$\widehat{G}_2(t, z) = \prod_{i:Z_i=z, \tilde{T}_{i:n} \leq t} \left\{1 - \frac{(1 - \delta_{i:n})}{n - i + 1}\right\}$$

Table 2.1 Summary of simulation study for estimating $P(T > 70) = 62\%$ based on right censored data where $P(C \leq 70,\ T > C) = 60\%$. In setting 1 both the survival time distribution and the censoring time distribution depend on a binary covariate. In setting 2 only the survival time distribution depends on a binary covariate

Setting	Estimate	Bias (%)	Variance (%)	MSE (%)
Censoring dependent	$\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$	5.0118	0.309	0.560
on covariate	$\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_2)$	−0.0366	0.266	0.266
Censoring independent	$\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$	−0.0228	0.286	0.286
of covariate	$\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_2)$	−0.0369	0.261	0.261

is the stratified Kaplan-Meier estimate for the censored times conditional on the strata defined by $Z = z$. We report averaged small sample bias and mean squared errors across 2000 simulated data sets.

Table 2.1 shows the results for sample size 200. In the first setting there is a large bias in the marginal Kaplan-Meier estimate whereas $\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_2)$ is less biased. In addition, the variance of the marginal Kaplan-Meier estimate is bigger. In the second setting the marginal Kaplan-Meier IPCW estimate is no longer biased. The same holds for the stratified Kaplan-Meier IPCW estimate. However, the marginal Kaplan-Meier IPCW estimate still has a larger variance than the stratified Kaplan-Meier IPCW estimate (see Table 2.1).

Figure 2.1 illustrates the difference between the estimators as a function of the sample size. We see that the MSE can be large and this is due to a large bias as can be seen from the data in Table 2.1. The left panel of the figure indicates that the difference in MSE decreases with increasing sample size. However, Fig. 2.2 reveals that the relative advantage of the stratified Kaplan-Meier IPCW estimate does note decrease with increasing sample size. The figure also shows that the magnitude of the relative gain in MSE depends on the predictiveness of the covariate and on the amount of censoring.

2.7 Non-Markov Illness-Death Model Without Recovery

The illness-death model without recovery has important biostatistical applications, for example in oncology. In this section we make the connection with the Kaplan-Meier integral. We therefore consider a stochastic process $(X_t)_{t\in[0,\infty)}$ which has state space $\{0, 1, 2\}$, right-continuous sample paths, initial state 0, $P(X_0 = 0) = 1$, intermediate state 1 and absorbing state 2. This process describes an illness-death model without recovery when also the probability of a recovery event is zero, i.e.,

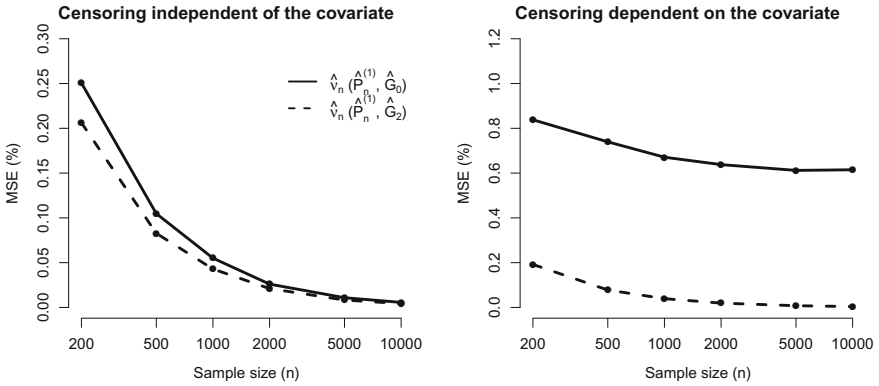


Fig. 2.1 Estimation of S(70). Mean squared error as a function of sample size for Kaplan-Meier integral ($\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$, *solid line*) and IPCW estimator based on stratified Kaplan-Meier for censoring time distribution ($\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_2)$, *dashed line*). In the *left panel* the log-hazard ratio of a binary covariate on survival is 3 and on censoring is 0. In the *right panel* the log-hazard ratio of a binary covariate on survival is 3 and on censoring is 1.2. Data show averages across 2000 simulation runs for each sample size

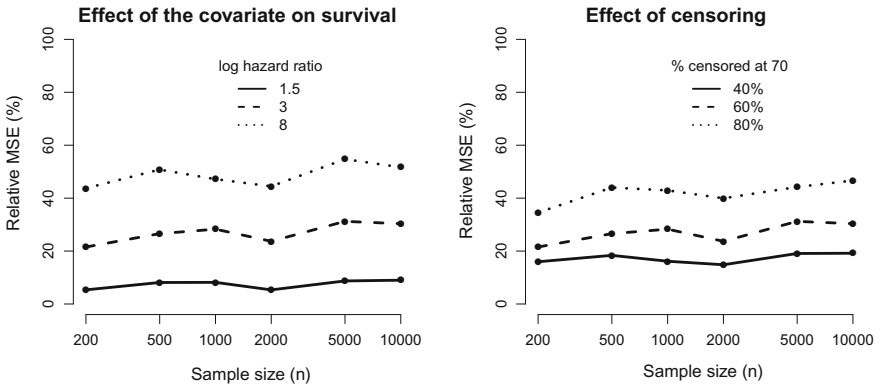


Fig. 2.2 Estimation of S(70). Ratio of mean squared error for Kaplan-Meier integral ($\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_0)$) and IPCW estimator based on stratified Kaplan-Meier for censoring time distribution ($\hat{\nu}_n(\widehat{\mathrm{P}}_n^{(1)}, \widehat{G}_2)$). In both panels the log-hazard ratio of a binary covariate on censoring hazard is 0. In the *left panel* the effect of the binary covariate on survival hazard is varied and in the *right panel* the percentage censored is varied. Data show averages across 2000 simulation runs for each sample size

when $P(X(t) = 0|X(s) = 1) = 0$ for all $s \leq t$. The process can equivalently be described by a pair of random variables

$$T_0 = \inf\{t \ : \ X_t \neq 0\} \text{ and } T = \inf\{t \ : \ X_t = 2\}$$

so that T_0 is the waiting time in the initial state, $X_{T_0} \in \{1, 2\}$, and T the time until the absorbing state is reached. The process passes through the intermediate state 1, if and only if $T_0 < T$, and $T_0 = T$ if the process does not pass through the intermediate state. Our aim is to estimate the transition probabilities between state $l \in \{0, 1\}$ and state $j \in \{1, 2\}$

$$P_{lj}(s, t) = P(X_t = j \mid X_s = l) \tag{2.9}$$

for pairs of time points (s, t) that satisfy $s \leq t$.

Based on right censored data of the illness-death process Meira-Machado et al. (2006) derive an estimator for (2.9) starting with the following representations:

$$\begin{aligned} P_{01}(s, t) &= \frac{P(s < T_0 \leq t, t < T)}{P(T_0 > s)}, \\ P_{11}(s, t) &= \frac{P(T_0 \leq s, t < T)}{P(T > s) - P(T_0 > s)}. \end{aligned} \tag{2.10}$$

The challenge in estimating the right hand sides in (2.10) stems from the numerators, while straightforward Kaplan-Meier estimation applies to estimating $P(T_0 > s)$ and $P(T > s)$. For the numerators, Meira-Machado et al. (2006) apply Stute's Kaplan-Meier integral with 'covariate' $Z = T_0$. Allignol et al. (2014) showed that the estimator of Meira-Machado et al. (2006) can alternatively be derived from a suitably defined competing risks process and they also obtain an IPCW representation of the estimator of Meira-Machado et al. (2006) for $P_{01}(s, t)$ in a similar fashion as we have for the Kaplan-Meier integral in Sect. 2.2. In bivariate (T_0, T)-time several IPCW estimators are available, and Allignol et al. (2014) also discuss an IPCW estimator which uses the estimate of the survival function of the censored times suggested by Tsai and Crowley (1998). This results in a simplified estimator which could easily be extended to left-truncated data. Unfortunately, the Tsai and Crowley (1998) approach is not applicable for estimating $P_{11}(s, t)$.

In what follows we discuss the Kaplan-Meier-integral based estimator of $P_{11}(s, t)$ from the IPCW-perspective. For this we express $P(T_0 \leq s, t < T)$ as a special case of (2.1):

$$P(T_0 \leq s, t < T) = \int I(z \leq s, y > s) \, \mathrm{P}^{T_0, T}(\mathrm{d}z, \mathrm{d}y).$$

For estimation we assume i.i.d. replications $(\tilde{T}_{0i}, \tilde{T}_i, \Delta_i)$, $i = 1, \ldots n$, where $\tilde{T}_{0i} = \min(T_{0i}, C_i)$, $\tilde{T}_i = \min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$. It is convenient to introduce counting processes

$$N(u) = \sum_{i=1}^{n} I\{i \ : \ \tilde{T}_i \leq u, \Delta_i = 1\},$$
$$N^*(u) = \sum_{i=1}^{n} I\{i \ : \ \tilde{T}_i \leq u, \Delta_i = 1, T_{0i} \leq s, T_i > t\},$$
$$Y(u) = \sum_{i=1}^{n} I\{i \ : \ \tilde{T}_i \geq u\}.$$

Straightforward algebra shows that the estimator of Meira-Machado et al. (2006) for $P(T_0 \leq s, t < T)$ equals

$$\sum_u \prod_v \left(1 - \frac{\Delta N(v)}{Y(v)}\right) \frac{\Delta N^*(u)}{Y(u)}, \tag{2.11}$$

where both the sum and the product in (2.11) are over all observed unique times to the absorbing state and ΔN and ΔN^* denote the increments of the counting processes. Since $\prod_v (1 - \frac{\Delta N(v)}{Y(v)})$ is a standard Kaplan-Meier estimator, the IPCW-representations discussed earlier give rise to different possible IPCW-variants of (2.11),

$$\frac{1}{n} \sum_u \left(\hat{P}_a(C \geq u)\right)^{-1} \Delta N^*(u),$$

where $\hat{P}_a(C \geq \cdot)$ is some consistent estimator of the censoring survival function. Recall that in bivariate time there are several possible Kaplan-Meier-type estimators of $P(C \geq \cdot)$, simple choices only using either $\{\tilde{T}_{0i} \ : \ T_{0i} > C_i, i = 1, \ldots n\}$ or $\{\tilde{T}_i \ : \ T_i > C_i, i = 1, \ldots n\}$. Using representation (2.10), we may estimate $P_{11}(s, t)$ by

$$\frac{\frac{1}{n} \sum_u \left(\hat{P}_a(C \geq u)\right)^{-1} \Delta N^*(u)}{\frac{|\{i : \tilde{T}_i > s\}|}{n \hat{P}_b(C > s)} - \frac{|\{i : \tilde{T}_{0i} > s\}|}{n \hat{P}_c(C > s)}}, \tag{2.12}$$

where $\hat{P}_b(C \geq \cdot)$ and $\hat{P}_c(C \geq \cdot)$ are some consistent estimators of the censoring survival function. Because $P_{11}(s, t)$ conditions on being in state 1 at time s, the idea is now to estimate the censoring survival function using the censoring times of the

subjects that are uncensored by time s and are in the intermediate state at the end of followup. In order to formalize this, introduce

$$Y(u;s) = \sum_{i=1}^{n} I\{i \,:\, T_{0i} < s, \tilde{T}_i \geq u\},\ u > s,$$
$$N(u;s) = \sum_{i=1}^{n} I\{i \,:\, T_{0i} < s, \tilde{T}_i \geq s, \tilde{T}_i \leq u, \Delta_i = 1\},\ u > s,$$
$$N^C(u;s) = \sum_{i=1}^{n} I\{i \,:\, T_{0i} < s, \tilde{T}_i \geq s, \tilde{T}_i \leq u, \Delta_i = 0\},\ u > s.$$

In words, $Y(u;s)$ is the number of individuals at risk of absorption at $u-$ in the subset of the data of subjects who are in the intermediate state and uncensored at time s with associated counting process of observed absorption event $N(u;s)$. N^C is the censoring counting process in this data subset. Note that N^* only counts events in the data subset at hand.

Now, define the following estimator of $P(C \geq u)$, $u > s$,

$$\begin{aligned}\tilde{P}(C \geq u;s) &= \tilde{P}(C \geq u \,|\, C > s)\tilde{P}(C > s) \\ &= \prod_{v \in (s,u)} \left(1 - \frac{\Delta N^C(v;s)}{Y(v;s) - \Delta N(v;s)}\right) \tilde{P}(C > s), \qquad (2.13)\end{aligned}$$

where the product in the last display is over all unique jump times of $N^C(\cdot;s)$ and $\tilde{P}(C > s)$ is some consistent estimator of $P(C > s)$.

Using (2.13) in (2.12) (and the same $\tilde{P}(C > s)$ also for $\hat{P}_b$ and $\hat{P}_c$) leads to the estimator

$$\hat{P}_{11}(s,t) = \sum_u \prod_v \left(1 - \frac{\Delta N(v;s)}{Y(v;s)}\right) \frac{\Delta N^*(u;s)}{Y(u;s)}. \qquad (2.14)$$

We note four important facts about $\hat{P}_{11}(s,t)$. Firstly, the estimator is similar to (2.11) but evaluated in the data subset 'in the intermediate state 1 at time s and under observation at s'. Secondly, this data subsetting accounts for the conditioning on $X_s = 1$, and such data subsetting is, in biostatistics, known as *landmarking* (e.g., Anderson et al. 2008; van Houwelingen and Putter 2012). Thirdly, the new estimator (2.14) is just the right-hand limit of the standard Aalen-Johansen estimator of a cumulative incidence function (irrespective of $X(t)$ being Markov or not) and inherits its asymptotic properties (e.g., Andersen et al. 1993, Sect. 4.4). And finally, data subsetting (or landmarking) can easily be extended to random left-truncation (delayed study entry). We illustrate this last aspect with a brief simulation study comparing the Aalen-Johansen estimator of $P_{11}(s,t)$ with the new $\hat{P}_{11}(s,t)$ in a left-truncated non-Markov illness-death model. Recall that the original estimator of Meira-Machado et al. (2006) has only been developed for right-censored data, but an

IPCW-perspective on Kaplan-Meier-integrals has led to an estimator that naturally accounts for left-truncation via landmarking.

To this end, consider n i.i.d. units under study with data $(L_i, \tilde{T}_{0i}, \tilde{T}_i, \Delta_i)$ as before but with the addition of left-truncation times L_i. We assume that (T_{0i}, T_i) is independent of (L_i, C_i) with $P(L_i < C_i) = 1$. We also assume that these n units are under study in the sense that $L_i < \tilde{T}_i$ for all i. In order to account for delayed study entry at time L_i, we re-define

$$Y(u; s) = \sum_{i=1}^{n} I\{i : T_{0i} < s, \tilde{T}_i \geq u, L_i < s\}, \ u > s,$$

and analogously for $N(u; s)$ and $N^C(u; s)$. Then $Y(u; s)$ still denotes the number of individuals at risk of absorption at $u-$ in the subset of subjects who are in the intermediate state and under observation at time s, but now in the presence of left-truncation.

Our simulation design is similar to the one of Meira-Machado et al. (2006). We simulate waiting times T_0 in the initial state from an exponential distribution with parameter $0.039 + 0.026$ and entries into the intermediate state, $X_{T_0} = 1$, with binomial probability $0.039/(0.039 + 0.026)$. For individuals moving through the intermediate state, we set $T = 4.0 \cdot T_0$, making the model non-Markov. Random right-censoring times were drawn from an exponential distribution with parameters 0.013 and 0.035, respectively, and random left-truncation was simulated from a skew normal distribution with location parameter -5, scale 10 and shape 10. We report averages of 1000 simulation runs per scenario, each with a simulated sample size of 200 units.

Table 2.2 shows bias (negative values indicate underestimation) and empirical variance of our new estimator (2.13) and the standard Aalen-Johansen estimator for $P_{11}(s, t)$ (Table 2.3),

$$\prod_{u \in (s,t]} \left(1 - \frac{|\{i : L_i < u = T_i \leq C_i, T_{0i} < T_i\}|}{|\{i : L_i < u \leq \tilde{T}_i, T_{0i} < u\}|}\right)$$

for $s = 25$. In the scenarios considered, the new estimator underestimates and the Aalen-Johansen estimator over-estimates the true probability. The absolute bias in general favours the new estimator, save for early time points and in particular with more pronounced censoring. The empirical variance of the Aalen-Johansen estimator tends to be smaller save for later time points, one possible explanation being that the new estimator uses less data.

2.8 Discussion

The Kaplan-Meier integral can be written as an inverse of probability of censoring weighted estimator for which the weights are estimated with the usual Kaplan-Meier

Table 2.2 Simulation results for estimating $P_{11}(25, t)$ from left-truncated and right-censored non-Markovian data

	censoring hazard 0.013				censoring hazard 0.035			
t	$\hat{P}_{11}(25, t)$		Aalen-Johansen		$\hat{P}_{11}(25, t)$		Aalen-Johansen	
	Bias	Variance	Bias	Variance	Bias	Variance	Bias	Variance
30	−0.0063	0.0029	0.0052	0.0023	−0.1090	0.0311	0.0066	0.0041
40	−0.0089	0.0065	0.0376	0.0047	−0.1101	0.0386	0.0404	0.0094
50	−0.0093	0.0081	0.0818	0.0056	−0.1049	0.0419	0.0877	0.0136
60	−0.0072	0.0082	0.1266	0.0060	−0.0999	0.0415	0.1341	0.0172
70	−0.0100	0.0077	0.1650	0.0061	−0.0974	0.0344	0.1691	0.0221
80	−0.0077	0.0064	0.2019	0.0058	−0.0728	0.0235	0.2132	0.0270
90	−0.0044	0.0037	0.2350	0.0055	−0.0378	0.0115	0.2530	0.0328

Table 2.3 True values of $P_{11}(25, t)$ to be estimated in the simulation study

t	30	40	50	60	70	80	90
$P_{11}(25, t)$	0.8890	0.6930	0.5256	0.3843	0.2649	0.1623	0.0744

method for the censoring times. With this representation the large sample properties of the Kaplan-Meier integral and various modifications can be directly derived with the functional delta method. We further showed in Sect. 2.3 that the conditions imposed by Stute (1993, 1996, 1999) and followers (e.g. Orbe et al. 2003; De Uña Álvarez and Rodriguez-Campos 2004) are practically equivalent to assuming that the censoring is independent of the survival time and of the covariates. Then we showed that it can be advantageous to derive estimators under the conditional independence assumption allowing that the censoring distribution depends on covariates. This improves efficiency and simultaneously reduces the risk of a large sample bias (Robins and Rotnitzky 1992). Our empirical results illustrate the potential bias and the inefficiency of the Kaplan-Meier integral in a specific setting (Table 2.1).

However, in real data applications there is a tradeoff between the simplicity of weighting all the uncensored observations with the Kaplan-Meier for the censoring times and the potential advantages obtained with a working regression model for the conditional censoring distribution. For example in a multi-state framework it is possible to define consistent IPCW estimators for transition probabilities by using the marginal Kaplan-Meier for the censoring (see e.g. Meira-Machado et al. 2006). But, this approach implies that every censored process is weighted unconditional on the state which is occupied at the censoring time. On the other hand, the methods comprised in van der Laan et al. (2002); Van der Laan and Robins (2003) show how to derive more efficient estimators based on an estimate of the survival function of the censored times conditional on the history of the multi-state process and other covariates. In Sect. 2.7, we have exploited this to derive a new estimator of a transition

probability in a non-Markovian illness-death model. Starting with an estimator based on Kaplan-Meier integrals and using the IPCW principle, we also extended the estimator for the case of right-censored and left-truncated data.

Stute's theory of Kaplan-Meier integrals has arguably not entered the mainstream literature on survival analysis, at least not the more biostatistically oriented one, notable exceptions also including Orbe et al. (2002). On the other hand, Kaplan-Meier integrals may form the basis for attacking complex survival models and finding efficient estimators, which we have illustrated for the important illness-death model. We believe that the theory deserves more attention, another possible field of application being competing risks models with a continuous mark (e.g. Gilbert et al. 2008).

References

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Statist.*, 22:1299–1327.

Akritas, M. G. (2000). The central limit theorem under censoring. *Bernoulli*, pages 1109–1120.

Allignol, A., Beyersmann, J., Gerds, T., and Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 20:495–513.

Andersen, P. and Perme, M. (2008). Inference for outcome probabilities in multi-state models. *Lifetime Data Analysis*, 14(4):405–431.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer, New York.

Anderson, J., Cain, K., and Gelber, R. (2008). Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. *Journal of Clinical Oncology*, 26(24):3913–3915.

Begun, J. M., Hall, W. J., Huang, W.-M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11:432–452.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 24:1713–1723.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins.

De Uña Álvarez, J. and Rodriguez-Campos, M. (2004). Strong consistency of presmoothed Kaplan-Meier integrals when covariables are present. *Statistics*, 38:483–496.

de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-markov illness-death model: A comparative study. *Biometrics*, 71(2):364–375.

Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Gerds, T. A. (2002). *Nonparametric efficient estimation of prediction error for incomplete data models*. PhD thesis, Albert-Ludwig Universität Freiburg.

Gilbert, P. B., McKeague, I. W., and Sun, Y. (2008). The 2-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy. *Biostatistics*, 9(2):263–276.

Gill, R. D. (1980). Censoring and stochastic integrals. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

Gill, R. D., Van der Laan, M. J., and Robins, J. M. (1995). Coarsening at random: Characterizations, conjectures and counter-examples. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer Lecture Notes in Statistics.

Graf, E., Schmoor, C., Sauerbrei, W. F., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statist. Med.*, 18:2529–2545.

Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inference based on incomplete observations in disease state models. *Biometrics*, 47:595–605.

Hudson, H. M., Lô, S. N., John Simes, R., Tonkin, A. M., and Heritier, S. (2014). Semiparametric methods for multistate survival models in randomised trials. *Statistics in medicine*, 33(10):1621–1645.

Keiding, N. (1992). Independent delayed entry. In Klein, J. and Goel, P., editors. *Survival analysis: state of the art, Kluwer, Dordrecht*, pages 309–326.

Malani, H. M. (1995). A modification of the redistribution to the right algorithm using disease markers. *Biometrika*, 82:515–526.

Meira-Machado, L., de Uña Álvarez, J., and Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-markov illness-death model. *Lifetime Data Analysis*, 12(3):325–344.

Neuhaus, G. (2000). A method of constructing rank tests in survival analysis. *Journal of Statistical Planning and inference*, 91(2):481–497.

Orbe, J., Ferreira, E., and Núñez-Antón, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in medicine*, 21(22):3493–3510.

Orbe, J., Ferreira, E., and Nunez-Anton, V. (2003). Censored partial regression. *Biostatistics*, 4:109–121.

Reeds, J. A. (1976). *On the definition of von Mises Functionals*. PhD thesis, Havard University, Cambridge, Massachusetts.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology, Methodological Issues*, pages 297–331. Birkh"auser, Boston.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82:805–820.

Satten, G. and Datta, S. (2001). The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.*, 45:89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23:461–71.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 9:1089–1102.

Tsai, W. and Crowley, J. (1998). A note on nonparametric estimators of the bivariate survival function under univariate censoring. *Biometrika*, 85(3):573–580.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci.*, 72:20–22.

Van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.

van der Laan, M. J., Hubbard, A. E., and Robins, J. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association*, 97:494–507.

Van der Vaart, A. W. (1991). On differentiable functionals. *Ann. Statist.*, 19:178–204.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van Houwelingen, J. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

Semi-parametric Random Censorship Models

3

Gerhard Dikta

3.1 Introduction

In lifetime studies, the occurrence of incomplete observations due to some type of censoring is the rule rather than the exception, and we have to take care of this in our data analysis. Different censoring mechanisms can naturally arise from study design, and one type frequently used and widely accepted in practice is the random censorship model (RCM).

This model is described by two independent sequences of independent and identically distributed (i.i.d.) random variables: the survival times $X_1, \ldots, X_n$ and the censoring times $Y_1, \ldots, Y_n$. Based on these two sequences, the observations are given by $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$, where $Z_i = \min(X_i, Y_i)$ and δ_i indicates whether the observation time Z_i is a survival time ($\delta_i = 1$) or a censoring time ($\delta_i = 0$).

We assume here that all random variables are defined over some probability space $(\Omega, \mathscr{A}, \mathbb{P})$, and denote the distribution functions (d.f.) of X, Y, and Z by F, G, and H, respectively. Furthermore, we assume that F, G, and therefore H, are continuous.

Nonparametric statistical inference of F under the RCM usually rests upon the time-honored Kaplan-Meier (KM) product limit estimator, see Kaplan and Meier (1958), defined by

$$1 - F_n^{KM}(t) = \prod_{i:\, Z_i \le t} \Big(1 - \frac{\delta_i}{n - R_{i,n} + 1}\Big),$$

G. Dikta (✉)
Fachhochschule Aachen, Heinrich-Mußmann-Str. 1, 52428 Jülich, Germany
e-mail: dikta@fh-aachen.de

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_3

where $R_{i,n}$ denotes the rank of Z_i among the $Z-$sample. Let $Z_{1:n}, \ldots, Z_{n:n}$ denote the ordered $Z-$sample and $\delta_{[1:n]}, \ldots, \delta_{[n:n]}$ the associated censoring indicators. The mass attached by F_n^{KM} to $Z_{i:n}$ is given by

$$W_{i,n}^{KM} = F_n^{KM}(Z_{i:n}) - F_n^{KM}(Z_{i-1:n}) = \frac{\delta_{[i:n]}}{n-i+1} \prod_{k=1}^{i-1} \left(1 - \frac{\delta_{[k:n]}}{n-k+1}\right). \quad (3.1)$$

Obviously, F_n^{KM} puts mass only on the uncensored observations. Efron (1967) pointed out that the mass attached by F_n^{KM} increases from the smallest to the largest uncensored observation while the amount of increase between two uncensored observations depends on the number of censored observations between them.

When many observations are censored, F_n^{KM} will only have a few jumps with increasing sizes and one might not be satisfied with the accuracy of F_n^{KM}. In such a situation, and if a complete parametric model assumption for F is too restrictive, we can try a semi-parametric extension of the RCM.

Under this approach, a mild parametric assumption is added to RCM to define the semi-parametric random censorship model (SRCM). Precisely, it is assumed that the conditional expectation of δ given the observation time $Z = z$,

$$m(z) = \mathbb{E}(\delta \mid Z = z) = \mathbb{P}(\delta = 1 \mid Z = z),$$

belongs to a parametric family

$$m(z) = m(z, \theta_0),$$

where $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,k}) \in \Theta \subset \mathbb{R}^k$. Essentially, we assume a binary regression model. Together with the SRCM, the semi-parametric estimator

$$1 - F_n^{SE1}(t) = \prod_{i:\, Z_i \le t} \left(1 - \frac{m(Z_i, \theta_n)}{n - R_{i,n} + 1}\right) \quad (3.2)$$

was introduced in Dikta (1998). Here, θ_n is the maximum likelihood estimator (MLE) of θ_0, that is, the maximizer of the (partial) likelihood function

$$l_n(\theta) = \prod_{i=1}^{n} m(Z_i, \theta)^{\delta_i} (1 - m(Z_i, \theta))^{1-\delta_i}.$$

This estimator puts mass on all observations $Z_{i:n}$ according to

$$W_{i,n}^{SE1} = \frac{m(Z_{i:n}, \theta_n)}{n-i+1} \prod_{k=1}^{i-1} \left(1 - \frac{m(Z_{k:n}, \theta_n)}{n-k+1}\right). \quad (3.3)$$

If the largest observation is censored, the total mass attached by F_n^{KM} is less than one. In this case, the KM-estimator is only a sub-d.f. But also F_n^{SE1} has this defect if

$m(Z_{n:n}, \theta_n) \neq 1$. In some cases, this insufficiency could produce misleading results. As an example, assume that we want to estimate $\mathbb{E}(X) = \int x\, F(dx)$, the expected survival time. We can use the plug-in estimator $\int x\, F_n^{KM}(dx)$ or if the SRCM can be assumed $\int x\, F_n^{SE1}(dx)$. If the last observation is censored, $\int x\, F_n^{KM}(dx)$ will underestimate $\mathbb{E}(X)$ since the largest observation gets no weight. Also $\int x\, F_n^{SE1}(dx)$ has this artificial bias if F_n^{SE1} is only a sub-d.f.

Modifications of the KM-estimator are considered in the literature which extends F_n^{KM} to a real d.f. A discussion of some extensions are given in Wellner (1985) and Chen et al (1982). As an example, Efron's self-consistent version of the KM-estimator is a real d.f., see Efron (1967, Theorem 7.1). Under the SRCM, a modification of F_n^{SE1} to a real d.f. is given by

$$1 - F_n^{SE}(t) = \prod_{i:\, Z_i \le t} \Big(1 - \frac{m(Z_i, \theta_n)}{n - R_{i,n} + m(Z_i, \theta_n)}\Big), \tag{3.4}$$

which puts the mass

$$W_{i,n}^{SE} = \frac{m(Z_{i:n}, \theta_n)}{n - i + m(Z_{i:n}, \theta_n)} \prod_{k=1}^{i-1} \Big(1 - \frac{m(Z_{k:n}, \theta_n)}{n - k + m(Z_{k:n}, \theta_n)}\Big) \tag{3.5}$$

on the observation $Z_{i:n}$, for $i = 1, \ldots, n$, see Dikta et al (2016). Note that F_n^{SE} is a real d.f. if $m(Z_{n:n}, \theta_n) \neq 0$.

It is the purpose of this article to review some probabilistic results of the semi-parametric estimators under the SRCM. Further important applied statistical oriented issues are not discussed here. In Sect. 3.2, we will outline an approach to derive survival time estimators in general. Sect. 3.3, focuses on semi-parametric integrals and the 4th section on some goodness-of-fit tests to check the parametric assumption of the SRCM. Sect. 3.5, deals with the bootstrap under the SRCM.

3.2 Deriving Survival Time Estimators Under RCM and SRCM

To motivate this approach, we recall a derivation of the KM-estimator introduced in the survey paper of Gill and Johansen (1990). Set $H^1(s) = \mathbb{P}(\delta = 1\,,\ Z \le s)$ and $\bar{V}(s) = 1 - V(s)$, for an arbitrary d.f. V, to derive for $t \ge 0$ under the RCM and the assumed continuity of the d.f.s that

$$\bar{H}(t) = \bar{F}(t)\, \bar{G}(t) \quad \text{and} \quad H^1(t) = \int_0^t \bar{G}(s)\, F(ds).$$

The last equality states that $\bar{G}$ is a Radon-Nikodym derivative of H^1 with respect to F. Consequently, we get for $t \ge 0$

$$F(t) = \int_0^t \frac{\bar{H}(s)}{\bar{H}(s)}\, F(ds) = \int_0^t \frac{\bar{F}(s)}{\bar{H}(s)}\, \bar{G}(s)\, F(ds) = \int_0^t \frac{\bar{F}(s)}{\bar{H}(s)}\, H^1(ds),$$

the identifying Volterra type integral equation for F. As indicated in Gill and Johansen (1990), this integral equation can be used to derive the KM-estimator, see Kaplan and Meier (1958), by applying an explicit Euler scheme for the approximated integral equation.

To open up the identifying equation for some additional assumptions extending the RCM, we continue as in Dikta et al (2016). Since

$$H^1(t) = \int_0^t m(s)\, H(ds),$$

m is a Radon-Nikodym derivative of H^1 with respect to H and we can modify the identifying equation to get

$$F(t) = \int_0^t \frac{\bar{F}(s)}{\bar{H}(s)}\, m(s) H(ds).$$

Let $\hat{F}$ denote a generic estimator of F and H_n the empirical d.f. of H to get the corresponding estimating equation

$$\hat{F}(t) = \int_0^t \left(\bar{F}(s)\big/\bar{H}(s)\right)_n m_n(s)\, H_n(ds),$$

where $(\bar{F}(s)\big/\bar{H}(s))_n$ and $m_n(s)$ are some estimators of $\bar{F}(s)/\bar{H}(s)$ and $m(s)$, respectively. Thus

$$\hat{F}(Z_{i:n}) = \hat{F}(Z_{i-1:n}) + \int_{]Z_{i-1:n}, Z_{i:n}]} \left(\bar{F}(s)\big/\bar{H}(s)\right)_n m_n(s)\, H_n(ds), \tag{3.6}$$

for $i = 1, \ldots, n$, where we set $Z_{0:n} = 0$ and $\hat{F}(0) = 0$.

Now substitute the integrand $(\bar{F}(s)\big/\bar{H}(s))_n$ with the constant $\bar{\hat{F}}(Z_{i-1:n})\big/\bar{H}_n(Z_{i-1:n})$ (explicit Euler scheme) to get

$$\hat{F}(Z_{i:n}) = \hat{F}(Z_{i-1:n}) + \frac{\bar{\hat{F}}(Z_{i-1:n})\, m_n(Z_{i:n})}{n - i + 1}$$

and, after some basic rearrangements,

$$\bar{\hat{F}}(Z_{i:n}) = \prod_{k=1}^{i} \Big(1 - \frac{m_n(Z_{k:n})}{n-k+1}\Big) \tag{3.7}$$

for $i = 1, \ldots, n$.

If $i < n$, we can also use $\bar{\hat{F}}(Z_{i:n}) \big/ \bar{H}_n(Z_{i:n})$ to substitute $(\bar{F}(s) \big/ \bar{H}(s))_n$ in Eq. (3.6) (implicit Euler scheme) and derive

$$\hat{F}(Z_{i:n}) = \hat{F}(Z_{i-1:n}) + \frac{\bar{\hat{F}}(Z_{i:n})\, m_n(Z_{i:n})}{n-i}$$

to obtain

$$\bar{\hat{F}}(Z_{i:n}) = \prod_{k=1}^{i} \Big(1 - \frac{m_n(Z_{k:n})}{n-k+m_n(Z_{k:n})}\Big) \tag{3.8}$$

for $i = 1, \ldots, n$. Note that this equation holds for $i < n$ but it can be extended to $i = n$ if $m_n(Z_{n:n}) > 0$.

Finally, to get the specific $\hat{F}$, we have to specify the estimator $m_n(Z_{k:n})$ in the two Eqs. (3.7) and (3.8), respectively.

In the absence of any further information about m (besides RCM), we only know that $\mathbb{E}(\delta_{[k:n]} | Z_{k:n} = x) = m(x)$, see Stute and Wang (1993, Lemma 2.1), and therefore $\delta_{[k:n]}$ is the only possible candidate for $m_n(Z_{k:n})$. With this substitution, Eqs. (3.7) and (3.8) are the KM-estimator.

A pre-smoothed version of the KM-estimator can be obtained from these two equations if a non-parametric estimator of $m(Z_{k:n})$ is plugged in for $m_n(Z_{k:n})$. But to use such a non-parametric estimator, we have to know that m is a smooth function, thus we need some additional assumptions. The pre-smoothed KM-estimator based on Eq. (3.7) was introduced by Ziegler (1995), see also Cao et al (2003). To the best of our knowledge, results about the corresponding pre-smoothed version based on Eq. (3.8) are not available yet.

If we know that m belongs to a parametric family, we are in the SRCM and use $m(Z_{k:n}, \theta_n)$ for $m_n(Z_{k:n})$. Equations (3.7) and (3.8) then yield the semi-parametric estimators given in Eqs. (3.2) and (3.4).

3.3 Integral Estimators

In data analysis, we often have to estimate some specific parameters of an underlying d.f. which can be expressed by an integral of a Borel-measurable function φ with respect to the underlying d.f. F, that is, $\int \varphi \, dF$. Some candidates for φ which are of particular interest in statistics, are discussed in Stute and Wang (1993). If all of our observations would be uncensored, the empirical d.f. F_n would be plugged

in to get $\int \varphi \, dF_n$ as an estimator of $\int \varphi \, dF$. In this case, the strong law of large numbers (SLLN) and the central limit theorem (CLT) guarantee strong consistency and asymptotic normality of the plug-in estimator. In this section, we will discuss the corresponding results for $\int \varphi \, dF_n^{KM}$, $\int \varphi \, dF_n^{SE1}$, and $\int \varphi \, dF_n^{SE}$. Note that

$$\int \varphi \, dF_n^{KM} = \sum_{i=1}^{n} \varphi(Z_{i:n}) W_{i,n}^{KM},$$

where $W_{i,n}$ is defined by (3.1). The semi-parametric integrals are defined similarly, see (3.3) and (3.5).

Since all three estimators distribute mass only on the observations Z, none of them can be used to estimate $F(t)$ if $t > \tau_H$, where

$$\tau_H = \inf\{x \, : \, H(x) = 1\}$$

is the rightmost point of the support of H. Consequently, we can only expect to estimate the restricted integral

$$\int_0^{\tau_H} \varphi \, dF$$

with these estimators. However, if $\tau_F \leq \tau_H$, the restricted integral coincides with $\int \varphi \, dF$.

3.3.1 Strong Consistency

The most general result with respect to the choice of φ for strong consistency of KM-integrals is given in Stute and Wang (1993, Theorem 1.1). Note that this theorem is not restricted to continuous F and G. The proof is mainly based on Stute and Wang (1993, Lemma 2.2) which states that for continuous H and $\varphi \geq 0$

$$\Big(\int \varphi \, dF_n^{KM}, \mathscr{F}_n^{KM} \Big)_{n \geq 1}$$

is a reversed-time supermartingale, where

$$\mathscr{F}_n^{KM} = \sigma\big(Z_{i:n}, \delta_{[i:n]}, 1 \leq i \leq n, Z_{n+1}, \delta_{n+1}, \ldots\big).$$

This guarantees, among other things, the almost sure (a.s.) convergence of the KM-integral to a random variable S according to Neveu (1975, Proposition V-3–11). The identification of the limit S is also based on a reversed-time supermartingale approach, where $m(t)$, the conditional expectation of δ given $Z = t$ is crucial.

The semi-parametric counterpart, restricted to continuous H, can be found in Dikta (2000, Theorem 1.1). Under a mild moment condition, strong consistency of the MLE θ_n, and some local smoothness of the parametric model for m

$$\int \varphi \, dF_n^{SE1} \longrightarrow \int_0^{\tau_H} \varphi \, dF, \quad \text{a.s.}$$

as $n \to \infty$, is derived in this theorem. The proof is based on the reversed-time supermartingale approach introduced by Stute and Wang (1993) in the context of the KM-integral estimator. However, since the MLE θ_n is used under SRCM to obtain the estimate of m, a direct reversed supermartingale approach seems not to be applicable. Instead, it is shown that for every $\varepsilon > 0$, $\int \varphi \, dF_n^{SE1}$ can be enclosed a.s., for large n, between two random sequences $\xi_n(\varepsilon)$ and $\eta_n(\varepsilon)$ which converge a.s. to constant limits $U(\varepsilon)$ and $O(\varepsilon)$ such that

$$U(\varepsilon) \le \int_0^{\tau_H} \varphi \, dF \le O(\varepsilon)$$

and $O(\varepsilon) - U(\varepsilon) \to 0$, as $\varepsilon \to 0$. The convergence of the sequences ξ_n and η_n is based on Dikta (2000, Lemma 2.1), saying that for $\varphi \ge 0$ and every Borel-measurable function $q : \mathbb{R} \ni t \to q(t) \in [0, 1]$

$$\Big(\sum_{i=1}^{n} \varphi(Z_{i:n}) W_{i,n}(q), \mathscr{F}_n \Big)_{n \ge 1} \text{ and } \Big(\sum_{i=1}^{n} \varphi(Z_{i:n}) \bar{W}_{i,n}(q), \mathscr{F}_n \Big)_{n \ge 1},$$

are reversed-time supermartingales, where $\mathscr{F}_n = \sigma\big(Z_{i:n}, 1 \le i \le n, Z_{n+1}, \dots\big)$ and

$$W_{i,n}(q) = \frac{q(Z_{i:n})}{n-i+1} \prod_{k=1}^{i-1} \Big(1 - \frac{q(Z_{k:n})}{n-k+1}\Big), \; \bar{W}_{i,n}(q) = \frac{1}{n-i+1} \prod_{k=1}^{i-1} \Big(1 - \frac{q(Z_{k:n})}{n-k+1}\Big).$$

Strong consistency of the second semi-parametric integral, that is,

$$\int \varphi \, dF_n^{SE} \longrightarrow \int_0^{\tau_H} \varphi \, dF, \quad \text{a.s.}$$

is given in Dikta et al. (2016, Theorem 2.7). As we already mentioned in the introduction, F_n^{SE} is a real d.f. which is an obvious advantage over F_n^{SE1} here. The proof of this theorem relies on the a.s. convergence of $\int \varphi \, dF_n^{SE1}$ and on

$$\Big| \int \varphi \, dF_n^{SE} - \int \varphi \, dF_n^{SE1} \Big| \longrightarrow 0, \quad \text{a.s.},$$

as $n \to \infty$. To establish the last result, the equality

$$\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i = \sum_{i=1}^{n} (a_i - b_i) \Big(\prod_{k=1}^{i-1} a_k \prod_{k=i+1}^{n} b_k \Big) \tag{3.9}$$

is used. This formula appears in elementary proofs of the Lindeberg-Feller CLT and holds for any sequences of complex numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$.

If we specify $\varphi(x) = 1(x \leq t)$, where $1(x \in A)$ denotes the indicator function corresponding to the set A, the strong consistency results discussed here are the point-wise a.s. convergence of $F_n^{KM}(t)$, $F_n^{SE1}(t)$, and $F_n^{SE}(t)$ towards $F(t)$, for $0 \leq t < \tau_H$. This point-wise convergence can be extended to Glivenko-Cantelli (uniform) convergence by standard arguments, see Loève (1977, p. 21). Further generalized uniform convergence results may be established from these strong consistent integral estimators in connection with Stute (1976).

3.3.2 Asymptotic Normality

Under some moment conditions, Stute (1995, Theorem 1.1) states an asymptotic linear representation of the KM-integral. According to the CLT, this leads directly to the asymptotic normality of the KM-integral estimator. Precisely,

$$n^{1/2}\Big(\int \varphi \, dF_n^{KM} - \int_0^{\tau_H} \varphi \, dF\Big) \longrightarrow \mathcal{N}(0, \sigma_{KM}^2),$$

in distribution, as $n \to \infty$, where

$$\sigma_{KM}^2 = \mathrm{VAR}\big(\varphi(Z)\gamma_0(Z)\delta + \gamma_1(Z)(1-\delta) - \gamma_2(Z)\big).$$

The precise definitions of γ_0, γ_1 and γ_2 are given in Stute (1995) and are omitted here.

Under SRCM and some regularity assumptions and moment conditions, a corresponding asymptotic linear representation is derived in Dikta et al. (2005, Theorem 2.1) for the first semi-parametric integral estimator to obtain

$$n^{1/2}\Big(\int \varphi \, dF_n^{SE1} - \int_0^{\tau_H} \varphi \, dF\Big) \longrightarrow \mathcal{N}(0, \sigma_{SE}^2),$$

in distribution, as $n \to \infty$, where

$$\sigma_{SE}^2 = \mathrm{VAR}\Big(\varphi(Z)\gamma_0(Z)m(Z,\theta_0) + \gamma_1(Z)(1 - m(Z(\theta_0))) - \gamma_2(Z) - K(Z,\delta)\big(\gamma_3(Z) - \gamma_4(Z)\big)\Big).$$

The definition of γ_3, γ_4, and K can be looked up there.

A general comparison of the two variances under a correctly specified parametric model for m shows that $\sigma_{SE}^2 \leq \sigma_{KM}^2$, where equality will be the exception, see Dikta et al. (2005, Corollary 2.5). To see this, we redo the essential part of the proof here for the special case that θ_0 is one dimensional. As pointed out in Dikta et al. (2005, (4.8)),

$$\sigma_{KM}^2 - \sigma_{SE}^2 = \mathbb{E}\big(m(Z)\bar{m}(Z)(A^2(Z) - B^2(Z))\big),$$

where $\bar{m}(z) = 1 - m(z)$, $m(z) = m(z, \theta_0)$, and

$$m(z)\bar{m}(z)B(z) = \frac{Dm(z,\theta_0)}{\sigma^2} \int A(x) Dm(x,\theta_0)\, H(dx). \tag{3.10}$$

Here, $\sigma^2 = \mathbb{E}\big([Dm(Z,\theta_0)/\sqrt{m(Z,\theta_0)\bar{m}(Z,\theta_0)}]^2\big)$, where $Dm(z,\theta_0)$ denotes the derivative of $m(z,\theta)$ at θ_0. The precise definitions of A and B can be found in Dikta et al. (2005, (2.4)), but they are not relevant here.

According to (3.10),

$$\begin{aligned}\mathbb{E}\big(m(Z)\bar{m}(Z)B^2(Z)\big) &= \mathbb{E}\Big(\frac{(Dm(Z,\theta_0))^2}{\sigma^4 m(Z)\bar{m}(Z)}\Big(\int A(x)Dm(x,\theta_0)\, H(dx)\Big)^2\Big) \\ &= \sigma^{-2}\Big(\int \sqrt{m(x)\bar{m}(x)}A(x)\,\frac{Dm(x,\theta_0)}{\sqrt{m(x)\bar{m}(x)}}\, H(dx)\Big)^2.\end{aligned}$$

Recall the definition of σ^2 to get, according to Cauchy-Schwarz's inequality, that the last term is less or equal to $\int m(x)\bar{m}(x)A^2(x)\, H(dx)$ which shows that $\sigma^2_{SE} \le \sigma^2_{KM}$. Furthermore, $\sigma^2_{SE} = \sigma^2_{KM}$ can only occur if there is equality in the application of Cauchy-Schwarz's inequality. Compare Shorack and Wellner (1986, p. 843) to see that this can only appear in an exceptional case.

3.3.3 Efficiency

Since F_n^{SE1} incorporates the additional parametric model information which can not be used for F_n^{KM} under the RCM, the variance comparison in the last section is unfair and the result is not surprising. In other words, F_n^{SE1} would be useless if no gain in efficiency compared to F_n^{KM} could be achieved. To evaluate the quality of the semi-parametric integral estimator, one has to compare it with other possible estimators under SRCM. Dikta (2014, Corollary 3.11) states the result of such a comparison. There it is shown that the semi-parametric integral estimator is asymptotically efficient with respect to the class of all regular estimators of $\int 1(0 \le x \le \tau_H)\varphi(x)\, F(dx)$ under the SRCM. Note that this result holds if the correct parametric model is used. Compare also Wellner (1982) for the asymptotic efficiency of the KM-estimator within the class of all regular estimating sequences under the RCM.

Asymptotic linearity, normality, and efficiency have been obtained for $\int \varphi\, dF_n^{SE1}$. But an application of (3.9) also shows that

$$n^{1/2}\Big(\int \varphi\, dF_n^{SE1} - \int \varphi\, dF_n^{SE}\Big) \longrightarrow 0, \quad \text{in probabiliy, as } n \to \infty,$$

see Dikta et al. (2016, Theorem 2.2). Therefore, these results are the same for both semi-parametric integral estimators.

3.4 Testing the Parametric Assumption of SRCM

Under SRCM, $(\delta_1, Z_1), \ldots, (\delta_n, Z_n)$ can be interpreted as observations from a binary regression model (BRM). Some popular examples of BRMs are discussed in textbooks on generalized linear models (GLM), e.g. in McCullagh and Nelder (1989). Further examples of parametric models for m can be found in Dikta (1998) or can be constructed based on Eq. (3.3) of that paper which relates m to the two hazard rates corresponding to F and G.

Once a parametric model for m is specified, one should examine its validity by applying a goodness-of-fit (GOF) test, that is, one has to test the null hypothesis

$$H_0 : m(\cdot) \in \mathcal{M} \quad \text{versus} \quad H_1 : m(\cdot) \notin \mathcal{M},$$

where $\mathcal{M} = \{m(\cdot, \theta \mid \theta \in \Theta)\}$ specifies the parametric model for m.

To obtain an universal approach for GOF tests in the BRM setup, Dikta et al (2006) adapt the ideas outlined in Stute (1997); Stute et al (1998a, b). In Stute (1997), a marked empirical process (MEP), that is, the cusum process based on the residuals of some parametric regression models, is studied. Stute (1997, Corollary 1.3) shows that the MEP converges in distribution to a centered Gaussian process. The continuous mapping theorem then guarantees that critical values for Kolmogorov-Smirnov (KS) and Cramér-von Mises (CvM) type GOF tests which are based on this MEP can be approximated by the distribution of corresponding functionals applied to the limit process of the MEP. Unfortunately, the limit distributions of the MEP and of these functionals are complicated and model depending, thus not distribution free. In Stute et al (1998b), this problem is tackled by replacing the MEP with its innovation martingale to get asymptotically distribution free KS and CvM tests. A wild bootstrap approach is also applicable to handle this problem properly, compare Stute et al (1998a).

The MEP for our BRM is defined by

$$R_n(x) = n^{-1/2} \sum_{i=1}^{n} (\delta_i - m(Z_i, \theta_n)) 1(Z_i \leq x), \quad 0 \leq x \leq \infty.$$

Based on this process, the corresponding KS and CvM test statistics are given by

$$D_n = \sup_{0 \leq x \leq \infty} |R_n(x)| \text{ and } W_n = \int R_n^2(x)\, H_n(dx),$$

respectively. Stute (1997, Corollary 1.3) can be applied to derive the convergence in distribution of R_n to a centered Gaussian process R in $D[0, \infty]$, the space of càdlàg functions. Among other things, the covariance function of R depends on the model $\mathcal{M}$, see Dikta et al. (2006, (10)) for the concrete covariance structure of the limiting process.

Instead of an innovation martingale approach, a model-based bootstrap (MBB) technique is applied there to handle the model dependence of the limit distributions.

Compared to the innovation martingale technique, MBB has the charm of simplicity and can easily be implemented.

Since a p-value of a statistical test is always calculated or approximated under the null distribution, the underlying resampling of the bootstrap data should reflect this, regardless whether the original observations are generated under the null hypothesis or the alternative. To guarantee this, the underlying resampling should always be done under the null hypothesis or as close as possible to it. In the BRM setup, δ has a Bernoulli distribution with success parameter $m(z)$ if $Z = z$ is observed. Under MBB, the bootstrap data $(Z_1^*, \delta_1^*), \ldots, (Z_n^*, \delta_n^*)$ are i.i.d., where

$$Z_i^* = Z_i, \quad \delta_i^* \sim \text{Bernoulli}(m(Z_i^*, \theta_n)).$$

Let θ_n^* be the MLE of the bootstrap sample and define the corresponding bootstrap version of the MEP by

$$R_n^*(x) = n^{-1/2} \sum_{i=1}^{n} (\delta_i^* - m(Z_i^*, \theta_n^*)) 1(Z_i^* \leq x), \quad 0 \leq x \leq \infty.$$

Obviously, the bootstrap data are generated under the null hypothesis even if the original data are from the alternative. Dikta et al. (2006, Theorem 2, Remark 2) show that R_n^* tends to the same Gaussian process R as R_n does under the null hypothesis. Even under the alternative, if $m(\cdot, \theta_0)$ is interpreted as the projection of $m(\cdot)$ onto $\mathscr{M}$ with respect to the Kullback-Leibler geometry, R_n^* tends to this limit process in distribution with probability 1. Overall this guarantees that the distribution of KS and CvM statistics based on R_n^* can be used to obtain approximated p-values for the original KS and CvM test.

3.5 Bootstrapping Under SRCM

The validation method discussed in the last section is based on the MEP, a cusum residual process. The corresponding KS and CvM tests are completely specified by the chronology of the residuals while concrete time-stamps, given by the ordered $Z-$sample, do not influence these statistics directly. Bootstrapping of the MEP in this scenario mainly has to reflect the heteroscedastic nature of the BRM residuals; whereas the time-stamps can be kept fixed.

This situation changes, when we want to approximate the distribution of the process

$$\alpha_n^1(t) = n^{1/2}\big(F_n^{SE1}(t) - F(t)\big), \qquad 0 \leq t \leq \tau < \tau_H$$

with a bootstrap approach. Now there is no cusum residual process involved anymore and both, the $Z-$ and $m(Z)-$sample, carry the important part of the information described by the process α_n^1. A resampling procedure which takes care of this situation is the two-stage model-based bootstrap (TMBB) introduced in Subramanian and

Zhang (2013). Under TMBB, the bootstrap data $(Z_1^*, \delta_1^*), \ldots, (Z_n^*, \delta_n^*)$ are i.i.d., where

$$Z_i^* \sim H_n, \quad \delta_i^* \sim \text{Bernoulli}(m(Z_i^*, \theta_n)).$$

Note that in the first step, the $Z-$data are resampled according to H_n, like in the classical bootstrap of the empirical process, while in the second step the MBB is used. Based on this dataset, the MLE θ_n^* is calculated to derive

$$1 - F_n^{SE1*}(t) = \prod_{i:\, Z_{i:n}^* \le t} \Big(1 - \frac{m(Z_{i:n}^*, \theta_n^*)}{n - i + 1}\Big),$$

where $Z_{1:n}^* \le \cdots \le Z_{n:n}^*$ denotes the ordered Z^*- sample. The corresponding bootstrap version of α_n^1 is then defined by

$$\alpha_n^{1*}(t) = n^{1/2}\big(F_n^{SE1*}(t) - F_n^{SE1}(t)\big), \qquad 0 \le t \le \tau < \tau_H.$$

In Subramanian and Zhang (2013, Theorem 3) it is shown that with probability 1 α_n^{1*} tends to the same Gaussian process as α_n^1. Based on this result, a simultaneous confidence band (SCB) for F is constructed. Furthermore, in a simulation study, this SCB is compared to SCBs which are based on other procedures. In the concluding discussion, the authors pointed out that their numerical studies show that the TMBB based SCB "performs as well as or better than competing SCBs whether or not there is a parametric misspecification."

Similar results for the TMBB based bootstrap version of the process

$$\alpha_n(t) = n^{1/2}\big(F_n^{SE}(t) - F(t)\big), \qquad 0 \le t \le \tau < \tau_H$$

are not studied yet.

In the case of the Kaplan-Meier estimator, there are two resampling plans discussed in the literature. In Reid (1981), the bootstrap data are generated as an i.i.d. sample from F_n^{KM}. Efron (1981) suggested to generate the bootstrap sample by drawing uniformly from the original data $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ with replacement. As pointed out by Akritas (1986), only the latter approach can by used to construct SCBs. Under the first resampling plan, the limit process obtained for the bootstrap does not match the limit of the Kaplan-Meier process.

References

Akritas MG (1986) Bootstrapping the Kaplan-Meier estimator. J Amer Statist Assoc 81(396):1032–1038, URL http://links.jstor.org/sici?sici=0162-1459(198612)81:396<1032:BTKE>2.0.CO;2-Q&origin=MSN

Cao R, Janssen P, López de Ullibarri I, Veraverbeke N (2003) Presmoothed kaplan-meier and nelson-aalen estimators. Tech. rep., Reports in Statistics and Operations Research, University of Santiago de Compostela, URL http://eio.usc.es/pub/reports/report0301.pdf

Chen YY, Hollander M, Langberg NA (1982) Small-sample results for the Kaplan-Meier estimator. J Amer Statist Assoc 77(377):141–144, URL http://links.jstor.org/sici?sici=0162-1459(198203)77:377<141:SRFTKE>2.0.CO;2-U&origin=MSN

Dikta G (1998) On semiparametric random censorship models. J Statist Plann Inference 66(2):253–279

Dikta G (2000) The strong law under semiparametric random censorship models. J Statist Plann Inference 83(1):1–10

Dikta G (2014) Asymptotically efficient estimation under semi-parametric random censorship models. J Multivariate Anal 124:10–24, DOI:10.1016/j.jmva.2013.10.002, URL http://dx.doi.org/10.1016/j.jmva.2013.10.002

Dikta G, Ghorai J, Schmidt C (2005) The central limit theorem under semiparametric random censorship models. J Statist Plann Inference 127(1-2):23–51

Dikta G, Kvesic M, Schmidt C (2006) Bootstrap approximations in model checks for binary data. J Amer Statist Assoc 101(474):521–530

Dikta G, Reißel M, Harlaß C (2016) Semi-parametric survival function estimators deduced from an identifying volterra type integral equation. Journal of Multivariate Analysis 147:273 – 284, DOI:10.1016/j.jmva.2016.02.008, URL http://www.sciencedirect.com/science/article/pii/S0047259X16000300

Efron B (1967) The two sample problem with censored data. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health, University of California Press, Berkeley, Calif., pp 831–853, URL http://projecteuclid.org/euclid.bsmsp/1200513831

Efron B (1981) Censored data and the bootstrap. J Amer Statist Assoc 76(374):312–319, URL http://links.jstor.org/sici?sici=0162-1459(198106)76:374<312:CDATB>2.0.CO;2-2&origin=MSN

Gill RD, Johansen S (1990) A survey of product-integration with a view toward application in survival analysis. Ann Statist 18(4):1501–1555, DOI:10.1214/aos/1176347865, URL http://dx.doi.org/10.1214/aos/1176347865

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Amer Statist Assoc 53:457–481

Loève M (1977) Probability theory. I, 4th edn. Springer-Verlag, New York-Heidelberg, graduate Texts in Mathematics, Vol. 45

McCullagh P, Nelder JA (1989) Generalized linear models. Monographs on Statistics and Applied Probability, Chapman & Hall, London, DOI:10.1007/978-1-4899-3242-6, URL http://dx.doi.org/10.1007/978-1-4899-3242-6, second edition [of MR0727836]

Neveu J (1975) Discrete-parameter martingales, revised edn. North-Holland Publishing Co., Amsterdam-Oxford; American Elsevier Publishing Co., Inc., New York, translated from the French by T. P. Speed, North-Holland Mathematical Library, Vol. 10

Reid N (1981) Estimating the median survival time. Biometrika 68(3):601–608, DOI:10.1093/biomet/68.3.601, URL http://dx.doi.org/10.1093/biomet/68.3.601

Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York

Stute W (1976) On a generalization of the Glivenko-Cantelli theorem. Z Wahrscheinlichkeitstheorie und Verw Gebiete 35(2):167–175

Stute W (1995) The central limit theorem under random censorship. Ann Statist 23(2):422–439

Stute W (1997) Nonparametric model checks for regression. Ann Statist 25(2):613–641, DOI:10.1214/aos/1031833666, URL http://dx.doi.org/10.1214/aos/1031833666

Stute W, Wang JL (1993) The strong law under random censorship. Ann Statist 21(3):1591–1607

Stute W, González Manteiga W, Presedo Quindimil M (1998a) Bootstrap approximations in model checks for regression. J Amer Statist Assoc 93(441):141–149, DOI:10.2307/2669611, URL http://dx.doi.org/10.2307/2669611

Stute W, Thies S, Zhu LX (1998b) Model checks for regression: an innovation process approach. Ann Statist 26(5):1916–1934, DOI:10.1214/aos/1024691363, URL http://dx.doi.org/10.1214/aos/1024691363

Subramanian S, Zhang P (2013) Model-based confidence bands for survival functions. J Statist Plann Inference 143(7):1166–1185

Wellner JA (1982) Asymptotic optimality of the product limit estimator. Ann Statist 10(2):595–602

Wellner JA (1985) A heavy censoring limit theorem for the product limit estimator. Ann Statist 13(1):150–162, DOI:10.1214/aos/1176346583, URL http://dx.doi.org/10.1214/aos/1176346583

Ziegler S (1995) Ein modifizierter Kaplan-Meier Schätzer. University of Giessen, diploma thesis, supervisor: W. Stute

Nonparametric Estimation of an Event-Free Survival Distribution Under Cross-Sectional Sampling

4

Jacobo de Uña-Álvarez

4.1 Introduction

In Survival Analysis and other fields, a target of much interest is the event-free survival. This function evaluates along time the probability of surviving without undergoing a certain intermediate event. The intermediate event may represent a post-operatory complication (like infection), a recurrence of a disease, and so on; in these biomedical examples, the event-free survival reports the survival probability but restricted to the healthy population, being referred to as infection-free, disease-free, or recurrence-free survival.

Cross-sectional samplings or prevalence studies are often applied due to their simplicity relative to prospective or incidence designs (see e.g. Wang 1991; Fluss et al. 2013). Under cross-sectional sampling, only individuals in progress (that is, alive) at the cross-section date are recruited and, therefore, the survival times are left-truncated by the recruitment times. Besides, right-censoring will often appear due to limitations in the follow-up of the individuals, withdrawals, deaths unrelated to the disease of interest, etc. See Fig. 4.1 for a graphical description. In this setting, nonparametric estimation of the (total) survival has been deeply investigated along the last three decades. The product-limit estimator with left-truncated and right-censored data (Tsai et al. 1987), which extends the time-honoured Kaplan-Meier estimator to the truncated setting, is consistent under some identifiability assumptions; these include independence between the truncation-censoring times and the survival times, and support conditions to ensure the availability of sampling information on the lifetime of interest. See Tsai et al. (1987), Wang (1991), Stute (1993), Gijbels and Wang (1993), Zhou and Yip (1999) or, more recently, Stute and Wang (2008) for results

J. de Uña-Álvarez (✉)
Department of Statistics and OR & CINBIO, University of Vigo, Vigo, Spain
e-mail: jacobo@uvigo.es

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_4

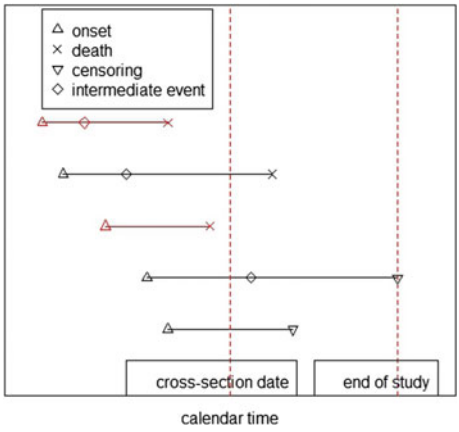


Fig. 4.1 Left-truncated and right-censored cross-sectional survival data with an intermediate event. Red segments are unobserved

and access to the related literature. However, estimation of the event-free survival under cross-sectional sampling has not been investigated in much detail.

The main point to introduce a suitable estimator for the event-free survival under the described cross-sectional sampling scenario is to note that the left-truncation time acts on the total survival, T^0 say, rather than on the event-free lifetime, Z^0 say. That is, it is possible to recruit individuals with Z^0 smaller than the left-truncation time, as long as T^0 is larger than the latter (see Fig. 4.1). This makes a difference with respect to the standard setting with left-truncated and right-censored data, and the left-truncation times must be handled in a different, particular manner for the construction of the estimator. Still, estimation of the joint distribution of (Z^0, T^0), from which an estimator of the event-free survival can be immediately obtained by taking the corresponding marginal, has received some attention in the recent years. We review this approach in Sect. 4.2, and we discuss its main theoretical and practical limitations.

The rest of the paper is organized as follows. In Sect. 4.2 we introduce the needed notations and we review existing estimators for the event-free survival. In Sect. 4.3 a new nonparametric estimator for the event-free survival is introduced, and some asymptotic results are established. In Sect. 4.4 a comparative numerical study is conducted. Some conclusions and final remarks are given in Sect. 4.5.

4.2 Notations and Existing Estimators

The basic model for the cross-sectional sampling scenario described in the Introduction is as follows. We observe $(L, Z, \delta, T, \Delta)$ if and only if $L \leq T$ where $Z = \min(Z^0, C)$, $\delta = I(Z^0 \leq C)$, $T = \min(T^0, C)$, $\Delta = I(T^0 \leq C)$. Here, C is the potential right-censoring time and L is the left-truncation time (time from onset to cross-section). Therefore, with this notation, Z and T are the censored versions

of Z^0 and T^0, while δ and Δ are their corresponding censoring indicators. Since $P(Z^0 \leq T^0) = 1$, it happens that $\Delta = 1$ implies $\delta = 1$; in words, Z^0 is uncensored whenever T^0 is.

We assume that (L, C) and (Z^0, T^0) are independent, which is a standard requirement with left-truncated and right-censored data. This independence assumption just means that the cross-sectional sampling and the follow-up of the patients are unrelated to the disease under investigation. However, we allow C and L to be dependent. Note that, under cross-sectional sampling, a positive correlation between C and L is generally present, since L is a portion of the censoring time. Indeed, it is often the case (at least for a subpopulation) that $C = L + \tau$ for some constant τ which represents the maximum follow-up time after recruitment and, therefore, the pair (L, C) falls on a line with positive probability. We naturally assume $P(L \leq C) = 1$ too (that is, censoring may only occur after recruitment), from which the no-truncation event $\{L \leq T\}$ may be rewritten as $\{L \leq T^0\}$. We also assume $\alpha = P(L \leq T) > 0$. The standard model for random right-censorship is obtained as a particular case by taking $P(L = 0) = 1$.

We first introduce an estimator for $F_{Z^0T^0}(z, t) = P(Z^0 \leq z, T^0 \leq t)$. Let $(L_i, Z_i, \delta_i, T_i, \Delta_i)$, $1 \leq i \leq n$, be the available data; these are, iid copies with the distribution of $(L, Z, \delta, T, \Delta)$ conditionally on $L \leq T$. We will see that a natural estimator for $F_{Z^0T^0}(z, t)$ is the one that weights the indicator $I(Z_i \leq z, T_i \leq t)$ by the jump at time T_i of Tsai et al. (1987)'s product-limit estimator for the total survival $S_{T^0}(t) = P(T^0 \geq t)$.

Consider the joint subdistribution function of the observed (Z, T)'s with $\Delta = 1$, that is, $F^{1*}_{ZT}(z, t) = P(Z \leq z, T \leq t, \Delta = 1 | L \leq T)$. We have:

$$F^{1*}_{ZT}(z, t) = \int_0^t \alpha^{-1} P(L \leq v \leq C) F_{Z^0T^0}(z, dv)$$

where we have used the independence between (L, C) and (Z^0, T^0). Introduce $K_T(v) = P(L \leq v \leq T | L \leq T)$. It is easily seen that $K_T(v) = \alpha^{-1} P(L \leq v \leq C) S_{T^0}(v)$, from which $P(L \leq v \leq C) = \alpha K_T(v)/S_{T^0}(v)$ whenever $S_{T^0}(v) > 0$. Therefore,

$$F^{1*}_{ZT}(z, t) = \int_0^t \frac{K_T(v)}{S_{T^0}(v)} F_{Z^0T^0}(z, dv).$$

Provided that $K_T(v) > 0$ on the support of T^0 we thus have

$$F_{Z^0T^0}(z, t) = \int_0^t \frac{S_{T^0}(v)}{K_T(v)} F^{1*}_{ZT}(z, dv).$$

This equation suggests the estimator:

$$\hat{F}_{Z^0T^0}(z, t) = \int_0^t \frac{\hat{S}_{T^0}(v)}{\hat{K}_T(v)} \hat{F}^{1*}_{ZT}(z, dv) = \sum_{i=1}^n \frac{\hat{S}_{T^0}(T_i)\Delta_i}{n\hat{K}_T(T_i)} I(Z_i \leq z, T_i \leq t), \quad (4.1)$$

where $\hat{S}_{T^0}(t)$ and $\hat{K}_T(t)$ are, respectively, the product-limit estimator of $S_{T^0}(t)$ under left-truncation and right-censoring (Tsai et al. 1987), and the sampling proportion of data satisfying $L_i \leq t \leq T_i$, and where

$$\hat{F}^{1*}_{ZT}(z,t) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leq z, T_i \leq t, \Delta_i = 1).$$

Explicitly, and assuming no ties among the T_i's for simplicity,

$$\hat{S}_{T^0}(t) = \prod_{T_i<t}[1 - \frac{\Delta_i}{n\hat{K}_T(T_i)}], \qquad \hat{K}_T(t) = \frac{1}{n}\sum_{i=1}^{n} I(L_i \leq t \leq T_i). \tag{4.2}$$

It can be seen that, indeed, $W_i \equiv \hat{S}_{T^0}(T_i)\Delta_i/n\hat{K}_T(T_i) = -d\hat{S}_{T^0}(T_i)$. Thus, (4.1) weights the observed pairs (Z_i, T_i) through the jumps of the estimator for the marginal survival of T^0. The estimator (4.1) can be regarded as a special case of the product-limit integral in Sánchez-Sellero et al. (2005) for the indicator function $\varphi(u,v) = I(u \leq z, v \leq t)$, where Z^0 plays the role of the covariate in that paper. These authors investigated asymptotic properties of general product-limit integrals by assuming the independence between the truncation and censoring variables; however, this independence assumption can be removed (see e.g. de Uña-Álvarez and Veraverbeke 2017), which is very important in our setting, as discussed above.

The distribution of Z^0 can be estimated by taking the marginal of (4.1) corresponding to Z^0, $\hat{F}_{Z^0}(z) = \hat{F}_{Z^0T^0}(z,\infty)$. This product-limit integral-type estimator presents however two drawbacks. First, it gives no mass to the uncensored Z_i's with censored T_i, which may result in a loss of efficiency. Actually, an ideal estimator should reduce in the non-truncated setting to the standard Kaplan-Meier estimator, which is efficient, and this is not the case for $\hat{F}_{Z^0}(z)$. Second, and more important, consistency of the Z^0-marginal is ensured only when the censoring support contains the lifetime support, which is often unrealistic in practice. To be more specific, the estimator $\hat{F}_{Z^0}(z)$ converges in general to $F^{b_T}_{Z^0}(z) = P(Z^0 \leq z, T^0 \leq b_T)$, where b_T denotes the upper bound of the support of T and, consequently, $\hat{F}_{Z^0}(z)$ underestimates the target. In practice, this systematic bias is more visible at the right tail of the distribution, due to the positive correlation between Z^0 and T^0. Despite of this, $\hat{F}_{Z^0}(z)$ is recommended if there is no censoring; in such a setting, the aforementioned limitations vanish and, indeed, $\hat{F}_{Z^0}(z)$ can be introduced as a nonparametric maximum-likelihood estimator for $F_{Z^0}(z) = F_{Z^0T^0}(z,\infty)$ in that case.

4.3 A New Estimator

In this Section we introduce an estimator of the marginal distribution of Z^0 alternative to $\hat{F}_{Z^0}(z)$, which somehow generalizes the Kaplan-Meier estimator to our truncated

setting. Introduce $S_{Z^0}(z) = P(Z^0 \geq z) = 1 - F_{Z^0}(z^-)$. Note that the estimator of $S_{Z^0}(z)$ based on the approach in Sect. 4.2 is just $\hat{S}_{Z^0}(z) = 1 - \hat{F}_{Z^0}(z^-)$.

Recall $K_T(v) = P(L \leq v \leq T | L \leq T)$ and introduce the analogue for Z, namely $K_Z(v) = P(L \leq v \leq Z | L \leq T)$. Since $K_T(v) = \alpha^{-1} P(L \leq v \leq C) S_{T^0}(v)$ and $K_Z(v) = \alpha^{-1} P(L \leq v \leq C) S_{Z^0}(v)$, we have

$$S_{Z^0}(v) = S_{T^0}(v) \frac{K_Z(v)}{K_T(v)}.$$

This suggest the estimator

$$\hat{S}^*_{Z^0}(v) = \hat{S}_{T^0}(v) \frac{\hat{K}_Z(v)}{\hat{K}_T(v)} \tag{4.3}$$

where $\hat{S}_{T^0}(v)$ and $\hat{K}_T(v)$ are defined in (4.2), and where $\hat{K}_Z(v) = n^{-1} \sum_{i=1}^n I(L_i \leq v \leq Z_i)$. It is easily seen that (4.3) takes values in the [0, 1] interval. Note that the fraction in (4.3) is itself a sampling proportion, specifically the proportion of cases satisfying $v \leq Z_i$ among those with $L_i \leq v \leq T_i$; and, therefore, one has indeed $\hat{S}^*_{Z^0}(v) \leq \hat{S}_{T^0}(v)$. Also, (4.3) is consistent since it is a function of consistent estimators; see the Theorem below. In the case with no truncation, it is easily seen that the estimator (4.3) reduces to the ordinary Kaplan-Meier estimator of $S_{Z^0}(v)$ but for a factor which is the rate of two different estimators of the survival function of the censoring time $S_C(v) = P(C \geq v)$: the one based on the $(Z_i, 1 - \delta_i)$'s, and the one based on the $(T_i, 1 - \Delta_i)$'s. This suggests that $\hat{S}^*_{Z^0}(v)$ may be almost efficient at least when there is no truncation or when truncation is light.

Let $[a_\xi, b_\xi]$ denote the support of a given random variable ξ. Introduce $F^{1*}_T(t) = F^{1*}_{ZT}(\infty, t)$. We will refer to the following conditions:

C1. The random variables Z^0, T^0, L and C are continuous

C2. (L, C) and (Z^0, T^0) are independent

C3. $P(L \leq C) = 1$ and $\inf_{a_T \leq v \leq b} P(L \leq v \leq C) > 0$ for some $b < b_T$

Condition C1 avoids the discussion of possible ties among the observed data. Stute and Wang (2008) gives a proper treatment of ties under random left-truncation and similar arguments could be applied here. However, for simplicity of exposure we assume continuity throughout. Both the independence assumption C2 and condition $P(L \leq C) = 1$ appearing in C3 have been already discussed at the beginning of Sect. 4.2. Condition C3 was used in de Uña-Álvarez and Veraverbeke (2017) to deal with cross-sectional data; it implies $a_L \leq a_T$, which ensures that S_{T^0} can be identified on its whole support. Besides, under C3 we have $K_T(v) > 0$ for $v \in [a_T, b]$, which serves to control the denominator $\hat{K}_T(v)$ appearing in (4.3), as indicated in the proof below. On the other hand, C3 immediately gives

$$\int_{a_T}^b K_T(v)^{-3} dF^{1*}_T(v) < \infty$$

for some $b < b_T$. Under this integrability condition, Zhou and Yip (1999) derived an almost sure rate of convergence of $O(n^{-1} \log \log n)$ for the remainder in the

asymptotic representation for $\hat{S}_{T^0}$ as a sum of independent and identically distributed (iid) random variables, see their Theorems 2.1 and 2.2. This representation is used in the proof of our Theorem.

Theorem *Under C1-C3 we have, uniformly on* $a_T \leq z \leq b < b_T$,

$$\hat{S}^*_{Z^0}(z) - S_{Z^0}(z) = \frac{1}{n}\sum_{i=1}^n \psi_i(z) + R_n(z)$$

where the $\psi_i(z)$*'s are zero-mean iid random variables, and where* $\sup_{a_T \leq z \leq b} |R_n(z)| = O(n^{-1}\log\log n)$ *with probability 1.*

Proof By adding and substracting terms in an obvious manner, we obtain

$$\begin{aligned}\hat{S}^*_{Z^0}(z) - S_{Z^0}(z) \quad \sim \quad & \frac{K_Z(z)}{K_T(z)}\{\hat{S}_{T^0}(z) - S_{T^0}(z)\} \\ & + \frac{S_{T^0}(z)}{K_T(z)}\{\hat{K}_Z(z) - K_Z(z)\} \\ & + \frac{K_Z(z)S_{T^0}(z)}{K_T(z)^2}\{\hat{K}_T(z) - K_T(z)\}.\end{aligned}$$

We then apply Theorem 2.2 in Zhou and Yip (1999) and standard results for the sample means $\hat{K}_T(z)$ and $\hat{K}_Z(z)$ to get the iid representation and the order for the remainder. For this, note that condition C3 ensures that the denominator $K_T(z)$ remains bounded away from zero uniformly on the interval $[a_T, b]$. ■

Remark Explicitly, the iid representation $n^{-1}\sum_{i=1}^n \psi_i(z)$ in the Theorem is given by (see Theorem 2.2 in (Zhou and Yip, 1999))

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \psi_i(z) = & -S_{T^0}(z)\frac{K_Z(z)}{K_T(z)}\int_{a_T}^z \frac{d(\hat{F}_T^{1*} - F_T^{1*})(v)}{K_T(v)} \\ & + S_{T^0}(z)\frac{K_Z(z)}{K_T(z)}\int_{a_T}^z \frac{\hat{K}_T(v) - K_T(v)}{K_T(v)^2} dF_T^{1*}(v) \\ & + \frac{S_{T^0}(z)}{K_T(z)}\{\hat{K}_Z(z) - K_Z(z)\} \\ & + \frac{K_Z(z)S_{T^0}(z)}{K_T(z)^2}\{\hat{K}_T(z) - K_T(z)\}\end{aligned}$$

where $\hat{F}_T^{1*}(v) = \hat{F}_{ZT}^{1*}(\infty, v)$. That is, the $\psi_i(z)$'s are given by

$$\begin{aligned}\psi_i(z) = &-S_{T^0}(z)\frac{K_Z(z)}{K_T(z)}\left\{\frac{I(T_i \le z)\Delta_i}{K_T(T_i)} - \int_{a_T}^{z}\frac{dF_T^{1*}(v)}{K_T(v)}\right\}\\ &+S_{T^0}(z)\frac{K_Z(z)}{K_T(z)}\left\{\int_{\max(a_T, L_i)}^{\min(z, T_i)}\frac{dF_T^{1*}(v)}{K_T(v)^2} - \int_{a_T}^{z}\frac{dF_T^{1*}(v)}{K_T(v)}\right\}\\ &+\frac{S_{T^0}(z)}{K_T(z)}\{I(L_i \le z \le Z_i) - K_Z(z)\}\\ &+\frac{K_Z(z)S_{T^0}(z)}{K_T(z)^2}\{I(L_i \le z \le T_i) - K_T(z)\}.\end{aligned}$$

■

The estimator $\hat{S}_{Z^0}^*(z)$ is non-monotone along time z. A monotone estimator can be constructed from $\hat{S}_{Z^0}^*(z)$ by considering its monotonized version $\hat{S}_{Z^0}^{*,m}(z) = \inf_{v \le z}\hat{S}_{Z^0}^*(v)$. On the other hand, unlike for the product-limit type estimator $\hat{S}_{Z^0}(v)$, the jump points of (4.3) are not restricted to the Z_i's, but they involve the T_i's and the L_i's too. This may create some technical difficulties in the computation of the estimator for small sample sizes. For example, although $n\hat{K}_T(T_i) \ge 1$ is always guaranteed (thus allowing for the construction of (4.1)), $\hat{K}_T(v) = 0$ may happen for specific v's. In practice, we suggest to take the value $1/n$ for $\hat{K}_T(v)$ in (4.3) whenever this denominator becomes zero. On the other hand, when $n\hat{K}_T(T_i) = 1$, Tsai et al. (1987)'s product-limit estimator $\hat{S}_{T^0}(t)$ is zero for $t > T_i$, a situation which may occur even for small T_i's. This problem of possible 'holes' in the data is provoked by left-truncation, and has received some attention in the recent literature (Strzalkowska-Kominiak and Stute 2010). The issue of 'holes' affect both the product-limit integral type estimator and the new estimator introduced in this paper.

4.4 Simulation Study

In this Section we conduct a simulation study to investigate the finite sample performance of the several estimators reviewed along Sects. 4.2 and 4.3. To this end, we independently draw $Z^0 \sim Exp(3)$, $V^0 \sim Exp(1)$ and $\gamma \sim Ber(0.7)$, and then compute $T^0 = Z^0 + \gamma V^0$. This simulates an exponential event-free survival, with a 70% of the individuals undergoing the intermediate event, after which the residual lifetime is independently distributed according to another exponential distribution. We then draw an independent left-truncation time $L \sim U(0, b_L)$ with $b_L = 2$ (56% of truncated data), so we keep the datum (L, Z^0, T^0) only when $L \le T^0$. The uniform distribution for the left-truncation time is often used to simulate processes in the steady (or stationary) state, under which the incidence rate is constant (see Asgharian

and Wolfson 2005; Fluss et al. 2013). In this scenario, the probability of sampling a lifetime $T^0 = t$ is proportional to its length t (the so-called length-biased sampling).

For the censoring time we draw $\nu \sim Ber(0.5)$ and we compute $C = \nu C^{adm} + (1-\nu)C^{rdm}$, where $C^{adm} = L + \tau$ is an administrative potential censoring time due to the end of study, which occurs τ time units after recruitment; and where $C^{rdm} = L + U(0, \tau)$ is a random potential censoring time representing lost to follow-up issues. Note that, in our simulated setting, 50% of the censoring times correspond to administrative censoring. For τ we take values $\{0.5, \infty\}$ to introduce both the censored and the uncensored cases. In the censored case ($\tau = 0.5$), the censoring rates for T^0 and Z^0 are 68% and 14% respectively. The observable variables are finally computed as $Z = \min(Z^0, C)$, $\delta = I(Z^0 \leq C)$, $T = \min(T^0, C)$, and $\Delta = I(T^0 \leq C)$.

For sample sizes $n = 250, 500$ (after truncation), we generate 1,000 Monte Carlo trials and we compute the bias, the standard deviation (SD), and the mean squared error (MSE) of the estimators for the event-free survival $S_{Z^0}(z)$ at the quartiles of the $Exp(3)$ distribution, namely $z = 0.0959, 0.2310, 0.4621$ (Q_1, Q_2 and Q_3 in Tables). The results are given in Tables 4.1 and 4.2.

From Table 4.1, corresponding to the uncensored case, we see that the product-limit integral-type estimator (PLI) $\hat{S}_{Z^0}$ has a smaller MSE compared to that of the new estimator $\hat{S}^*_{Z^0}$, basically due to the smaller standard deviation of the former. The differences are more clear at the left tail of Z^0 and small sample size. Thus, the new estimator is not optimal in this case. The MSE of both estimators decrease with an increasing sample size.

The situation changes in the censored case (Table 4.2). From Table 4.2 it is seen that the PLI type estimator has a systematic bias which does not decrease as n increases. This bias is larger at the right tail of Z^0, where $1 - F^{b_T}_{Z^0}(z)$ (the truncated limit of $\hat{S}_{Z^0}(z)$, see Sect. 4.2) deviates more from the target $S_{Z^0}(z)$. As a consequence, the MSE of the PLI estimator is much larger than that of the new estimator at the right tail (between 2.5 and 3.7 times the MSE of the new estimator). Still, $\hat{S}_{Z^0}$ is more accurate than $\hat{S}^*_{Z^0}$ at the left tail, because of the relatively small impact of the bias. The new estimator performs well at the three quartiles of Z^0, the bias being of a smaller order of magnitude compared to the standard deviation in all the cases.

To complete our study, in Table 4.3 we consider the censored case ($\tau = 0.5$) but with no truncation ($b_L = 0$). This is interesting to study the behavior of the PLI type and new estimators relative to that of the standard Kaplan-Meier estimator, which is efficient in that situation. The censoring rates on T^0 and Z^0 are 72% and 39% respectively in this case. From Table 4.3 we see that the new estimator is competitive, exhibiting a MSE which is never above 1.2 times that of the Kaplan-Meier estimator. The results of the PLI type estimator are disappointing, with a MSE which is between 16 and 213 times that of the new estimator.

Table 4.1 Bias, standard deviation (SD), and mean squared error (MSE) of $\hat{S}_{Z^0}$ (PLI) and $\hat{S}^*_{Z^0}$ (New) at the quartiles of Z^0 along 1,000 Monte Carlo trials, with $\tau = \infty$ (no censoring) and $b_L = 2$

		$n = 250$		$n = 500$	
		PLI	New	PLI	New
Q_1	Bias	0.0004	0.0053	0.0068	0.0077
	SD	0.1041	0.1236	0.0735	0.0843
	MSE	0.0108	0.0153	0.0055	0.0072
Q_2	Bias	−0.0006	−0.0012	0.0037	0.0043
	SD	0.0790	0.0925	0.0558	0.0648
	MSE	0.0062	0.0085	0.0031	0.0042
Q_3	Bias	−0.0005	−0.0003	0.0026	0.0028
	SD	0.0453	0.0515	0.0316	0.0362
	MSE	0.0021	0.0026	0.0010	0.0013

Table 4.2 Bias, standard deviation (SD), and mean squared error (MSE) of $\hat{S}_{Z^0}$ (PLI) and $\hat{S}^*_{Z^0}$ (New) at the quartiles of Z^0 along 1,000 Monte Carlo trials, with $\tau = 0.5$ (censored case) and $b_L = 2$

		$n = 250$		$n = 500$	
		PLI	New	PLI	New
Q_1	Bias	0.0164	0.0045	0.0236	0.0075
	SD	0.1110	0.1256	0.0783	0.0851
	MSE	0.0126	0.0158	0.0067	0.0073
Q_2	Bias	0.0339	−0.0010	0.0382	0.0040
	SD	0.0942	0.0950	0.0668	0.0662
	MSE	0.0100	0.0090	0.0059	0.0044
Q_3	Bias	0.0559	0.0002	0.0563	0.0025
	SD	0.0693	0.0561	0.0489	0.0390
	MSE	0.0079	0.0032	0.0056	0.0015

4.5 Main Conclusions

In this paper the problem of estimating an event-free survival function from cross-sectional data has been studied. Two different nonparametric estimators have been considered. The first one is a special case of the product-limit integrals for left-truncated and right-censored data investigated in Sánchez-Sellero et al. (2005). The second estimator is new (for the best of our knowledge), and it is simply defined as a time-varying portion of Tsai et al. (1987)'s product-limit estimator; asymptotic results for this new estimator can be established in a straightforward way. While the PLI type estimator is, in general, systematically biased, the new estimator is

Table 4.3 Bias, standard deviation (SD), and mean squared error (MSE) of $\hat{S}_{Z^0}$ (PLI), $\hat{S}^*_{Z^0}$ (New), and Kaplan-Meier estimator (KME) at the quartiles of Z^0 along 1,000 Monte Carlo trials, with $\tau = 0.5$ and $b_L = 0$ (no truncation)

		$n = 250$			$n = 500$		
		PLI	New	KME	PLI	New	KME
Q_1	Bias	0.1102	0.0014	−0.0007	0.1109	0.0003	−0.0008
	SD	0.0242	0.0287	0.0283	0.0171	0.0207	0.0203
	MSE	0.0127	0.0008	0.0008	0.0126	0.0004	0.0004
Q_2	Bias	0.2342	0.0007	−0.0012	0.2353	0.0010	−0.0003
	SD	0.0319	0.0345	0.0336	0.0212	0.0254	0.0240
	MSE	0.0559	0.0012	0.0011	0.0558	0.0006	0.0006
Q_3	Bias	0.3822	0.0009	−0.0002	0.3851	0.0018	0.0010
	SD	0.0364	0.0361	0.0325	0.0246	0.0267	0.0238
	MSE	0.1474	0.0013	0.0011	0.1489	0.0007	0.0006

consistent. This has been investigated both theoretically and through simulations. The simulation study conducted in this paper suggests that the PLI type estimator may be recommended in the special case of no censoring but that, in the censored case, it should not be used. The simulations suggest that there is some ground for improvements of the new estimator particularly at the left tail of the event-free survival time, where it can be beated by the PLI type estimator when there is no censoring, or when the censoring is light.

The cross-sectional sampling scenario considered in this paper results in left-truncation on the total survival time T^0. Different situations when sampling prevalent cases are possible. For example, Chang and Tzeng (2006) considered an alternative sampling procedure in which the recruited individuals are those with event-free survival time Z^0 larger than the left-truncation time. In this alternative sampling scheme S_{Z^0} can be estimated through the standard Tsai et al. (1987)'s product-limit estimator applied to the event-free survival times; however, for the estimation of S_{T^0}, specific estimators must be derived.

In some applications with cross-sectional data, the assumption of uniformly distributed left-truncation times is plausible. This is the case under the so-called steady state (e.g. Fluss et al. 2013). When information on the truncation distribution is available, improved estimators of the event-free survival can be constructed. For example, with uniform truncation, Tsai et al. (1987)'s product-limit estimator $\hat{S}_{T^0}(v)$ in (4.3) can be replaced by the nonparametric maximum-likelihood estimator in Asgharian and Wolfson (2005) to obtain a better estimator for the event-free survival. Properties of such alternative estimator are still unexplored.

Acknowledgements The author thanks a referee for helpful comments. Work supported by the Grant MTM2014-55966-P of the Spanish Ministerio de Economía y Competitividad.

References

Asgharian M, Wolfson DB (2005) Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from censored prevalent cohort data. Annals of Statistics 33, 2109–2131.

Chang SH, Tzeng SJ (2006) Nonparametric estimation of sojourn time distributions for truncated serial event data a weight-adjusted approach. Lifetime Data Analysis 12, 53–67.

de Uña-Álvarez J, Veraverbeke N (2017) Copula-graphic estimation with left-truncated and right-censored data. Statistics 51, 387–403.

Fluss R, Mandel M, Freedman LS, Weiss IS, Zohar AE, Haklai Z, Gordond ES, Simchena E (2013) Correction of sampling bias in a cross-sectional study of post-surgical complications. Statistics in Medicine, 32, 2467–2478.

Gijbels I, Wang JL (1993) Strong representations of the survival function estimator for truncated and censored data with applications. Journal of Multivariate Analysis 47, 210–229.

Sánchez-Sellero C, González-Manteiga W, Van Keilegom I (2005) Uniform representation of product-limit integrals with applications. Scandinavian Journal of Statistics 32, 563–581.

Strzalkowska-Kominiak E, Stute W (2010) On the probability of holes in truncated samples. Journal of Statistical Planning and Inference 140, 1519–1528.

Stute W (1993) Almost sure representations of the product-limit estimator for truncated data. Annals of Statistics 21, 146–156.

Stute W, Wang JL (2008) The central limit theorem under random truncation. Bernoulli 14, 604–622.

Tsai WY, Jewell NP, Wang MC (1987) A note on the product-limit estimator under right censoring and left truncation. Biometrika 74, 883–886.

Wang MC (1991) Nonparametric estimation from cross-sectional survival data. Journal of the American Statistical Association 86 130–143.

Zhou Y, Yip PSF (1999) A strong representation of the product-limit estimator for left truncated and right censored data. Journal of Multivariate Analysis 69, 261–280.

Part II
Model Checks

On the Asymptotic Efficiency of Directional Models Checks for Regression

5

Miguel A. Delgado and Juan Carlos Escanciano

JEL classification: C12 · C14

5.1 Introduction

Let (Y, X) be a bivariate random vector with probability measure $\mathbb{P}$ and regression function $r(x) := \mathbb{E}(Y|X = x)$. In a landmark paper, Stute (1997) introduced omnibus, smooth and directional tests of the null hypothesis

$$H_0 : r \in \mathcal{M}_0, \tag{5.1}$$

where $\mathcal{M}_0$ is a family of regression functions in $\mathbb{R}$ linear in parameters, i.e.

$$\mathcal{M}_0 = \left\{ m : m(x) = \boldsymbol{\beta}^T \mathbf{g}(x) : \boldsymbol{\beta} \in \mathbb{R}^k \right\},$$

for a known k-dimensional vector of measurable functions $\mathbf{g} : \mathbb{R} \to \mathbb{R}^k$, where henceforth $\mathbf{a}^T$ denotes the transpose of the vector $\mathbf{a}$. The discussion in what follows is also valid for models that are non-linear in parameters and satisfy standard regularity conditions.

Research funded by the Spanish Plan Nacional de I+D+I, reference number ECO2014-55858-P. This article is dedicated to Winfried Stute in his 70th birthday.

M.A. Delgado (✉)
Universidad Carlos III de Madrid, Madrid, Spain
e-mail: miguelangel.delgado@uc3m.es

J.C. Escanciano
Indiana University, Bloomington, USA

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_5

Stute's (1997) directional tests specify a local alternative of the form

$$H_{1n} : r \in \mathcal{M}_{1n}, \tag{5.2}$$

where

$$\mathcal{M}_{1n} = \left\{ m : m(x) = \beta^T \mathbf{g}(x) + \frac{d(x)}{\sqrt{n}} : \beta \in \mathbb{R}^k \right\},$$

and d is a known measurable function, indicating the direction of departure from $\mathcal{M}_0$.

The main contribution of this article is to show that Stute's (1997) directional test is efficient in a semiparametric sense. We formalize the directional testing problem as a one-sided parametric testing problem within a semiparametric model. Asymptotically Uniformly Most Powerful (AUMP) tests in this context have been defined in Choi et al. (1996, Sect. 3, Theorem 1) as tests that are asymptotically equivalent to the canonical efficient score, suitably standardized. The main result of this article shows that, under conditional homoskedasticity, Stute's (1997) directional test is AUMP. We also show that Stute's directional test is asymptotically equivalent to a standard $t-$ratio test under homoskedasticity. We study the heteroskedastic case, and show that the directional functional likelihood ratio test based on the CUSUM of (conditionally) standarized residuals (Stute et al. 1998) is AUMP, and is asymptotically equivalent to the t-ratio using the generalized least squares estimator. In summary, we show that Stute's (1997) directional tests, which were motivated from the functional likelihood ratio approach of Grenander (1950), are also asymptotically efficient in a semiparametric sense.

The rest of the article is organized as follows. Section 5.2 introduces Stute's (1997) directional test as a functional likelihood ratio test based on the CUSUM of residuals. Section 5.3 formalizes this testing problem as a parametric testing problem within a semiparametric model, and discusses AUMP tests. Section 5.4 contains the main results of the article, which include proving the efficiency of Stute's directional test and its relation with the more standard t–ratio tests.

5.2 Stute's Directional Test

Assume that Y is square integrable and that the random vector $\mathbf{g}(X)$ is linearly independent, in the sense that $\mathbb{E}\left[\mathbf{g}(X)\mathbf{g}(X)^T\right]$ is non-singular. Then, H_0 is satisfied iff $r = \beta_0^T \mathbf{g}$, where $\beta_0^T \mathbf{g}(X)$ is the best linear predictor of Y given $\mathbf{g}(X)$, i.e. under H_0

$$\beta_0 = \mathbb{E}\left[\mathbf{g}(X)\mathbf{g}(X)^T\right]^{-1} \mathbb{E}\left[\mathbf{g}(X)Y\right]. \tag{5.3}$$

Stute (1997) characterized H_0 as

$$\int_{\{X \leq x\}} Y d\mathbb{P} = \int_{\{X \leq x\}} \boldsymbol{\beta}_0^T \mathbf{g}(X) d\mathbb{P} \text{ a.s.} \tag{5.4}$$

Define

$$R_0(x) := \frac{1}{\sigma} \int_{\{X \leq x\}} \left(Y - \boldsymbol{\beta}_0^T \mathbf{g}(X)\right) d\mathbb{P} = \frac{1}{\sigma} \mathbb{E}\left[\left(Y - \boldsymbol{\beta}_0^T \mathbf{g}(X)\right) 1_{\{X \leq x\}}\right], \tag{5.5}$$

where $1_{\{A\}}$ denotes the indicator function of the event A and $\sigma^2 := \mathbb{E}\left(Y - \boldsymbol{\beta}_0^T \mathbf{g}(X)\right)^2$. In view of (5.4), for a suitable norm $\|\cdot\|$, H_0 can be expressed as

$$H_0 : \|R_0\| = 0. \tag{5.6}$$

Omnibus tests are consistent in the direction of any nonparametric alternative such that $\|R_0\| > 0$.

Given a random sample of (Y, X) of size n, $\{Y_i, X_i\}_{i=1}^n$, a scale invariant sample analog of $R_0(x)$ is

$$R_n(x) := \frac{1}{n} \sum_{i=1}^n \frac{\varepsilon_{ni}}{\widehat{\sigma}} 1_{\{X_i \leq x\}},$$

where $\varepsilon_{in} := Y_i - \boldsymbol{\beta}_n^T \mathbf{g}(X_i)$ are residuals from the ordinary least squares (OLS) estimator, i.e.

$$\boldsymbol{\beta}_n = \left[\sum_{i=1}^n \mathbf{g}(X_i)\mathbf{g}(X_i)^T\right]^{-1} \sum_{i=1}^n \mathbf{g}(X_i) Y_i$$

and $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n \varepsilon_{in}^2$ estimates σ^2.

Stute (1997) shows that, under H_0 and uniformly in $x \in \mathbb{R}$,

$$R_n = R_n^1 + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{5.7}$$

with

$$R_n^1(x) = \frac{1}{n} \sum_{i=1}^n \frac{\varepsilon_i}{\sigma} w(X_i, x),$$

$\varepsilon_i = Y_i - \boldsymbol{\beta}_0^T \mathbf{g}(X_i)$, and

$$w(z, x) = 1_{\{z \leq x\}} - \mathbb{E}\left[\mathbf{g}(X)^T 1_{\{X \leq x\}}\right] \left(\mathbb{E}\left[\mathbf{g}(X)\mathbf{g}^T(X)\right]\right)^{-1} \mathbf{g}(z).$$

Consider $\eta_n(x) = \sqrt{n}R_n(x)$ as a random element of the infinite-dimensional Hilbert space $\mathcal{L}_F^2$ of measurable real-valued functions on $\mathbb{R}$ that are square integrable with respect to F, the cumulative distribution function of X, with corresponding norm

$$\|h\|_{\mathcal{L}_F^2} = \sqrt{\langle h, h\rangle},$$

and inner product $\langle h_1, h_2\rangle = \mathbb{E}\left[h_1(X)h_2(X)\right]$. Using (5.7), and applying a central limit theorem (CLT) for Hilbert spaces (e.g. Kundu et al. 2000), under H_0

$$\eta_n \to_d \eta_\infty,$$

where η_∞ is a Gaussian element of $\mathcal{L}_F^2$ with mean zero and covariance kernel

$$C(x_1, x_2) := \mathbb{E}\left[\eta_\infty(x_1)\eta_\infty(x_2)\right] = \mathbb{E}\left[\frac{\varepsilon^2 w(X, x_1)w(X, x_2)}{\sigma^2}\right],$$

where $\varepsilon = Y - \beta_0^T \mathbf{g}(X)$. Henceforth, "$\to_d$" means convergence in distributions of sequences of random variables, random vectors, or random elements in $\mathcal{L}_F^2$, and "$\to_p$" means convergence in probability. Therefore, by the continuous mapping theorem, under H_0, $\|\eta_n\|^2_{\mathcal{L}_F^2} \to_d \|\eta_\infty\|^2_{\mathcal{L}_F^2}$. The distribution of $\|\eta_\infty\|^2_{\mathcal{L}_F^2}$ is unknown, but critical values of the test $\Psi_n(c) = 1_{\left\{\|\eta_n\|^2_{\mathcal{L}_F^2} > c\right\}}$ can be estimated using bootstrap (see Stute et al. 1998).

Under H_{1n},

$$\eta_n \to_d \eta_\infty + \delta,$$

where

$$\delta(x) := \mathbb{E}\left[d(X)w(X, x)\right]. \tag{5.8}$$

Hence, by the continuous mapping theorem $\|\eta_n\|^2_{\mathcal{L}_F^2} \to_d \|\eta_\infty + \delta\|^2_{\mathcal{L}_F^2}$, respectively. Therefore, $\Psi_n(c)$ is able to detect local alternatives converging at the parametric rate.

Introduce the Fredholm integral operator $\mathcal{K}$, where for any generic function $h \in \mathcal{L}_F^2$,

$$\mathcal{K}h(\cdot) = \int h(x)C(x, \cdot)F(dx).$$

Henceforth, we drop the region of integration for simplicity of notation. Since $\mathcal{K}$ is a compact, linear and positive operator, it has countable spectrum $\{\lambda_j, \varphi_j\}_{j=1}^\infty$, where $\{\lambda_j\}_{j=1}^\infty$ are real-valued, positive, with $\lambda_j \downarrow 0$, and $\{\varphi_j\}_{j=1}^\infty$ are such that $\mathcal{K}\varphi_j = \lambda_j\varphi_j$, for all $j \in \mathbb{N}$. Let $\xi_j := \lambda_j^{-1/2}\langle\eta_\infty, \varphi_j\rangle$, $j \in \mathbb{N}$, be the so-called principal components of η_∞.

Hence, $\{\xi_j\}_{j\geq 0}$ are iid as standard normals and $\eta_\infty \stackrel{d}{=} \sum_{j=1}^{\infty} \xi_j \lambda_j^{1/2} \varphi_j$, where "$\stackrel{d}{=}$" means equality in distribution. Then, after choosing some $m > 0$, the statistic

$$N_{n,m} = \sum_{j=1}^{m} \xi_{nj}^2,$$

with $\xi_{nj} := \lambda_j^{-1/2} \langle \eta_n, \varphi_j \rangle$ forms a basis for a Neyman-type smooth test.

Denote by $\mathbb{Q}$ the limiting probability measure of η_n, and $\mathbb{Q}_0$ and $\mathbb{Q}_1$ the corresponding probability measures under H_0 and H_1, respectively; i.e. the distributions of η_∞ and $\eta_\infty + \delta$. Assume that

$$\sum_{j=0}^{\infty} \frac{\langle \delta, \varphi_j \rangle^2}{\lambda_j} < \infty, \tag{5.9}$$

so that $\mathbb{Q}_1$ is absolutely continuous with respect to $\mathbb{Q}_0$, as shown by Grenander (1950) and Skorohod (1974, Chap. 16, Theorem 2), with Radom-Nikodyn derivative

$$h \in \mathcal{L}_F^2 \longmapsto \frac{d\mathbb{Q}_1}{d\mathbb{Q}_0}(h) = \exp\left\{ - \int A(x) \left[h(x) - \frac{\delta(x)}{2} \right] F(dx) \right\}, \tag{5.10}$$

with

$$A(x) = - \sum_{j=0}^{\infty} \frac{\langle \delta, \varphi_j \rangle}{\sqrt{\lambda_j}} \varphi_j.$$

Then, each test of H_0 in the direction H_{1n}, which is based on η_n, asymptotically becomes one of testing the simple hypothesis $\tilde{H}_0 : \mathbb{Q} = \mathbb{Q}_0$ versus $\tilde{H}_1 : \mathbb{Q} = \mathbb{Q}_1$ in the exponential model (5.10) with $\mathbb{Q}$ the limiting distribution of η_n. By the Neyman-Pearson Lemma, the optimal test rejects $\tilde{H}_0$ in favor of $\tilde{H}_1$ if and only if, for suitable critical value c_1,

$$- \int A(x) \left[\eta_\infty(x) - \frac{\delta(x)}{2} \right] F(dx) \geq c_1,$$

or equivalently, for a suitable critical value c_2,

$$T_\infty = \sum_{j=1}^{\infty} \frac{\langle \delta, \varphi_j \rangle \langle \eta_\infty, \varphi_j \rangle}{\lambda_j} \geq c_2.$$

Notice that T_∞ is a normal *r.v.* with $\mathbb{E}(T_\infty) = 0$, under H_0, and $Var(T_\infty) = \sum_{j=0}^{\infty} \langle \delta, \varphi_j \rangle^2 / \lambda_j$, which can be estimated from data.

Stute (1997, page 633) suggested a directional test, rejecting H_0 against H_{1n} for large values of

$$T_{n,m_n} = \sum_{j=1}^{m_n} \frac{\langle \delta, \varphi_j \rangle \langle \eta_n, \varphi_j \rangle}{\lambda_j}, \tag{5.11}$$

where m_n is a tuning parameter that goes to infinity as n goes to infinity.

The directional test proposed by Stute (1997) was later extended by Stute et al. (1998), Boning and Sowell (1999), Bischoff and Miller (2000) and Escanciano (2009) to other regression settings. Applications to conditional distributions were given in Delgado and Stute (2008), and to tests for correct specification of the covariance structure of a linear process in Delgado et al. (2005). The optimality of directional tests in a very general framework of semiparameric moment restrictions, which includes the one from this article, have been previously studied in Escanciano (2012).

5.3 Efficient Semiparametric Test

This section formalizes the directional testing problem as a parametric test within a semiparametric model. Then, it discusses the asymptotically efficient test for the semiparametric problem.

Let $\mathbb{P}$ be absolutely continuous with respect to a σ-finite measure μ. Consider the following nonparametric family of error's probability density functions (p.d.f.) with respect to μ,

$$\mathcal{F} = \left\{ f : f \geq 0, \int f d\mu = 1, \int e f(e, X) \mu(de, X) = 0 \text{ a.s.} \right\}.$$

Define $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \gamma)$ and $\mathbf{h}^T = (\mathbf{g}^T, d)$, and assume $\mathbb{E}\left[\mathbf{h}(X)\mathbf{h}(X)^T\right]$ is non-singular. Define the semiparametric class of models

$$\mathcal{P}_1 := \left\{ \mathbb{P}_{\boldsymbol{\theta}, f} : \frac{d\mathbb{P}_{\boldsymbol{\theta}, f}}{d\mu}(y, x) = f\left(y - \boldsymbol{\theta}^T \mathbf{h}(x), x\right) : \boldsymbol{\theta} \in \mathbb{R}^{k+1}, f \in \mathcal{F} \right\}.$$

Henceforth, for a generic parameter, the subscript zero denotes that the parameter is evaluated under the distribution that generates the data, i.e. $\mathbb{P}$. Define $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \gamma_0)^T = \mathbb{E}\left[\mathbf{h}(X)\mathbf{h}(X)^T\right]^{-1} \mathbb{E}\left[\mathbf{h}(X)Y\right]$. That is, $\boldsymbol{\beta}_0$ corresponds to (5.3) when $\gamma_0 = 0$.

Then, Stute's (1997) directional test is simply a parametric test for

$$H_0 : \gamma_0 = 0 \text{ vs } H_1 : \gamma_0 > 0, \tag{5.12}$$

within the semiparametric class of models defined by $\mathcal{P}_1$ (i.e. with the maintained hypothesis that $\mathbb{P} \in \mathcal{P}_1$). Although we term directional tests as parametric, we note

that they involve unknown infinite-dimensional nuisance parameters under both the null and alternative hypotheses, namely $\upsilon := (\boldsymbol{\beta}, f) \in \mathbb{R}^k \times \mathcal{F}$.

The concept of optimality that we use is well explained in Choi et al. (1996, Sect. 3). These authors show that a test $\Upsilon_n(c) = 1_{\{T_n > c\}}$ of asymptotic level α is *asymptotically uniformly most powerful* for testing (5.12), in short AUMP(α), if for every $\upsilon_0 := (\boldsymbol{\beta}_0, f_0) \in \mathbb{R}^k \times \mathcal{F})$ under H_0,

$$T_n = \zeta_{n\gamma} + o_{\mathbb{P}}(1),$$

where $\zeta_{n\gamma}$ is the standardized canonical effective score test statistic

$$\zeta_{n\gamma} := \frac{1}{\sqrt{n\sigma_\gamma^2}} \sum_{i=1}^{n} \dot{\ell}_\gamma^*(Z_i),$$

where $\dot{\ell}_\gamma^*$ is the efficient score defined below and $\sigma_\gamma^2 := Var(\dot{\ell}_\gamma^*)$ is the efficient information. Define the marginal class of semiparametric models with γ fixed at γ_0 by $\mathcal{P}_{\gamma_0} := \{\mathbb{P}_{(\gamma_0,\upsilon)} : \upsilon = (\boldsymbol{\beta}, f) \in \mathbb{R}^k \times \mathcal{F}\}$, and let $\dot{\mathcal{P}}_{\gamma_0}$ be the tangent space of $\mathcal{P}_{\gamma_0}$ at $\mathbb{P}_{(\gamma_0,\upsilon_0)}$, i.e. the closed linear span of scores (derivatives of log-likelihood ratios in many cases) passing through the semiparametric model $\mathbb{P}_{(\gamma_0,\upsilon_0)}$. Given the score $\dot{\ell}_\gamma$ in the marginal family $\mathcal{P}_{\upsilon_0} = \{\mathbb{P}_{(\gamma,\upsilon_0)} : \gamma \in \mathbb{R}\}$, we define the efficient score $\dot{\ell}_\gamma^*$ as the orthogonal mean square projection of the score $\dot{\ell}_\gamma$ onto the orthocomplement of $\dot{\mathcal{P}}_2$. We show below that for our semiparametric testing problem the efficient score is

$$\dot{\ell}_\gamma^*(Z) = \varepsilon \frac{\left(d(X) - \boldsymbol{\pi}_0^T \mathbf{g}(X)\right)}{\tau^2(X)}, \tag{5.13}$$

where $\boldsymbol{\pi}_0 = \left(\mathbb{E}\left[\tau^{-2}(X)\,\mathbf{g}(X)\mathbf{g}(X)\right]\right)^{-1} \mathbb{E}\left[\tau^{-2}(X)\,\mathbf{g}(X)d(X)\right]$, with $\tau^2(X) := \mathbb{E}\left(\varepsilon^2 \middle| X\right)$, and corresponding efficient information

$$\sigma_\gamma^2 = \mathbb{E}\left[\frac{\left(d(X) - \boldsymbol{\pi}_0^T \mathbf{g}(X)\right)^2}{\tau^2(X)}\right]. \tag{5.14}$$

The presence of heterokedasticity, i.e. $\tau^2(X) \neq \sigma^2$ with positive probability, introduces an infinite-dimensional nuisance parameter in the efficient score, which substantially complicates the implementation of efficient inference, see Robinson (1987).

Efficient scores and efficient informations have been extensively discussed for regression problems; see e.g. Chamberlain (1987), Newey (1990) and Bickel et al. (1993), among many others. We compute $\dot{\ell}_\gamma^*$ for our semiparametric regression testing problem using parametric submodels. A parametric submodel in our problem has a parametric density

$$f_\rho\left(y - \boldsymbol{\beta}^T \mathbf{g}(x) - \gamma d(x), x\right) \tag{5.15}$$

depending on finite-dimensional parameters $(\boldsymbol{\beta}, \gamma, \rho)$, where ρ is a scalar parameter in a neighborhood of zero and $f_\rho \in \mathcal{F}$. The parametric submodel is regular, in the sense of satisfying a classical mean square differentiability property, and passes through the truth, meaning that $f_0\left(y - \boldsymbol{\beta}_0^T \mathbf{g}(x) - \gamma_0 d(x), x\right)$ is the true density that generated the data. An example of f_ρ is

$$f_\rho(e, x) = f_0(e, x)(1 + \rho a(e, x)),$$

where $a(e, x)$ is a (bounded) measurable square integrable function satisfying

$$\int a(e, x) f_0(e, x)\, d\mu(e, x) = 0.$$

Compute the scores for $(\boldsymbol{\beta}, \gamma, \rho)$ in (5.15) under $H_0 : \gamma_0 = 0$ as

$$\begin{aligned}
\dot{\ell}_{\boldsymbol{\beta}}(z) &= \frac{f_0^{(1)}\left(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x\right)}{f_0\left(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x\right)} \mathbf{g}(x) \equiv b_\beta(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x)\mathbf{g}(x), \\
\dot{\ell}_\gamma(z) &= \frac{f_0^{(1)}\left(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x\right)}{f_0\left(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x\right)} d(x) \equiv b_\gamma(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x) d(x), \\
\dot{\ell}_{\boldsymbol{\rho}}(z) &= \left.\frac{\partial \log f_\rho\left(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x\right)}{\partial \rho}\right|_{\rho=0} \equiv a(y - \boldsymbol{\beta}_0^T \mathbf{g}(x), x),
\end{aligned}$$

and $f_0^{(1)}(e, \cdot) = \partial f_0(e, \cdot)/\partial e$. The condition $f_\rho \in \mathcal{F}$ implies by differentiation

$$\int e a(e, X) f_0(e, X) \mu(de, X) = 0 \text{ a.s.}$$

This means that the set of scores for the infinite-dimensional parameter includes zero mean and square integrable $a(e, x)$ such that

$$\mathbb{E}[\varepsilon a(\varepsilon, X) | X] = 0 \text{ a.s.}$$

The set of orthogonal functions to such $a's$ necessarily are functions of the form $\varepsilon s(X)$ for a measurable function s. Thus, the projection of $\dot{\ell}_\gamma(Z)$ onto such set is some $\varepsilon s_\gamma(X)$, for a function s_γ such that for all measurable functions s

$$\mathbb{E}\left[b_\gamma(\varepsilon, X) d(X) \varepsilon s(X)\right] = \mathbb{E}\left[\varepsilon s_\gamma(X) \varepsilon s(X)\right].$$

Solving for s_γ in this equation, we find

$$s_\gamma(x) = \frac{\mathbb{E}\left[\varepsilon b_\gamma(\varepsilon, X) \middle| X = x\right] d(x)}{\tau^2(x)} = \frac{d(x)}{\tau^2(x)},$$

where we have used that

$$\mathbb{E}\left[\varepsilon b_\gamma(\varepsilon, X)\middle| X = x\right] = \frac{\int e f_0^{(1)}(e, x)\,\mu(de, x)}{\int f_0(e, x)\,\mu(de, x)} = 1.$$

It remains to project $\varepsilon s_\gamma(X)$ onto the orthocomplement of the space generated by $\dot{\ell}_\beta$. Similarly to what we did for γ, the orthogonal projection of $\dot{\ell}_\beta$ onto the space orthogonal to scores of the infinite-dimensional parameter is

$$\varepsilon s_\beta(X) = \varepsilon \frac{\mathbf{g}(X)}{\tau^2(X)}.$$

Then, by standard least squares theory

$$\begin{aligned}\dot{\ell}^*_\gamma(Z) &= \varepsilon s_\gamma(X) - \mathbb{E}\left[\varepsilon s_\gamma(X)\varepsilon s_\beta(X)\right]\left(\mathbb{E}\left[\varepsilon s_\beta(X)\varepsilon s_\beta(X)\right]\right)^{-1}\varepsilon s_\beta(X)\\ &= \varepsilon\frac{\left(d(X) - \boldsymbol{\pi}_0^T\mathbf{g}(X)\right)}{\tau^2(X)}.\end{aligned}$$

Then, a test $\Upsilon_n(c) = 1_{\{T_n > c\}}$ is AUMP(α), if it is asymptotically of level α and under H_0,

$$T_n = \zeta_{n\gamma} + o_{\mathbb{P}}(1),$$

where

$$\zeta_{n\gamma} := \frac{1}{\sqrt{n\sigma_\gamma^2}}\sum_{i=1}^{n}\left(\varepsilon_i\frac{\left(d(X_i) - \boldsymbol{\pi}_0^T\mathbf{g}(X_i)\right)}{\tau^2(X_i)}\right). \tag{5.16}$$

5.4 Efficiency of Stute's Directional Test

5.4.1 The Homoskedastic Case

This Section shows that Stute's directional test based on T_{n,m_n} is AUMP(α) under homoskedasticity. We relate Stute's directional test with the standard t-test, and we extend these results to conditional heteroskedasticity of unknown form.

The first step in this analysis consists of deriving the asymptotic equivalence of T_{n,m_n} and its asymptotic approximation T^1_{n,m_n}, defined as (cf. 5.7)

$$T^1_{n,m_n} := \sum_{j=1}^{m_n}\frac{\langle\delta, \varphi_j\rangle\langle\sqrt{n}R^1_n, \varphi_j\rangle}{\lambda_j}.$$

To that end, we require the following mild condition on m_n.

Assumption 1 Assume that $m_n \to \infty$ as $n \to \infty$ and either (i) (5.9) holds and $\sum_{j=1}^{m_n} \lambda_j^{-1} = o(n)$; or (ii)

$$\sum_{j=0}^{\infty} \frac{\langle \delta, \varphi_j \rangle^2}{\lambda_j^2} < \infty. \tag{5.17}$$

Since the eigenvalues λ_j are unknown, it is hard to evaluate this assumption. However, for univariate regressors we expect $\lambda_j = O(j^{-2})$, see Stute (1997 , p. 621), so that $\sum_{j=1}^{m_n} \lambda_j^{-1} = O(m_n^3)$, and hence Assumption 1(i) would require $m_n^3/n \to 0$. Assumption 1(ii) strengthens the key condition (5.9) required for absolute continuity, but relaxes the rate conditions on m_n ($m_n \to \infty$ as $n \to \infty$ arbitrarily in this case.) An important new implication of our results below is that a sufficient condition for Stute's absolutely continuity assumption (5.9) is simply $\mathbb{E}\left[d^2(X)\right] < \infty$. This follows from Parseval's identity and Fubini's Theorem, since

$$\begin{aligned} \frac{\langle \delta, \varphi_j \rangle}{\sqrt{\lambda_j}} &= \frac{1}{\sqrt{\lambda_j}} \int \int d(\bar{x}) w(\bar{x}, x) \varphi_j(x) F(d\bar{x}) F(dx) \\ &= E[d(X)\psi_j(X)]. \end{aligned} \tag{5.18}$$

where

$$\psi_j(X) := \frac{\langle w(X, \cdot), \varphi_j \rangle}{\sqrt{\lambda_j}}, \tag{5.19}$$

and hence

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{\langle \delta, \varphi_j \rangle^2}{\lambda_j} &= \sum_{j=0}^{\infty} \left(\mathbb{E}[d(X)\psi_j(X)]\right)^2 \\ &\leq \mathbb{E}[d^2(X)]. \end{aligned}$$

Alternatively, the stronger (5.17) requires

$$\sum_{j=0}^{\infty} \frac{\langle \delta, \varphi_j \rangle^2}{\lambda_j^2} = \sum_{j=0}^{\infty} \frac{\left(\mathbb{E}[d(X)\psi_j(X)]\right)^2}{\lambda_j} < \infty,$$

i.e. the Fourier coefficients of d with respect to $\{\psi_j\}_{j=1}^{\infty}$ decay sufficiently fast. This is a mild "smoothness" condition on d.

Proposition 1 *Under Assumption 1, as* $n \to \infty$

$$T_{n,m_n} = T_{n,m_n}^1 + o_{\mathbb{P}}(1).$$

Proof Note that

$$R_n(x) = \frac{1}{n\sigma}\sum_{i=1}^{n}\varepsilon_i w_n(X_i, x), \tag{5.20}$$

where

$$\begin{aligned} w_n(z,x) &:= \frac{\sigma}{\widehat{\sigma}}\left\{1_{\{z\le x\}} - \mathbf{g}^T(z)\Sigma_n^{-1}G_n(x)\right\},\\ \Sigma_n &:= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(X_i)^T\mathbf{g}^T(X_i),\\ G_n(x) &:= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(X_i)^T 1_{\{X_i\le x\}}. \end{aligned}$$

Similarly, define $G(x) := \mathbb{E}\left[\mathbf{g}(X)^T 1_{\{X\le x\}}\right]$ and $\Sigma := \mathbb{E}\left[\mathbf{g}(X)\mathbf{g}^T(X)\right]$. Then, write

$$\begin{aligned} T_{n,m_n} - T^1_{n,m_n} &= \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left(\sum_{j=1}^{m_n}\frac{\langle\delta,\varphi_j\rangle\langle w_n(X_i,\cdot) - w(X_i,\cdot),\varphi_j\rangle}{\lambda_j}\right)\\ &= \left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\mathbf{g}(X_i)^T\right)\Sigma^{-1}\left(\sum_{j=1}^{m_n}\frac{\langle\delta,\varphi_j\rangle\langle G - G_n,\varphi_j\rangle}{\lambda_j}\right)^T\\ &\quad + \left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\mathbf{g}(X_i)^T\right)\left(\Sigma^{-1} - \Sigma_n^{-1}\right)\left(\sum_{j=1}^{m_n}\frac{\langle\delta,\varphi_j\rangle\langle G_n,\varphi_j\rangle}{\lambda_j}\right)^T + o_{\mathbb{P}}(1)\\ &= o_{\mathbb{P}}(1), \end{aligned}$$

since, under Assumption 1 and (5.9)

$$\begin{aligned} \left|\sum_{j=1}^{m_n}\frac{\langle\delta,\varphi_j\rangle\langle G - G_n,\varphi_j\rangle}{\lambda_j}\right| &\le \|G - G_n\|\left(\sum_{j=1}^{m_n}\lambda_j^{-1}\right)^{1/2}\left(\sum_{j=1}^{m_n}\frac{\langle\delta,\varphi_j\rangle^2}{\lambda_j}\right)^{1/2}\\ &= o_{\mathbb{P}}(n^{-1/2})o(n^{1/2})O(1) = o_{\mathbb{P}}(1). \end{aligned}$$

The same holds true under (5.17). ■

By Proposition 1 we can focus in what follows on T^1_{n,m_n}. Then, plugging R^1_n in T^1_{n,m_n} and replacing the orders of summation, we can write

$$\begin{aligned} T^1_{n,m_n} &= \frac{1}{\sigma\sqrt{n}}\sum_{j=1}^{m_n}\sum_{i=1}^{n}\varepsilon_i\frac{\langle\delta,\varphi_j\rangle\langle w(X_i,\cdot),\varphi_j\rangle}{\lambda_j}\\ &= \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{m_n}\frac{\langle\delta,\varphi_j\rangle\langle w(X_i,\cdot),\varphi_j\rangle}{\lambda_j}\\ &= \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i s_n(X_i), \end{aligned} \tag{5.21}$$

where

$$s_n := \sum_{j=1}^{m_n} \frac{\langle \delta, \varphi_j \rangle \langle w(X_i, \cdot), \varphi_j \rangle}{\lambda_j}.$$

In the next Theorem we show that s_n converges in mean square to a limit s_∞. Then, we show that $\sigma^{-1}\varepsilon_i s_\infty(X_i)$ is proportional to the efficient score for the semiparametric testing problem (5.12), as derived in the previous section but assuming, as in Stute (1997),

Assumption 2 $\tau^2(X) = \sigma^2$ a.s.
That is, we show

$$\widehat{s}^{-1} T^1_{n,m_n} \equiv \frac{1}{n\sigma\widehat{s}} \sum_{i=1}^n \varepsilon_i s_n(X_i) = \frac{1}{n\sigma_\gamma} \sum_{i=1}^n \varepsilon_i \frac{\left(d(X_i) - \boldsymbol{\pi}_0^T \mathbf{g}(X_i)\right)}{\sigma^2} + o_{\mathbb{P}}(n^{-1/2}), \tag{5.22}$$

where $\widehat{s}^2 = n^{-1} \sum_{i=1}^n s_n^2(X_i)$ and σ_γ^2 was defined in (5.14). This asymptotic equivalence in (5.22) is the main result of this paper.

Theorem 2 *Under Assumptions 1 and 2, Stute's (1997) directional test is AUMP(α), i.e. (5.22) holds.*

Proof We show that $s_n(X_i)$ converges in mean square error to the function $s_\infty(X_i) = d(X_i) - \boldsymbol{\pi}_0^T \mathbf{g}(X_i)$, i.e.

$$\|s_n - s_\infty\| \to 0 \text{ as } n \to \infty.$$

By (5.18),

$$s_n(X_i) = \sum_{j=1}^{m_n} E[d(X)\psi_j(X)]\psi_j(X_i)$$

is the Fourier expansion of $d(X)$ in the basis $\{\psi_j\}_{j=1}^\infty$. By Kress (1999, Theorem 15.16) $\{\psi_j\}$ is a basis that spans the orthocomplement of $\ker(\mathcal{T}) = \left\{f \in \mathcal{L}_F^2 : \mathcal{T}s = 0\right\}$, where $\mathcal{T}$ is the linear operator

$$\mathcal{T}a(x) := \mathbb{E}[w(X, x)a(X)] \qquad x \in \mathbb{R}^p, a \in \mathcal{L}_F^2.$$

Hence s_n converges in mean square error to

$$s_\infty(X_i) = d(X_i) - \Pi_{\ker(\mathcal{T})} d(X_i).$$

It is straightforward to show that

$$\ker(\mathcal{T}) = span\left\{\mathbf{g}(X_i)\right\}.$$

Then, we conclude that

$$\frac{1}{n\sigma}\sum_{i=1}^{n}\varepsilon_i s_n(X_i) = \frac{1}{n\sigma}\sum_{i=1}^{n}\varepsilon_i\left(d(X_i) - \boldsymbol{\pi}_0^T\mathbf{g}(X_i)\right) + o_{\mathbb{P}}(n^{-1/2}),$$

and

$$\begin{aligned}\widehat{s}^2 &= \mathbb{E}[s_\infty^2(X_i)] + o_{\mathbb{P}}(1)\\ &= \sigma^2\sigma_\gamma^2 + o_{\mathbb{P}}(1).\end{aligned}$$

Thus,

$$\frac{1}{n\sigma\widehat{s}}\sum_{i=1}^{n}\varepsilon_i s_n(X_i) = \frac{1}{n\sigma_\gamma}\sum_{i=1}^{n}\varepsilon_i\frac{\left(d(X_i) - \boldsymbol{\pi}_0^T\mathbf{g}(X_i)\right)}{\sigma^2} + o_{\mathbb{P}}(n^{-1/2}).$$

■

Remark 3 From the proof of Theorem 2 we see that m_n plays the role of a "bandwidth" (number of terms in a series expansion, more precisely) for estimating the score $s_\infty(\cdot)$. The optimal bandwidth choice is $m_n = \infty$.

Incidentally, a test that is also AUMP for the homoskedastic case is the classical t-test. The t-test rejects H_0 for large values of

$$t_n = \frac{\gamma_n}{s.e(\gamma_n)},$$

where γ_n is the OLS estimator of γ_0 and $s.e(\gamma_n)$ its standard error. Straightforwardly, it is shown that under H_0,

$$t_n = \frac{1}{\sqrt{n}\sigma_\gamma}\sum_{i=1}^{n}\varepsilon_i\frac{\left(d(X_i) - \boldsymbol{\pi}_0^T\mathbf{g}(X_i)\right)}{\sigma^2} + o_{\mathbb{P}}(1).$$

5.4.2 The Heteroskedastic Case

In the heterokedastic case Stute's directional test is not asymptotically equivalent under H_0 to the standardized canonical effective score test given in (5.16). Therefore, it is not efficient. In this section we show how Stute's (1997) directional test needs to be implemented in the conditional heteroskedastic case to achieve asymptotic efficiency.

Stute et al. (1998) considered the martingale part of the CUSUM process

$$\tilde{R}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i - \tilde{\boldsymbol{\beta}}_n^T \mathbf{g}(X_i)}{\tau(X_i)} 1_{\{X_i \leq x\}},$$

where $\tilde{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\pi}_0$. To achieve efficiency in the directional tests below, it is crucial to use the infeasible GLS estimator

$$\tilde{\boldsymbol{\beta}}_n = \left[\sum_{i=1}^{n} \frac{\mathbf{g}(X_i)\mathbf{g}(X_i)^T}{\tau^2(X_i)}\right]^{-1} \sum_{i=1}^{n} \frac{\mathbf{g}(X_i)Y_i}{\tau^2(X_i)}.$$

Stute et al. (1998) developed omnibus, smooth and directional tests of H_0, where $\tau(\cdot)$ is estimated using smoothers. Under H_0, $\tilde{\eta}_n = \sqrt{n}\tilde{R}_n$ satisfies

$$\tilde{\eta}_n \to_d \tilde{\eta}_\infty,$$

and under H_{1n},

$$\tilde{\eta}_n \to_d \tilde{\eta}_\infty + \tilde{\delta},$$

with $\tilde{\eta}_\infty$ a Gaussian process with zero mean and covariance function

$$\tilde{C}(x_1, x_2) := \mathbb{E}\left[\tilde{\eta}_\infty(x_1)\tilde{\eta}_\infty(x_2)\right] = \mathbb{E}\left[\tilde{w}(X, x_1)\tilde{w}(X, x_2)\right],$$

where

$$\tilde{w}(z, x) = 1_{\{z \leq x\}} - \mathbb{E}\left[\frac{\mathbf{g}(X)}{\tau(X)}^T 1_{\{X \leq x\}}\right]\left(\mathbb{E}\left[\frac{\mathbf{g}(X)\mathbf{g}^T(X)}{\tau^2(X)}\right]\right)^{-1} \frac{\mathbf{g}(z)}{\tau(z)}$$

and $\tilde{\delta}(x) = \mathbb{E}\left[\tilde{w}(X, x)\, d(X)\, /\tau(X)\right]$. The functional likelihood ratio based on $\tilde{\eta}_n$ is

$$\tilde{T}_\infty = \sum_{j=1}^{\infty} \frac{\left\langle \tilde{\delta}, \tilde{\varphi}_j \right\rangle \left\langle \tilde{\eta}_\infty, \tilde{\varphi}_j \right\rangle}{\tilde{\lambda}_j},$$

where $\{\tilde{\lambda}_j, \tilde{\varphi}_j\}_{j=1}^{\infty}$ is the spectrum of $\tilde{\eta}_\infty$. The sample analog of $\tilde{T}_\infty$ is

$$\tilde{T}_{n,m_n} := \sum_{j=1}^{m_n} \frac{\left\langle \tilde{\delta}, \tilde{\varphi}_j \right\rangle \left\langle \tilde{\eta}_n, \tilde{\varphi}_j \right\rangle}{\tilde{\lambda}_j}.$$

Next Theorem establishes the semiparametric efficiency of the functional likelihood ratio test based on $\tilde{T}_{n,m_n}$ in the heteroskedastic case.

Theorem 4 *Under Assumption 1 with δ replaced by $\tilde{\delta}$, the directional test based on $\tilde{T}_{n,m_n}$ is AUMP.*

Proof The proof is identical to that of Theorem 2, but replacing Y, $\mathbf{g}(x)$, and $d(x)$ by $Y/\tau(x)$, $\mathbf{g}(x)/\tau(x)$, and $d(x)/\tau(x)$, respectively.

Another efficient test is based on the infeasible GLS estimator of γ_0, given by

$$\tilde{\gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\varepsilon_{ni} \left(d(X_i) - \boldsymbol{\pi}_n^T \mathbf{g}(X_i) \right)}{\tau^2(X_i)},$$

with

$$\boldsymbol{\pi}_n = \left[\sum_{i=1}^{n} \frac{\mathbf{g}(X_i)\mathbf{g}(X_i)^T}{\tau^2(X_i)} \right]^{-1} \sum_{i=1}^{n} \frac{\mathbf{g}(X_i) d(X_i)}{\tau^2(X_i)}.$$

The corresponding *t—ratio* statistic for the significance of $d(X)$ is the generalized least squares *t—ratio*

$$\tilde{t}_n = \frac{\tilde{\gamma}_n}{\hat{\sigma}_\gamma},$$

where

$$\hat{\sigma}_\gamma^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(d(X_i) - \boldsymbol{\pi}_n^T \mathbf{g}(X_i) \right)^2}{\tau^2(X_i)},$$

and which satisfies $\tilde{t}_n = \zeta_{n\gamma} + o_{\mathbb{P}}(1)$ under H_0.

Feasible versions of the tests above are constructed by replacing $\tau^2(\cdot)$ with a smooth estimator of it, e.g. the *k—nn* suggested by Robinson (1987) for semiparametric GLS. Using the *k—nn* estimator and applying Robinson's (1987) results, we obtain that feasible versions are asymptotically equivalent to the infeasible ones under H_0 and H_{1n}. Hence, following the results above it is shown that the feasible test is also AUMP.

5.5 Conclusions

This article has shown the efficiency of Stute's (1997) directional test in a semiparametric sense under conditional homoskedasticity. It has shown its asymptotic equivalence to the classical *t—ratio* test in that setting. For conditional heteroskedasticity of unknown form, a directional test based on the conditionally standardized CUSUM process of Stute et al. (1998) with an efficient estimator for the slope coefficients is shown to be semiparametrically efficient, and asymptotic equivalent to the *t—ratio* test based on the generalized least squares estimator. Thus, directional tests, originally derived as functional likelihood ratio tests, are shown to be efficient in a more traditional sense in a semiparametric one-sided testing problem (cf. Choi et al. 1996, Sect. 3).

Two-sided and/or multidimensional versions of the proposed semiparametric tests are also available. The results of this article also show that the corresponding versions of directional tests have certain optimality properties in these settings. For example, a two-sided version of Stute's (1997) directional test rejects the null for large absolute values of T_{n,m_n} in (5.11), and our results imply that this test is AUMP within the class of unbiased tests (cf. Choi et al. 1996, Sect. 4). Multidimensional tests correspond to a multivariate d, and the corresponding Stute's (1997) directional tests are AUMP within a class of invariant test defined in Choi et al. (1996, Sect. 5).

Another important contribution of Stute (1997) was the development of Neyman's smooth tests for regression models based on the principal components of the CUSUM process. Combining the results of the present article with those of Escanciano (2009), it is shown that Stute-Neyman's smooth tests are also AUMP and invariant tests for an implicitly defined multidimensional d, precisely the vector with components $\{\psi_j(X)\}_{j=1}^m$ in (5.19) or their heteroskedastic version. Details of this result are beyond the scope of this paper, but we refer to Escanciano (2009). Related Neyman's smooth tests for regression with errors independent of covariates and multidimensional d have been obtained by Inglot and Ledwina (2006). Neyman's smooth tests for regression under conditional mean independence of errors and covariates are classical score tests based on least squares estimates, or based on generalized least squares estimators under conditional heteroskedasticity. They are called Lagrange Multiplier tests in econometrics. These tests are optimal in a semiparametric sense discussed in Choi et al. (1996, Sect. 5), they are easy to interpret and are a compromise between the directional tests and omnibus tests of Stute (1997), which have been so fundamental in the development of model checks for regression.

References

BICKEL, P.J., C.A. KLAASEN, Y. RITOV AND J.A.WELLNER (1993): *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Univ. Press, Baltimore.

BISCHOFF, W. AND MILLER, F. (2000): "Asymptotically Optimal Tests and Optimal Designs for Testing the Mean in Regression Models with Applications to Change-Point Problems," *Annals of the Institute of Statistical Mathematics,* 52, 658–679.

BONING, B., AND F. SOWELL (1999): "Optimality for the Integrated Conditional Moment Test," *Econometric Theory,* 15, 710–718.

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics,* 34, 305–334.

CHOI, S., W.J.HALL AND A. SCHICK (1996): "Asymptotically Uniformly Most Powerful Tests in Parametric and Semiparametric Models," *The Annals of Statistics,* 24, 841–861.

DELGADO, M.A., J.HIDALGO AND C.VELASCO (2005): "Distribution Free Goodness-of-fit Tests for Linear Processes," *The Annals of Statistics,* 33, 2568–2609.

DELGADO, M.A. AND W. STUTE (2008): "Distribution-Free Specification Tests of Conditional Models," *Journal of Econometrics,* 143, 37–55.

ESCANCIANO, J. C. (2009): "On the Lack of Power of Omnibus Specification Tests," *Econometric Theory,* 25, 162–194.

ESCANCIANO, J. C. (2012): " Semiparametric Efficient Tests," Unpublished manuscript.

GRENANDER, U. (1950): "Stochastic Processes and Statistical Inference," *Arkiv for Matematik,* 1, 195–277.

INGLOT, T. AND T. LEDWINA (2006): "A Data-Driven Score Test for Homocedastic Linear Regression: Asymptotic Results," *Probability and Mathematical Statistics* , 26, 41–61.

KRESS, R. (1999): *Linear Integral Equations*. Springer.

KUNDU, S., MAJUMDAR, S. AND K. MUKHERJEE (2000): "Central Limit Theorems revisited," *Statistics & Probability Letters*, 47, 265–275.

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* , 5, 99–135.

ROBINSON, P.M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica,* 55, 875–891.

SKOROHOD, A. V. (1974): *Integration in Hilbert space*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 79, Springer-Verlag, New York.

STUTE, W. (1997): "Nonparametric Model Checks for Regression," *The Annals of Statistics,* 25, 613–641.

STUTE, W., GONZÁLEZ- MANTEIGA, W. AND M. PRESEDO- QUINDIMIL (1998): "Bootstrap Approximations in Models Checks for Regression," *Journal of the American Statistical Association*, 93, 141–199.

STUTE, W., S.THIES, AND L.X.ZHU (1998): "Model Checks for Regression: An Innovation Process Approach," *The Annals of Statistics,* 26, 1916–1934.

Goodness–of–Fit Test for Stochastic Volatility Models

6

Wenceslao González-Manteiga, Jorge Passamani Zubelli, Abelardo Monsalve-Cobis and Manuel Febrero-Bande

6.1 Introduction

Understanding and quantifying volatility is one of the main challenges in present–day financial analysis. This is thoroughly justified by its impact in pricing and risk management, among other applications. For instance, it is crucial in pricing and hedging derivatives a good model selection.

However, this is not an easy task to accomplish. The volatility of a process is not directly observed and thus needs to be estimated by some indirect process. In addition, the term *volatility* has different meanings depending on the discipline or field of study, which has given rise to quite a few different definitions of volality throughout the literature (see Ghyles et al. (1996), Shephard (2005) and references therein for a sample of the various approaches and historical background).

Nevertheless, there is a key feature of volatility which somehow unifies the alternative approaches: volatility refers to a measure of variation or oscillation of the observed quantity time series. Intuitively, higher volatility acts as if time would be

W. González-Manteiga (✉) · M. Febrero-Bande
Faculty of Mathematics, University of Santiago de Compostela, Santiago, Spain
e-mail: wenceslao.gonzalez@usc.es

M. Febrero-Bande
e-mail: manuel.febrero@usc.es

J.P. Zubelli
Institute for Pure and Applied Mathematics (IMPA), Rio de Janeiro, Brazil
e-mail: zubelli@gmail.com

A. Monsalve-Cobis
University Centroccidental Lisandro Alvarado, Barquisimeto, Venezuela
e-mail: amonsalve@ucla.edu.ve

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_6

running faster and more information is being added to the observed system (see Shephard (2005) and references therein).

In financial applications, the classical work of Markowitz (1952) connects the volatility directly with investment strategies and risk management. The seminal papers by Black and Scholes (1973) and Merton (1973) in option pricing make use of a powerful simplifying assumption. Namely, that the underlying asset follows a geometric Brownian motion (GBM) with drift

$$dr_t = \mu r_t dt + \sigma r_t dW_t \ , \tag{6.1}$$

where dW_t refers to the differential of the Wiener process, r_t denotes the asset price and σ and μ are constant values. However, this hypothesis has been deeply scrutinized and questioned in the literature. For instance, even in the case of US stock returns, departures from GBM have been well–documented (see Campbell et al. (1997), Sect. 9.3.6).

In addition, the unpredictability and evidence of non–stationarity of the volatility in financial time series under different scales has been well documented in the literature and goes back to Mandelbrot (1963) and Officer (1973). This naturally leads to the proposal of more general models than (6.1), such as

$$dr_t = m(r_t)dt + \sigma(r_t)dW_t, \tag{6.2}$$

where the drift m and the volatility σ are now dependent on the underlying asset, r_t. Equation (6.2) can be analyzed as a parametric model by assuming that

$$dr_t = m(r_t, \theta)dt + \sigma(r_t, \theta)dW_t, \tag{6.3}$$

where the functional form for m an σ are well–defined within a certain class that depends on an unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$ with d a positive integer. Equation (6.3) allows for the representation of a fairly broad family of financial models. See for example Andersen and Lund (1997) and Hull and White (1987).

A plethora of volatility definitions and indices arises when volatility models are formulated in a discrete time scale. For instance, a great deal of attention has been paid to models such as the autoregressive conditional heteroskedasticity (ARCH) model by Engle (1982) and its generalizations. The need for continuous models is obvious and crucial for comparisons, model simulations and ultimately pricing and risk-management. As it can be seen in Shephard (2005), there have been efforts in both simulation and inference methods on continuous–time stochastic volatility models. Nevertheless, to the best of our knowledge, the joint use of goodness–of–fit tests and Kalman filtering techniques has not been explored.

In this work, financial models will be considered as continuous in time and described by stochastic differential equations whose coefficients are to be determined parametrically. In particular, the focus will be placed on models for an observed quantity r_t given by the stochastic differential equation

$$\begin{aligned} dr_t &= m_1(r_t, \theta)dt + \sigma_t \upsilon_1(r_t, \theta)dW_{1,t} \\ dg(\sigma_t) &= m_2(g(\sigma_t), \vartheta)dt + \upsilon_2(g(\sigma_t), \vartheta)dW_{2,t} \end{aligned} \tag{6.4}$$

where g, m_1, υ_1, m_2 and υ_2 are known functions; $\Phi = (\theta, \vartheta) \in \mathbb{R}^d$ is an unknown vector parameter (to be estimated); σ_t^2 is the unobserved volatility and $W_{1,t}$ y $W_{2,t}$ are (possibly correlated) Brownian motions.

As remarked in Campbell et al. (1997) there are many open issues in statistical inference for continuous–time processes with discretely sampled data. For instance, Aït-Sahalia (1993) proposes a nonparametric estimator of the diffusion coefficient (assuming some constraints on the drift). Genon-Catalot et al. (1999) introduce appropriate and explicit functions of the observations to replace either the log-likelihood or the score function. Aït-Sahalia and Kimmel (2007) developed an alternative method that employs maximum likelihood, using closed form approximations to the true (but unknown) likelihood function. Specifically, for Model (6.3), the goodness–of–fit testing problem has been discussed by Dette and von Lieres und Wilkau (2003), Dette et al. (2006) and Monsalve-Cobis et al. (2011).

For the stochastic volatility model in Eq. (6.4), most of the existing methods for goodness–of–fit testing are not directly applicable due the fact that the volatility is not directly observed, but there have been some approaches for testing its components. For example, Lin et al. (2013) propose a goodness–of–fit test for the volatility distribution in (6.4), based on the deviation between the empirical characteristic function and its parametric counterparts.

In this work, a goodness–of–fit test based on the empirical process is proposed. First, a discretized version of Model (6.4) is considered. Then, Kalman filtering techniques are applied to obtain the associated state space model. Finally, the ideas described in Monsalve-Cobis et al. (2011) for the construction of some generalized statistical tests are applied to this context. Thus, the goal is to introduce a goodness–of–fit test for the (parametric) drift and volatility functions in those models with a stochastic volatility component. Calibration of the tests is done using bootstrap procedures (see Rodriguez and Ruiz (2012) and Monsalve-Cobis et al. (2011)).

This article is organized as follows: the continuous time stochastic volatility models are presented in Sect. 6.2, discussing the corresponding state space structure. In Sect. 6.3, the new goodness–of–fit tests for the drift and the volatility is introduced. Section 6.4 is devoted to the bootstrap strategy used for calibration. Finally, in Sect. 6.5, some preliminary simulation results are provided, jointly with a real data application of the tests, dealing with interbank intereset rates in the Eurozone.

6.2 The Stochastic Volatility Model

Consider the stochastic volatility Model (6.4), where g, m_1, m_2, υ_1 and υ_2 are known real valued functions satisfying certain regularity conditions in order to ensure the existence and uniqueness of the solution of the underlying stochastic differential equations (see Genon-Catalot et al. (1999) and Lin et al. (2013)). The coefficients in (6.4) depend on the unknown parameters $\Phi = (\theta, \vartheta) \in \Theta \in \mathbb{R}^d$, and therefore, different models can be generated for stochastic volatility by choosing different parametric forms for the functions g, m_1, m_2, υ_1 and υ_2. The developments presented in this paper will be focused on a widely studied model, which has been used in several financial applications: the *CKLS* model proposed by Andersen and Lund (1997). This model incorporates the volatility as a non observable stochastic factor, being an extension of the *CKLS* model introduced by Chan et al. (1992). The specification proposed by Andersen and Lund (1997) assumes mean reversion -both at the level of the interest rate and at the volatility (in log scale). More concretely:

$$dr_t = \kappa_1(\mu - r_t)dt + \sigma_t r_t^{\gamma} dW_{1,t}$$

$$d\log(\sigma_t^2) = \kappa_2(\alpha - \log(\sigma_t^2))dt + \xi dW_{2t},$$

where W_{1t} and W_{2t} are independent Brownian motions, and $\alpha, \kappa_1, \kappa_2, \mu, \gamma$ and ξ are the unknown parameters.

It should be also noted that it is not unusual to find in Model (6.4) a correlation between r_t and σ_t as a consequence of the corresponding Brownian processes. In that case, the following kind of dependence structure can be used,

$$dW_{1t} = \rho dW_{2t} + \sqrt{1-\rho^2} dW_{3t},,$$

with W_{2t} and W_{3t} independent Brownian motions. However, along this paper, ρ will be set to 0.

Although model in Eq. (6.4) specifies a proper framework for continuous time financial process analysis, in practice, the phenomena associated to such processes is just observed at discrete time points. Hence, discretized versions of continuous time models must be considered for application in practice. For that purpose, assume that the process $\{r_t : 0 \le t \le T\}$ is observed at discrete equally spaced times $t_i = i\Delta$, $i = 0, 1, \ldots, n$, with a fixed $\Delta > 0$ within an observation window $[0, n\Delta = T]$, which increases as n grows. Then, the discrete time version of Model (6.4) can be formulated as

$$r_{t_{i+1}} - r_{t_i} = m_1(r_{t_i}, \theta)\Delta + \sigma_{t_i}\upsilon_1(r_{t_i}, \theta)\left(W_{1,t_{i+1}} - W_{1,t_i}\right)$$

$$g(\sigma_{t_{i+1}}) - g(\sigma_{t_i}) = m_2(g(\sigma_{t_i}), \vartheta)\Delta + \upsilon_2(g(\sigma_{t_i}), \vartheta)\left(W_{2,t_{i+1}} - W_{2,t_i}\right)$$

and taking into account the properties of the Brownian motion, the process can be expressed as

$$\begin{aligned} \frac{y_{t_i}}{\Delta} &= m_1(r_{t_i}, \theta) + \sigma_{t_i} \upsilon_1(r_{t_i}, \theta) \Delta^{-1/2} \varepsilon_{1,t_i} \\ g(\sigma_{t_{i+1}}) - g(\sigma_{t_i}) &= m_2(g(\sigma_{t_i}), \vartheta)\Delta + \upsilon_2(g(\sigma_{t_i}), \vartheta)\sqrt{\Delta}\varepsilon_{2,t_i}, \end{aligned} \tag{6.5}$$

where, $y_{t_i} = r_{t_{i+1}} - r_{t_i}$, and $\{\varepsilon_{1,t_i}, \varepsilon_{2,t_i}\}$ are two independent random variables with distribution $N(0, 1)$, for $i = 1, \ldots, n$.

An important issue when analyzing the behaviour of the aforemetioned processes is the large sample scheme, since there is not a unique way of defining it. The most natural approach considered in practice consists in taking Δ (spacing between two consecutive observations) as fixed and let the number of observations n grow (see Kessler (2000) and Iacus (2008), for some examples). However, there are other alternatives, as the one considered by Genon-Catalot et al. (1999), where the sampling distance $\Delta = \Delta_n$ goes to zero whereas the window $n\Delta_n$ goes to infinity. The main goal of the different observation schemes is related to keeping the asymptotic properties of the estimators and to allow the use of statistical inference methods (see Lin et al. (2013)).

6.2.1 State Space Model

The estimation of stochastic volatility models turns out to be a complex problem, partly motivated by the estimation of the transition density funciton of r_t (the state variable), which is itself a difficult task, even under closed formulations. In addition, the state variables that determine the volatility are not directly observable. Thus, the estimation for such a function just from information of the underlying process in its essence calls for the use of filtering techniques. With this purpose, Kalman filtering techniques are applied to obtain the state space representation (6.5) of the model in (6.4). Taking $x_{t_i} = g(\sigma_{t_i})$, with g strictly monotonic and after some algebraic manipulations,

$$\begin{aligned} \frac{y_{t_i}}{\Delta} &= m_1(r_{t_i}, \theta) + g^{-1}(x_{t_i}) \upsilon_1(r_{t_i}, \theta) \Delta^{-1/2} \varepsilon_{1,t_i} \\ x_{t_{i+1}} &= x_{t_i} + m_2(x_{t_i}, \vartheta)\Delta + \upsilon_2(x_{t_i}, \vartheta)\sqrt{\Delta}\varepsilon_{2,t_i}. \end{aligned} \tag{6.6}$$

The main goal of this representation is to capture the dynamics of the observable variables y_{t_i} and r_{t_i}, in terms of the unobservable σ_{t_i}. It is important to stress that, for convenience, the state space model is required to fall within the class of linear

state space models. This is achieved considering, for example, $g(y) = \log(y^2)$. Thus, Model (6.5) with $g(\cdot) = 2\log(\cdot)$ can be written as Eq. (6.6):

$$\frac{y_{t_i}}{\Delta} = m_1(r_{t_i}, \theta) + \sigma_{t_i} \upsilon_1(r_{t_i}, \theta)\Delta^{-1/2}\varepsilon_{1,t_i}$$

$$\log(\sigma^2_{t_{i+1}}) = \log(\sigma^2_{t_i}) + m_2(\log(\sigma^2_{t_i}), \vartheta)\Delta + \upsilon_2(\log(\sigma^2_{t_i}), \vartheta)\sqrt{\Delta}\varepsilon_{2,t_i}.$$

Following the derivation in Harvey et al. (1994), denote by e_{t_i} the error obtained from the equation

$$e_{t_i} = \frac{y_{t_i}}{\Delta} - m_1(r_{t_i}, \theta) = \sigma_{t_i} \upsilon_1(r_{t_i}, \theta)\Delta^{-1/2}\varepsilon_{1,t_i} \ ,$$

which gives:

$$\log(e^2_{t_i}) = \log(\sigma^2_{t_i}) + 2\log(\upsilon_1(r_{t_i}, \theta)) - \log(\Delta) + \log(\varepsilon^2_{1,t_i})$$

Now, taking $u_{t_i} = \log(e^2_{t_i})$, and $x_{t_{i+1}} = \log(\sigma^2_{t_{i+1}})$, the following state space model is obtained:

$$\begin{aligned} u_{t_i} &= x_{t_i} + 2\log(\upsilon_1(r_{t_i}, \theta)) + \eta_{t_i} - \kappa \\ x_{t_{i+1}} &= x_{t_i} + m_2(x_{t_i}, \vartheta)\Delta + \upsilon_2(x_{t_i}, \vartheta)\sqrt{\Delta}\varepsilon_{2,t_i} \end{aligned} \tag{6.7}$$

with $\eta_{t_i} = -\log(\Delta) + \log(\varepsilon^2_{1,t_i}) + \kappa$ and $\kappa = \log(\Delta) - \mathbb{E}\left[\log(\varepsilon^2_{1,t_i})\right]$. The parameter estimation of $\Phi = (\theta, \vartheta)$ can be obtained by maximum likelihood, computing the likelihood from the innovations $\eta_{t_1}, \ldots, \eta_{t_n}$.

In the sequel, the estimation can be obtained using Kalman filters considering a mixture of Gaussian variables to approximate the non–Gaussian errors, but other alternatives are also possible. With respect to this issue, note that innovation errors in the previous state space model are not Gaussian. If ε^2_{1,t_i} follows a lognormal distribution, then the state space model presents Gaussian errors, and it can be estimated using basic Kalman filter techniques. Unfortunately, under the assumption of normality for ε_{1,t_i}, the variable ε^2_{1,t_i} has a χ^2 distribution with one degree of freedom, and the density under the logarithmic transformation is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(e^x - x)}, \qquad -\infty < x < \infty,$$

with mean -1.2704 and variance $\pi^2/2$. It is clear the need of applying methodology that allows to obtain an equation involving the non observable variable x_{t_i}, with Gaussian mixture distributions. Therefore, writing the observation equation in Model (6.7) with $u_{t_i} = \log(e^2_{t_i})$ as

$$u_{t_i} = x_{t_i} + 2\log(\upsilon_1(r_{t_i}, \theta)) + \eta_{t_i} - \kappa$$

where η_{t_i} is zero-mean noise, the assumption of a normal mixture distribution will be considered. In particular, for a mixture of two distributions:

$$\eta_{t_i} - \kappa = I_{t_i} z_{t_i 0} + (1 - I_{t_i}) z_{t_i 1},$$

where I_{t_i} is an iid process such that $\mathbb{P}\{I_{t_i} = 0\} = \pi_0$, $\mathbb{P}\{I_{t_i} = 1\} = \pi_1$, $(\pi_0 + \pi_1 = 1)$, $z_{t_i 0} \sim iid\ N(0, \sigma_0^2)$, and $z_{t_i 1} \sim iid\ N(\mu_1, \sigma_1^2)$. The advantage of such procedure hinges upon the use of normality.

The estimation of the model parameters is performed by maximum likelihood, being the log–likelihood function to optimize:

$$\log L(\Phi) = \sum_{i=1}^{n} \log \left(\sum_{j=0}^{1} \pi_j f_j(t_i | t_i - 1) \right)$$

where the transition density $f_j(t_i | t_i - 1)$ is approximated by a normal or normal mixture density, with parameters given by the filter. For details see, for example, Shumway and Stoffer (2011), Sects. 6.8 and 6.9.

An alternative method for the estimation of the stochastic volatility can be found in Aït-Sahalia and Kimmel (2007). In this reference, maximum likelihood is also used but considering numerical approximations of the true likelihood. In order to consider positive correlation between Brownian motions, the methods introduced by Sect. 6.7–Shumway and Stoffer (2011) and Nisticò (2007), also based on Kalman filtering techniques, could be considered.

6.3 GOF-Tests

A generalization of the goodness–of–fit test proposed in Monsalve-Cobis et al. (2011) for the stochastic volatility Model (6.4) will be presented in this section. The proposal follows the methodology developed by Stute (1997) for the regression context, based on empirical residual processes. The goal in this work is to compare the parametric form of the drift functions and the volatility for the model under consideration, establishing as null hypothesis:

$$\mathscr{H}_{0m} : m_1 \in \{m_1(\cdot, \theta) : \theta \in \Theta\} \tag{6.8}$$

for the parametric form of the drift function, and

$$\mathscr{H}_{0v} : \upsilon_1 \in \{\upsilon_1(\cdot, \theta) : \theta \in \Theta\} \tag{6.9}$$

for the parametric form of the volatility. The construction of the test statistic and the testing procedure will be described in the next sections.

6.3.1 Drift Function Test

Assume that $\hat{\Phi} = (\hat{\theta}, \hat{\vartheta})$ is an appropriate estimator (satisfying a root-n consistency condition) of the true parameter $\Phi = (\theta, \vartheta)$ of the stochastic volatility model. The test statistic for assessing the parametric form of the drift function, under the assumption that $\mathscr{H}_{0v}$ given by (6.9) holds, is based on the empirical process:

$$D_n(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}_{\{r_{t_i} \leq r\}} \left\{ \frac{y_{t_i}}{\Delta} - m_1(r_{t_i}, \hat{\theta}) \right\}, \quad \text{with } r \in \mathbb{R},$$

being $\mathbf{1}_{\{\cdot\}}$ the indicator function. For constructing a test statistic, a continuous functional $\Psi(\cdot)$ of the empirical process can be considered. In general, such a test statistic will be defined as $T_n = \Psi(D_n)$ and the null hypothesis $\mathscr{H}_{0m}$ is rejected if $T_n > c_{1-\alpha}$ where $c_{1-\alpha}$ satisfies

$$\mathbb{P}\{T_n > c_{1-\alpha} | \mathscr{H}_{0m}\} = \alpha \, .$$

Two examples of such a test statistic are the following

$$T_n^{KS} = \sup_r |D_n(r)|, \qquad \text{and} \qquad T_n^{CvM} = \int_{\mathbb{R}} D_n(r)^2 F_n(dr)$$

being the first one a Kolmogorov-Smirnov (KS) type test and the second one a Cramér-von Mises (CvM) statistic. In the previous formulation, F_n denotes the empirical distribution of $\{r_{t_i}\}_{i=1}^n$. Along the text, $T_n = T_n^{KS}$ or $T_n = T_n^{CvM}$ will be used to indicate the specific statistics under consideration.

6.3.2 Volatility Function Test

Focusing now on the volatility component, and similarly to the ideas presented for the test designed for the drift function, assume that $\hat{\Phi} = (\hat{\theta}, \hat{\vartheta})$ is an appropriate estimator of the true parameter $\Phi = (\theta, \vartheta)$ in the volatility model. The goodness–of–fit test for the parametric form of the volatility function under the assumption that $\mathscr{H}_{0m}$ given by (6.8) holds, is based on the empirical process:

$$V_n(r, x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}_{\{r_{t_i} \leq r, \hat{\sigma}_{t_i}^2 \leq x\}} \left\{ \left(\frac{y_{t_i}}{\Delta} - m_1(r_{t_i}, \hat{\theta}) \right)^2 - \frac{\hat{\sigma}_{t_i}^2 \upsilon_1^2(r_{t_i}, \hat{\theta})}{\Delta} \right\}, \quad \text{with } r, x \in \mathbb{R}.$$

with $\hat{\sigma}_t^2$ an estimate of the volatility. As before, a continuous functional $\Psi(\cdot)$ of the empirical process can be considered to define, in general, the test statistics $U_n = \Psi(V_n)$. Similarly, the null hypothesis $\mathscr{H}_{0\upsilon}$ is rejected if $U_n > c_{1-\alpha}$ where $c_{1-\alpha}$ is the critical value for the α–level test:

$$\mathbb{P}\{U_n > c_{1-\alpha} | \mathscr{H}_{0\upsilon}\} = \alpha \, ,$$

Again, Kolmogorov-Smirnov (KS) and Cramér-von Mises (CvM) statistics can be expressed as

$$U_n^{KS} = \sup_{r,x} |V_n(r,x)|, \qquad \text{and} \qquad U_n^{CvM} = \int\int_{\mathbb{R}^2} (V_n(r,x))^2 \, F_n(dr,dx)$$

where F_n is the empirical distribution of $\{r_{t_i}, \hat{\sigma}_{t_i}^2\}$. $U_n = U_n^{KS}$ or $U_n = U_n^{CvM}$ will be used to denote the corresponding statistics.

In the definition of $V_n(r,x)$, a parametric model is assumed for the drift m_1. If such a model in $\mathscr{H}_{0m}$ is not specified, a nonparametric estimator for m_1 must be used. In that case, the problem is that: $\mathbb{E}\left[y_{t_i}/\Delta | r_{t_i}\right] = m_1(r_{t_i},\theta) + \upsilon_1(r_{t_i},\theta)\Delta^{-1/2}$ $\mathbb{E}\left[g^{-1}(x_{t_i})\epsilon_{1,t_i}|r_{t_i}\right]$ and additional assumptions on the stochastic volatility would be necessary to obtain a consistent estimator of m_1 (for example, using a kernel estimation). Clearly, this is still an open problem and more research is needed in this direction.

Both for the drift and the volatility tests, the critical values under the null hypothesis, denoted by $c_{1-\alpha}$, must be determined. For that purpose, the distribution of the processes T_n and U_n must be specified, which turns out to be difficult in general. Alternatively, approximations of such critical values by means of bootstrap techniques can be considered for testing purposes. A bootstrap approximation will be introduced in the next section.

6.4 Bootstrap Approximations

A bootstrap algorithm will be presented for approximating the critical values of the proposed test statistics. The procedure is based on the generation of an artificial sample with the same characteristics of the initial one. From such a sample, critical values are estimated as follows.

First, let $\{(r_t^*)\}$ be an artificial process (to be defined later in detail) and let $\hat{\Phi}^* = (\hat{\theta}^*, \hat{\vartheta}^*)$ be a parameter estimator obtained from such process. Then, the bootstrap versions D_n and V_n are given by:

$$D_n^*(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}_{\{r_{t_i}^* \le r\}} \left\{ \frac{y_{t_i}^*}{\Delta} - m_1(r_{t_i}^*, \hat{\theta^*}) \right\}, \quad \text{with } r \in \mathbb{R}.$$

$$V_n^*(r,x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}_{\{r_{t_i}^* \le r, \hat{\sigma}_{t_i}^{*2} \le x\}} \left\{ \left(\frac{y_{t_i}^*}{\Delta} - m_1(r_{t_i}^*, \hat{\theta^*}) \right)^2 - \frac{\hat{\sigma}_{t_i}^{*2} \upsilon_1^2(r_{t_i}^*, \hat{\theta^*})}{\Delta} \right\}, \quad \text{with } r, x \in \mathbb{R}.$$

The critical value $c_{1-\alpha}$ will be approximated by its bootstrap counterpart, $c_{1-\alpha}^*$, so that

$$\mathbb{P}^*\{T_n^* > c_{1-\alpha}^*\} = \alpha, \quad \mathbb{P}^*\{U_n^* > c_{1-\alpha}^*\} = \alpha,$$

where $\mathbb{P}^*$ denotes the probability measure associated to the bootstrap with

$$T_n^{*KS} = \sup_r |D_n^*(r)|, \quad \text{or} \quad T_n^{*CvM} = \int_{\mathbb{R}} D_n^*(r)^2 F_n(dr)$$

and

$$U_n^{*KS} = \sup_{r,x} |V_n^*(r,x)|, \quad \text{or} \quad U_n^{*CvM} = \int\int_{\mathbb{R}^2} V_n^*(r,x)^2 F_n(dr,dx)$$

In practice,

$$c_{1-\alpha}^* = T_n^{*\lceil B(1-\alpha)\rceil}, \quad \text{or} \quad c_{1-\alpha}^* = U_n^{*\lceil B(1-\alpha)\rceil}$$

that is, the $\lceil B(1-\alpha)\rceil$-th order statistic calculated on the B bootstrap replicates $T_n^{*j} = T_n^*$ ($U_n^{*j} = U_n^*$), $1 \le j \le B$. The empirical p-value for the bootstrap sample can be calculated as

$$\frac{\sharp\{T_n^{*j} > T_n\}}{B} \quad \text{or} \quad \frac{\sharp\{U_n^{*j} > U_n\}}{B},$$

that is to say, the p-value is taken as the fraction of values from the bootstrap versions T_n^* (U_n^*) exceeding the value of T_n (U_n). It seems clear that appropriate (consistent) parametric estimates for the stochastic volatility model parameters is crucial for the procedure. In the following section, some aspects concerning the characteristics of the artificial sample used in the bootstrap procedure implementation will be described.

6.4.1 Bootstrap Resampling

For the construction of the bootstrap sample, the state space model structure must be taken into account. This feature will be illustrated for Model (6.7). A crucial condition is that such model presents Gaussian errors, or at least, that such errors are approximately Gaussian distributed (which must be checked using some statistical procedure). Under these premises, the bootstrap sample can be generated as follows:

1. Let $\hat{\Phi} = (\hat{\theta}, \hat{\vartheta})$ be the estimator of the true parameter Φ, obtained by maximum likelihood. That is:
$$\hat{\Phi} = \arg\max_{\Phi} L(\Phi),$$
assuming that the errors follow Gaussian mixture distribution.
2. Following Shumway and Stoffer (2011) Sect. 6.7, apply the Kalman filter equations to obtain a bootstrap resample $\left\{\left(u_{t_i}^*, x_{t_i}^*, r_{t_i}\right)\right\}$ where the $\left\{r_{t_i}\right\}$ remain fixed.
3. Once the bootstrap resample is obtained, estimate the corresponding parameters $\hat{\Phi}^* = (\hat{\theta}^*, \hat{\vartheta}^*)$ associated to the state space model by maximum likelihood, based on $L^*(\Phi)$.
4. Get the bootstrap versions of the aforementioned processes $D_n^*(r,x)$ and $V_n^*(r,x)$, for $x, r \in \mathbb{R}$, and T_n^*, (U_n^*).

5. Repeat Steps 2–4 B times and get copies T_n^{*j}, (U_n^{*j}), for $j = 1, 2, \ldots, B$,
6. Finally, compute the bootstrap approximations of the critical values

$$\hat{c}^*_{1-\alpha} = T_n^{*\lceil B(1-\alpha)\rceil} \quad \text{or} \quad \hat{c}^*_{1-\alpha} = U_n^{*\lceil B(1-\alpha)\rceil}.$$

6.5 Some Applications

The performance of the testing procedure introduced in this work is illustrated in this section. First, some preliminary simulation results are shown, considering a previously studied model from the financial literature and showing the procedure perfomance. A real data example is also provided. The dataset gathers interest rate curves at the European markets, the EURIBOR®-(Euro Interbank Offered Rate).

As an example of artificial data, consider Model (1) given in (Monsalve-Cobis et al. 2011, Eq. (35)):

$$dr_t = (0.0408 - 0.5921 r_t)dt + \sigma_t r_t^{1.4999} dW_{1,t}$$

where the deterministic value $\sigma_t = \sqrt{1.6704}$ in the previous reference is replaced by a stochastic volatility model with $d(\log \sigma_t^2) = \omega dW_{2,t}$ being ω an unknown parameter.

Equation (6.7) adapted for this model is given by:

$$\begin{aligned} u_{t_i} &= x_{t_i} + \gamma_0 \log(r_{t_i}^2) - \log \Delta - 1.2704 + \eta_{t_i} \\ x_{t_{i+1}} &= x_{t_i} + \omega\sqrt{\Delta}\varepsilon_{2,t_i} \end{aligned}$$

with $\gamma_0 = 1.4999$, η_{t_i} a random variable following the centered density given in Sect. 6.2 and ε_{2,t_i} distributed as a standard normal.

The null hypothesis $\mathscr{H}_{0v} : \upsilon_1(r_t, \theta) = r_t^{1.4999}$ was tested under the assumption that the drift is completely known and with $\omega = 0.0046$ in the simulated model. For simulations, $\{r_t : 0 \le t \le T\}$ was observed at discrete equally spaced times $t_i = i\Delta$, $i = 0, 1, \ldots, n = 300$ with $\Delta = 1/52$. The Kolmogorov–Smirnov U_n^{KS} statistic was applied for testing in 100 trials. The distribution of the test under the null was calibrated by the suggested bootstrap resampling with $B = 1000$. For the resampling, the density of η_{t_i} was simulated using a mixture of seven normal densities as described in Kim et al. (1998). Table 6.1 shows the empirical power for the levels $\alpha = 0.1, 0.05$ obtained for the null and for the alternatives $\upsilon_1(r_t, \theta) = r_t^{\gamma}$ with $\gamma = 1.4999$ (the null) and $\gamma = 1.25$ and $\gamma = 1.0$.

Needless to say that this is a very simple example and more research is necessary about the theoretical and practical behaviour of the different tests and the simulation in more complex models. Even in this simple case, the optimization procedure of the Kalman filter is quite demanding in computing time. That effect is multiplied here by the number of bootstrap replicates. So, a revision of the (possibly high time consuming) steps involved in the procedure (design of Kalman filter, optimization

Table 6.1 Empirical power for the null ($\gamma = 1.4999$) and for the two alternatives ($\gamma = 1.25, 1.0$)

	$\gamma = 1.4999$	$\gamma = 1.25$	$\gamma = 1.00$
$\alpha = 0.10$	0.11	0.39	0.89
$\alpha = 0.05$	0.08	0.24	0.77

techniques, constraints for the parameters, …) is required in order to get better calibration levels in a more extensive study.

As far as the real data set is concerned, the interest rate curves of EURIBOR, representing the rates at which different interbank Euro denominated deposits, with distinct maturities, are offered within Eurozone. Such maturities for the *EURIBOR* time series are 1, 2, and 3 weeks, and $1, 2, \ldots, 12$ months, being the EURIBOR time frequency a daily scale. For the analysis, the data is divided in two observed time series:

- Previously to the crisis: From October 15th, 2001 till March 31st, 2006
- During the crisis: From January 2nd, 2008 till November 30th, 2011.

Figures 6.1 and 6.2 display the graphical evolution of the EURIBOR series during the above mentioned periods. As null hypotheses, for the goodness–of–fit tests for the drift and the volatility, a *CKLS* formulation incorporating the stochastic volatility model proposed in Andersen and Lund (1997) is considered:

$$\begin{aligned} dr_t &= \kappa_1(\mu - r_t)dt + \sigma_t r_t^{\gamma} dW_{1,t} \\ d\log(\sigma_t^2) &= \kappa_2(\alpha - \log(\sigma_t^2))dt + \xi dW_{2t} \end{aligned}$$

where W_{1t} and W_{2t} are independent Brownian motions. The Euler scheme is applied to discretize the model and to obtain the first order approximation

$$\begin{aligned} r_{t_i+1} - r_{t_i} &= \kappa_1(\mu - r_{t_i})\Delta + \sigma_{t_i} r_{t_i}^{\gamma} \sqrt{\Delta}\varepsilon_{1,t_i} \\ \log(\sigma_{t_i+1}^2) - \log(\sigma_{t_i}^2) &= \kappa_2(\alpha - \log(\sigma_{t_i}^2))\Delta + \xi\sqrt{\Delta}\varepsilon_{2,t_i} \end{aligned}$$

where $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are independent normal random variables $N(0, 1)$ with fixed Δ and weekly frequency (hence, $\Delta = 1/52$). The corresponding general Model (6.7) is, in this case:

$$u_t = x_t + 2\gamma \log(r_t) - 1.27 + \zeta_t - \log(\Delta)$$

$$x_t = \phi_0 + \phi_1 x_{t-1} + \xi\sqrt{\Delta}\varepsilon_{2,t}$$

where

- $\phi_0 = (1 - \kappa_2\Delta)$, $\phi_1 = \kappa_2\alpha\Delta$
- $\zeta_t = \log(\varepsilon_{1,t}^2) + 1.27$

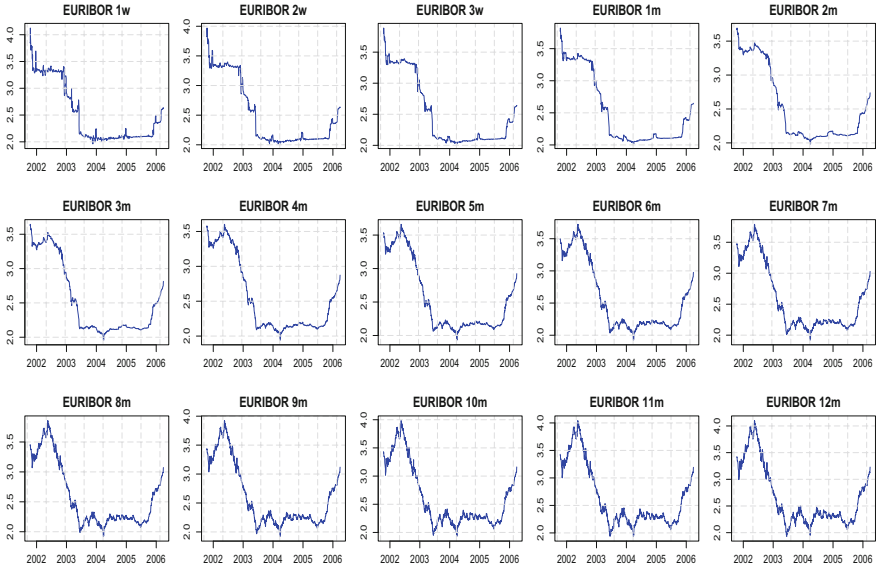


Fig. 6.1 Time series for the interbank deposits in the Eurozone for the period 2001–2006 with maturities of 1, 2, and 3 weeks; and 1, 2, . . . , 12 months

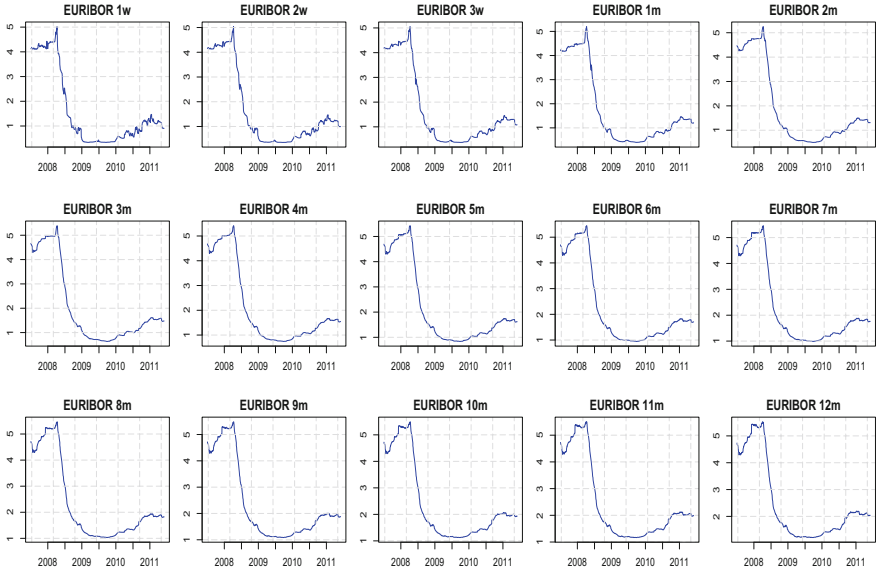


Fig. 6.2 Time series for the interbank deposits in the Eurozone for the period 2008–2011 with maturities of 1, 2, and 3 weeks; and 1, 2, . . . , 12 months

Table 6.2 p-values associated to the goodness–of–fit test for the drift and volatility functions of the stochastic volatility model adjusted to the EURIBOR series *before* the crisis

Maturity	GOF-Drift		GOF-Volatility	
	$\hat{p}_{value}$	$D_n(r,x)$	$\hat{p}_{value}$	$V_n(r,x)$
1 week	0.036	28.634	0.126	3.717
2 week	0.660	21.603	0.098	2.586
3 week	0.754	27.816	0.984	0.017
1 month	0.752	37.673	0.998	0.003
2 month	0.194	57.085	0.996	0.002
3 month	0.054	51.173	0.990	0.003
4 month	0.186	38.473	0.990	0.004
5 month	0.246	38.926	0.982	0.007
6 month	0.442	33.302	0.990	0.009
7 month	0.238	35.567	0.948	0.013
8 month	0.132	38.118	0.870	0.019
9 month	0.088	38.478	0.766	0.026
10 month	0.050	39.023	0.652	0.033
11 month	0.076	39.375	0.528	0.037
12 month	0.034	42.505	0.440	0.050

- $u_t = \log(e_t^2)$ and $e_t = Y_t/\Delta - m_1(r_t, \theta) = Y_t/\Delta - \kappa_1\,(\mu - r_t)$
- $x_t = \log(\sigma_t^2)$.

Based on the above state space model, the proposed tests are applied. Model parameters are estimated by maximum likelihood and Kalman filtering procedures are applied to obtain the non observable variable x_t, required for performing the tests. $B = 500$ bootstrap copies are generated to approximate the distribution of the corresponding processes involved in the tests contruction and estimate the empirical p-values. The results collected in Tables 6.2 and 6.3 were obtained applying the resampling scheme described in Sect. 6.4. It can be noted that the p-values associated to the tests for the drift and the volatility of the EURIBOR series, except for a few cases of maturity, *do not* reject the null hypothesis for the periods before the crisis and during the crisis. Therefore, the *CKLS* model, incorporating the volatility factor, is capable of characterizing such series. It is important to emphasize that the *CKLS* model considered in Monsalve-Cobis et al. (2011), without taking into account the stochastic volatility, was rejected in a conclusive way for the volatility component.

In the light of the results, incorporating a stochastic model for the volatility function in a more flexible way seems to allow for a more effective characterization of the EURIBOR series.

Table 6.3 p-values associated to the goodness–of–fit test for the drift and volatility functions of the stochastic volatility model adjusted to the EURIBOR series *during* the crisis

Maturity	GOF-Drift		GOF-Volatility	
	$\hat{p}_{value}$	$D_n(r,x)$	$\hat{p}_{value}$	$V_n(r,x)$
1 week	0.222	25.947	0.460	0.198
2 week	0.898	19.519	0.616	0.159
3 week	0.748	23.504	0.402	0.171
1 month	0.996	16.458	0.870	0.129
2 month	0.828	20.524	0.544	0.177
3 month	0.794	19.166	0.596	0.258
4 month	0.476	21.579	0.336	0.310
5 month	0.040	25.021	0.072	0.418
6 month	0.140	27.131	0.096	0.468
7 month	0.436	23.855	0.976	0.013
8 month	0.094	29.190	0.104	0.506
9 month	0.064	28.640	0.066	0.869
10 month	0.056	27.945	0.118	0.656
11 month	0.006	28.735	0.156	0.646
12 month	0.166	26.243	0.098	0.767

Acknowledgements The work by Wenceslao González–Manteiga and Manuel Febrero–Bande was partially supported by grant MTM2013-41383-P from Ministerio de Economía y Competitividad, Spain. The work by Jorge P. Zubelli was supported by CNPq through grants 302161/2003-1 and 474085/2003-1, and from FAPERJ through the programs *Cientistas do Nosso Estado* and *Pensa Rio*. The authors want to thank the work by two anonymous referees for their constructive comments and the special collaboration of the Prof. Rosa Crujeiras Casais for her suggestions/improvements in the final version of this paper.

References

Aït-Sahalia, Y., 1993. Nonparametric Functional Estimation with Applications to Financial Models. Ph.D. thesis. Department of Economics Massachusetts Institute of Technology.

Aït-Sahalia, Y., Kimmel, R., 2007. Maximum likelihood estimation of stochastic volatility models. Journal of Financial Economics 83, 413–452.

Andersen, T.G., Lund, J., 1997. Estimating continuous-time stochastic volatility models of the short-term interest rate. Journal of Econometrics 77, 343–377.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–54.

Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press.

Chan, K.C., Karolyi, G.A., Longstaff, F.A., Sanders, A.B., 1992. An empirical comparison of alternative models of the short-term interest rate. Journal of Finance 47, 1209–1227.

Dette, H., Podolskij, M., Vetter, M., 2006. Estimation on integrated volatility in continuous-time financial models with applications to goodness-of-fit testing. Scandinavian Journal of Statistics 33, 259–278.

Dette, H., von Lieres und Wilkau, C., 2003. On a test for a parametric form of volatility in continuous time financial models. Finance and Stochastics 7, 363–384.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations. Econometrica 50, 987–1007.

Genon-Catalot, V., Jeantheau, T., Laredo, C., 1999. Parameter estimation for discretely observed stochastic volatility models. Bernoulli 5, 855–872.

Ghyles, E., Harvey, A.C., Renault, E., 1996. Stochastic volatility, in: Statistical Methods in Finance. North Holland, pp. 119–191.

Harvey, A., Ruiz, E., Shephard, N., 1994. Multivariate stochastic variance models. Review of Economic Studies 61, 247–264.

Hull, J., White, A., 1987. The pricing of options on assets with stochastic volatilities. Journal of Finance 42, 281–300.

Iacus, S.M., 2008. Simulation and Inference for stochastics Differencial Equations. Springer.

Kessler, M., 2000. Simple and explicit estimating functions for a discretely observed diffusion process. Scandinavian Journal of Statistics 27, 65–82.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. The Review of Economic Studies 65, 361–393.

Lin, L.C., Lee, S., Guo, M., 2013. Goodness-of-fit test for stochastic volatility models. Journal of Multivariate Analysis 116, 473–498.

Mandelbrot, B., 1963. The variation of certain speculative prices. Journal of Business 36, 394–419.

Markowitz, H., 1952. Portfolio selection. The Journal of Finance 7, 77–91.

Merton, R., 1973. The theory of rational option pricing. Bell Journal of Economics 4, 141–183.

Monsalve-Cobis, A., González-Manteiga, W., Febrero-Bande, M., 2011. Goodness of fit test for interest rates models: an approach based on empirical processes. Computational Statistics & Data Analysis 25, 3073–3092.

Nisticò, S., 2007. Measurement Errors and the Kalman Filter: A Unified Exposition. LL EE Working Document 45. Luiss Lab of European Economics.

Officer, R.R., 1973. The variability of the market factor of the New York Stock Exchange. The Journal of Business 46, 434–453.

Rodriguez, A., Ruiz, E., 2012. Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. Computational Statistics & Data Analysis 56, 62–74.

Shephard, N., 2005. Stochastic volatility. Working Paper. Nuffield College (University of Oxford).

Shumway, R., Stoffer, D., 2011. Time Series Analysis and Its Applications: With R Examples. Springer Texts in Statistics. 3rd edition.

Stute, W., 1997. Nonparametric model checks for regression. Annals of Statistics 25, 613–641.

A Review on Dimension-Reduction Based Tests For Regressions

7

Xu Guo and Lixing Zhu

7.1 Introduction

Regression analysis is a powerful tool to describe the relationship between response and predictors. In practice, parametric regression models such as linear regression models are widely used due to their simplicity and interpretability. When a parametric regression model is correctly fitted by the data, further statistical analysis can be easily and accurately elaborated with good explanations. However, such further analysis and interpretation could be misleading when the model does not fit the data well. It is therefore necessary to assess the suitability of a parametric model and this is often performed through a goodness-of-fit test. A practical example is production theory in economics, in which the Cobb-Douglas function is commonly used to describe the linear relationship between the log-inputs, such as labor and capital, and the log-output. However, this function may not well describe the relationship. To avoid model mis-specification, Kumbhakar et al. (2007) suggested a semi-parametric regression model to fit data. Zhang and Wu (2011) and Lin et al. (2014) respectively

L. Zhu—This article is dedicated to Winfried Stute for his 70th birthday. The authors wish him uncountable years of good health. The authors thank the Editor and two referees for their constructive comments and suggestions, which led to the substantial improvement of an early manuscript. The research described here was supported by a grant from the University Grants Council of Hong Kong, Hong Kong, Natural Science Foundation of China (NSFC11671042, 11601227) and Natural Science Foundation of Jiangsu Province, China (BK20150732).

X. Guo · L. Zhu
School of Statistics, Beijing Normal University, Beijing, China

L. Zhu (✉)
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
e-mail: lzhu@hkbu.edu.hk

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_7

worked on tests of parametric functional form in nonstationary time series and fixed effects panel data models. In summary, when describing the regression relationship between responses and predictors, we may have to make a choice between the simple but fragile parametric model and the flexible but more complicated model. This is a typical scenario in the initial steps of statistical analysis. In this review, we focus on independent data.

There are a number of methods available for testing a parametric regression model against a general nonparametric model. Examples include the following. Härdle and Mammen (1993) considered the L_2 distance between the null parametric regression and the alternative nonparametric regression as the base of their test statistic construction. Zheng (1996) proposed a quadratic form of the conditional moment test that was also independently developed by Fan and Li (1996). Dette (1999) developed a test based on the difference between variance estimates under the null and alternative models. See also Fan et al. (2001) and Zhang and Dette (2004). We call these tests local smoothing tests as they require nonparametric local smoothing methods. Another main class of tests is based on empirical processes. For instance, Bierens (1982; 1990) suggested some tests that are based on weighted residual moment with characteristic function weights. Stute (1997) introduced a nonparametric principal component decomposition based on a residual marked empirical process. Inspired by the Khmaladze transformation used in goodness-of-fitting for distributions, Stute et al. (1998b) first developed the innovation martingale approach to obtain distribution-free tests. Khmaladze and Koul (2009) studied the goodness-of-fit problem for errors in nonparametric regression. We call this class of tests global smoothing tests because they are actually based on the weighted averages of residuals and averaging itself is a global smoothing step. We may pay special attention to González-Manteiga and Crujeiras (2013a; 2013b) who provided a comprehensive review of the literature on the lack-of-fit and goodness-of-fit testing for regression models with modest number of covariates. In the present review, we focus on dimension reduction type tests.

The rest of this review is organized as follows. In Sect. 7.2, some basic ideas of constructing tests are briefly reviewed. Section 7.3 presents various tests designed for avoiding the curse of dimensionality. In Sect. 7.4, the important Stute's contributions other than the results of dimension reduction nature are very briefly reviewed. Finally some conclusions and discussions are presented in Sect. 7.5.

7.2 Basic Ideas of Test Statistics Construction

We first review local smoothing tests and global smoothing tests that do not involve special dimension reduction strategies.

7.2.1 Local Smoothing Based Tests

Consider a regression model $Y = m(X) + \epsilon$ with $\{(X_i, Y_i)\}_{i=1}^n$ being a random sample of (X, Y). Here Y is a scalar response, X is a predictor vector of p-dimension and $m(\cdot)$ is an unknown function. For a given parametric function $g(\cdot, \cdot)$, the goal is to test the null hypothesis:

$$H_0 : P(m(X) = g(X, \theta_0)) = 1, \tag{7.1}$$

for some unknown $\theta_0 \in \Theta$ for a parameter space Θ in the Euclidean space $\mathbb{R}^d$ against the alternative hypothesis

$$yH_1 : P(m(X) = g(X, \theta)) < 1, \tag{7.2}$$

for any $\theta \in \Theta$. In the following, denote $\epsilon_0 = Y - g(X, \theta_0)$, the random error. Under the null hypothesis, the conditional expectation of ϵ_0 given X is zero: $E(\epsilon_0|X) = 0$, while under the alternative, it is not zero.

Local smoothing tests are based on estimating $E(\epsilon_0|X)$ which requires local smoothing methods such as the Nadaraya-Watson kernel estimator. Let the local weight function be

$$W_{ni}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^n K_h(x - X_j)},$$

where $K_h(\cdot) = K(\cdot/h)/h^p$ with $K(\cdot)$ being a kernel function and h is the bandwidth.

Härdle and Mammen's (1993) test is based on the L_2 distance between parametric and nonparametric estimators. The test statistic has the form:

$$T_{HM} = \int \left(\sum_{i=1}^n W_{ni}(x) \hat{\epsilon}_{0i} \right)^2 \omega(x) dx. \tag{7.3}$$

Here $\hat{\epsilon}_{0i} = Y_i - g(X_i, \hat{\theta})$ and $\omega(\cdot)$ is some positive weight function. $\hat{\epsilon}_{0i}$ is the residual with an estimator $\hat{\theta}$ of θ_0. The estimator $\hat{\theta}$, e.g. the nonlinear least squares estimator, can be $\sqrt{n}$-consistent.

It can be shown that the limiting null distribution of T_{HM} is

$$\begin{aligned} & nh^{p/2} \left(T_{HM} - (nh^p)^{-1} \int K^2(x) dx \int \frac{\sigma^2(x)\omega(x)}{f(x)} dx \right) \\ & \Rightarrow N\left(0, 2 \int (K * K)^2 dx \int \frac{\sigma^4(x)\omega^2(x)}{f^2(x)} dx \right), \end{aligned} \tag{7.4}$$

here $f(x)$ is the density of the predictor vector X, $\sigma^2(x) = Var(Y|X = x)$ is the conditional variance, the symbol $*$ denotes the convolution operator and $K * K(x) = \int K(t)K(x - t)dt$. As the significance level cannot be well maintained when the

limiting null distribution is used to determine critical values, Monte Carlo approximation/bootstrap to its sampling null distribution is required.

Zheng (1996) developed a quadratic conditional moment test which was also independently proposed by Fan and Li (1996). The test statistic is a consistent estimate of $E[\epsilon_0 E(\epsilon_0|X) f(X)\omega(X)]$, which is zero under the null hypothesis and positive under the alternative hypotheses. The test statistic is defined by

$$T_{ZH} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(X_i - X_j)\hat{\epsilon}_{0i}\hat{\epsilon}_{0j}\omega(X_i). \tag{7.5}$$

Under the null hypothesis, the asymptotic normality is provided as:

$$nh^{p/2} T_{ZH} \Rightarrow N\left(0, 2\int K^2(x)dx \int \sigma^4(x)\omega^2(x)f^2(x)dx\right). \tag{7.6}$$

A main advantage of T_{ZH}, compared to T_{HM}, is the asymptotic unbiasedness and then with no need of bias-correction.

By noticing that $E\left(\left[\epsilon_0^2 - (\epsilon_0 - E(\epsilon_0|X))^2\right]\omega(x)\right)$ is zero under H_0, Dette (1999) introduced a test statistic based on the difference between the error variance estimates under the null and alternative hypotheses. The test statistic is defined as

$$T_{DE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i, \hat{\theta}))^2\omega(X_i) - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2\omega(X_i). \tag{7.7}$$

Here $\hat{m}(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i$ is the nonparametric estimator of the regression function. Denote $K^{2*} = 2K - K * K$. Then the asymptotic distribution of the variance difference statistic T_{DE} is

$$\begin{aligned} &nh^{p/2}\left(T_{DE} - (nh^p)^{-1}K^{2*}(0)\int \sigma^2(x)\omega(x)dx\right) \\ &\Rightarrow N\left(0, 2\int K^{2*}(x)dx \int \sigma^4(x)\omega^2(x)dx\right). \end{aligned} \tag{7.8}$$

Similarly as T_{HM}, this test also has a bias term diverging to infinity. For the comparison of these above three methods, see Zhang and Dette (2004).

Inspired by the classical likelihood ratio test, Fan et al. (2001) suggested a generalized likelihood ratio test which resembles the F-test construction for regression models. A significant property of this test is that the limiting null distribution does not depend on nuisance functions, exhibiting what is known as Wilks phenomenon. For more details, see also Fan and Jiang (2007).

Koul and Ni (2004) studied a class of minimum distance tests for multidimensional covariates and heteroscedaticity. These tests are based on certain minimized L_2 distances between a nonparametric regression function estimator and the parametric model being fitted. Different bandwidths for the estimation of the numerator and

denominator in the nonparametric regression function estimator are adopted. Compare with Härdle and Mammen (1993), Koul and Ni (2004)'s tests do not require the null regression function, $g(\cdot, \cdot)$, to be twice continuously differentiable. Thus, they are relatively broadly applicable. Readers may also refer to Koul and Song (2009).

The tests in this class are called local-smoothing tests. The monograph by Hart (1997) that is a comprehensive reference collected many local smoothing tests.

7.2.2 Empirical Process-Based Tests

Another class of tests for model checking is based on the estimator of the integrated function $\mathcal{I}(x) = \int_{-\infty}^{x}(m(t) - g(t, \theta_0))dF(t) = E((Y - g(X, \theta_0))$ $I(X \leq x))$, where $I(\cdot)$ is the indicator function. We can estimate $\mathcal{I}(x)$ by an empirical process defined as

$$\mathcal{I}_n(x) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i, \hat{\theta}))I(X_i \leq x).$$

Based on this empirical process, we can construct Cramér-von Mises or Kolmogorov-Smirnov type tests. Early studies for these types of test statistics are due to Bierens (1982), Su and Wei (1991) and Stute (1997). However, for composite null hypothesis H_0, the above defined test statistics depend on the parametric form of $g(X, \theta)$ and also $\hat{\theta}$, and are not distribution-free. Inspired by the Khmaladze transformation used in goodness of fit for distributions, Stute et al. (1998b) developed the innovation martingale approach to obtain some distribution-free tests in the case of one-dimensional predictor. Khmaladze and Koul (2009) further studied the goodness-of-fit problem for errors in nonparametric regressions.

Van Keilegom et al. (2008) considered continuous functionals of the distance between the empirical distribution of the residuals under the null and the alternative hypotheses. The test statistic that is based on this distance needs the independence between the error term and the predictors. Huskova and Meintanis (2009) considered an alternative route based on the characteristic function requiring weaker conditions than their analogues using empirical distributions. For these two methods, see also Dette et al. (2007) and Huskova and Meintanis (2010).

Compared with other local smoothing tests, these tests can either avoid the selection of bandwidth, or at least depend less on the bandwidth. We call these tests global smoothing tests. These tests are also called empirical process-based tests in the literature. We use the term "global smoothing" because the integral/average over all indices is a global smoothing procedure. Another advantage of global smoothing tests is that the tests in this class can detect local alternatives converging to the null hypothesis with the rate of $n^{-1/2}$, whereas for local smoothing based tests, the optimal rate is $n^{-1/2}h^{-p/4}$. However, we should also mention that the higher detecting rate of empirical process based tests do not mean they can generally have larger power compared with the local smoothing tests that can be more sensitive

to oscillating/high-frequency models. Global smoothing tests generally yield low powers against high-frequency alternatives, see Fan and Li (2000) and Guo et al. (2016). Thus the two schools of local and global smoothing tests could be viewed as complementing each other.

7.3 Tests Designed to Avoid the Curse of Dimensionality

As commented above, existing local smoothing tests could be more powerful for detecting high-frequency regression models. However, a very obvious and serious shortcoming is that these methods suffer severely from dimensionality due to the inevitable use of multivariate nonparametric function estimation. Under the corresponding null hypotheses, existing local smoothing test statistics converge to their limits at the rate $O(n^{-1/2}h^{-p/4})$, which can be very slow when p is large and selecting a proper bandwidth is also an issue. Therefore, the significance level very often cannot be maintained when used with moderate sample size. This problem has been acknowledged in the literature. Thus, even though their limiting null distributions are given, there are still a number of local smoothing tests that use the wild bootstrap or Monte Carlo approximation to help determine critical values (or p values). Examples include Härdle and Mammen (1993), Delgado and González-Manteiga (2001) and Dette et al. (2007). In contrast, most of the existing global smoothing methods depend on high-dimensional stochastic processes (see, e.g., Stute et al. 1998a). Their power performance often drops significantly as p increases due to the data sparseness in high-dimensional space. A recent reference is Guo et al. (2016).

These difficulties lead to different modifications of the previous methods in order to avoid the curse of dimensionality. In the following three subsections, we will respectively focus on projection-pursuit based tests, sufficient dimension-reduction based tests and other relevant tests. For projection-pursuit based tests, the original covariates X are first projected onto to be $\beta^\top X$ for all $||\beta|| = 1$ in the unit sphere $\mathbb{S}^p = \{\beta \in \mathbb{R}^p : ||\beta|| = 1\}$. Here "$\|a\|$" is the Euclidean norm of vector a. The test statistics are then constructed based on $\beta^\top X$. Since there are infinitely many β satisfying $||\beta|| = 1$, the resulting tests are either supremum or integral over all $\beta \in \mathbb{S}^p$. Both supremum and integral can be approximated by using Monte Carlo approximation in general. Instead of considering infinitely many directions β in $\mathbb{S}^p$, sufficient dimension-reduction based tests can automatically adopt the dimension reduction structure of X under the null and alternative hypotheses. It is worth pointing out that projection-pursuit based tests can handle more general hypotheses than those sufficient dimension-reduction based tests. For example, projection-pursuit based tests can handle general nonlinear regression models as hypothetical models. But the latter can be more efficient than the former when a dimension reduction structure does exist. More specifically, sufficient dimension-reduction based tests can fully utilize the dimension reduction structures in the respective hypothetical models such as the one specified in the null hypothesis (7.13) in Sect. 7.3.1 below. However, due to the involvement of infinite directions, projection-pursuit based tests are more computa-

tional intensive and the powers of these tests are often lower. For more details about the computational burden of the projection-pursuit based tests, readers may refer to Zhu et al. (2017). Other tests designed to avoid the curse of dimensionality are also reviewed.

7.3.1 Projection-Pursuit Based Tests

To the best of our knowledge, the first effort in this direction would be traced back to Zhu and An (1992), which was motivated by the projection pursuit technique (see a review paper by Huber (1985)). They proposed the following test statistic:

$$K_n = \arg \inf_{||\beta||=1} \hat{S}_n(\beta)/\hat{\sigma}^2. \tag{7.9}$$

Here $\hat{\epsilon}_{0i} = Y_i - g(X_i, \hat{\theta}_i)$, $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \hat{\epsilon}_{0i}^2$ and

$$\hat{S}_n(\beta) = n^{-1}\sum_{i=1}^n (\hat{\epsilon}_{0i} - \hat{m}_\beta^i(X_i))^2;$$
$$\hat{m}_\beta^i(x) = \sum_{j\neq i} K(\beta^\top(x - X_j))\hat{\epsilon}_{0j} / \sum_{j\neq i} K(\beta^\top(x - X_j)).$$

The authors proved that in probability under H_0, $K_n \to k = 1$, whereas under H_1, $K_n \to k < 1$. They reject H_0 if $K_n < 1 - n^{-C/2}$. Here $C > 0$ is a pre-specified constant.

Xia (2009) considered some more general null hypotheses. To be precise, the null hypothesis can be other more complex models, such as partial linear model, single index model, etc. The alternative hypothesis is that the assumed model in the null hypothesis does not hold. Xia (2009) projected the covariate X in the direction of $\beta = \beta_1$ such that β_1 (with $||\beta_1|| = 1$) minimizes $E(\epsilon_0 - E(\epsilon_0|\beta^\top X))^2 = E(\epsilon_0 - m_\beta(X))^2$ over all β. This enables us to construct a test statistic as

$$T_n = S_n / TSS_n. \tag{7.10}$$

Here $TSS_n = n^{-1}\sum_{i=1}^n (\hat{\epsilon}_{0i} - \bar{\epsilon})^2$ with $\bar{\epsilon} = n^{-1}\sum_{i=1}^n \hat{\epsilon}_{0i}$,

$$S_n = \frac{1}{n}\sum_{i=1}^n (\hat{\epsilon}_{0i} - \hat{m}_{\hat{\beta}_i}(X_i))^2,$$

$$\hat{\beta}_i = \arg \min_{||\beta||=1} \sum_{j\neq i} (\hat{\epsilon}_{0j} - \hat{m}_\beta^i(X_j))^2,$$

where $\hat{\epsilon}_{0i}$ and $\hat{m}_\beta^i(x)$ are defined before. The null hypothesis is rejected if $T_n < 1$. Xia (2009)'s work enhanced and extended Zhu and An's (1992) test. Note that for

these two tests, the sampling and limiting null distributions are not given and the probabilities of rejection under the null hypothesis tend to zero in the large sample sense. Thus, they can not be used to test significance at any nominal level.

Later, Zhu and Li (1998) gave the following lemma.

Lemma 1 *A necessary and sufficient condition for H_0 to hold is that for any vector $\beta \in \mathbb{R}^p$ with $||\beta|| = 1$,*

$$E(\epsilon_0|\beta^\top X) = 0 \text{ a.s. for some } \theta_0 \in \Theta.$$

The authors noted that H_0 is equivalent to

$$\int_{\mathbb{S}^p} R(\beta) d\mu(\beta) = 0,$$

where $R(\beta) = E[E(\epsilon_0|\beta^\top X)]^2$ and $\mu(\cdot)$ is the uniform distribution on $\mathbb{S}^p$. They proposed using an unweighed integral of expectations conditional on single linear indices for checking a linear regression model. They did not give much distributional details of this test. Instead, their test was based on the empirical version of the above integral plus a directional test with the form $n^{-1}\sum_{i=1}^n \hat{\epsilon}_{0i}\phi(||X_i||)$, where $\phi(\cdot)$ is the univariate standard normal density. Thus their test is actually a combination of local smoothing test and a directional test. They found that the limiting null distribution of the test statistic can be determined by the latter term.

From a lemma similar to the above, Escanciano (2006) realized that consistent tests for H_0 can be based on one-dimensional projections. This makes the above idea much more implementable. To be precise, the following equivalence holds:

$$H_0 \Leftrightarrow E[\epsilon_0 I(\beta^\top X \le t)] = 0 \text{ a.s. for any } ||\beta|| = 1, t \in \mathbb{R}.$$

This leads to an empirical process

$$R_n(\beta, t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - g(X_i, \hat{\theta})) I(\beta^\top X_i \le t)$$

indexed in β and t. Specifically, the Kolmogorov-Smirnov and the Cramrvon Mises tests can be extended to this setting, with the following statistics:

$$\begin{aligned} T_{nKS} &= \sup_t \sup_{||\beta||=1} |n^{-1/2} \sum_{i=1}^n (Y_i - g(X_i, \hat{\theta})) I(\beta^\top X_i \le t)|, \\ T_{nCM} &= \int_{||\beta||=1} \int_{-\infty}^{\infty} [n^{-1/2} \sum_{i=1}^n (Y_i - g(X_i, \hat{\theta})) I(\beta^\top X_i \le t)]^2 dF_{n\beta}(t) d\omega(\beta). \end{aligned} \tag{7.11}$$

Here $F_{n\beta}$ is the empirical distribution of $\{\beta^\top X_i\}_{i=1}^n$ and ω is a weight function over the projection direction.

Compared with Zhu and Li (1998), Escanciano (2006) proved weak convergence of the related empirical process and the corresponding test statistics. Further, power study under local alternatives converging to the null at the rate of order $1/\sqrt{n}$ was also investigated in his paper. A bootstrap approximation is implemented to determine critical values. Lavergne and Patilea (2008; 2012) also used projections to construct test statistics to improve on the performance of local smoothing tests, particularly, Zheng (1996)'s test.

Lavergne and Patilea (2012) gave the following lemma that can be deduced from Lemma 1.

Lemma 2 *A necessary and sufficient condition for H_0 to hold is that for any vector $\beta \in \mathbb{R}^p$ with $||\beta|| = 1$,*

$$E[\epsilon_0 E(\epsilon_0|\beta^\top X) f_\beta(\beta^\top X)] = 0 \text{ a.s. for some } \theta_0 \in \Theta$$

where $f_\beta(\cdot)$ is the density of $\beta^\top X$.

Based on this lemma, the authors first defined

$$Q_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{\epsilon}_{0i} \hat{\epsilon}_{0j} \frac{1}{h} K(\beta^\top (X_i - X_j)),$$

as an estimator of $E[\epsilon_0 E(\epsilon_0|\beta^\top X) f_\beta(\beta^\top X)]$. This statistic is the one studied by Zheng (1996) applied to the index $\beta^\top X$. The resulting test statistic is as follows:

$$T_{LP1} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{\epsilon}_{0i} \hat{\epsilon}_{0j} \frac{1}{h} \int_{||\beta||=1} K(\beta^\top (X_i - X_j)) d\beta. \tag{7.12}$$

Lavergne and Patilea (2012) proved that under certain conditions, the test statistic under the null hypothesis converges to its limit at a faster rate of order $O(n^{-1/2}h^{-1/4})$ rather than $O(n^{-1/2}h^{-p/4})$, and it is consistent against any global alternative hypothesis and can detect local alternatives distinct from the null at the rate of order $O(n^{-1/2}h^{-1/4})$. This improvement is significant particularly when p is large because the new test does behave like a local smoothing test as if X was one-dimensional. Thus in theory, the test could well maintain the significance level with better power performance than Zheng's test. Lavergne and Patilea (2008) considered a test based on $Q_n(\hat{\beta}_n)$, with

$$\hat{\beta}_n = \arg \max_{||\beta||=1} nh^{1/2} Q_n(\beta) - \alpha_n I(\beta \neq \beta^*)$$

where β^* represents a favored direction and α_n is a slowly diverging penalty sequence. Their procedure allows incorporation of some information on the preferred single-index alternative, as defined through β^*, but introduces a supplementary user chosen parameter α_n.

Ma et al. (2014) investigated the integrated conditional moment test for partially linear single index models incorporating dimension-reduction. Conde-Amboage et al. (2015) studied the lack of fit test for quantile regression models with high-dimensional covariates by following the concepts proposed by Escanciano (2006).

7.3.2 Sufficient Dimension-Reduction Based Tests

For models with dimension reduction structure, Guo et al. (2016) developed a dimension reduction model-adaptive approach to avoid the dimensionality difficulty. The main idea is to fully utilize the dimension reduction structure of X under the null hypothesis, but to adapt the alternative model such that the test is still omnibus. To achieve this goal, sufficient dimension reduction (SDR) technique is adopted.

Consider the hypotheses as follows:

$$\begin{aligned} H_0: &\quad \exists \beta_0 \in \mathbb{R}^p, \theta_0 \in \mathbb{R}^d, \text{ such that, } P(m(X) = g(\beta_0^T X, \theta_0)) = 1; \\ H_1: &\quad \nexists \beta \in \mathbb{R}^p, \nexists \theta \in \mathbb{R}^d, \text{ such that, } P(m(X) = g(\beta^T X, \theta)) < 1. \end{aligned} \tag{7.13}$$

Here $g(\cdot, \cdot)$ is a known parametric function, β_0 is a p-dimensional unknown index vector and θ_0 is a d-vector of parameters. Note that, for any $p \times p$ orthonormal matrix B, $G(X) = G(BB^T X) := \tilde{G}(B^T X)$. Based on this observation, consider a parsimonious alternative model that is widely used in SDR:

$$Y = G(B^T X) + \eta, \tag{7.14}$$

where B is a $p \times q$ orthonormal matrix with q orthogonal columns for an unknown number q with $1 \le q \le p$, G is an unknown smooth function and $E(\eta|X) = 0$. When $q = p$, this model is a purely nonparametric regression model. This implies that there always exists a $p \times q$ matrix $B, 1 \le q \le p$, such that $m(X) = E(Y|B^\top X)$. Under the null hypothesis, $q = 1$ and then $B = \beta_0/||\beta_0||$, and under the alternative, $q \ge 1$. Let $\epsilon_0 = Y - g(\beta_0^\top X, \theta_0)$. Then, under H_0,

$$E\{\epsilon_0 E(\epsilon_0|B^\top X) W(B^\top X)\} = E\{E^2(\epsilon_0|B^\top X) W(B^\top X)\} = 0, \tag{7.15}$$

where $W(X)$ is some positive weight function that is discussed below.

Under H_1, $E(\epsilon_0|B^\top X) = E(Y|B^\top X) - g(\beta_0^\top X, \theta_0) \neq 0$ and thus

$$E\{\epsilon_0 E(\epsilon_0|B^\top X) W(B^\top X)\} = E\{E^2(\epsilon_0|B^\top X) W(B^\top X)\} > 0. \tag{7.16}$$

The empirical version of the left hand side in (7.15) can be used as a test statistic, and H_0 will be rejected for large values of the test statistic. A non-standardized test statistic is defined by

$$V_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{\epsilon}_{0i} \hat{\epsilon}_{0j} K_h\{\hat{B}(\hat{q})^\top (X_i - X_j)\}. \tag{7.17}$$

Here, $\hat{\epsilon}_{0j} = Y_j - g(\hat{\beta}^\top X_j, \hat{\theta})$, $\hat{\beta}$ and $\hat{\theta}$ are the commonly used least squares estimates of β_0 and θ_0, $\hat{B}(\hat{q})$ is a sufficient dimension reduction estimate with an estimated structural dimension $\hat{q}$ of q, $K_h(\cdot) = K(\cdot/h)/h^{\hat{q}}$ with $K(\cdot)$ being a $\hat{q}$-dimensional kernel function and h being a bandwidth.

Remark 1 The test statistic suggested by Zheng (1996) is

$$T_{ZH} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{\epsilon}_{0i} \hat{\epsilon}_{0j} \tilde{K}_h(X_i - X_j). \tag{7.18}$$

Here $\tilde{K}_h(\cdot) = \tilde{K}(\cdot/h)/h^p$ with $\tilde{K}(\cdot)$ being a p-dimensional kernel function. Comparing Eq. (7.17) with Eq. (7.18), there are two main differences. First, V_n uses $\hat{B}(\hat{q})^\top X$ rather than X itself in T_{ZH} and applies $\hat{q}$th order kernel $K_h(\cdot)$ instead of the pth order kernel $\tilde{K}_h(\cdot)$. This reduces the dimension p down to $\hat{q}$. Second, under H_0, we will show that with a probability going to one, $\hat{q} = 1$, and $\hat{B}(\hat{q}) \to \beta_0/||\beta_0||_2$, and further using the normalizing constant $nh^{1/2}$ to get that $nh^{1/2}V_n$ has a finite limit. Under the alternative model (7.14), we will show that $\hat{q} = q \geq 1$ with a probability going to one and $\hat{B}(\hat{q}) \to BC$ for a $q \times q$ orthonormal matrix C. Further, comparing with projection-pursuit based tests, it fully uses the dimension reduction structure under the null hypothesis and thus only one projection needs to be used. The significance level can be easily maintained, see Guo et al. (2016). The theoretical results show that the test is consistent against any global alternative and can detect local alternatives distinct from the null at the rate of order $O(n^{-1/2}h^{-1/4})$. This improvement is significant particularly when p is large.

As estimating the matrix B and its dimension is crucial, the authors adopted two popular sufficient dimension reduction methods, discretization-expectation estimation (DEE, Zhu et al. 2010) and minimum average variance estimation (MAVE, Xia et al. 2002) to estimate the matrix B. Moreover the authors applied two BIC criteria to determine the dimension q.

This methodology can be readily applied to many other testing methods and problems. Recently, Niu et al. (2016) developed a model-adaptive enhancement of the nonparametric generalized likelihood ratio (GLR) test. Fan et al. (2001) proposed the following GLR test:

$$\Lambda_n = \frac{n}{2} \log \frac{RSS_0}{RSS_1} \approx \frac{n}{2} \frac{RSS_0 - RSS_1}{RSS_1}.$$

Here $RSS_0 = n^{-1} \sum_{i=1}^{n} (Y_i - g(\hat{\beta}^\top X_i, \hat{\theta}))^2$; $RSS_1 = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{m}(X_i))^2$. When the dimension of predictors is high, it can not well control type I error and has very low power. Another drawback of the original GLR is that there is a bias term in its limiting null distribution which can cause the test not to well control type I error and thus bootstrap or Monte Carlo approximation for critical value determination is

required. In a similar framework to Guo et al. (2016), Niu et al. (2016) replaced the above RSS_1 by

$$\widetilde{\mathrm{RSS}}_1 = \sum_{i=1}^{n} \left| [Y_i - \hat{G}(\hat{B}(\hat{q})^\top X_i)][Y_i - g(\hat{\beta}^\top X_i, \hat{\theta})] \right|,$$

where, $|\cdot|$ denotes the absolute value and the leave-one out kernel estimator $\hat{G}(\hat{B}(\hat{q})^\top X)$ of $G(B^\top X)$ is applied:

$$\hat{G}(\hat{B}(\hat{q})^\top X_i) = \frac{\sum_{j\neq i}^{n} K\{\hat{B}(\hat{q})^\top (X_i - X_j)/h\} Y_j}{\sum_{j\neq i}^{n} K\{\hat{B}(\hat{q})^\top (X_i - X_j)/h\}}. \tag{7.19}$$

Finally, a test statistic that is asymptotically unbiased is defined as:

$$\tilde{T}_n = \frac{n}{2} \frac{\mathrm{RSS}_0 - \widetilde{\mathrm{RSS}}_1}{\widetilde{\mathrm{RSS}}_1}. \tag{7.20}$$

The proposed test statistic still possesses the Wilks phenomenon, and behaves like a test with only one covariate. Thus the null distribution of $\tilde{T}_n$ converges weakly to its limit at a much faster rate. Moreover, this test is much more sensitive to alternative models than the original nonparametric GLR test. In fact, the test can detect the local alternatives distinct from the null at the rate of order $n^{-1/2}h^{-1/4}$ while the GLR test can only detect those converging to the null at the rate of order $n^{-1/2}h^{-p/4}$.

Note that both Guo et al. (2016) and Niu et al. (2016) are improvements over classical local smoothing tests. On the other hand, Tan et al. (2017) proposed a projection-based adaptive-to-model approach to improve the performance of global based tests when the number of covariates is greater than one. The procedure is an interesting combination of the projection-pursuit based and sufficient dimension-reduction based tests. First, under the null hypothesis,

$$E[\epsilon_0 I(\beta_0^\top X \leq t)] = 0.$$

Here $\epsilon_0 = Y - g(\beta_0^\top X, \theta_0)$ is defined as before. For the parametric single-index model (7.13), Stute and Zhu (2002) considered the following empirical process

$$R_n(t) = n^{-1/2} \sum_{i=1}^{n} (Y_i - g(\hat{\beta}^\top X_i, \hat{\theta})) I(\hat{\beta}^\top X_i \leq t).$$

Further, they recommended the martingale transformation to derive the asymptotically distribution-free property. In this test, $\hat{\beta}$ is an estimate of β_0 that is the index under the null hypothesis. Therefore, it is a directional test rather than an omnibus test. Guo et al. (2016) provided a simple example to show this property. The resulting test statistic by Escanciano (2006) involves all directions β in $\mathbb{S}^p$ and thus, the omnibus property can be held.

To benefit from Stute and Zhu's (2002) idea, but make a test omnibus, Tan et al. (2017) considered the following strategy. According to lemma 1, under the alternative hypothesis, for an $\alpha \in \mathbb{R}^q$ with $||\alpha|| = 1$,

$$E[\epsilon_0 I(\alpha^\top B^\top X \le t)] \neq 0.$$

The authors then used SDR to estimate B and its dimension q. An adaptive-to-model residual marked empirical process was suggested by the authors as follows:

$$V_n(t, \hat{\alpha}) = n^{-1/2} \sum_{i=1}^{n} [Y_i - g(\hat{\beta}^\top X_i, \hat{\theta})] I(\hat{\alpha}^\top \hat{B}(\hat{q}) X_i \le t),$$

$$V_n(t) = \sup_{\hat{\alpha} \in \mathbb{S}^{\hat{q}}} V_n(t, \hat{\alpha}),$$

where $\mathbb{S}^{\hat{q}} = \{\hat{\alpha} \in R^{\hat{q}} : \|\hat{\alpha}\| = 1, \hat{\alpha}_1 \ge 0\}$ with $\hat{\alpha}_1$ being the first component of $\hat{\alpha}$. The significant feature of this projection-based test is as follows. Under the null hypothesis, $\mathbb{S}^{\hat{q}}$ has only one element $\hat{\alpha} = 1$ and thus the supremum over all elements in $\mathbb{S}^{\hat{q}}$ is $V_n(t, 1)$ that is asymptotically equal to Stute and Zhu's (2002) test. Therefore, a martingale transformation that was proposed by Stute et al. (1998b) is easily performed to derive the asymptotically distribution-free property. Under the alternative hypothesis, the test is still omnibus. This is a substantial improvement over Stute and Zhu's (2002) test.

Zhu et al. (2017) considered a more general null hypothesis of the partially parametric single-index model:

$$Y = g(\beta_0^\top X, W, \theta_0) + \epsilon. \tag{7.21}$$

Here (X, W) is the covariate vector in $\mathbb{R}^{p+s}$, $g(\cdot)$ is a known smooth function that depends not only on the index $\beta^\top X$ but also on the covariate W and the error ϵ follows a continuous distribution and is independent with the covariates (X, W). The model (7.21) reduces to the parametric single-index model in the absence of the covariate W and to the general parametric model in the absence of the index $\beta^\top X$. This structure is often meaningful as in many applications, p is often large while s is not. The alternative model is taken as:

$$Y = G(B^\top X, W) + \eta,$$

where B is a $p \times q$ orthonormal matrix with q orthogonal columns for an unknown number q with $1 \le q \le p$ and $G(\cdot)$ is an unknown smooth function. For identifiability consideration, assume that the matrix B satisfies $B^\top B = I_q$. This model covers many popularly used models in the literature such as the single-index models

with $B = \beta$, the multi-index models with the absence of W, and partial single-index models with the mean function $g_1(\beta^\top X) + g_2(W)$. Similarly as before, let $\epsilon_0 = Y - g(\beta_0^\top X, W, \theta_0)$. Under the null model, we have for all (t, ω)

$$E\left[\epsilon_0 I\{(B^\top X, W) \le (t, \omega)\}\right] = 0,$$

and under the alternative models, for some (t, ω)

$$E\left[\epsilon_0 I\{(B^\top X, W) \le (t, \omega)\}\right] \neq 0.$$

A residual-marked empirical process is defined as

$$V_n(t, \omega) = n^{-1/2} \sum_{i=1}^{n} (Y_i - g(\hat{\beta}^\top X_i, W_i, \hat{\theta})) I\{(\hat{B}(\hat{q})^\top X_i, W_i) \le (t, \omega)\}, \quad (7.22)$$

where $\hat{\beta}$ and $\hat{\theta}$ are the nonlinear least squares estimates of β and θ, respectively. Note that in this model, there is a covariate W that is not in the projected subspace spanned by $B^\top X$ and thus, the sufficient dimension reduction methods for estimating B and q described before does not work. Thus, the partial discretization-expectation estimation (Feng et al. 2013) is used to identify/estimate the matrix B. The resulting estimate $\hat{B}(\hat{q})$ is called the partial sufficient dimension reduction estimate of B with the structural dimension estimate $\hat{q}$ of q where the ridge-type eigenvalue ratio (Xia et al. 2015) is applied to estimate q.

Some other recent developments in this direction can be mentioned. Zhu et al. (2016) proposed a dimension reduction adaptive nonparametric test for heteroskedasticity in nonparametric regression model. Zhu and Zhu (2016) developed a dimension-reduction based adaptive-to-model test for significance of a subset of covariates in the context of a nonparametric regression model. Niu and Zhu (2016) developed a test statistic that is robust against outliers. Koul et al. (2016) provided some useful dimension-reduction based tests for parametric single-index regression models when covariates are measured with error and validation data are available. For further details, readers can refer to the aforementioned references.

7.3.3 Some Other Tests

Lavergne et al. (2015) considered testing the significance of a subset of covariates in a nonparametric regression. The null hypothesis is as follows:

$$H_0 : P(E[Y|W, X] = E[Y|W]) = 1. \quad (7.23)$$

Under this hypothesis, the covariates $X \in \mathbb{R}^p$ are redundant for modelling the relationship between the response and the covariates. This hypothesis is equivalent to

$$H_0 : P(E[u|W, X] = 0) = 1 \quad (7.24)$$

where $u = Y - E[Y|W]$. They give the following lemma to characterize the null hypothesis H_0 using a suitable unconditional moment equation.

Lemma 3 *Let* (W_1, X_1, u_1) *and* (W_2, X_2, u_2) *be two independent copies of* (W, X, u) *and* $v(W)$ *a strictly positive function on the support of* W *such that* $E[u^2v^2(W)] < \infty$, *and* $K(\cdot)$ *and* $\psi(\cdot)$ *are even functions with (almost everywhere) positive Fourier integrable transforms. Define*

$$I(h) = E[u_1u_2v(W_1)v(W_2)h^{-s}K((W_1 - W_2)/h)\psi(X_1 - X_2)].$$

Then for any $h > 0$, $I(h) \geq 0$ *and*

$$E[u|W, X] = 0\, a.s. \Leftrightarrow I(h) = 0 \Leftrightarrow \lim_{h\to 0} I(h) = 0.$$

Suppose that a random sample $\{(Y_i, W_i, X_i), 1 \leq i \leq n\}$ from (Y, W, X) is available. A test statistic is defined as

$$I_n = \frac{2}{(n-1)^2n(n-1)} \sum_a \sum_{k\neq i} \sum_{l\neq j} (Y_i - Y_k)(Y_j - Y_l)L_{nik}L_{njl}K_{nij}\psi_{ij}. \quad (7.25)$$

Here $L_{nik} = g^{-s}L((W_i - W_k)/g)$, $K_{nij} = h^{-s}K((W_i - W_j)/h)$, $\psi(X_i - X_j)$, $L(\cdot)$ and $K(\cdot)$ are two kernel function with bandwidths g and h, respectively.

A remarkable feature of the proposed test is that this is a combination of global and local smoothing test procedure with no local smoothing relative to the covariates X. Unlike the results in Fan and Li (1996) and Lavergne and Vuong (2000) who used a multidimensional smoothing kernel $h^{-(s+p)}\tilde{K}((W_i - W_j)/h, (X_i - X_j)/h)$ over (W, X), it is showed that I_n has a weak limit by multiplying $nh^{s/2}$ rather than $nh^{(s+p)/2}$. This provides a much faster convergence rate than those in Fan and Li (1996) and Lavergne and Vuong (2000).

The choice of the function $\psi(\cdot)$ is flexible. There are many functions that possess an almost everywhere positive and integrable Fourier transform. Examples include (products of) the triangular, normal, Laplace, logistic and Student density. This idea has been applied by Patilea and his coauthors to address the issue of lack-of-fit testing for a parametric quantile regression (Maistre et al. 2017) and nonparametric model checks for single index regression (Maistre and Patilea 2017).

For the regression model $Y = m(X) + \epsilon$, Bierens (1982; 1990) tested the null hypothesis that the regression function $m(X) = g(X, \theta_0)$ almost surely for some θ_0. The following result holds

$$E(\epsilon_0|X) = 0 \Leftrightarrow E(\epsilon_0 \exp(it^\top X)) = 0,$$

where $\epsilon_0 = Y - g(X, \theta_0)$ and $i = \sqrt{-1}$ denotes the imaginary unit. Unlike Stute (1997) who used indicator function $I(X \leq t)$, Bierens (1982) used the characteristic weight function $\omega(X, x) = \exp(it^\top X)$ in the above equivalence between the conditional moment and the unconditional moment. Further, when the p-variate normal

density function is used as the integration function for the argument t, the Cramër-von Mises type test statistic is defined:

$$CvM_{n,exp} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\epsilon}_{0i} \hat{\epsilon}_{0j} \exp(-\frac{1}{2}||X_i - X_j||^2).$$

This is a typical global smoothing test. Compared with the indicator weight used in Stute (1997) and his followers, the characteristic weight function, which is based on one-dimensional projections, is less sensitive to the dimension p. This can be seen more clearly in the formulation of $CvM_{n,exp}$. The dimensionality has little impact on the Euclidean distance between X_i and X_j. On the contrast, if we use $I(X_i \leq X_j)$ or $X_i - X_j$ to express distance between X_i and X_j, the data can be sparse when p is large. Thus this type of test statistics can also be used to avoid the curse of dimensionality.

Another idea is to use weighted residual process. By assuming that ϵ is independent with X, under the null hypothesis $E(\Sigma^{-1}(X - E(X))|\epsilon_0) = 0$ is equivalent to $E(\Sigma^{-1}(X - E(X))I(\epsilon_0 \leq t)) = 0$. Here Σ is the covariance matrix of X. Zhu (2003) then proposed the following weighted residual process

$$I_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{\Sigma}^{-1/2}(X_j - \bar{X}) I(\hat{\epsilon}_{0j} \leq t),$$

where $\hat{\Sigma}$ is the sample covariance matrix of X_i's. Let

$$T_n = \sup_{||\beta||=1} \beta^\top [\frac{1}{n} \sum_{i=1}^{n} I_n(X_i) I_n^\top(X_i)] \beta.$$

The null hypothesis is rejected for large value of T_n.

Stute et al. (2008) further generalized this idea and proposed the following empirical process

$$R_n(t) = n^{-1/2} \sum_{i=1}^{n} (g(X_i) - \bar{g}) I(\hat{\epsilon}_{0i} \leq t),$$

with $\bar{g} = n^{-1} \sum_{i=1}^{n} g(X_i)$. By suitable selection of weight function vector $g(\cdot)$, the power of the test can be much enhanced. The selection was discussed in Stute et al. (2008). These two tests can also greatly avoid the curse of dimensionality. But they are not omnibus tests either.

7.4 Stute's Contributions on Model Checking

In the model checking field for regressions, Stute has several fundamental contributions. In this section, we give a very brief summary for his main results that may not be of dimension reduction nature, but stimulate the research in this area.

For local smoothing tests, Stute and González-Manteiga (1996) proposed a test for linearity. The test is based on the comparison of a nearest neighbor estimator and a parametric estimator of the regression function. Instead of Nadaraya-Watson kernel estimator that was used in many local smoothing tests, Stute and González-Manteiga (1996) adopted the symmetrized nearest neighbor estimators of $m(\cdot)$ as studied in Stute (1984).

He then devoted more efforts on developing global smoothing tests. A seminal work of the empirical process-based methodology that can be used to construct global smoothing tests is Stute (1997). A nonparametric principal component analysis was studied in detail, which stimulated many following researches in constructing global smoothing tests. From this work, we can see clearly that empirical process-based tests are usually not distribution-free asymptotically. Stute et al. (1998b) then skillfully introduced the innovation process approach to model checking to obtain an asymptotically distribution-free test in the case of one-dimensional covariate. The innovation process approach was first proposed by Khamlazde (1981) in the theory of goodness-of-fit for distributions and is now well known as the Khamaladze martingale transformation. As such an innovation process approach is in general not easy to be applied to the cases with multidimensional covariates, more practical implementation method is to use a wide bootstrap. Stute et al. (1998a) formally proved that the wild bootstrap yields a consistent approximation of the null distribution of the residual marked process-based test. They also demonstrated that the residual-based bootstrap is consistent only when the errors are homoscedastic. The wild bootstrap has become a standard method to approximate limiting null distributions in this area. One of the earlier references for this approach is Stute et al. (1993) for goodness-of-test of distribution functions.

Besides the above fundamental developments, Stute also makes great contributions on model checking with other types of data structure and regression models. For time series data, Koul and Stute (1999) developed some residual marked process-based tests for autoregressive models. Koul et al. (2005) proposed diagnostic tests for self-exciting threshold autoregressive models. Later on, Stute et al. (2006) investigated model checking for higher order autoregressive models. For censored data, Stute et al. (2000) also considered the empirical process-based tests. Stute and Zhu (2005) proposed certain score-type test statistics to check the suitability of semiparametric single index models. The tests can detect Pitman alternatives at a $\sqrt{n}$-rate, and also peak alternatives.

Some other contributions should also be mentioned. Ferreira and Stute (2004) introduced tests for equality of two regression curves when the inputs are driven by a time series. The basic construction is based on empirical process of the time series

marked by the difference in the pertaining dependent variables. Srihera and Stute (2010) constructed tests based on weighted differences of two regression curves computed at selected points.

7.5 Conclusion and Discussion

In this brief review, we focus on dimension-reduction based tests for regression models with independent data. We have not included methods for other types of data, such as time-series data, although there are numerous proposals in the literature. On the other hand, how to deal with data with divergent dimension or ultra-high dimension is of interest and importance. We hope this review can give readers a relative clear introduction of tests in the literature to avoid the curse of dimensionality. As this is an important research topic that deserves more further studies in the future for high, and ultra-high dimension paradigms.

References

Bierens, H. J. (1982). Consistent model especification tests. *Journal of Econometrics*, **20**, 105–134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Ecomometrica*, **58**, 1443–1458.

Conde-Amboage, M. and González-Manteiga, W. (2015). A lack-of-fit test for quantile regression models with high-dimensional covariates. *Computational Statistics & Data Analysis*, **88**, 128–138.

Delgado, M. A. and Gonzáles-Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics*, **29**, 1469–1507.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimates. *Annals of Statistics*, **27**, 1012–1050.

Dette, H., Neumeyer, N. and Van Keilegom, I. (2007). A new test for the parametric form of the variance function in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 903–917.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, **22**, 1030–1051.

Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *Test*, **16**, 409–444.

Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, **29**, 153–193.

Fan, Y. and Li, Q. (1996). Consistent model specifiation tests: omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865–890.

Fan, Y. and Li, Q. (2000). Consistent Model Specification Tests: Kernel-Based Tests Versus Bierens' ICM Tests. *Econometric Theory*, **16**, 1016–1041.

Feng, Z., Wen, X., Yu Z. and Zhu, L. X. (2013). On partial sufficient dimension reduction with applications to partially linear multi-index models. *Journal of the American Statistical Association*, **501**, 237–246.

Ferreira, E. and Stute, W. (2004). Testing for differences between conditional means in a time series context. *Journal of the American Statistical Association*, **99**, 169–174.

González-Manteiga, W. and Crujeiras, R. M. (2013a). An updated review of Goodness-of-Fit tests for regression models. *Test*, **22**, 361–411.

González-Manteiga, W. and Crujeiras, R. M. (2013b). Rejoinder on: An updated review of Goodness-of-Fit tests for regression models. *Test*, **22**, 442–447.

Guo, X., Wang, T. and Zhu, L. X. (2016). Model checking for parametric single-index models: a dimension reduction model-adaptive approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**, 1013–1035.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1926–1947.

Hart, J. (1997). *Nonparametric smoothing and lack-of-fit tests.* Springer, Berlin.

Huber, J. P. (1985). Projection Pursuit, *Annals of Statistics*, **13**, 435–475.

Huskova, M. and Meintanis, S. (2009). Goodness-of-fit tests for parametric regression models based on empirical characteristic functions. *Kybernetika*, **45**, 960–971.

Huskova, M. and Meintanis, S. (2010). Test for the error distribution in nonparametric possibly heterocedastic regression models. *Test*, **19**, 92–112.

Khmaladze, E. V. (1981). Martingale Approach in the Theory of Goodness-of-fit Tests. *Theor. Prob. Appl.*, **26**, 240–257.

Khmaladze, E. V. and Koul, H. L. (2009). Goodness-of-fit problem for errors in nonparametric regression: distribution free approach. *Annals of Statistics*, **37**, 3165–3185.

Koul, H. L. and Ni, P. P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference*, **119**,109–141.

Koul, H. L. and Song, W. X. (2009). Minimum distance regression model checking with Berkson measurement errors. *Annals of Statistics*, **37**, 132–156.

Koul, H. L. and Stute, W. (1999). Nonparametric model checks for time series. *Annals of Statistics*, **27**, 204–236.

Koul, H. L., Stute, W. and Li, F. (2005). Model diagnosis for setar time series. *Statistica Sinica*, **15**, 795–817.

Koul, H. L., Xie, C. L. and Zhu, L. X. (2016). An adaptive-to-model test for parametric single-index errors-in-variables models. Submitted.

Kumbhakar, S. C., Park, B. U., Simar, L. and Tsionas, E. G. (2007), Nonparametric stochastic frontiers: a local likelihood approach. *Journal of Econometrics*, **137**, 1–27.

Lavergne, P., Maistre, S. and Patilea, V. (2015). A significance test for covariates in nonparametric regression. *Electronic Journal of Statistics*, **9**, 643–678.

Lavergne, P. and Patilea, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics*, **143**, 103–122.

Lavergne, P. and Patilea, V. (2012). One for all and all for One: regression checks with many regressors. *Journal of Business & Economic Statistics*, **30**, 41–52.

Lavergne, P. and Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory*, **16**, 576–601.

Lin, Z. J., Li, Q. and Sun. Y. G. (2014), A consistent nonparametric test of parametric regression functional form in fixed effects panel data models. *Journal of Econometrics*, **178(1)**, 167–179.

Ma, S. J., Zhang. J., Sun, Z. H. and Liang, H. (2014). Integrated conditional moment test for partially linear single index models incorporating dimension-reduction. *Electronic Journal of Statistics*, **8**, 523–542.

Maistre, S., Lavergne, P. and Patilea, V. (2017). Powerful nonparametric checks for quantile regression. *Journal of Statistical Planning and Inference*, **180**, 13–29.

Maistre, S. and Patilea, V. (2017). Nonparametric model checks of single-index assumptions. *Statistica Sinica*, Online.

Niu, C. Z., Guo, X. and Zhu, L. X. (2016). Enhancements of nonparametric generalized likelihood ratio test: bias-correction and dimension reduction. Working paper.

Niu, C. Z. and Zhu, L. X. (2016). A robust adaptive-to-model enhancement test for parametric single-index models. Working paper.

Srihera, R. and Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis*, **101**, 2039–2059.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Annals of Statistics*, **12**, 917–926.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, **25**, 613–641.

Stute, W. and Gonzáles-Manteiga, W. (1996). NN goodness-of-fit tests for linear models. *Journal of Statistical Planning and Inference*, **53**, 75–92.

Stute, W., Gonzáles-Manteiga, W. and Presedo-Quindimil, M. (1993). Boostrap based goodness-of-fit tests. *Metrika*, **40**, 243–256.

Stute, W., Gonzáles-Manteiga, W. and Presedo-Quindimil, M. (1998a). Bootstrap approximation in model checks for regression. *Journal of the American Statistical Association*, **93**, 141–149.

Stute, W., Gonzáles-Manteiga, Sánchez-Sellero, C. (2000). Nonparametric model checks in censored regression. *Communication in Statistics-Theory and Methods*, **29**, 1611–1629.

Stute, W., Presendo-Quindimil, M., Gonzáles-Manteiga, W. and Koul, H. L. (2006). Model checks for higher order time series. *Statistics & Probability Letters*, **76**, 1385–1396.

Stute, W., Thies, S. and Zhu, L. X. (1998b). Model checks for regression: An innovation process approach. *Annals of Statistics*, **26**, 1916–1934.

Stute, W., Xu, W.L. and Zhu, L. X. (2008). Model diagnosis for parametric regression in high dimensional spaces. *Biometrika*, **95**, 451–467.

Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, **29**, 535–546.

Stute, W. and Zhu, L. X. (2005). Nonparametric checks for single-index models. *Annals of Statistics*, **33**, 1048–1083.

Su, J. Q. and Wei, L. J. (1991). A lack of fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, **86**, 420–426.

Tan, F. L., Zhu, X. H. and Zhu, L. X. (2017). A projection-based adaptive-to-model test for regressions. *Statistica Sinica*, Online.

Van Keilegom, I., Gonzáles-Manteiga, W. and Sánchez Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test*, **17**, 401–415.

Xia, Q., Xu, W. L. and Zhu. L. X. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica*, **25**, 1025–1044.

Xia, Y. C. (2009). Model check for multiple regressions via dimension reduction. *Biometrika*, **96**, 133–148.

Xia, Y. C., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 363–410.

Zhang, C. and Dette, H. (2004). A power comparison between nonparametric regression tests. *Statistics and Probability Letters*, **66**, 289–301.

Zhang, T. and Wu, W. B. (2011). Testing parametric assumptions of trends of a nonstationary time series. *Biometrika*, **98**, 599–614.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**, 263–289.

Zhu, L. P., Wang, T., Zhu, L.X. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295–304.

Zhu, L. X. (2003), Model checking of dimension-reduction type for regression. *Statistica Sinica*, **13**, 283–296.

Zhu, L. X. and An, H. Z. (1992). A test method for nonlinearity of regression model. *Journal of Math.*, **12**, 391–397. In Chinese

Zhu, L. X. and Li, R. (1998). Dimension-reduction type test for linearity of a stochastic model. *Acta Mathematicae Applicatae Sinica*, **14**, 165–175.

Zhu, X. H., Guo, X. and Zhu, L. X. (2017). An adaptive-to-model test for partially parametric single-index models. *Statistics and Computing*, **27**, 1193–1204.

Zhu, X. H., Chen, F., Guo, X. and Zhu, L. X. (2016). Heteroscedasticity checks for regression models: A dimension-reduction based model adaptive approach. *Computational Statistics & Data Analysis*, **103**, 263–283.

Zhu, X. H. and Zhu, L. X. (2016). Dimension-reduction based significance testing in nonparametric regression. Working paper.

Part III
Asymptotic Nonparametric Statistics and Change-Point Problems

Asymptotic Tail Bounds for the Dempfle-Stute Estimator in General Regression Models

8

Dietmar Ferger

8.1 Introduction and Main Results

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of a vector $(X, Y) \in \mathbb{R}^2$ defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ such that Y has finite expectation. Then the regression function $m(x) := \mathbb{E}(Y|X = x), x \in \mathbb{R}$, exists and admits the representation

$$Y = m(X) + \varepsilon \quad \text{with}\, \mathbb{E}(\varepsilon|X) = 0\, \text{a.s.} \tag{8.1}$$

If

$$Q(x, \cdot) := \mathbb{P}(Y \in \cdot|X = x), \ x \in \mathbb{R},$$

denotes the conditional distribution of Y given $X = x$ then the distribution of (X, Y) is uniquely determined by

$$\mathbb{P} \circ (X, Y)^{-1} = \mathbb{P} \circ X^{-1} \otimes Q$$

and the regression function m by

$$m(x) = \int y \ Q(x, dy). \tag{8.2}$$

Dempfle and Stute (2002) consider the special case, where

$$m(x) = a1_{\{x \le \theta\}} + b1_{\{x > \theta\}} \tag{8.3}$$

D. Ferger (✉)
Department of Mathematics, Technische Universität Dresden,
Zellescher Weg 12-14, 01069 Dresden, Germany
e-mail: dietmar.ferger@tu-dresden.de

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_8

is a unit step function with jump at point $\theta \in \mathbb{R}$ and levels $a \neq b \in \mathbb{R}$. Here, the parameters θ, a and b are unknown. For the estimation of θ they propose an estimator $\hat{\theta}_n$, which can be rewritten as

$$\hat{\theta}_n \in \operatorname*{argmax}_{t\in\mathbb{R}} \sum_{i=1}^{n} 1_{\{X_i \le t\}}(Y_i - \bar{Y}_n) \quad \text{with } \bar{Y}_n := n^{-1} \sum_{i=1}^{n} Y_i. \tag{8.4}$$

Relation (8.4) means that $\hat{\theta}_n$ is a random variable, which maximizes the *marked empirical distribution function*

$$E_n(t) := \sum_{i=1}^{n} 1_{\{X_i \le t\}}(Y_i - \bar{Y}_n).$$

Let $X_{1:n} \le \ldots \le X_{n:n}$ be the order statistics of $X_1, \ldots, X_n$ with pertaining concomitants $Y_{[1:n]}, \ldots, Y_{[n:n]}$. If

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \le x\}}$$

denotes the empirical distribution function of the X-sample then

$$E_n(t) := \sum_{i=1}^{n} 1_{\{X_i \le t\}}(Y_i - \bar{Y}_n) = \sum_{i=1}^{nF_n(t)} (Y_{[i:n]} - \bar{Y}_n). \tag{8.5}$$

It follows that E_n is a step function with jumps exactly at the points $X_{i:n}$, $1 \le i \le n$, which vanishes outside of the interval $[X_{1:n}, X_{n:n})$. Put

$$S_k := E_n(X_{k:n}), \quad 1 \le k \le n,$$

and

$$\lambda_n := \min\{1 \le l \le n : S_l = \max_{1\le k\le n} S_k\}.$$

Then the measurable choice

$$\hat{\theta}_n := X_{\lambda_n:n} \tag{8.6}$$

gives a maximizing point of the marked empirical process E_n and thus an explicit estimator for θ. Notice that the (simple) computational formula (8.6) is valid for every data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ in $\mathbb{R}^2$ no matter of which type the data are. In particular, the Y_i can be $0-1$ variables or ties in the X-sample are permitted. However, if there are no ties then S_k further simplifies to

$$S_k = \sum_{i=1}^{k} (Y_{[i:n]} - \bar{Y}_n).$$

Dempfle and Stute (2002) prove the following theorem. Here, they make assumptions on the distribution function F of X and on the conditional variance

$$V(x) := \operatorname{Var}(Y|X = x) = \int (y - m(x))^2 \, Q(x, dy), \quad x \in \mathbb{R}. \tag{8.7}$$

In the sequel $\varepsilon > 0$ denotes a generic constant.

Theorem 1 (Dempfle-Stute) *Assume that m is of type (8.3), where $a > b$. If*
(F) F is continuously differentiable in $[\theta - \varepsilon, \theta + \varepsilon]$ with $F'(\theta) > 0$
and V is bounded on $\mathbb{R}$, then

$$n(\hat{\theta}_n - \theta) = O_{\mathbb{P}}(1),$$

that is

$$\lim_{y \to \infty} \limsup_{n \to \infty} \mathbb{P}(n|\hat{\theta}_n - \theta| \geq y) = 0.$$

We will generalize and extend this result in many respects. Consider an arbitrary conditional distribution $Q(x, \cdot)$ with induced regression function (8.2) of the following type: the graph of m runs above a mean level $\bar{m}$ to the left of θ and it runs below that level on the right-hand side. More precisely, let

$$\bar{m} := \mathbb{E}(Y) = \int_{\mathbb{R}} m(x) F(dx)$$

be the mean output. Then we assume that

(C1) $m(x) > \bar{m} \quad \forall\, x \in [\theta - \varepsilon, \theta)$ and $m(x) < \bar{m} \quad \forall\, x \in (\theta, \theta + \varepsilon]$,
(C2) $m(x) \geq \bar{m} \quad \forall\, x < \theta - \varepsilon$ and $m(x) \leq \bar{m} \quad \forall\, x > \theta + \varepsilon$.

Similarly as (F) the requirement (C1) is a *local condition*. It can be rewritten in a closed form as

$$\operatorname{sign}(x - \theta)(m(x) - \bar{m}) < 0 \quad \forall\, x \in [\theta - \varepsilon, \theta + \varepsilon] \backslash \{\theta\}.$$

That means m must be strictly separated from $\bar{m}$ in a local punctured neighborhood of θ and outside of that region it may touch it. Notice that there is no continuity assumption on m. So, there might be a jump at θ or alternatively there is a continuous crossing. For the unit-step function m in (8.3) the value of $\bar{m}$ is equal to $aF(\theta) + b(1 - F(\theta))$, whence m satisfies the above conditions as long as $0 < F(\theta) < 1$, which is clearly implied by (F).

Theorem 2 *If* $(C1)$ *and* $(C2)$ *hold and* F *is strictly increasing in* $[\theta-\varepsilon, \theta+\varepsilon]$, *then*

$$\hat{\theta}_n \to \theta \quad a.s. \tag{8.8}$$

Proof First, observe that $\hat{\theta}_n \in \operatorname{argmax}_{t\in\mathbb{R}} n^{-1}E_n(t)$, since the positive factor n^{-1} leaves the maximizing point unchanged. Now,

$$\rho_n(t) := n^{-1}E_n(t) = H_n(t) - \bar{Y}_n F_n(t), \tag{8.9}$$

with

$$H_n(t) = n^{-1}\sum_{i=1}^{n} 1_{\{X_i\le t\}}Y_i.$$

By the Glivenko-Cantell Theorem F_n converges to F uniformly on $\mathbb{R}$ with probability one. As Stute (1997) points out we also have that

$$\sup_{t\in\mathbb{R}} |H_n(t) - H(t)| \to 0 \quad \text{a.s.},$$

where

$$H(t) = \int_{-\infty}^{t} m(x)F(dx).$$

The Strong Law of Large Numbers ensures that $\bar{Y}_n \to \bar{m}$ a.s. To sum up, we obtain from (8.9) that

$$\sup_{t\in\mathbb{R}} |\rho_n(t) - \rho(t)| \to 0 \quad \text{a.s.},$$

where

$$\rho(t) := H(t) - \bar{m}F(t) = \int_{-\infty}^{t} (m(x)-\bar{m})\ F(dx). \tag{8.10}$$

From Lemma C in the Appendix we know that ρ is right-continuous with left-hand limits (in short: rcll or cadlag) with *well-separated* supremizing point θ. Thus we can apply Theorem 3.3 of Ferger (2015) to get the desired result (8.8). □

Notice that after all the *global* behavior of the function ρ in (8.10) is essential for the consistency of $\hat{\theta}_n$. Using the consistency we will demonstrate how the *local* behavior of ρ determines the tail probabilities of $\hat{\theta}_n - \theta$.

In the sequel we focus on two different types of regression functions m. The first type has a finite jump at θ and is continuous in an arbitrary small region on the left and right of θ. Formally, it is required that

(J) The left and right limits $m(\theta-)$ and $m(\theta+)$ exist and are finite with

$$m(\theta-) > \bar{m} > m(\theta+)$$

and m is continuous in a punctured neighborhood $[\theta-\varepsilon, \theta+\varepsilon]\backslash\{\theta\}$.

The second type is a locally smooth function.

(S) m is continuous and decreasing in $[\theta-\varepsilon, \theta+\varepsilon]$. Moreover, there exist sequences $\alpha_n \to \infty$ and $\beta_n \to \infty$ and functions $\phi, \varphi : (0,\infty) \to (0,\infty)$ such that

$$\liminf_{n\to\infty} \sqrt{n/\alpha_n}\big(m(\theta) - m(\theta + u/\alpha_n)\big) \geq \phi(u) \quad \forall\, u > 0, \tag{8.11}$$

$$\liminf_{n\to\infty} \sqrt{n/\beta_n}\big(m(\theta - u/\beta_n) - m(\theta)\big) \geq \varphi(u) \quad \forall\, u > 0. \tag{8.12}$$

The requirements (8.11) and (8.12) can be considered as generalizations of differentiability. They are comparable to those given by Smirnov (1952). However, if m is (continuously) differentiable at θ with negative derivative, then $\alpha_n = \beta_n = n^{1/3}$ and $\phi(u) = \varphi(u) = -m'(\theta)u$. A huge class of non-differentiable m is given in Example 1 below.

Under (S) and (F) we obtain from Fatou's Lemma that

$$\liminf_{n\to\infty} \sqrt{n\alpha_n}\big(\rho(\theta) - \rho(\theta + u/\alpha_n)\big) \geq F'(\theta) \int_0^u \phi(s)ds =: \psi(u) \quad \forall\, u \geq 0, \tag{8.13}$$

$$\liminf_{n\to\infty} \sqrt{n\beta_n}\big(\rho(\theta) - \rho(\theta + u/\beta_n)\big) \geq F'(\theta) \int_0^{|u|} \varphi(s)ds =: \psi(u) \quad \forall\, u \leq 0. \tag{8.14}$$

Similarly, under (J) and (F) it follows that

$$\liminf_{n\to\infty}\ n\big(\rho(\theta) - \rho(\theta + u/n)\big) \geq \psi(u) \quad \forall\, u \in \mathbb{R}, \tag{8.15}$$

$$\psi(u) := F'(\theta) \cdot \begin{cases} (\bar{m} - m(\theta+))\, u, & u \geq 0 \\ (m(\theta-) - \bar{m})\, |u|, & u \leq 0. \end{cases} \tag{8.16}$$

Thus (8.15) corresponds to (8.13) and (8.14) with $\alpha_n = \beta_n = n$. So, even though the regression functions m in (J) and (S) look completely different they share comparable features as far as the induced function ρ is concerned.

The following example is an adaption of Knight's example (6) in Knight (1998).

Example 1 Let a, b, α, β be positive constants. We assume that m satisfies

$$m(x) - m(\theta) = \begin{cases} -a(x-\theta)^{\alpha} L(x-\theta), & x \in (\theta, \theta+\varepsilon] \\ b|x-\theta|^{\beta} l(|x-\theta|), & x \in [\theta-\varepsilon, \theta) \end{cases}$$

where L and l are continuous and slowly varying functions at 0 such that m is decreasing. In this case m meets (S) with

$$\alpha_n = n^{\frac{1}{1+2\alpha}} L^*(n), \quad \phi(u) = au^{\alpha}, \quad \beta_n = n^{\frac{1}{1+2\beta}} l^*(n), \quad \varphi(u) = bu^{\beta},$$

where L^* and l^* are slowly varying at infinity. (This can be verified by the representation theorem for slowly varying functions of Karamata, see Bojanic and Seneta (1971).) Notice that ϕ and φ do not depend on L or l, respectively. For example, if $L(x) = \log(1/x)$ then we can take

$$\alpha_n = (1+2\alpha)^{-2/(1+2\alpha)}\, n^{1/(1+2\alpha)} \log(n)^{2/(1+2\alpha)}.$$

The corresponding ψ in (8.13) and (8.14) is given by

$$\psi(u) = F'(\theta) \cdot \begin{cases} \frac{a}{1+\alpha} u^{1+\alpha}, & u \geq 0 \\ \frac{b}{1+\beta} |u|^{1+\beta}, & u \leq 0. \end{cases}$$

Our main result in the next theorem gives the asymptotic upper and lower tail bounds of $\hat{\theta}_n - \theta$ in terms of ψ-integrals.

Theorem 3 *Assume that (C1), (C2) and (F) are true and that V is continuous in a local punctured neighborhood $[\theta-\varepsilon, \theta+\varepsilon]\backslash\{\theta\}$ with finite limits $V(\theta+)$ and $V(\theta-)$. Let constants C_+, C_-, D_+, D_- and C be given by*

$$C_{\pm} = F'(\theta)\{108\, V(\theta\pm) + 144\, m(\theta\pm)^2 + 9(|\bar{m}|+1)^2\},$$

$$D_{\pm} = 2\, \frac{108\, V(\theta\pm) + 144\, m(\theta\pm)^2 + 9(|\bar{m}|+1)^2}{F'(\theta)(m(\theta\pm) - \bar{m})^2}$$

and $C = \max\{D_+, D_-\}$.

- *If (J) holds, then*

$$\limsup_{n\to\infty} \mathbb{P}(n(\hat{\theta}_n - \theta) \geq y) \leq C_+\{\, y\, \psi(y)^{-2} + \int_y^{\infty} \psi(s)^{-2} ds\} \quad \forall\, y > 0, \tag{8.17}$$

$$\limsup_{n\to\infty} \mathbb{P}(n(\hat{\theta}_n - \theta) \leq -y) \leq C_-\{\, y\, \psi(-y)^{-2} + \int_{-\infty}^{-y} \psi(s)^{-2} ds\} \quad \forall\, y > 0 \tag{8.18}$$

with ψ given in (8.16). In particular,

$$\limsup_{n\to\infty} \mathbb{P}(n|\hat{\theta}_n - \theta| \geq y) \leq C\ y^{-1} \quad \forall\ y > 0. \tag{8.19}$$

- *If (S) holds and $\mathbb{E}(Y^2) < \infty$, then*

$$\limsup_{n\to\infty} \mathbb{P}(\alpha_n(\hat{\theta}_n - \theta) \geq y) \leq C_+\{\ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds\} \quad \forall\ y > 0, \tag{8.20}$$

$$\limsup_{n\to\infty} \mathbb{P}(\beta_n(\hat{\theta}_n - \theta) \leq -y) \leq C_-\{\ y\ \psi(-y)^{-2} + \int_{-\infty}^{-y} \psi(s)^{-2} ds\} \quad \forall\ y > 0 \tag{8.21}$$

with ψ given in (8.13) and (8.14).

Proof First, we find $r \in (0, \varepsilon]$ such that all local conditions are fulfilled simultaneously in the neighborhood $[\theta - r, \theta + r]$ or $[\theta - r, \theta + r]\backslash\{\theta\}$, respectively. Let us begin with the simple inequality

$$\mathbb{P}(\hat{\theta}_n - \theta \geq d) \leq \mathbb{P}(d \leq \hat{\theta}_n - \theta \leq r) + \mathbb{P}(|\hat{\theta}_n - \theta| > r), \quad d \leq r. \tag{8.22}$$

In view of (8.15) we put $\alpha_n := n$ if (J) holds. This makes us to treat the cases (S) and (J) jointly. Henceforth, for given $y > 0$ we define

$$d := d_n(y) := y\ \alpha_n^{-1}. \tag{8.23}$$

Then the left side in (8.22) is equal to the upper tail probability in (8.20) or (8.17), respectively. Since $d_n(y) \to 0$ as $n \to \infty$ there is an integer $n_0 = n_0(y, r)$ such that $d_n(y) \leq r$ for all $n \geq n_0$. By Theorem 2 the second probability on the right side in (8.22) converges to zero as $n \to \infty$. Thus it remains to bound the first probability on the right side in (8.22) with d in (8.23) for $n \geq n_0$.

Observe that

$$\{d \leq \hat{\theta}_n - \theta \leq r\} \subseteq \bigcup_{d\leq u\leq r} \{\rho_n(\theta + u) - \rho_n(\theta) \geq 0\} =: E_n(y, r). \tag{8.24}$$

To see this basic relation (8.24), assume that $\rho_n(\theta + u) < \rho_n(\theta)$ for all $u \in [d, r]$. Then $u = \hat{\theta}_n - \theta$ yields $\rho_n(\hat{\theta}_n) < \rho_n(\theta)$, which is a contradiction to $\hat{\theta}_n \in \operatorname{argmax}_{t\in\mathbb{R}} \rho_n(t)$.

We decompose the process ρ_n into

$$\rho_n(t) = U_n(t) + V_n(t) + Z_n(t) + \rho(t) \tag{8.25}$$

with

$$U_n(t) = H_n(t) - H(t),$$

$$V_n(t) = -\bar{Y}_n(F_n(t) - F(t))$$

and

$$Z_n(t) = (\bar{m} - \bar{Y}_n)F(t).$$

Thus the increments of ρ_n can be written as

$$\begin{aligned}\rho_n(\theta+u) - \rho_n(\theta) &= \{U_n(\theta+u) - U_n(\theta)\} + \{V_n(\theta+u) - V_n(\theta)\} \\ &\quad + (\bar{m} - \bar{Y}_n)\{F(\theta+u) - F(\theta)\} - \{\rho(\theta) - \rho(\theta+u)\}.\end{aligned}$$

By Lemma C in the Appendix

$$\rho(\theta) - \rho(\theta+u) > 0 \quad \forall\, 0 \neq |u| \leq r,$$

hence we can conclude that

$$\begin{aligned}\mathbb{P}(E_n(y,r)) &\leq \mathbb{P}\Big(\sup_{d\leq u\leq r} \frac{|U_n(\theta+u) - U_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \geq \frac{1}{3}\Big) \\ &\quad + \mathbb{P}\Big(\sup_{d\leq u\leq r} \frac{|V_n(\theta+u) - V_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \geq \frac{1}{3}\Big) \\ &\quad + \mathbb{P}\Big(|\bar{Y}_n - \bar{m}| \sup_{d\leq u\leq r} \frac{|F(\theta+u) - F(\theta)|}{\rho(\theta) - \rho(\theta+u)} \geq \frac{1}{3}\Big) \\ &=: P_{1,n}(y) + P_{2,n}(y) + P_{3,n}(y).\end{aligned} \tag{8.26}$$

Since

$$U_n(t) = n^{-1}\sum_{i=1}^{n} 1_{\{X_i\leq t\}}(Y_i - m(X_i)) + n^{-1}\sum_{i=1}^{n} 1_{\{X_i\leq t\}}m(X_i) - H(t) =: S_n(t) + L_n(t),$$

it follows that

$$\begin{aligned}P_{1,n}(y) &= \mathbb{P}\Big(\sup_{d\leq u\leq r} \frac{|U_n(\theta+u) - U_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \geq \frac{1}{3}\Big) \\ &\leq \mathbb{P}\Big(\sup_{d\leq u\leq r} \frac{|S_n(\theta+u) - S_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \geq \frac{1}{6}\Big) \\ &\quad + \mathbb{P}\Big(\sup_{d\leq u\leq r} \frac{|L_n(\theta+u) - L_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \geq \frac{1}{6}\Big) =: q_{1,n}(y) + q_{2,n}(y).\end{aligned} \tag{8.27}$$

Set

$$\begin{aligned}\Delta_n(u) &:= S_n(\theta+u) - S_n(\theta)\\ &= n^{-1}\sum_{i=1}^{n} 1_{\{\theta < X_i \le \theta+u\}}(Y_i - m(X_i)) \\ &= n^{-1}\sum_{i=1}^{n} 1_{\{\theta < X_{i:n} \le \theta+u\}}(Y_{[i:n]} - m(X_{i:n}))\\ &= n^{-1}\sum_{nF_n(\theta)<i\le nF_n(\theta+u)} (Y_{[i:n]} - m(X_{i:n})).\end{aligned} \tag{8.28}$$

For the investigation of

$$q_{1,n}(y) = \mathbb{P}(\sup_{d\le u\le r}\frac{|S_n(\theta+u)-S_n(\theta)|}{\rho(\theta)-\rho(\theta+u)} \ge \frac{1}{6}) = \mathbb{P}(\sup_{d\le u\le r}\frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)} \ge \frac{1}{6}) \tag{8.29}$$

we need to introduce several quantities. Define

$$\begin{aligned}\Omega_n &:= \bigcup_{i=1}^{n}\{X_{i:n} \in (\theta+d, \theta+r]\} = \{\exists 1 \le i \le n : X_{i:n} - \theta \in (d, r]\},\\ T_n &:= \{\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 \le \ldots \le x_n\},\\ R_n &:= \{\underline{x}_n \in T_n : \exists 1 \le i \le n : x_i - \theta \in (d, r]\},\\ a(\underline{x}_n) &:= \min\{1 \le i \le n : x_i - \theta \in (d, r]\}, \quad \underline{x}_n \in R_n,\\ b(\underline{x}_n) &:= \max\{1 \le i \le n : x_i - \theta \in (d, r]\}, \quad \underline{x}_n \in R_n,\\ c(\underline{x}_n) &:= \sum_{i=1}^{n} 1_{\{x_i \le \theta\}},\\ \underline{X}_n &:= (X_{1:n}, \ldots, X_{n:n}),\\ \underline{Y}_n &:= (Y_{[1:n]}, \ldots, Y_{[n:n]}),\\ \mu_n &:= \mathbb{P} \circ \underline{X}_n^{-1}.\end{aligned}$$

On the event Ω_n the process $\{|\Delta_n(u)| : d \le u \le r\}$ is piecewise constant with jumps exactly at the points $U_i := X_{i:n} - \theta,\ A \le i \le B$, where $A := a(\underline{X}_n)$ and $B := b(\underline{X}_n)$. So, U_A and U_B are the smallest and largest jump-point in $(d, r]$, respectively. Since $u \mapsto (\rho(\theta) - \rho(\theta+u))^{-1}$ is strictly decreasing and continuous

on $[-r, r]$ by Lemma C in the Appendix, the process $\{\frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)} : d \le u \le r\}$ has its supremizing points in the set $\{U_i : A \le i \le B\} \cup \{d\}$. Thus, on Ω_n

$$\begin{aligned}&\sup_{d\le u\le r} \frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)}\\ &\le \max\{\max_{A\le l\le B} \frac{|\Delta_n(U_l)|}{\rho(\theta)-\rho(X_{l:n})}, \max_{A\le l\le B} \frac{|\Delta_n(U_l-)|}{\rho(\theta)-\rho(X_{l:n})}, \frac{|\Delta_n(d)|}{\rho(\theta)-\rho(\theta+d)}\}.\end{aligned} \tag{8.30}$$

On the complement Ω_n^c the process $\{\frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)} : d \le u \le r\}$ has no jumps at all. It is continuous and strictly decreasing, whence

$$\sup_{d\le u\le r} \frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)} = \frac{|\Delta_n(d)|}{\rho(\theta)-\rho(\theta+d)}.$$

In view of (8.29) and (8.30) we therefore obtain that

$$\begin{aligned}q_{1,n}(y) &= \mathbb{P}(\{\sup_{d\le u\le r} \frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)} \ge \frac{1}{6}\} \cap \Omega_n) + \mathbb{P}(\{\sup_{d\le u\le r} \frac{|\Delta_n(u)|}{\rho(\theta)-\rho(\theta+u)} \ge \frac{1}{6}\} \cap \Omega_n^c)\\ &\le \mathbb{P}(\{\max_{A\le l\le B} \frac{|\Delta_n(U_l)|}{\rho(\theta)-\rho(X_{l:n})} \ge \frac{1}{6}\} \cap \Omega_n) + \mathbb{P}(\{\max_{A\le l\le B} \frac{|\Delta_n(U_l-)|}{\rho(\theta)-\rho(X_{l:n})} \ge \frac{1}{6}\} \cap \Omega_n)\\ &+ \mathbb{P}(\frac{|\Delta_n(d)|}{\rho(\theta)-\rho(\theta+d)} \ge \frac{1}{6})\\ &=: a_{1,n}(y) + a_{2,n}(y) + a_{3,n}(y).\end{aligned} \tag{8.31}$$

Observe that $nF_n(\theta + U_l) = nF_n(X_{l:n}) = l$ upon noticing that there are no ties in the interval $(d, r]$ by the (local) continuity of F. Moreover $C := c(\underline{X}_n) = nF_n(\theta)$. Thus

$$\Delta_n(U_l) = n^{-1} \sum_{C<i\le l} Y_{[i:n]} - m(X_{i:n})$$

and so

$$\max_{A\le l\le B} \frac{|\Delta_n(U_l)|}{\rho(\theta)-\rho(X_{l:n})} = n^{-1} \max_{A\le l\le B} \frac{|\sum_{C<i\le l} Y_{[i:n]} - m(X_{i:n})|}{\rho(\theta)-\rho(X_{l:n})}.$$

For $\underline{x}_n \in R_n$ and $\underline{y}_n \in \mathbb{R}^n$ we introduce

$$H(\underline{x}_n, \underline{y}_n) := 1_{A(\underline{x}_n, \underline{y}_n)},$$

where the indicator set is defined as

$$A(\underline{x}_n, \underline{y}_n) := \{\max_{a\le l\le b} \frac{|\sum_{i=c+1}^{l} y_i - m(x_i)|}{\rho(\theta)-\rho(x_l)} \ge \frac{n}{6}\}$$

with $a := a(\underline{x}_n)$, $b := b(\underline{x}_n)$ and $c := c(\underline{x}_n)$. Since $\Omega_n = \underline{X}_n^{-1}(R_n)$ conditioning on $\underline{X}_n$ yields

$$\begin{aligned} a_{1,n}(y) &= \mathbb{E}(1_{\Omega_n} H(\underline{X}_n, \underline{Y}_n)) = \mathbb{E}(1_{\Omega_n} \mathbb{E}(H(\underline{X}_n, \underline{Y}_n)|\underline{X}_n)) \\ &= \int_{R_n} \mathbb{E}(H(\underline{X}_n, \underline{Y}_n)|\underline{X}_n = \underline{x}_n)\mu_n(d\underline{x}_n) \end{aligned} \tag{8.32}$$

with integrand

$$I(\underline{x}_n) := \mathbb{E}(H(\underline{X}_n, \underline{Y}_n)|\underline{X}_n = \underline{x}_n) = \int_{\mathbb{R}^n} H(\underline{x}_n, \underline{y}_n)\mathbb{P}(\underline{Y}_n \in d\underline{y}_n|\underline{X}_n = \underline{x}_n).$$

Lemma 2.1 of Stute and Wang (1993) implies that

$$\mathbb{P}(\underline{Y}_n \in d\underline{y}_n|\underline{X}_n = \underline{x}_n) = \bigotimes_{i=1}^{n} Q(x_i, dy_i),$$

hence

$$I(\underline{x}_n) = \int_{\mathbb{R}^n} H(\underline{x}_n, \underline{y}_n) \bigotimes_{i=1}^{n} Q(x_i, dy_i).$$

For each fixed $\underline{x}_n \in R_n$ let $Z_1, \ldots, Z_n$ be independent random variables defined w.l.o.g. on $(\Omega, \mathscr{A}, \mathbb{P})$ with $Z_i \sim Q(x_i, \cdot)$, $1 \le i \le n$, i.e.

$$\bigotimes_{i=1}^{n} Q(x_i, \cdot) = \mathbb{P} \circ (Z_1, \ldots, Z_n)^{-1}.$$

Then

$$I(\underline{x}_n) = \mathbb{E}(H(\underline{x}_n, Z_1, \ldots, Z_n)) = \mathbb{P}(\max_{a \le l \le b} \frac{|\sum_{i=c+1}^{l} Z_i - m(x_i)|}{\rho(\theta) - \rho(x_l)} \ge \frac{n}{6}).$$

Notice that $\xi_i := Z_i - m(x_i)$ are centered with variance $V(x_i)$, because

$$\mathbb{E}(Z_i) = \int_{\mathbb{R}} y \; Q(x_i, dy) = m(x_i) \quad \text{by (8.2)}$$

and

$$Var(\xi_i) = \mathbb{E}((Z_i - m(x_i))^2) = \int_{\mathbb{R}} (y - m(x_i))^2 \; Q(x_i, dy) = V(x_i) \quad \text{by (8.7)}.$$

Thus by a simple index-transformation

$$I(\underline{x}_n) = \mathbb{P}\Big(\max_{a-c\le k\le b-c} \frac{|\sum_{i=1}^{k}\xi_{c+i}|}{\rho(\theta)-\rho(x_{c+k})} \ge \frac{n}{6}\Big).$$

Since $(\rho(\theta)-\rho(x_{c+k}))^{-1}$, $a-c\le k\le b-c$ is a decreasing sequence of positive weights as a consequence of $\underline{x}_n \in R_n$, we may apply the Hájek-Rényi inequality to get

$$\begin{aligned} I(\underline{x}_n) &\le \frac{36}{n^2}\left\{(\rho(\theta)-\rho(x_a))^{-2}\sum_{i=1}^{a-c}V(x_{c+i}) + \sum_{i=a-c+1}^{b-c}(\rho(\theta)-\rho(x_{c+i}))^{-2}V(x_{c+i})\right\} \\ &= \frac{36}{n^2}\left\{(\rho(\theta)-\rho(x_a))^{-2}\sum_{i=c+1}^{a}V(x_i) + \sum_{i=a+1}^{b}(\rho(\theta)-\rho(x_i))^{-2}V(x_i)\right\} \\ &= \frac{36}{n^2}\left\{(\rho(\theta)-\rho(x_a))^{-2}\sum_{i=c+1}^{a-1}V(x_i) + \sum_{i=a}^{b}(\rho(\theta)-\rho(x_i))^{-2}V(x_i)\right\} \\ &= \frac{36}{n^2}\left\{I_1(\underline{x}_n)+I_2(\underline{x}_n)\right\} \end{aligned} \tag{8.33}$$

with

$$I_1(\underline{x}_n) = (\rho(\theta)-\rho(x_a))^{-2}\sum_{i=c+1}^{a-1}V(x_i)$$

and

$$I_2(\underline{x}_n) = \sum_{i=a}^{b}(\rho(\theta)-\rho(x_i))^{-2}V(x_i). \tag{8.34}$$

By (8.32) and (8.33) we arrive at

$$a_{1,n}(y) \le \frac{36}{n^2}\{\int_{R_n} I_1(\underline{x}_n)\,\mu_n(d\underline{x}_n) + \int_{R_n} I_2(\underline{x}_n)\,\mu_n(d\underline{x}_n)\}. \tag{8.35}$$

By Change of Variable this gives for the first integral

$$\int_{R_n} I_1(\underline{x}_n)\,\mu_n(d\underline{x}_n) = \mathbb{E}(1_{\Omega_n}(\rho(\theta)-\rho(X_{A:n}))^{-2}\sum_{i=C+1}^{A-1}V(X_{i:n})). \tag{8.36}$$

It follows from the definition of A that $\theta+d < X_{A:n} \le \theta+r$ on the event Ω_n. By Lemma C in the Appendix ρ is strictly decreasing on $[\theta+d,\theta+r]$, whence we can conclude that

$$\mathbb{E}(1_{\Omega_n}(\rho(\theta)-\rho(X_{A:n}))^{-2}\sum_{i=C+1}^{A-1}V(X_{i:n}))$$
$$\leq(\rho(\theta)-\rho(\theta+d))^{-2}\mathbb{E}(1_{\Omega_n}\sum_{i=1}^{n}1_{\{C+1\leq i\leq A-1\}}V(X_{i:n})). \tag{8.37}$$

Using the minimality property of A and the definition of C one easily verifies that on the event Ω_n the following equivalence holds:

$$C+1\leq i\leq A-1 \quad\Leftrightarrow\quad \theta<X_{i:n}\leq\theta+d.$$

Thus the expectation on the left side in (8.37) simplifies to

$$\mathbb{E}(1_{\Omega_n}\sum_{i=1}^{n}1_{\{\theta<X_{i:n}\leq\theta+d\}}V(X_{i:n}))$$
$$\leq\mathbb{E}(\sum_{i=1}^{n}1_{\{\theta<X_{i:n}\leq\theta+d\}}V(X_{i:n}))=\mathbb{E}(\sum_{i=1}^{n}1_{\{\theta<X_i\leq\theta+d\}}V(X_i)) \tag{8.38}$$
$$=n\int_{(\theta,\theta+d]}V(x)F(dx). \tag{8.39}$$

Here, the equation in (8.38) is simply the commutative law. Combining (8.36), (8.37) and (8.39) we arrive at

$$\int_{R_n}I_1(\underline{x}_n)\,\mu_n(d\underline{x}_n)\leq n\,(\rho(\theta)-\rho(\theta+d))^{-2}\int_{(\theta,\theta+d]}V(x)F(dx). \tag{8.40}$$

For the second integral in (8.35) the Change of Variable gives

$$\int_{R_n}I_2(\underline{x}_n)\,\mu_n(d\underline{x}_n)=\mathbb{E}(1_{\Omega_n}\sum_{i=1}^{n}1_{\{A\leq i\leq B\}}(\rho(\theta)-\rho(X_{i:n}))^{-2}V(X_{i:n})). \tag{8.41}$$

Recall the definition of A and B to see that $\Omega_n\cap\{A\leq i\leq B\}\subseteq\{d<X_{i:n}-\theta\leq r\}$. Therefore and because $V\geq 0$ by (8.7), we can infer that the expectation on the right side in (8.41) is less than or equal to

$$\mathbb{E}(\sum_{i=1}^{n}1_{\{\theta+d<X_{i:n}\leq\theta+r\}}(\rho(\theta)-\rho(X_{i:n}))^{-2}V(X_{i:n}))$$
$$=\mathbb{E}(\sum_{i=1}^{n}1_{\{\theta+d<X_i\leq\theta+r\}}(\rho(\theta)-\rho(X_i))^{-2}V(X_i))$$
$$=n\int_{(\theta+d,\theta+r]}(\rho(\theta)-\rho(x))^{-2}V(x)F(dx). \tag{8.42}$$

Consequently, by (8.41) and (8.42) it follows that

$$\int_{R_n} I_2(\underline{x}_n)\ \mu_n(d\underline{x}_n) \le n \int_{(\theta+d,\theta+r]} (\rho(\theta)-\rho(x))^{-2} V(x) F(dx).$$

Together with (8.35) and (8.40) we can conclude that

$$\begin{aligned} a_{1,n}(y) &\le 36\, n^{-1} (\rho(\theta)-\rho(\theta+d))^{-2} \int_{(\theta,\theta+d]} V(x) F(dx) \\ &\quad + 36\, n^{-1} \int_{(\theta+d,\theta+r]} (\rho(\theta)-\rho(x))^{-2} V(x) F(dx). \end{aligned} \tag{8.43}$$

To bound the probability $a_{2,n}(y)$ in (8.31) we can proceed in the same way as above, because

$$\max_{A\le l\le B} \frac{|\Delta_n(U_l-)|}{\rho(\theta)-\rho(X_{l:n})} = n^{-1} \max_{A\le l\le B} \frac{|\sum_{C<i\le l-1} Y_{[i:n]} - m(X_{i:n})|}{\rho(\theta)-\rho(X_{l:n})}.$$

The conditional argument yields that

$$a_{2,n}(y) \le \frac{36}{n^2}\{\int_{R_n} I_1(\underline{x}_n)\ \mu_n(d\underline{x}_n) + \int_{R_n} I_2^*(\underline{x}_n)\ \mu_n(d\underline{x}_n)\} \tag{8.44}$$

with

$$I_2^*(\underline{x}_n) = \sum_{i=a}^{b-1} (\rho(\theta)-\rho(x_{i+1}))^{-2} V(x_i).$$

Integration leads to

$$\begin{aligned} \int_{R_n} I_2^*(\underline{x}_n)\ \mu_n(d\underline{x}_n) &= \mathbb{E}(1_{\Omega_n} \sum_{i=1}^n 1_{\{A\le i\le B-1\}} (\rho(\theta)-\rho(X_{i+1:n}))^{-2} V(X_{i:n})) \\ &\le \mathbb{E}(1_{\Omega_n} \sum_{i=1}^n 1_{\{\theta+d<X_{i:n}\le\theta+r\}} (\rho(\theta)-\rho(X_{i+1:n}))^{-2} V(X_{i:n})) \\ &\le \mathbb{E}(1_{\Omega_n} \sum_{i=1}^n 1_{\{\theta+d<X_{i:n}\le\theta+r\}} (\rho(\theta)-\rho(X_{i:n}))^{-2} V(X_{i:n})) \\ &\le \mathbb{E}(\sum_{i=1}^n 1_{\{\theta+d<X_i\le\theta+r\}} (\rho(\theta)-\rho(X_i))^{-2} V(X_i)) \\ &= n \int_{(\theta+d,\theta+r]} (\rho(\theta)-\rho(x))^{-2} V(x) F(dx). \end{aligned} \tag{8.45}$$

To see (8.45) observe that $X_{i+1:n} \geq X_{i:n}$ and $V \geq 0$ by (8.7) as well as ρ is (strictly) decreasing on $[\theta, \theta + r]$. Consequently, by (8.40) and (8.44) we arrive at

$$\begin{aligned} a_{2,n}(y) &\leq 36\, n^{-1}(\rho(\theta) - \rho(\theta + d))^{-2} \int_{(\theta,\theta+d]} V(x)F(dx) \\ &\quad + 36\, n^{-1} \int_{(\theta+d,\theta+r]} (\rho(\theta) - \rho(x))^{-2} V(x)F(dx). \end{aligned} \tag{8.46}$$

For the third probability $a_{3,n}(y)$ in (8.31) recall that by (8.28)

$$\begin{aligned} a_{3,n}(y) &= \mathbb{P}\Big(\frac{|\Delta_n(d)|}{\rho(\theta) - \rho(\theta + d)} \geq \frac{1}{6}\Big) \\ &= \mathbb{P}\Big(\Big|\sum_{i=1}^{n} 1_{\{\theta < X_i \leq \theta + d\}}(Y_i - m(X_i))\Big| \geq \frac{1}{6} n(\rho(\theta) - \rho(\theta + d))\Big). \end{aligned}$$

The summands $\eta_i := 1_{\{\theta < X_i \leq \theta + d\}}(Y_i - m(X_i)) = 1_{\{\theta < X_i \leq \theta + d\}}\varepsilon_i$, $1 \leq i \leq n$, are i.i.d. and centered, because in view of (8.1)

$$\mathbb{E}(1_{\{\theta < X \leq \theta + d\}}\varepsilon) = \mathbb{E}(1_{\{\theta < X \leq \theta + d\}}\mathbb{E}(\varepsilon | X)) = 0.$$

Again, by conditioning on X

$$\begin{aligned} Var(\eta_i) = \mathbb{E}(\eta_i^2) &= \mathbb{E}(1_{\{\theta < X \leq \theta + d\}}\mathbb{E}((Y - m(X))^2 | X)) \\ &= \mathbb{E}(1_{\{\theta < X \leq \theta + d\}} V(X)) \\ &= \int_{(\theta,\theta+d]} V(x)F(dx). \end{aligned}$$

Thus the Tschebyscheff-inequality guarantess that

$$a_{3,n}(y) \leq 36\, n^{-1}(\rho(\theta) - \rho(\theta + d))^{-2} \int_{(\theta,\theta+d]} V(x)F(dx).$$

From (8.31), (8.43) and (8.46) we finally obtain

$$\begin{aligned} q_{1,n}(y) &\leq 108\, n^{-1}(\rho(\theta) - \rho(\theta + d))^{-2} \int_{(\theta,\theta+d]} V(x)F(dx) \\ &\quad + 72\, n^{-1} \int_{(\theta+d,\theta+r]} (\rho(\theta) - \rho(x))^{-2} V(x)F(dx) \\ &= 108\{\sqrt{n\alpha_n}(\rho(\theta) - \rho(\theta + y/\alpha_n))\}^{-2} \int_0^y V(\theta + s/\alpha_n)F'(\theta + s/\alpha_n)ds \\ &\quad + 72 \int_y^{\alpha_n r} \{\sqrt{n\alpha_n}(\rho(\theta) - \rho(\theta + s/\alpha_n))\}^{-2} V(\theta + s/\alpha_n)F'(\theta + s/\alpha_n)ds, \end{aligned} \tag{8.47}$$

where the equality (8.47) is simply a consequence of $d = y\alpha_n^{-1}$ and the substitution $s = \alpha_n(x-\theta)$. Thus Fatou's Lemma in combination with (8.13) and (8.15) yields that

$$\begin{aligned}\limsup_{n\to\infty} q_{1,n}(y) &\le c_1\ y\ \psi(y)^{-2} + c_2 \int_y^\infty \psi(s)^{-2} ds \\ &\le c_1 \left\{ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds \right\} \end{aligned} \tag{8.48}$$

with $c_1 = 108\ V(\theta+)F'(\theta)$ and $c_2 = 72\ V(\theta+)F'(\theta)$.

Next, we turn our attention to the second probability $q_{2,n}(y)$ in (8.27). It involves the increments of the process

$$L_n(t) = n^{-1} \sum_{i=1}^n 1_{\{X_i \le t\}} m(X_i) - H(t) =: \hat{L}_n(t) - H(t).$$

By Theorem A in the Appendix the Doob-Meyer decomposition of the process

$$\hat{L}_n(t) = n^{-1} \sum_{i=1}^n 1_{\{X_i \le t\}} m(X_i)$$

admits the representation

$$L_n(t) = M_n(t) - D_n(t),\ t \in \mathbb{R},$$

where $(M_n(t) : t \in \mathbb{R})$ is a centered rcll martingale with respect to some filtration $(\mathscr{F}_n(t) : t \in \mathbb{R})$ specified in Theorem A, and

$$D_n(t) = \int_{(-\infty,t]} \frac{F_n(x-) - F(x-)}{1 - F(x-)} m(x) F(dx).$$

It follows that

$$\begin{aligned} q_{2,n}(y) &= \mathbb{P}\Big(\sup_{d\le u\le r} \frac{|L_n(\theta+u) - L_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{6}\Big) \quad \text{by definition} \\ &\le \mathbb{P}\Big(\sup_{d\le u\le r} \frac{|M_n(\theta+u) - M_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{12}\Big) \\ &+ \mathbb{P}\Big(\sup_{d\le u\le r} \frac{|D_n(\theta+u) - D_n(\theta)|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{12}\Big) =: b_{1,n}(y) + b_{2,n}(y). \end{aligned} \tag{8.49}$$

In order to bound the probability $b_{1,n}(y)$ we introduce the process

$$\overline{M}_n(u) := M_n(\theta+u) - M_n(\theta), \quad d \le u \le r.$$

Observe that $(\overline{M}_n(u) : d \le u \le r)$ is likewise a centered rcll martingale with respect to the filtration $(\mathscr{F}_n(\theta+u) : d \le u \le r)$. Thus we may apply the extended Birnbaum-Marshall inequality, confer Theorem B in the Appendix, and get

$$\begin{aligned} b_{1,n}(y) &= \mathbb{P}(\sup_{d\le u\le r} \frac{|\overline{M}_n(u)|}{\rho(\theta)-\rho(\theta+u)} \ge \frac{1}{12}) \\ &\le \mathbb{P}(\sup_{d\le u\le r} (\rho(\theta)-\rho(\theta+u))^{-2}\,\overline{M}_n(u)^2 \ge \frac{1}{144}) \\ &\le 144\left\{\int_d^r (\rho(\theta)-\rho(\theta+u))^{-2} G_n(du) + (\rho(\theta)-\rho(\theta+d))^{-2} G_n(d)\right\} \end{aligned} \tag{8.50}$$

upon noticing that $(\overline{M}_n(u)^2, d \le u \le r)$ is a nonnegative submartingale. Here,

$$G_n(u) = \mathbb{E}[\overline{M}_n(u)^2] = \mathbb{E}[(M_n(\theta+u) - M_n(\theta))^2] = \mathbb{E}[M_n(\theta+u)^2] - \mathbb{E}[M_n(\theta)^2], \tag{8.51}$$

where the last equality holds by the martingale property. Thus we have to compute $\mathbb{E}[M_n(t)^2]$, $t \in \mathbb{R}$. For that purpose notice that by Theorem A

$$\begin{aligned} M_n(t) &= \hat{L}_n(t) - \int_{(-\infty,t]} \frac{1-F_n(x-)}{1-F(x-)}\, m(x) F(dx) \\ &= n^{-1}\sum_{i=1}^n [1_{\{X_i\le t\}}\, m(X_i) - \int_{(-\infty,t]} \frac{1_{\{X_i\ge x\}}}{1-F(x-)}\, m(x) F(dx)] \\ &= n^{-1}\sum_{i=1}^n \rho_i(t), \end{aligned}$$

where

$$\rho_i(t) := 1_{\{X_i\le t\}}\, m(X_i) - \int_{(-\infty,t]} \frac{1_{\{X_i\ge x\}}}{1-F(x-)}\, m(x) F(dx)\ , 1 \le i \le n,$$

are i.i.d. copies of $M_1(t)$ and hence in particular are centered. Consequently,

$$\mathbb{E}[M_n(t)^2] = n^{-1}\mathbb{E}[M_1(t)^2] \tag{8.52}$$

with

$$\begin{aligned} \mathbb{E}[M_1(t)^2] &= \mathbb{E}[1_{\{X\le t\}}\, m(X)^2] + \mathbb{E}[(\int_{(-\infty,t]} \frac{1_{\{X\ge x\}}}{1-F(x-)}\, m(x) F(dx))^2] \\ &\quad - 2\mathbb{E}[1_{\{X\le t\}}\, m(X) \int_{(-\infty,t]} \frac{1_{\{X\ge x\}}}{1-F(x-)}\, m(x) F(dx)]. \end{aligned} \tag{8.53}$$

Using Fubini's Theorem for the computation of the second and third expectation in (8.53) we arrive after some straightforward calculations at

$$\mathbb{E}[M_1(t)^2] = K(t) - \int_{(-\infty,t]} \frac{F(\{x\})}{1-F(x-)}\, m(x)^2 F(dx), \tag{8.54}$$

where

$$K(t) = \int_{(-\infty,t]} m(x)^2 F(dx).$$

It follows from (8.51)–(8.54) that

$$\begin{aligned} G_n(u) &= n^{-1}[\int_{(\theta,\theta+u]} m(x)^2 F(dx) - \int_{(\theta,\theta+u]} \frac{F(\{x\})}{1-F(x-)}\, m(x)^2 F(dx)] \\ &= n^{-1} \int_{(\theta,\theta+u]} m(x)^2 F(dx) \\ &= n^{-1} \int_{(0,u]} m(\theta+x)^2 F'(\theta+x)dx, \end{aligned}$$

since F' is continuous on $[\theta - r, \theta + r]$ by (F). Thus for the integral in (8.50) we obtain

$$\int_d^r (\rho(\theta) - \rho(\theta+u))^{-2} G_n(du) = n^{-1} \int_d^r (\rho(\theta) - \rho(\theta+u))^{-2} m(\theta+u)^2 F'(\theta+u)du$$

and for the second summand

$$(\rho(\theta) - \rho(\theta+d))^{-2} G_n(d) = n^{-1} (\rho(\theta) - \rho(\theta+d))^{-2} \int_{(0,d]} m(\theta+x)^2 F'(\theta+x)dx.$$

Combine this with (8.50) to infer that

$$\begin{aligned} b_{1,n}(y) &\le 144 n^{-1} \int_d^r (\rho(\theta) - \rho(\theta+u))^{-2} m(\theta+u)^2 F'(\theta+u)du \\ &+ 144 n^{-1} (\rho(\theta) - \rho(\theta+d))^{-2} \int_{(0,d]} m(\theta+x)^2 F'(\theta+x)dx \\ &= 144 \int_y^{\alpha_n r} \{\sqrt{n\alpha_n}(\rho(\theta) - \rho(\theta+s/\alpha_n))\}^{-2} m(\theta+s/\alpha_n)^2 F'(\theta+s/\alpha_n))ds \\ &+ 144 \{\sqrt{n\alpha_n}(\rho(\theta) - \rho(\theta+y/\alpha_n))\}^{-2} \int_0^y m(\theta+s/\alpha_n)^2 F'(\theta+s/\alpha_n))ds. \end{aligned}$$

By Fatou's Lemma and (8.13) and (8.15) we obtain that

$$\limsup_{n\to\infty} b_{1,n}(y) \le d_1 \left\{ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds \right\} \tag{8.55}$$

with $d_1 = 144\ m(\theta+)^2 F'(\theta)$.

As to the second probability $b_{2,n}(y)$ in (8.49) observe that

$$\sup_{d\le u\le r}\frac{|D_n(\theta+u)-D_n(\theta)|}{\rho(\theta)-\rho(\theta+u)}\le c\,||F_n-F||\sup_{\theta+d\le u\le\theta+r}\frac{F(u)-F(\theta)}{\rho(\theta)-\rho(u)},\tag{8.56}$$

with $c=(1-F(\theta+r))^{-1}\sup_{\theta\le x\le\theta+r}|m(x)|$ and

$$||F_n-F||:=\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|.$$

Notice that $F(\theta+r)<1$ for r sufficiently small as a consequence of (F). From (J) or (S), respectively, we can infer that $\sup_{\theta\le x\le\theta+r}|m(x)|$ is finite, whence c is a finite constant. Put

$$l(u)=\frac{F(u)-F(\theta)}{\rho(\theta)-\rho(u)},\quad \theta<u\le\theta+r.$$

By (8.10)

$$\rho(\theta)-\rho(u)=\int_\theta^u \bar{m}-m(x)F(dx)=\int_\theta^u \bar{m}-m(x)F'(x)dx.\tag{8.57}$$

Therefore

$$\frac{d}{du}(\rho(\theta)-\rho(u))=(\bar{m}-m(u))F'(u)$$

and by the quotient rule

$$l'(u)=\frac{F'(u)\{\int_\theta^u \bar{m}-m(x)F(dx)-(\bar{m}-m(u))(F(u)-F(\theta))\}}{(\rho(\theta)-\rho(u))^2}.$$

From (S) we can infer that $\bar{m}-m(x)\le\bar{m}-m(u)$ for all $x\in[\theta,u]$ and thus

$$\int_\theta^u \bar{m}-m(x)F(dx)\le(\bar{m}-m(u))(F(u)-F(\theta)).$$

It follows that $l'(u)\le 0$ for all $u\in(\theta,\theta+r]$. Hence l is monotone decreasing on $(\theta,\theta+r]$ resulting in

$$\sup_{\theta+d\le u\le\theta+r}\frac{F(u)-F(\theta)}{\rho(\theta)-\rho(u)}\le\frac{F(\theta+d)-F(\theta)}{\rho(\theta)-\rho(\theta+d)}\le 2F'(\theta)\frac{d}{\rho(\theta)-\rho(\theta+d)}\tag{8.58}$$

for r sufficiently small. Consequently, by (8.56) and Massarts's (1990) inequality we arrive at

$$b_{2,n}(y) = \mathbb{P}(\sup_{d\le u\le r} \frac{|D_n(\theta+u) - D_n(\theta)|}{\rho(\theta)-\rho(\theta+u)} \ge \frac{1}{12})$$
$$\le 2\exp\{-C\ y^{-2}\alpha_n(\sqrt{n\alpha_n}(\rho(\theta)-\rho(\theta+y/\alpha_n))^2\}$$

with some finite positive constant C which could be specified. Infer from (8.13) that there exists an integer $n_1 = n_1(y)$ such that

$$\sqrt{n\alpha_n}(\rho(\theta)-\rho(\theta+y/\alpha_n) \ge \frac{1}{2}\psi(y) \quad \forall\ n \ge n_1 \tag{8.59}$$

and so

$$b_{2,n}(y) \le 2\exp\{-4\ C\ y^{-2}\alpha_n\ \psi(y)^2\} \quad \forall\ n \ge n_1.$$

This finally shows that under (S) we have

$$\lim_{n\to\infty} b_{2,n}(y) = 0 \quad \forall\ y > 0. \tag{8.60}$$

In case of (J) the above argument via monotonicity of the function l fails, because here we do not require that m is monotone decreasing. However, by (8.57)

$$\rho(\theta)-\rho(u) \ge \inf_{\theta\le x\le\theta+r}(\bar{m}-m(x))(F(u)-F(\theta)) \ge \frac{1}{2}(\bar{m}-m(\theta+))(F(u)-F(\theta)) \tag{8.61}$$

for r sufficiently small. Therefore, it follows from (8.56) and Massart's (1990) inequality that

$$b_{2,n}(y) \le 2\exp\{-D\ n\}$$

for some finite positive constant D, which guarantees that likewise under (J)

$$\lim_{n\to\infty} b_{2,n}(y) = 0 \quad \forall\ y > 0. \tag{8.62}$$

Summing up we obtain from (8.49), (8.55), (8.60) and (8.62) that

$$\limsup_{n\to\infty} q_{2,n}(y) \le d_1 \left\{ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds \right\} \tag{8.63}$$

with $d_1 = 144\ m(\theta+)^2 F'(\theta)$. Combine this with (8.27) and (8.48) to conclude that

$$\limsup_{n\to\infty} P_{1,n}(y) \le a_1 \left\{ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds \right\} \tag{8.64}$$

with finite constant $a_1 = c_1 + d_1 = F'(\theta)\{108\ V(\theta+) + 144\ m(\theta+)^2\}$.

We continue with the treatment of the second probability in (8.26) which is equal to

$$P_{2,n}(y) = \mathbb{P}(|\bar{Y}_n| \sup_{d\le u\le r} \frac{|F_n(\theta+u) - F(\theta+u) - (F_n(\theta) - F(\theta))|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{3}).$$

Since $|\bar{Y}_n| \le |\bar{m}| + 1$ on the event $\{|\bar{Y}_n - \bar{m}| \le 1\}$ a decomposition yields that

$$\begin{aligned} P_{2,n}(y) \le \mathbb{P}(\sup_{d\le u\le r} \frac{|F_n(\theta+u) - F(\theta+u) - (F_n(\theta) - F(\theta))|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{3(|\bar{m}|+1)}) \\ + \mathbb{P}(|\bar{Y}_n - \bar{m}| > 1) =: \tilde{q}_{2,n}(y) + p_n. \end{aligned} \tag{8.65}$$

By the Weak Law of Large Numbers

$$p_n = \mathbb{P}(|\bar{Y}_n - \bar{m}| > 1) \to 0, \ n \to \infty. \tag{8.66}$$

Notice that

$$\tilde{q}_{2,n}(y) = \mathbb{P}(\sup_{d\le u\le r} \frac{|F_n(\theta+u) - F(\theta+u) - (F_n(\theta) - F(\theta))|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{3(|\bar{m}|+1)})$$

is of the same type as probability $q_{2,n}(y)$ upon noticing that $L_n = F_n - F$, if $m = 1$. But Theorem A in particular holds for $m = 1$, and we can proceed in the same way as in the derivation of (8.63) to get

$$\limsup_{n\to\infty} \tilde{q}_{2,n}(y) \le \tilde{d}_1 \left\{ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds \right\} \tag{8.67}$$

with $\tilde{d}_1 = 9(|\bar{m}|+1)^2 F'(\theta)$. Thus (8.65)–(8.67) yield that

$$\limsup_{n\to\infty} P_{2,n}(y) \le \tilde{d}_1 \left\{ y\ \psi(y)^{-2} + \int_y^\infty \psi(s)^{-2} ds \right\}. \tag{8.68}$$

The third probability in (8.26) is equal to

$$P_{3,n}(y) = \mathbb{P}(|\bar{Y}_n - \bar{m}| \sup_{d\le u\le r} \frac{|F(\theta+u) - F(\theta)|}{\rho(\theta) - \rho(\theta+u)} \ge \frac{1}{3}).$$

It follows from (8.61) that, if (J) holds then

$$\sup_{d\le u\le r} \frac{|F(\theta+u) - F(\theta)|}{\rho(\theta) - \rho(\theta+u)} \le \frac{2}{\bar{m} - m(\theta+)},$$

whence by the Weak Law of Large Numbers

$$\lim_{n\to\infty} P_{3,n}(y) = 0 \quad \forall\ y > 0. \tag{8.69}$$

If (S) holds, then by (8.58)

$$\sup_{d\le u\le r}\frac{|F(\theta+u)-F(\theta)|}{\rho(\theta)-\rho(\theta+u)}\le 2\,F'(\theta)\frac{d}{\rho(\theta)-\rho(\theta+d)},$$

whence by the Tschebyscheff-inequality

$$\begin{aligned}
P_{3,n}(y) &\le \mathbb{P}(|\bar{Y}_n-\bar{m}|\ge \frac{1}{6\,F'(\theta)}\frac{\rho(\theta)-\rho(\theta+d)}{d})\\
&\le 36\,F'(\theta)^2 Var(Y)\,n^{-1}\,\frac{d^2}{(\rho(\theta)-\rho(\theta+d))^2}\\
&= 36\,F'(\theta)^2 Var(Y)y^2\,(\sqrt{n\alpha_n}(\rho(\theta)-\rho(\theta+y/\alpha_n)))^{-2}\,\alpha_n^{-1} \quad \text{since } d=y/\alpha_n\\
&\le 144\,F'(\theta)^2 Var(Y)y^2\,\psi(y)^{-2}\,\alpha_n^{-1} \quad \forall\, n\ge n_1 \qquad \text{by (59)}
\end{aligned}$$

and consequently

$$\lim_{n\to\infty} P_{3,n}(y)=0 \quad \forall\, y>0. \tag{8.70}$$

After all with (8.24), (8.26), (8.64), (8.68), (8.69) and (8.70) we arrive at

$$\limsup_{n\to\infty}\mathbb{P}(y/\alpha_n\le\hat{\theta}_n-\theta\le r)\le C_+\{\,y\,\psi(y)^{-2}+\int_y^\infty \psi(s)^{-2}ds\}\quad\forall\, y>0$$

for some $r>0$ sufficiently small. Hence, by (8.22) and Theorem 2 the upper tail bounds (8.17) and (8.20) follow immediately.

For the derivation of the lower tail bounds (8.18) and (8.21) we follow the same arguments. Here, one has to use the Doob-Meyer decomposition of the process

$$\bar{L}_n(t):=n^{-1}\sum_{i=1}^n 1_{\{X_i>t\}}m(X_i)$$

into a reverse martingale plus compensator, confer Theorem A in the Appendix. Finally, the explicit tail probability (8.19) results from (8.17) and (8.18) by an elementary integration of ψ given in (8.16). □

Observe that by (8.19) the sequence $n(\hat{\theta}_n-\theta)$ under (J) is stochastically bounded, i.e.,

$$\lim_{y\to\infty}\limsup_{n\to\infty}\mathbb{P}(n|\hat{\theta}_n-\theta|\ge y)=0. \tag{8.71}$$

Similarly, under (S) we obtain *one-sided stochastic boundedness*.

Corollary 1 *Let the preliminary assumptions of Theorem 3 be true and* $\mathbb{E}(Y^2)$ *be finite. If (S) holds with* ϕ *and* φ *monotone increasing, then*

$$\limsup_{n\to\infty} \mathbb{P}(\alpha_n(\hat{\theta}_n - \theta) \geq y) \leq L_+ y^{-1} \quad \forall\, y \geq 1 \tag{8.72}$$

$$\limsup_{n\to\infty} \mathbb{P}(\beta_n(\hat{\theta}_n - \theta) \leq -y) \leq L_- y^{-1} \quad \forall\, y \geq 1 \tag{8.73}$$

where L_+ *and* L_- *are finite constants. In particular,*

$$\lim_{y\to\infty} \limsup_{n\to\infty} \mathbb{P}(\alpha_n(\hat{\theta}_n - \theta) \geq y) = 0 \quad (stochastic\, boundedness\, from\, above) \tag{8.74}$$

and

$$\lim_{y\to\infty} \limsup_{n\to\infty} \mathbb{P}(\beta_n(\hat{\theta}_n - \theta) \leq -y) = 0 \quad (stochastic\, boundedness\, from\, below). \tag{8.75}$$

Proof For every $u \geq 1$ we have by (8.13) that

$$\begin{aligned}
\psi(u) &= F'(\theta)\{\int_0^{1/2} \phi(s)ds + \int_{1/2}^{u} \phi(s)ds\} \\
&\geq F'(\theta)\phi(1/2)(u - 1/2) && \text{since } \phi > 0 \text{ is monotone increasing} \\
&\geq 1/2 F'(\theta)\phi(1/2)\, u && \text{since } 1/2 \leq u/2,
\end{aligned}$$

whence

$$\psi(u)^{-2} \leq c\, u^{-2} \quad \forall\, u \geq 1$$

with finite constant c. Consequently, by elementary integration (8.20) yields (8.72), which in turn gives (8.74). Analogously, we obtain (8.73) and (8.75). □

Typically, we obtain sharper asymptotic tail bounds as in (8.72) and (8.73). Indeed, in the situation of Example 1 it follows from Theorem 3 that:

$$\limsup_{n\to\infty} \mathbb{P}(n^{\frac{1}{1+2\alpha}} L^*(n)(\hat{\theta}_n - \theta) \geq y) \leq K_+\, y^{-(1+2\alpha)} \quad \forall\, y > 0, \tag{8.76}$$

$$\limsup_{n\to\infty} \mathbb{P}(n^{\frac{1}{1+2\beta}} l^*(n)(\hat{\theta}_n - \theta) \leq -y) \leq K_-\, y^{-(1+2\beta)} \quad \forall\, y > 0, \tag{8.77}$$

where K_+ and K_- are finite constants.

8.2 Convergence in Distribution to a Random Closed Set

Recall that $\alpha_n = \beta_n = n$ under the assumption (J). For simplicitly, let us assume that (S) holds likewise with $\alpha_n = \beta_n$. To set a concrete example consider

$$m(x) - m(\theta) = -\text{sign}(x-\theta)|x-\theta|\log(|x-\theta|^{-1}), \quad x \in [\theta-\varepsilon, \theta+\varepsilon]\backslash\{\theta\},$$

which by Example 1 induces

$$\alpha_n = \beta_n = \frac{1}{\sqrt[3]{9}}\, n^{1/3}\log(n)^{2/3}.$$

We will show that $\alpha_n(\hat{\theta}_n - \theta)$ converges in distribution, where the limit variable in general is a random closed set and not only a random real point as in the classical theory. Here, our starting point is the rescaled process.

$$Z_n(t) := \gamma_n\{E_n(\theta + t/\alpha_n) - E_n(\theta)\}, \quad t \in \mathbb{R},$$

where γ_n is some appropriate sequence. By Lemma 2.2 (i) and (iii) in Ferger (2015) we obtain that

$$\alpha_n(\hat{\theta}_n - \theta) = \underset{t\in\mathbb{R}}{\text{argmax}}\ Z_n(t)$$

with Z_n a cadlag process. If we can show that

$$Z_n \xrightarrow{\mathscr{L}} Z \quad \text{in the Skorokhod space } D[-a,a] \quad \forall\, a > 0, \tag{8.78}$$

then stochastic boundedness (8.71) or (8.74) and (8.75), respectively, enables us to apply Theorem 3.11 of Ferger (2015). In combination with Theorem 3.13 in Ferger (2015) it ensures that

$$\limsup_{n\to\infty} \mathbb{P}(\alpha_n(\hat{\theta}_n - \theta) \in F) \le T_C(F) \quad \text{for all closed } F \subseteq \mathbb{R}, \tag{8.79}$$

where

$$C = A(Z) := \text{set of all supremizing points of } Z$$

and

$$T_C(F) = \mathbb{P}(C \cap F \neq \emptyset).$$

The set-function T_C is called *capacity functional of* C. According to a Theorem of Matheron there is an extension of T_C onto the Borel-σ algebra $\mathscr{B}(\mathbb{R})$ such that

$$T_C(B) = \mathbb{P}(C \cap B \neq \emptyset) \quad \text{for all Borel-sets } B \subseteq \mathbb{R}.$$

Moreover Choquet's Theorem states that T_C uniquely determines the distribution of the *random closed set* C. As a capacity functional T_C has many features in common with probability measures. On the other hand T_C is merely sub-additive and in general lacks additivity, whence it can be regarded as a generalization of a probability measure. (See, e.g., Matheron (1975), Molchanov (2005) or Nguyen (2006) for an introduction to the theory of random closed sets and capacity functionals.) Notice that (8.79) formally looks exactly like the equivalent characterization of weak convergence given by the Portmanteau-Theorem. In consideration of all these facts we take (8.79) as the basis for saying that the sequence $\alpha_n(\hat{\theta}_n - \theta)$ of random points on the real line *converges in distribution to the random closed set* C:

$$\alpha_n(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} C.$$

If (J) holds, then it turns out that the limit process Z in (8.78) is a two-sided compound Poisson process with $Z(t) \to -\infty$ a.s. as $|t| \to \infty$. Thus C is the finite union of disjoint compact intervals with probability one. If (S) holds (with the liminf replaced by lim and the inequality by equality) then Z is equal to a two-sided Brownian motion with a drift downwards. This process has almost surely a unique maximizing point τ, i.e., $C = \{\tau\}$. An application of Theorem 3.12 in Ferger (2015) gives traditional weak convergence:

$$\alpha_n(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} \tau.$$

Observe that Theorems 3.11 and 3.12 of Ferger (2015) contain two basic conditions, namely:

(1) $\alpha_n(\hat{\theta}_n - \theta)$ is stochastically bounded.
(2) Z_n converges weakly in the Skorokhod space.

Here, (1) is an essential part, which is covered by (8.71) or (8.74) and (8.75), respectively. A full treatment of (2), i.e., the functional limit theorem (8.78) will appear elsewhere, since it is beyond the scope of this paper. Moreover, we will derive limit theorems in case the sequences α_n and β_n are different.

8.3 Appendix

Recall the definitions of the marked empirical distribution function

$$\hat{L}_n(t) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \leq t\}} m(X_i), \quad t \in \mathbb{R}$$

and its reverse counterpart

$$\bar{L}_n(t) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i > t\}} m(X_i), \quad t \in \mathbb{R}.$$

The following theorem yields the Doob-Meyer decompositions of $\hat{L}_n$ and $\bar{L}_n$, where m needs not to be our regression function and the X_i may stem from any distribution function F.

Theorem A *Let $X_1, \ldots, X_n$ be i.i.d. with arbitrary distribution function F and let $m : \mathbb{R} \to \mathbb{R}$ be a measurable function, which at each point $t \in \mathbb{R}$ is either right-continuous with left-limit or left-continuous with right-limit. Then it holds:*

$$\hat{L}_n(t) = M_n(t) + A_n(t), \quad t \in \mathbb{R},$$

where

$$(M_n(t), \mathscr{F}_n(t) : t \in \mathbb{R})$$

is a centered rcll martingale with respect to the filtration

$$\mathscr{F}_n(t) = \sigma\big(\bigcup_{i=1}^{n} \sigma(\{\{X_i \le r\} : r \le t\})\big).$$

Moreover

$$A_n(t) = \int_{(-\infty, t]} \frac{1 - F_n(x-)}{1 - F(x-)} m(x) F(dx).$$

Similarly,

$$\bar{L}_n(t) = R_n(t) + B_n(t), \quad t \in \mathbb{R},$$

where

$$(R_n(t), \mathscr{G}_n(t) : t \in \mathbb{R})$$

is a centered rcll reverse martingale with respect to the filtration

$$\mathscr{G}_n(t) = \sigma\big(\bigcup_{i=1}^{n} \sigma(\{\{X_i > r\} : r \ge t\})\big).$$

Moreover

$$B_n(t) = \int_{(t,\infty)} \frac{F_n(x)}{F(x)} m(x) F(dx).$$

Proof See Ferger (2009) for the decomposition of $\hat{L}_n$. The process $\bar{L}_n$ can be treated analogously. □

The next result generalizes a martingale inequality, which goes back to Birnbaum and Marshall (1961), see also Shorack and Wellner (1986).

Theorem B *Let* $(S(u), \mathfrak{F}(u) : u \in [a,b])$, $a < b$, *be a submartingale with trajectories that are right-continuous with left limits. Let* $S(u)^+ := \max\{S(u), 0\}$ *and* $H(u) := \mathbb{E}(S(u)^+) < \infty$, $u \in [a,b]$. *Furthermore, let* $w : [a,b] \to (0,\infty)$ *be rcll and monotone decreasing. Then for all* $\lambda > 0$

$$P\left(\sup_{a \le u \le b} w(u)S(u) > \lambda\right) \le \lambda^{-1}\left(\int_a^b w(u)H(du) + w(a)H(a)\right).$$

Proof A slight modification of the proof of Lemma 3.3 in Ferger and Venz (2017) gives

$$P\left(\sup_{a \le u \le b} w(u)S(u) > \lambda\right) \le \lambda^{-1}\left(\int_a^b H(u)(-w)(du) + w(b)H(b)\right)$$

and integration by parts yields the assertion. □

The function ρ plays a crucial role. We summarize some of its properties.

Lemma C *Let*

$$\rho(t) = \int_{(-\infty,t]} (m(x) - \bar{m}) F(dx), \quad t \in \mathbb{R}.$$

Then the following statements hold:

(1) ρ is rcll.
(2) ρ is continuous at $t \Leftrightarrow m(t) = \bar{m}$ or F is continuous at t.
(3) ρ is monotone increasing on $(-\infty, \theta)$, if $m \ge \bar{m}$ on $(-\infty, \theta)$ and monotone decreasing on $[\theta, \infty)$, if $m \le \bar{m}$ on $[\theta, \infty)$.

(4) Assume (C1) and (C2) hold. If F is strictly monotone on $[\theta - \varepsilon, \theta + \varepsilon]$ then ρ is strictly monotone (increasing or decreasing, respectively) on $[\theta - \varepsilon, \theta + \varepsilon]$. In particular θ is the unique and *well-separated* supremizing point of ρ, i.e.,

$$\rho(\theta) > \sup\{\rho(t) : |t - \theta| \geq \eta\} \quad \forall\, \eta > 0.$$

Proof See Ferger (2009) or Ferger et al. (2012).

References

Birnbaum, Z., Marshall A.: Some multivariate Chebyshev inequalities with extensions to continuous parameter processes. Ann. Math. Statist. **32**, 687–703 (1961)

Bojanic, R., Seneta, E.: Slowly varying functions and asymptotic relations. J. Math. Anal. Appl. **34**, 302–315 (1971)

Dempfle, A., Stute, W.: Nonparametric estimation of a discontinuity in regression. Stat. Neerl. **56**, 233–242 (2002)

Ferger, D.: Stochastische Prozesse mit Strukturbrüchen. Dresdner Schriften zur Mathematischen Stochastik **7**, 1–123 (2009)

Ferger, D., Klotsche, J., Lüken, U.: Estimation and testing of crossing-points in fixed design regression. Stat. Neerl. **66**, 380–402 (2012)

Ferger, D.: Arginf-sets of multivariate cadlag processes and their convergence in hyperspace topologies. Theory Stoch. Process. **20 (36)**, 13–41 (2015)

Ferger, D., Venz, J.: Density estimation via best L^2-approximation on classes of step functions. Kybetnetika **53**, 198–219 (2017)

Knight, K.: Limiting distributions for L_1 regression estimators under general conditions. Ann. Statist. **26**, 755–770 (1998)

Massart, P.: The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. Ann. Probab. **18**, 1269–1283 (1990)

Matheron G.: Random Sets and Integral Geometry. John Wiley & Sons, New York, London, Sydney, Toronto (1975)

Molchanov I.: Theory of Random Sets. Springer-Verlag, London (2005)

Nguyen H.T.: An Introduction to Random Sets. Chapman & Hall/CRC, Boca Raton, London, New York (2006)

Shorack, G.R., Wellner, J.A.: Empirical Processes With Applications to Statistics. John Wiley & Sons, New York (1986)

Smirnov, N.V.: Limit distributions for the terms of a variational series. Amer. Math. Soc. Transl. Ser. (1) **11**, 82–143 (1952)

Stute, W.: Nonparametric model checks for regression. Ann. Statist. **25**, 613–641 (1997)

Stute, W., Wang, J.-L.: The strong law under random censorship. Ann. Statist. **21**, 1591–1607 (1993)

On Empirical Distribution Functions Under Auxiliary Information

9

Erich Haeusler

9.1 Introduction

Let $(X_i)_{i\geq 1}$ be a sequence of independent and identically distributed random variables with common continuous distribution function F. The nonparametric maximum likelihood estimator for estimating F based on the first n of these random variables is the classical empirical distribution function F_n and has been studied extensively.

Suppose now that auxiliary information about F of a nonparametric nature is available in the sense that for known measurable functions $g_1(x), \ldots, g_r(x)$ we have

$$E(g(X_1)) = 0\,, \tag{9.1}$$

where $g\,(x) = (g_1\,(x)\,, \ldots, g_r\,(x))$ for $x \in \mathbb{R}$. Examples are

(i) an integrable X_1 whose mean $E\,(X_1)$ has a known value $\mu \in \mathbb{R}$, where $g\,(x) = x - \mu$,

(ii) an X_1 for which it is known that a given $m \in \mathbb{R}$ is a median, i.e., $F\,(m) = \frac{1}{2}$; here $g\,(x) = 1_{(-\infty,m]}\,(x) - \frac{1}{2}$, or

(iii) a square integrable X_1 whose mean $E\,(X_1)$ has a known value $\mu \in \mathbb{R}$ and whose variance $\mathrm{Var}\,(X_1)$ has a known value $\sigma^2 \in (0, \infty)$, where g now is the $\mathbb{R}^2$-valued function $g\,(x) = \left(x - \mu, x^2 - \sigma^2 - \mu^2\right)$.

For model (9.1) (extended by an additional parameter θ which we will not consider here), using ideas from the concept of empirical likelihood as developed by Owen in (1988, 1990, 1991) (see also Owen 2001), Qin and Lawless in (1994) have derived the nonparametric maximum likelihood estimator $F_{n,g}$ for F, which turns out to be a

E. Haeusler (✉)
Mathematical Institute, University of Giessen, Arndtstrasse 2, 35392 Giessen, Germany
e-mail: erich.haeusler@math.uni-giessen.de

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_9

distribution function which puts random masses on the observed data $x_1, \ldots, x_n$ from $X_1, \ldots, X_n$. Qin and Lawless also derived asymptotic normality of the random variable $n^{1/2}\left(F_{n,g}(x) - F(x)\right)$ for fixed $x \in \mathbb{R}$, whereas Zhang in (1997) established the corresponding functional central limit theorem, i.e. weak convergence of the entire stochastic process $n^{1/2}\left(F_{n,g} - F\right) = \left(n^{1/2}\left(F_{n,g}(x) - F(x)\right)\right)_{x\in\mathbb{R}}$ towards an appropriate Gaussian process. For model (9.1), Theorem 3.3 in Zhang (1997) contains the following result: If

$$E\left(\|g(X_1)\|^3\right) < \infty, \tag{9.2}$$

where $\|\ \|$ denotes any norm in $\mathbb{R}^r$, and

$$\Sigma = E\left(g(X_1)\, g(X_1)^T\right) \quad \text{is positive definite}, \tag{9.3}$$

then

$$n^{1/2}\left(F_{n,g} - F\right) \xrightarrow{\mathcal{L}} W \quad \text{in} \quad D[-\infty, \infty] \quad \text{as } n \to \infty, \tag{9.4}$$

where $D[-\infty, \infty]$ is the Skorohod space over the compact time interval $[-\infty, \infty]$, where $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution and W is a centered Gaussian process with covariance function

$$\begin{aligned} E(W(x)\, W(y)) =& F(\min(x, y)) - F(x)\, F(y) \\ & - U(x)^T\, \Sigma^{-1} U(y) \quad \text{for } x, y \in \mathbb{R}, \end{aligned}$$

where $U(x) = E\left(g(X_1)\, 1_{\{X_1 \le x\}}\right)$. It is important to point out here that the third moment condition (9.2) is imposed in Zhang (1997) because there the more general model with the additional parameter θ mentioned earlier is considered for which this third moment condition is essential. If the more restricted model (9.1) is considered, as it is done in this note, then the functional central limit theorem (9.4) already holds if condition (9.2) is replaced by the weaker and somewhat more natural second moment condition

$$E\left(\|g(X_1)\|^2\right) < \infty. \tag{9.5}$$

This can be seen by an inspection of the proof of Theorem 3.3 in Zhang (1997) and is essential for the point we want to make later. Note that condition (9.5) is sufficient to guarantee the existence of the covariance matrix appearing in condition (9.3) which means that under (9.5) the only requirement in condition (9.3) is the positive definiteness of the matrix Σ.

If the functional central limit theorem (9.4) is compared to the classical Donsker functional central limit theorem

$$n^{1/2}(F_n - F) \xrightarrow{\mathcal{L}} Z \quad \text{in} \quad D[-\infty, \infty] \quad \text{as } n \to \infty \tag{9.6}$$

for the classical empirical distribution function F_n, where Z is a centered Gaussian process with covariance function

$$E\left(Z\left(x\right)Z\left(y\right)\right) = F\left(\min\left(x,\, y\right)\right) - F\left(x\right)F\left(y\right) \quad \text{for } x,\, y \in \mathbb{R}\,,$$

then clearly $\mathrm{Var}\left(Z\left(x\right)\right) \geq \mathrm{Var}\left(W\left(x\right)\right)$ for all $x \in \mathbb{R}$. Thus, the variances of the limit process Z are pointwise at least as big as the variances of the limit process W. More generally, if $\mathrm{Cov}\left[Y_1, \ldots, Y_m\right]$ denotes the covariance matrix of the random vector $\left(Y_1, \ldots, Y_m\right)$, then

$$\mathrm{Cov}\left[Z\left(x_1\right), \ldots, Z\left(x_m\right)\right] \geq \mathrm{Cov}\left[W\left(x_1\right), \ldots, W\left(x_m\right)\right] \quad \text{for all } x_1, \ldots, x_m \in \mathbb{R}$$

in the sense of the Loewner ordering for symmetric matrices (i.e., $B \geq A$ if $B - A$ is nonnegative definite). This entails, for example, that asymptotic bootstrap confidence bands for F based on $F_{n,g}$ tend to be smaller, and sometimes considerably smaller, than confidence bands based on F_n; see e.g. Haeusler and Plies (2000). Therefore, the modified distribution function $F_{n,g}$ has advantages over F_n in statistical applications, provided that, of course, the assumptions in model (9.1) are satisfied.

In order to further enlighten the role of the assumptions for (9.4) in model (9.1) we consider the examples (i) and (ii) from above in some detail. Let us begin with example (ii). Because the function $g\left(x\right) = 1_{(-\infty,m]}\left(x\right) - \frac{1}{2}$ is bounded, condition (9.5) holds for all continuous distribution functions F. Because g is one-dimensional, condition (9.3) reduces to $E\left(g\left(X_1\right)^2\right) > 0$, which is also clearly satisfied for all continuous F. Therefore, the functional central limit theorem (9.4) holds for all continuous distribution functions with known median m. In example (i) we will assume w.l.o.g. that $\mu = 0$ and will write $c\left(x\right) = x$ instead of $g\left(x\right) = x$ from now on. Clearly, condition (9.5) for the one-dimensional function c is equivalent to $E\left(X_1^2\right) < \infty$, i.e., to square integrability of X_1, whereas condition (9.3) is equivalent to $\mathrm{Var}\left(X_1\right) = E\left(X_1^2\right) = E\left(c\left(X_1\right)^2\right) > 0$, which is always satisfied for continuous F. Nevertheless, in example (i) the required auxiliary information about X_1 for (9.4) to hold is twofold: It has to be known that X_1 has mean zero and that it is not only integrable, but square integrable. Therefore, (9.4) provides no information about the asymptotic distributional behaviour of $n^{1/2}\left(F_{n,c} - F\right)$ if X_1 is integrable with known mean zero, but has infinite variance. It is the aim of this note to shed some light on this problem. For this, we will first give the complete definition of $F_{n,g}$ in Sect. 9.2. In Sect. 9.3 we will present two classes of centered random variables X_1 with infinite variance for which

$$n^{1/2}\left\|F_{n,c} - F_n\right\|_\infty = o_P\left(1\right) \quad \text{as } n \to \infty \tag{9.7}$$

holds true, where $\|\ \|_\infty$ denotes the supremum norm. From (9.6) and (9.7) it follows by Cramér's theorem that

$$n^{1/2}\left(F_{n,c} - F\right) \overset{\mathscr{L}}{\to} Z \quad \text{in} \quad D\left[-\infty, \infty\right] \quad \text{as } n \to \infty\,,$$

in contrast to (9.4). Consequently, for (9.4) to hold, square integrability of X_1 is essential, besides the knowledge of the mean of X_1. If the auxiliary information about X_1 is only integrability and a known mean, (9.4) need not to be true. (Of course, the statement of (9.4) itself requires already the existence of $\Sigma = \mathrm{Var}(X_1)$, i.e., square integrability of X_1).

9.2 The Definition of $F_{n,g}$ and $F_{n,c}$

Let F be a continuous distribution function and $X_1, \ldots, X_n$ be independent and identically distributed random variables with common distribution function F. Let $x_1, \ldots, x_n$ be an observed sample from $X_1, \ldots, X_n$, i.e., $x_i \in \mathbb{R}$ is a realization of the random variable X_i for every $i = 1, \ldots, n$. The nonparametric likelihood function of the sample $x_1, \ldots, x_n$ is, for every distribution function $\widetilde{F}$, given by

$$L\left(\widetilde{F}\right) = \prod_{i=1}^{n} \left(\widetilde{F}(x_i) - \widetilde{F}(x_i - 0)\right),$$

where $\widetilde{F}(x - 0)$ denotes the left-hand limit of $\widetilde{F}$ at $x \in \mathbb{R}$. The unique maximizer of the function L is the classical empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(x_i), \quad x \in \mathbb{R},$$

of the sample $x_1, \ldots, x_n$. To see this, note that the maximizer of L necessarily is of the form

$$\widetilde{F}(x) = \sum_{i=1}^{n} p_i 1_{(-\infty, x]}(x_i), \quad x \in \mathbb{R},$$

with $p_i > 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$ so that maximizing L reduces to maximizing

$$\widetilde{L}(p_1, \ldots, p_n) = \prod_{i=1}^{n} p_i$$

for these p_i. This easily leads to $p_i = \frac{1}{n}$ for $i = 1, \ldots, n$ and hence to F_n. This establishes F_n as the nonparametric maximum likelihood estimator for F.

Suppose now that we want to estimate F inside model (9.1). Then the estimator should satisfy the same restrictions given by (9.1) as F does. Consequently, the function L should be maximized over all distribution functions $\widetilde{F}$ satisfying the

restrictions imposed by (9.1). This leads to the maximization of $\widetilde{L}$ subject to the conditions

$$p_i > 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(x_i) = 0. \tag{9.8}$$

Note that the auxiliary information provided by (9.1) is taken into account through the equation $\sum_{i=1}^{n} p_i g(x_i) = 0$. As a consequence, it is no longer guaranteed that a maximizer of $\widetilde{L}$ exists for all samples $x_1, \ldots, x_n$. However, if 0 is an interior point of the convex hull of the points $g(x_1), \ldots, g(x_n)$ in $\mathbb{R}^r$, then a unique maximizer $(p_{n,1}, \ldots, p_{n,n})$ satisfying the conditions in (9.8) does exist, and it can be found by the method of Lagrange multipliers; see Owen (1990), p. 100, and Qin and Lawless (1994), p. 304/305. We have

$$p_{n,i} = \frac{1}{n\left(1 + t_n^T g(x_i)\right)} \quad \text{for } i = 1, \ldots, n.$$

Here t_n denotes the r-dimensional Lagrange multiplier and is a solution of the equation

$$\sum_{i=1}^{n} \frac{g(x_i)}{1 + t_n^T g(x_i)} = 0. \tag{9.9}$$

Because of $p_{n,i} < 1$ for $i = 1, \ldots, n$, the vector t_n belongs to the open convex set

$$M_n = \left\{ t \in \mathbb{R}^r : 1 + t^T g(x_i) > \frac{1}{n} \quad \text{for } i = 1, \ldots, n \right\}.$$

Moreover, Eq. (9.9) has exactly one solution in M_n provided that the $r \times r$-matrix $\sum_{i=1}^{n} g(x_i)\, g(x_i)^T$ is positive definite; see Owen (1990), p. 105, or use the mean value theorem for a direct proof. Notice that $\sum_{i=1}^{n} g(x_i)\, g(x_i)^T$ is always nonnegative definite by construction.

Consequently, under model (9.1) the appropriate empirical distribution function for the observed sample $x_1, \ldots, x_n$ is defined by

$$F_{n,g}(x) = \sum_{i=1}^{n} \frac{1}{n\left(1 + t_n^T g(x_i)\right)} 1_{(-\infty, x]}(x_i), \quad x \in \mathbb{R},$$

with t_n being the unique solution of (9.9) in M_n, provided that the sample satisfies the conditions

$$0 \text{ is an interior point of the convex hull of } \{g(x_1), \ldots, g(x_n)\} \subset \mathbb{R}^r \tag{9.10}$$

and

$$\sum_{i=1}^{n} g(x_i)\, g(x_i)^T \text{ is positive definite}. \tag{9.11}$$

Of course, not all samples $x_1, \ldots, x_n$ will usually satisfy these conditions. For a reasonable definition of $F_{n,g}$, we have to ensure at least that the probability that a sample $x_1, \ldots, x_n$ satisfies conditions (9.10) and (9.11) converges to one as the sample size n tends to infinity. To do this formally, we introduce the random events

$$A_n = \{0 \text{ is an interior point of the convex hull of } \{g(X_1), \ldots, g(X_n)\}\}$$

and

$$B_n = \left\{ \sum_{i=1}^{n} g(X_i)\, g(X_i)^T \text{ is positive definite} \right\}.$$

Then on $A_n \cap B_n$ the random distribution function

$$F_{n,g}(x) = \sum_{i=1}^{n} \frac{1}{n\left(1 + t_n^T g(X_i)\right)} 1_{(-\infty,x]}(X_i), \quad x \in \mathbb{R},$$

is well-defined by the requirement that the (random) vector $t_n \in \mathbb{R}^r$ is the unique vector in the (random) set

$$M_n = \left\{ t \in \mathbb{R}^r : 1 + t^T g(X_i) > \frac{1}{n} \quad \text{for } i = 1, \ldots, n \right\}$$

satisfying the equation

$$\sum_{i=1}^{n} \frac{g(X_i)}{1 + t_n^T g(X_i)} = 0.$$

Under (9.1) and (9.3) the proof of Lemma 2 in Owen (1990) adapted to the present situation gives, as $n \to \infty$,

$$P(A_n) \to 1, \tag{9.12}$$

whereas by (9.3) and (9.5) the law of large numbers implies

$$P(B_n) \to 1. \tag{9.13}$$

Hence under (9.3) and (9.5) the empirical distribution function $F_{n,g}$ is well-defined with probability converging to one.

Let us now discuss the definition of $F_{n,g}$ in the special case of example (i) with $\mu = 0$, i.e., in the notation of Sect. 9.1, the definition of $F_{n,c}$ if X_1 is integrable with mean zero, but need not be square integrable. Because of $g(x) = c(x) = x$, and observing that $X_1, \ldots, X_n$ are real-valued random variables, we see that the interior of the convex hull of $\{g(X_1), \ldots, g(X_n)\} = \{X_1, \ldots, X_n\}$ equals the open interval $\left(\min_{1 \le i \le n} X_i, \max_{1 \le i \le n} X_i\right)$ and hence $A_n = \left\{\min_{1 \le i \le n} X_i < 0 < \max_{1 \le i \le n} X_i\right\}$. Therefore, condition (9.12) now reads

$$P\left(\min_{1 \le i \le n} X_i < 0 < \max_{1 \le i \le n} X_i \right) \to 1 \quad \text{as } n \to \infty. \tag{9.14}$$

But for an integrable random variable X_1 with continuous distribution function F and mean zero we have $0 < F(0) < 1$ so that, as $n \to \infty$,

$$P\left(\min_{1\le i\le n} X_i \ge 0\right) = P\left(\bigcap_{i=1}^{n} \{X_i \ge 0\}\right) = P(X_1 \ge 0)^n = (1 - F(0))^n \to 0$$

and, similarly,

$$P\left(\max_{1\le i\le n} X_i \le 0\right) = F(0)^n \to 0\,.$$

Consequently, condition (9.14) and hence (9.12) is satisfied for an integrable X_1 with mean zero and a continuous distribution function. Condition (9.3), which was vital in the general case, is not needed. Moreover, using again $g(x) = c(x) = x$ and the fact that $X_1, \ldots, X_n$ are real-valued, we get $B_n = \left\{\sum_{i=1}^n X_i^2 > 0\right\}$. But $P(B_n) = 1$ for all $n \ge 1$ if $X_1, \ldots, X_n$ are independent and identically distributed with a continuous distribution function so that condition (9.13) is trivially satisfied. Note that, as before, condition (9.3) as well as the second moment condition (9.5) are not needed. Continuity of the distribution function alone is sufficient here. Thus we have seen that the two conditions (9.12) and (9.13) which are essential for a reasonable definition of $F_{n,c}$ are satisfied whenever X_1 is integrable with mean zero and has a continuous distribution function, i.e., in the setup of example (i).

The explicit definition of $F_{n,c}$ on $A_n \cap B_n$ is given by

$$F_{n,c}(x) = \sum_{i=1}^{n} \frac{1}{n(1 + t_n X_i)} 1_{(-\infty,x]}(X_i)\,, \quad x \in \mathbb{R}\,, \tag{9.15}$$

with t_n being the unique real number in the set

$$\begin{aligned} M_n &= \left\{t \in \mathbb{R} : 1 + tX_i > \frac{1}{n} \quad \text{for } i = 1, \ldots, n\right\} \\ &= \left(\left(\frac{1}{n} - 1\right)\frac{1}{\max_{1\le i\le n} X_i}, \left(\frac{1}{n} - 1\right)\frac{1}{\min_{1\le i\le n} X_i}\right) \end{aligned} \tag{9.16}$$

with

$$\sum_{i=1}^{n} \frac{X_i}{1 + t_n X_i} = 0\,. \tag{9.17}$$

The fact that in example (i) the set M_n equals the open interval given in (9.16) will be crucial in the next section.

9.3 The Asymptotic Distributional Behaviour of $F_{n,c}$ for Two Classes of Centered Random Variables with Infinite Variance

Recall that a real-valued positive measurable function L defined on some interval (x_0, ∞) is called slowly varying at infinity if

$$\frac{L(\lambda x)}{L(x)} \to 1 \quad \text{as } x \to \infty \quad \text{for all } \lambda \in (0, \infty) .$$

Because we will exclusively use functions which are slowly varying at infinity we will drop the specification "at infinity" in the sequel. A standard reference on slowly varying functions is the monograph (Bingham et al. 1987).

Now we can introduce the two classes of random variables which we will consider in this section:

Class 1 contains all continuous distribution functions F for which there exist constants $p_-, p_+ \in [1, 2)$, $-\infty < x_- < 0 < x_+ < \infty$ and slowly varying functions L_-, L_+ such that

$$F(x) = (-x)^{-p_-} L_-(-x) \quad \text{for all } x \in (-\infty, x_-)$$

and

$$1 - F(x) = x^{-p_+} L_+(x) \quad \text{for all } x \in (x_+, \infty) .$$

Moreover, we assume that a random variable X with distribution function F is integrable (which is automatically satisfied if $p_- > 1$ and $p_+ > 1$, but not if $p_- = 1$ or $p_+ = 1$) and has mean zero. Note that X always has infinite variance.

Class 2 contains all continuous distribution functions F for which there exist constants $-\infty < x_- < 0 < x_+ < \infty$ and a slowly varying function L such that

$$F(x) = x^{-2} L(-x) \quad \text{for all } x \in (-\infty, x_-)$$

and

$$1 - F(x) = x^{-2} L(x) \quad \text{for all } x \in (x_+, \infty) .$$

Moreover, we assume that a random variable X with distribution function F has mean zero (note that X is always integrable) and infinite variance (which is not automatically true but holds, for example, if $L(x) = 1$ for all large x).

Remark Recall that a positive measurable function f which is defined on an interval (x_0, ∞) is called regularly varying at infinity of index $\rho \in \mathbb{R}$ if $f(x) = x^\rho L(x)$ for $x \in (x_0, \infty)$ with some slowly varying function L; see Bingham et al. (1987), Theorem 1.4.1 and the Definition on p. 18. Thus, for F in Class 1 the tail-sum $1 - F(x) + F(-x)$ is the sum of two functions which are regularly varying at infinity of index $-p_+$ and $-p_-$, respectively, and therefore is regularly varying at

infinity of index $\max(-p_+, -p_-) = -\min(p_+, p_-)$ according to Bingham et al. (1987), Proposition 1.5.7 (iii). For all large $x \in \mathbb{R}$ we have

$$\frac{1-F(x)}{1-F(x)+F(-x)} = \frac{x^{-p_+}L_+(x)}{x^{-p_+}L_+(x)+x^{-p_-}L_-(x)} = \frac{L_+(x)}{L_+(x)+x^{p_+-p_-}L_-(x)}.$$

If $p_+ > p_-$, then for any positive $\varepsilon < \frac{1}{2}(p_+ - p_-)$ the Potter bounds on L_+ and L_- (see Bingham et al. (1987), Theorem 1.5.6 (i)) imply

$$L_+(x) \le A_\varepsilon x^\varepsilon \quad \text{and} \quad L_-(x) \ge B_\varepsilon x^{-\varepsilon} \quad \text{for all large } x \in \mathbb{R}$$

with positive constants A_ε and B_ε depending only on ε. Therefore, for all large $x \in \mathbb{R}$,

$$\frac{1-F(x)}{1-F(x)+F(-x)} \le \frac{L_+(x)}{x^{p_+-p_-}L_-(x)} \le \frac{A_\varepsilon}{B_\varepsilon} x^{-(p_+-p_-)+2\varepsilon} \to 0 \quad \text{as } x \to \infty$$

because $-(p_+ - p_-) + 2\varepsilon < 0$. An immediate consequence is

$$\frac{F(-x)}{1-F(x)+F(-x)} \to 1 \quad \text{as } x \to \infty.$$

For $p_+ < p_-$ a similar argument gives

$$\frac{1-F(x)}{1-F(x)+F(-x)} \to 1 \quad \text{and} \quad \frac{F(-x)}{1-F(x)+F(-x)} \to 0 \quad \text{as } x \to \infty.$$

Consequently, if $p_+ \neq p_-$, then the tail-balance condition (8.3.6) in Bingham et al. (1987) is satisfied, and by Theorem 8.3.1 (ii) in Bingham et al. (1987) we conclude that F belongs to the domain of attraction of a non-normal stable law; for a discussion of non-normal stable laws and their domains of attraction consult Bingham et al. (1987), Sects. 8.3.1 and 8.3.2.

If $p_+ = p_-$, then

$$\frac{1-F(x)}{1-F(x)+F(-x)} = \frac{L_+(x)}{L_+(x)+L_-(x)} = \frac{1}{1+L_-(x)/L_+(x)},$$

and the existence of the limit

$$\lim_{x\to\infty} \frac{1-F(x)}{1-F(x)+F(-x)} \quad (\text{in } [0,1])$$

is equivalent to the existence of the limit

$$\lim_{x\to\infty} \frac{L_-(x)}{L_+(x)} \quad (\text{in } [0,\infty]).$$

Since the last limit may or may not exist, F may or may not belong to the domain of attraction of a non-normal stable law.

For the remaining part of this section let $(X_i)_{i\geq 1}$ be a sequence of independent and identically distributed random variables with common distribution function F in Class 1 or Class 2. According to Sect. 9.2, the empirical distribution function $F_{n,c}$ based on $X_1, \ldots, X_n$ is well-defined with probability converging to one as n tends to infinity, and our main result in this note describes its distance to the classical empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i), \quad x \in \mathbb{R},$$

based on $X_1, \ldots, X_n$.

Theorem *For F in Class 1 or Class 2 we have*

$$n^{1/2}\left\|F_{n,c} - F_n\right\| = o_P(1) \quad \text{as } n \to \infty. \tag{9.7}$$

For the proof of this theorem we need two auxiliary results which will be formulated as lemmas.

Lemma 1 *Let F be a continuous distribution function for which there exist constants $p > 0$ and $0 < x_0 < \infty$ and a slowly varying function L such that*

$$1 - F(x) = x^{-p}L(x) \quad \text{for all } x \in (x_0, \infty). \tag{9.18}$$

If $(X_i)_{i\geq 1}$ is a sequence of independent and identically distributed random variables with common distribution function F, then, as $n \to \infty$,

$$\frac{\max_{1\leq i\leq n} X_i}{a_n} = O_P(1) \tag{9.19}$$

and

$$\frac{a_n}{\max_{1\leq i\leq n} X_i} = O_P(1), \tag{9.20}$$

where $a_n = n^{1/p}\widetilde{L}(n)$ for some slowly varying function $\widetilde{L}$ which depends only on the exponent p and the slowly varying function L.

Proof According to Assumption (9.18) the function $1 - F$ is regularly varying at infinity of index $-p$. Therefore, F is in the max-domain of attraction of the Fréchet extreme value distribution Φ_p (see e.g. Bingham et al. (1987), Theorem 8.13.2), which means that there exist constants a_n such that

$$\frac{\max_{1\leq i\leq n} X_i}{a_n} \xrightarrow{\mathscr{L}} Y_p \quad \text{as } n \to \infty,$$

where Y_p is a random variable with distribution Φ_p. This gives (9.19). By the continuous mapping theorem, the sequence $\left(a_n / \max_{1\le i\le n} X_i\right)_{n\ge 1}$ converges in distribution so that (9.20) holds as well. To complete the proof of the lemma, it therefore remains to determine the constants a_n. According to Bingham et al. (1987), Theorem 8.13.2, we can take

$$a_n = \inf\left\{x : 1 - F(x) \le \frac{1}{n}\right\} = \inf\left\{x : \frac{1}{1-F(x)} \ge n\right\} = \left(\frac{1}{1-F}\right)^{\leftarrow}(n) ,$$

where $f^{\leftarrow}(u) = \inf\{x : f(x) \ge u\}$ denotes the generalized inverse of a real-valued locally bounded function f with $f(x) \to \infty$ as $x \to \infty$. By Bingham et al. (1987), Proposition 1.5.7 (i), the function $\frac{1}{1-F}$ is regularly varying at infinity of index p so that by Bingham et al. (1987), Theorem 1.5.12, the function $\left(\frac{1}{1-F}\right)^{\leftarrow}$ is regularly varying at infinity of index $\frac{1}{p}$. Therefore, by Bingham et al. (1987), Theorem 1.4.1, we have

$$\left(\frac{1}{1-F}\right)^{\leftarrow}(x) = x^{1/p}\widetilde{L}(x)$$

for some slowly varying function $\widetilde{L}$ so that $a_n = n^{1/p}\widetilde{L}(n)$. Note that by (9.18) the right tail of F is completely determined by the exponent p and the slowly varying function L, and that $\left(\frac{1}{1-F}\right)^{\leftarrow}$ is a function of F. Therefore, $\widetilde{L}$ is completely determined by p and L, too. This completes the proof of the lemma. □

Since results for $\max_{1\le i\le n} X_i$ can be immediately transferred into results for $\min_{1\le i\le n} X_i$ through the relation

$$\min_{1\le i\le n} X_i = -\max_{1\le i\le n}(-X_i) ,$$

Lemma 1 yields the following

Corollary *Let F be a continuous distribution function for which there exist constants $p > 0$ and $-\infty < x_0 < 0$ and a slowly varying function L such that*

$$F(x) = (-x)^{-p} L(-x) \quad \text{for all } x \in (-\infty, x_0) .$$

If $(X_i)_{i\ge 1}$ is a sequence of independent and identically distributed random variables with common distribution function F, then, as $n \to \infty$,

$$\frac{\min_{1\le i\le n} X_i}{a_n} = O_P(1) \tag{9.21}$$

and

$$\frac{a_n}{\min_{1\le i\le n} X_i} = O_P(1) , \tag{9.22}$$

where $a_n = n^{1/p}\widetilde{L}(n)$ for some slowly varying function $\widetilde{L}$ which depends only on the exponent p and the slowly varying function L.

The next lemma is crucial for handling distribution functions from Class 2.

Lemma 2 *If $(X_i)_{i\geq 1}$ is a sequence of independent and identically distributed random variables with common distribution function in Class 2, then*

$$\left(\sum_{i=1}^{n} X_i^2\right)^{-1/2} \sum_{i=1}^{n} X_i = O_P(1) \quad as\ n \to \infty .$$

Proof We first show that any distribution function F from Class 2 is in the domain of attraction of the normal distribution; for a definition of this notion see e.g. Bingham et al. (1987), Sect. 8.3.1. For this, we will apply the normal convergence criterion in Bingham et al. (1987), Theorem 8.3.1 (i). Fix $x_0 \geq \max(x_+, -x_-)$ where x_+ and x_- are specified according to the definition of Class 2. For all $x \in (x_0, \infty)$, using the tail-symmetry of F, we decompose the truncated variance $V(x) = \int_{-x}^{x} t^2\, dF(t)$ of F as

$$V(x) = \int_{-x_0}^{x_0} t^2\, dF(t) + 2\int_{x_0}^{x} t^2\, dF(t) .$$

For the second summand on the right hand side we find, by an integration by parts,

$$\begin{aligned}\int_{x_0}^{x} t^2\, dF(t) &= 2\int_{x_0}^{x} t\,(1 - F(t))\, dt - x^2\,(1 - F(x)) + x_0^2\,(1 - F(x_0)) \\ &= 2\int_{x_0}^{x} \frac{L(t)}{t}\, dt - L(x) + x_0^2\,(1 - F(x_0)) .\end{aligned}$$

The first summand on the right hand side is slowly varying by Bingham et al. (1987), Proposition 1.5.9a, so that $V(x)$ is the sum of slowly varying functions and hence slowly varying by Bingham et al. (1987), Proposition 1.3.6 (iii). According to Bingham et al. (1987), Theorem 8.3.1 (i) this is equivalent to F being in the domain of normal attraction. Because X_1 has mean zero, Theorem 1.4 in Chistyakov and Götze (2004) implies that $\left(\sum_{i=1}^{n} X_i^2\right)^{-1/2} \sum_{i=1}^{n} X_i$ converges in distribution to a normal distribution and is therefore bounded in probability. □

Now we are prepared to give the

Proof of the Theorem Recall from Sect. 9.2 that on the event $A_n \cap B_n$ with

$$A_n = \left\{ \min_{1\leq i\leq n} X_i < 0 < \max_{1\leq i\leq n} X_i \right\} \quad \text{and} \quad B_n = \left\{ \sum_{i=1}^{n} X_i^2 > 0 \right\}$$

the empirical distribution function $F_{n,c}$ is defined by

$$F_{n,c}(x) = \sum_{i=1}^{n} \frac{1}{n\,(1 + t_n X_i)} 1_{(-\infty, x]}(X_i) , \quad x \in \mathbb{R} , \tag{9.15}$$

where t_n is the unique real number in

$$M_n = \left(\left(\frac{1}{n} - 1\right) \frac{1}{\max_{1 \le i \le n} X_i}, \left(\frac{1}{n} - 1\right) \frac{1}{\min_{1 \le i \le n} X_i}\right)$$

with

$$\sum_{i=1}^{n} \frac{X_i}{1 + t_n X_i} = 0\,. \tag{9.17}$$

As shown in Sect. 9.2 we have $P(A_n \cap B_n) \to 1$ as $n \to \infty$ so that for assertion (9.7) it plays no role how $F_{n,c}$ is defined on the complement of the event $A_n \cap B_n$. Therefore, we can and will ignore this complement completely and for simplicity assume w.l.o.g. from now on that $A_n \cap B_n$ is the whole sample space and that $F_{n,c}$ is defined by (9.15) and (9.17) on this whole sample space. Then

$$\begin{aligned} F_{n,c}(x) - F_n(x) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + t_n X_i} 1_{(-\infty,x]}(X_i) - \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(1 - t_n X_i + \frac{t_n^2 X_i^2}{1 + t_n X_i}\right) 1_{(-\infty,x]}(X_i) \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) \\ &= -\frac{1}{n} t_n \sum_{i=1}^{n} X_i 1_{(-\infty,x]}(X_i) + \frac{1}{n} t_n^2 \sum_{i=1}^{n} \frac{X_i^2}{1 + t_n X_i} 1_{(-\infty,x]}(X_i) \end{aligned}$$

and $1 + t_n X_i > \frac{1}{n}$ for $i = 1, \ldots, n$ so that the second summand on the right hand side of this chain of equations is nonnegative. Therefore, we get

$$n^{1/2} \sup_{x \in \mathbb{R}} \left|F_{n,c}(x) - F_n(x)\right| \le n^{-1/2} |t_n| \sum_{i=1}^{n} |X_i| + n^{-1/2} t_n^2 \sum_{i=1}^{n} \frac{X_i^2}{1 + t_n X_i}\,.$$

Moreover, by (9.17),

$$0 = \sum_{i=1}^{n} \frac{X_i}{1 + t_n X_i} = \sum_{i=1}^{n} \frac{1 + t_n X_i - t_n X_i}{1 + t_n X_i} X_i = \sum_{i=1}^{n} X_i - t_n \sum_{i=1}^{n} \frac{X_i^2}{1 + t_n X_i}$$

so that

$$t_n \sum_{i=1}^{n} \frac{X_i^2}{1 + t_n X_i} = \sum_{i=1}^{n} X_i\,. \tag{9.23}$$

This gives

$$n^{-1/2} t_n^2 \sum_{i=1}^{n} \frac{X_i^2}{1 + t_n X_i} = n^{-1/2} t_n \sum_{i=1}^{n} X_i \le n^{-1/2} |t_n| \sum_{i=1}^{n} |X_i|$$

so that

$$n^{1/2} \sup_{x \in \mathbb{R}} \left| F_{n,c}(x) - F_n(x) \right| \leq 2n^{-1/2} |t_n| \sum_{i=1}^{n} |X_i| = 2n^{1/2} |t_n| \frac{1}{n} \sum_{i=1}^{n} |X_i| .$$

Because of $E(|X_1|) < \infty$ the law of large numbers implies that $\frac{1}{n} \sum_{i=1}^{n} |X_i|$ converges almost surely and is therefore bounded in probability. Consequently, it is sufficient to show that

$$n^{1/2} t_n = o_P(1) \quad \text{as } n \to \infty . \tag{9.24}$$

Because of $t_n \in M_n$ we have

$$\left(\frac{1}{n} - 1\right) \frac{1}{\max_{1 \leq i \leq n} X_i} < t_n < \left(\frac{1}{n} - 1\right) \frac{1}{\min_{1 \leq i \leq n} X_i} .$$

Lemma 1 and its Corollary yield, by (9.20) and (9.22), as $n \to \infty$,

$$\frac{a_n}{\max_{1 \leq i \leq n} X_i} = O_P(1) \quad \text{and} \quad \frac{b_n}{\min_{1 \leq i \leq n} X_i} = O_P(1) \tag{9.25}$$

for $a_n = n^{1/p_+} \widetilde{L}_+(n)$ and $b_n = n^{1/p_-} \widetilde{L}_-(n)$ with p_+ and p_- as in the definition of Classes 1 and 2 and with slowly varying functions $\widetilde{L}_+$ and $\widetilde{L}_-$. Hence

$$\begin{aligned} |t_n| &\leq \frac{1}{\max_{1 \leq i \leq n} X_i} + \frac{1}{\left|\min_{1 \leq i \leq n} X_i\right|} \\ &\leq \frac{1}{\min(a_n, b_n)} \left(\frac{a_n}{\max_{1 \leq i \leq n} X_i} + \frac{b_n}{\left|\min_{1 \leq i \leq n} X_i\right|} \right) \\ &= \frac{1}{\min(a_n, b_n)} O_P(1) \end{aligned} \tag{9.26}$$

by (9.25).

For distribution functions from Class 1 it follows that, as $n \to \infty$,

$$n^{1/2} |t_n| \leq \frac{1}{\min\left(n^{-1/2} a_n, n^{-1/2} b_n\right)} O_P(1) .$$

Here

$$\min\left(n^{-1/2} a_n, n^{-1/2} b_n\right) = \min\left(n^{1/p_+ - 1/2} \widetilde{L}_+(n), n^{1/p_- - 1/2} \widetilde{L}_-(n)\right) \to \infty$$

as $n \to \infty$ because $1/p_+ - 1/2 > 0$, $1/p_- - 1/2 > 0$ and $\widetilde{L}_+, \widetilde{L}_-$ are slowly varying; see Bingham et al. (1987), Proposition 1.3.6 (v). This completes already the proof of (9.24) for distribution functions from Class 1.

To verify (9.24) also for distribution functions from Class 2 we use (9.23) to write

$$\sum_{i=1}^{n} X_i = t_n \sum_{i=1}^{n} \frac{X_i^2}{1+t_n X_i} = t_n \sum_{i=1}^{n} \frac{1+t_n X_i - t_n X_i}{1+t_n X_i} X_i^2$$
$$= t_n \sum_{i=1}^{n} X_i^2 - t_n^2 \sum_{i=1}^{n} \frac{X_i^3}{1+t_n X_i} .$$

Setting $V_n = \left(\sum_{i=1}^{n} X_i^2\right)^{1/2}$ we obtain

$$\left(n^{1/2} t_n\right)\left(n^{-1/2} V_n\right) = \frac{1}{V_n} \sum_{i=1}^{n} X_i + \frac{t_n^2}{V_n} \sum_{i=1}^{n} \frac{X_i^3}{1+t_n X_i} .$$

From $E\left(X_1^2\right) = \infty$ it follows that $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ converges to infinity almost surely so that

$$n^{-1/2} V_n = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2\right)^{1/2} \to \infty \quad \text{almost surely as } n \to \infty .$$

Therefore, to prove (9.24) it remains to show that

$$\frac{1}{V_n} \sum_{i=1}^{n} X_i + \frac{t_n^2}{V_n} \sum_{i=1}^{n} \frac{X_i^3}{1+t_n X_i} = O_P(1) \quad \text{as } n \to \infty . \tag{9.27}$$

The first summand on the left hand side of (9.27) is bounded in probability by Lemma 2. For the second summand we have (recall that $1 + t_n X_i$ is positive for $i = 1, \ldots, n$)

$$\left| \frac{t_n^2}{V_n} \sum_{i=1}^{n} \frac{X_i^3}{1+t_n X_i} \right| \le \frac{t_n^2}{V_n} \left(\max_{1 \le i \le n} |X_i| \right) \sum_{i=1}^{n} \frac{X_i^2}{1+t_n X_i} = \frac{t_n}{V_n} \left(\max_{1 \le i \le n} |X_i| \right) \sum_{i=1}^{n} X_i$$
$$\le |t_n| \max\left(\max_{1 \le i \le n} X_i, \left| \min_{1 \le i \le n} X_i \right| \right) \frac{1}{V_n} \left| \sum_{i=1}^{n} X_i \right| ,$$

where the equality follows from (9.23). Using (9.26), (9.19) and (9.21) as well as Lemma 2 this bound is less than or equal to

$$\frac{\max(a_n, b_n)}{\min(a_n, b_n)} O_P(1) .$$

But for distribution functions from Class 2 we have $a_n = n^{1/2} \widetilde{L}(n)$ and $b_n = n^{1/2} \widetilde{L}(n)$ for the same slowly varying function $\widetilde{L}$ because $F(x) = x^{-2} L(-x)$ and $1 - F(x) = x^{-2} L(x)$ for all $x \in \mathbb{R}$ with $|x|$ sufficiently large, i.e., the exponent $p = 2$ governing the tail decay of F is the same for both tails and so is the slowly varying function L. According to Lemma 1 and its Corollary this implies that the

slowly varying function in the definition of a_n and b_n depends only on $p = 2$ and L and therefore is the same in both cases. Thus $\max(a_n, b_n) = \min(a_n, b_n)$. This completes the proof of

$$\frac{t_n^2}{V_n} \sum_{i=1}^{n} \frac{X_i^3}{1 + t_n X_i} = O_P(1) \quad \text{as } n \to \infty$$

and of the theorem.

References

Bingham, N.H., Goldie, C.M., Teugels, J.L.: Regular Variation. Cambridge University Press, Cambridge etc. (1987)

Chistyakov, G.P., Götze, F.: Limit distributions of studentized means. Ann. Probab. **22**, 28–77 (2004)

Haeusler, E., Plies, C.: Bootstrapping empirical distributions under auxiliary information. In: Giné, E., Mason, D.M., Wellner, J.A. (eds.) High Dimensional Probability II, pp. 461–476. Birkhäuser, Boston (2000)

Owen, A.B.: Empirical likelihood ratio confidence intervals for a single functional. Biometrika **47**, 237–249 (1988)

Owen, A.B.: Empirical likelihood confidence regions. Ann. Statist. **18**, 90–120 (1990)

Owen, A.B.: Empirical likelihood for linear models. Ann. Statist. **19**, 1725–1747 (1991)

Owen, A.B.: Empirical Likelihood. Chapman and Hall/CRC, London/Boca Raton (2001)

Qin, J., Lawless, J.: Empirical likelihood and general estimating equations. Ann. Statist. **22**, 300–325 (1994)

Zhang, B.: Estimating a distribution function in the presence of auxiliary information. Metrika **46**, 221–244 (1997)

A Review and Some New Proposals for Bandwidth Selection in Nonparametric Density Estimation for Dependent Data

10

Inés Barbeito and Ricardo Cao

10.1 Introduction

This chapter deals with the widely known problem of data-driven choice of smoothing parameters in nonparametric density estimation, which is indeed an important research area in Statistics (see classical books such as Silverman (1986) and Devroye (1987), among others, for introductory background in the iid case). Our aim is to extend the critical review by Cao et al. (1994), for iid data, when dependence is assumed and to complete and update the review by Cao et al. (1993). Furthermore, some approaches regarding new techniques for bandwidth selection in density estimation are included. Subsequently, an extensive simulation study is carried out in order to check the good empirical behaviour of these bandwidth parameters, and also to compare them critically.

The current state of the art concerning nonparametric density estimation assuming independence has been extensively studied. A great amount of cross-validation methods proposed for iid data (see Rudemo 1982; Chow et al. 1983; Bowman 1984; Stone 1984; Marron 1985; Marron 1987; Hall 1983; Hall and Marron 1987a; Hall and Marron 1987b; Scott and Terrell 1987; Stute 1992; Feluch and Koronacki 1992) triggered the development of new techniques for the purpose of bandwidth selection. Apart from the plug-in procedures which have been proposed (see Park and Marron 1990; Hall and Marron 1991; Sheather and Jones 1991 or Jones et al. 1991),

I. Barbeito (✉) · R. Cao
Research Group MODES, CITIC, Faculty of Computer Science, Department of Mathematics, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain
e-mail: ines.barbeito@udc.es

R. Cao
e-mail: rcao@udc.es

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_10

there was also room for the development of bootstrap methods (see some remarkable approaches such as Taylor 1989; Hall 1990; Faraway and Jhun 1990; Léger and Romano 1990; Marron 1992 and Cao 1993). Some critical and extensive simulation studies were carried out in this context, we only refer to Park and Marron (1990), Cao et al. (1994) and Jones et al. (1996) for the sake of brevity.

Nonetheless, when the data are generated by a stochastic process observed in time, they will no longer be iid. In fact, very few papers have dealt with data-driven bandwidth selectors under stationarity for kernel density estimation. Only a few approaches appeared concerning this issue. Firstly, the classical cross-validation method was modified by Hart and Vieu (1990). Hall et al. (1995) also proposed an adaptation of the plug-in method when dependence is assumed, and a bandwidth parameter chosen by minimizing an asymptotic expression for the mean integrated squared error obtained by themselves was established. A deep simulation study in this context was carried out by Cao et al. (1993). Focusing on bootstrap procedures, a very recent development was proposed by Barbeito and Cao (2016). It consists in a smoothed version of the so-called stationary bootstrap by Politis and Romano (1994), more suitable when our aim is to estimate nonparametrically the density function. A closed expression for the smoothed stationary bootstrap version of the mean integrated squared error was also obtained by these authors, so Monte Carlo approximation is avoided.

Some review papers have already been published concerning bootstrap methods used in the dependent data setup, such as Cao (1999) or Kreiss and Paparoditis (2011). However, to the best of our knowledge, none have dealt with the specific problem of nonparametric density estimation, which is precisely the aim of this chapter. Section 10.2 presents an up-to-date review of the main methods and justify our choice of the bandwidths to be compared by simulation. Furthermore, the adaptation of two already existing methods is established: the modified cross-validation by Stute (1992), when dependence is considered, and the penalized cross-validation proposed by Estévez-Pérez et al. (2002) for hazard rate estimation, using it for density estimation. Similarly to the smoothed stationary bootstrap, a smoothed version of the moving blocks bootstrap (see Künsch 1989 and Liu and Singh 1992 for the unsmoothed case) is also established in Sect. 10.3. In addition, a closed expression for the smoothed moving blocks bootstrap version of the mean integrated squared error is presented in that section and a bootstrap bandwidth selector is proposed. The performance of those bandwidth parameters is analyzed via an extensive simulation study in Sect. 10.4, including some concluding remarks. Finally, an Appendix contains the proof of the result stated in Sect. 10.3.

10.2 A Critical Review of Smoothing Methods

We focus on the problem of estimating the density function in a nonparametric way. Let us consider a random sample, $(X_1, \ldots, X_n)$, coming from a population with density f. Throughout this chapter the kernel density estimator (see Parzen 1962;

Rosenblatt 1956) is studied,

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$, K is a kernel function (typically a probability density function) and $h = h_n > 0$ is the sequence of smoothing parameters or bandwidths. In fact, the choice of the bandwidth, h, is really important in order to obtain a correct density estimator, since it regulates the degree of smoothing applied to the data. From now on, the aim of this chapter is to select the bandwidth when dependent data are considered.

10.2.1 Cross-Validation Methods

Three cross-validation procedures are considered in the following. They are all related by the use of a leave-$(2l + 1)$-out device when computing the cross-validation function, which is intended to be minimized. Firstly, the well known leave-$(2l + 1)$-out cross-validation proposed by Hart and Vieu (1990) is studied. Thereupon, two adaptations to our setting of existing methods for hazard rate estimation with dependence and density estimation with iid data are established: the penalized cross-validation and the modified cross-validation.

10.2.1.1 Leave-$(2l + 1)$-Out Cross-Validation

This method (see Hart and Vieu 1990) is the adaptation to dependence of the classic leave-one-out cross-validation procedure for iid data proposed by Bowman (1984). Its aim is to minimize the cross-validation function (namely, CV_l) in order to obtain the optimal bandwidth parameter, where CV_l is given by:

$$CV_l(h) = \int \hat{f}^2(x)dx - \frac{2}{n}\sum_{j=1}^{n} \hat{f}_l^j(X_j),$$

being

$$\hat{f}_l^j(x) = \frac{1}{n_l}\sum_{i:|j-i|>l}^{n} \frac{1}{h}K\left(\frac{x - X_i}{h}\right),$$

and l is a sequence of positive integers known as the 'leave-out' sequence. It is also worth mentioning that n_l is chosen as follows:

$$n_l = \frac{\#\{(i, j) : |i - j| > l\}}{n}.$$

Finally, the leave-$(2l+1)$-out cross-validation bandwidth is defined as:

$$h_{CV_l} = \arg\min_{h>0} CV_l(h).$$

The asymptotic optimality of the method for a certain class of l, assuming some regularity conditions over the stationary process, is also stated by Hart and Vieu (1990). The convergence rates of h_{CV_l} are studied by Cox and Kim (1997), where regularity conditions are also assumed. It is worth mentioning that the regularity conditions imposed by Hart and Vieu (1990) demand short-range dependence of the underlying process. This issue is taken up by Hall et al. (1995) in their plug-in procedure (see Sect. 10.2.2) since they assume conditions of long-range dependence.

10.2.1.2 Penalized Cross-Validation

The penalized cross-validation (PCV) method was proposed by Estévez-Pérez et al. (2002) for hazard rate estimation under dependence in order to avoid undersmoothed estimations. As a consequence, they stated a penalization for the cross-validation bandwidth h_{CV_l}. In this chapter, we propose an adaptation to density estimation under dependence. It consists in adding to the value h_{CV_l}, obtained by means of Hart and Vieu (1990) cross-validation procedure, a parameter empirically chosen and somehow related with the estimated autocorrelation.

Specifically, the PCV bandwidth selector is

$$h_{PCV} = h_{CV_l} + \bar{\lambda},$$

where $\bar{\lambda}$ turns out to be

$$\bar{\lambda} = \left(0.8e^{7.9\hat{\rho}-1}\right) n^{-3/10} \frac{h_{CV_l}}{100},$$

and $\hat{\rho}$ is the estimated autocorrelation of order 1. In fact, as $\hat{\rho}$ increases, so does the bandwidth parameter, h_{PCV}. It is worth pointing out that h_{PCV} is obtained in such a way that it is still consistent, according to the consistency of both $\hat{\rho}$ and h_{CV_l}.

10.2.1.3 Modified Cross-Validation

An extension to dependent data of the modified cross-validation for iid data (see Stute 1992) is described now. In the independent case, the aim of this approach consists in avoiding undersmoothed estimations of the density function, considering a way which definitely differs from the usual. In this sense, this approach is based on a finite sample rather than an asymptotic argument, focusing on a statistic whose Hajek projection contains the unknown $\int \hat{f}_h(x) f(x) dx$, which takes part in the $ISE(h)$ expression. This is precisely the function studied by cross-validation procedures.

The main difference between the dependent case (SMCV) and the iid case (MCV) is the idea of leaving out $2l+1$ points (as in Hart and Vieu 1990) when computing the function intended to be minimized, $SMCV(h)$.

$$SMCV(h) = \frac{1}{nh}\int K^2(t)dt + \frac{1}{n(n-1)h}\sum_{i\neq j}\left[\frac{1}{h}\int K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx\right] - \frac{1}{nn_lh}\sum_{j=1}^{n}\sum_{i:|j-i|>l}^{n}\left[K\left(\frac{X_i-X_j}{h}\right) - dK''\left(\frac{X_i-X_j}{h}\right)\right],$$

where $d = \frac{1}{2}\int t^2K(t)dt$. Then the SMCV bandwidth selector is

$$h_{SMCV} = \arg\min_{h>0} SMCV(h).$$

As for the consistency of the method, similar results as those for the iid case can be obtained, assuming some regularity and moment conditions on the stochastic process.

10.2.2 Plug-In Method

Hall et al. (1995) proposed the plug-in when assuming dependence (see Sheather and Jones 1991 for the iid case). The key of this method is to minimize, in h, the $AMISE(h)$ expression obtained for dependent data, assuming that f is six times differentiable, and considering $R(f) = \int f(x)^2dx$, $R(f'') = \int f''(x)^2dx$, $R(f''') = \int f'''(x)^2dx$ and $\mu_k = \int z^kK(z)dz$. The AMISE expression is given by:

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2R(f'') - h^6\frac{1}{24}\mu_2\mu_4R(f''') + \frac{1}{n}\left(2\sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right)\int g_i(x,x)dx - R(f)\right), \quad (10.1)$$

where $g_i(x_1,x_2) = f_i(x_1,x_2) - f(x_1)f(x_2)$, and f_i is the density of (X_j, X_{i+j}).

Minimizing expression (10.1) in h leads to the plug-in bandwidth selector,

$$\hat{h} = \left(\frac{\hat{J}_1}{n}\right)^{1/5} + \hat{J}_2\left(\frac{\hat{J}_1}{n}\right)^{3/5},$$

where $\hat{J}_1$ is an estimator of $J_1 = \dfrac{R(K)}{\mu_2^2 R(f'')}$, and $\hat{J}_2$ is an estimator of $J_2 = \dfrac{\mu_4 R(f''')}{20\mu_2 R(f'')}$.

Thereupon, Hall et al. (1995) propose to replace directly $R(f'')$ and $R(f''')$ by their respective estimators, that is, $\hat{I}_2$ and $\hat{I}_3$, that can be obtained as follows:

$$\hat{I}_k = 2\hat{\theta}_{1k} - \hat{\theta}_{2k}, k = 2, 3$$

where $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$ are the respective estimators of $\theta_{1k} = \int \left(\mathbb{E}(\hat{f}_1)\right) f^{(k)}$, $\theta_{2k} = \int \left(\mathbb{E}(\hat{f}_1^{(k)})\right)^2$, $k = 1, 2, 3$, and, $\hat{f}_1$ is the nonparametric Parzen-Rosenblatt density estimator obtained with a kernel K_1 and a bandwidth h_1. The upcoming expressions are obtained for $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$:

$$\hat{\theta}_{1k} = 2\left(n(n-1)h_1^{2k+1}\right)^{-1} \sum\sum_{1\le i<j\le n} K_1^{(2k)}\left(\frac{X_i - X_j}{h_1}\right),$$

$$\hat{\theta}_{2k} = 2\left(n(n-1)h_1^{2(k+1)}\right)^{-1} \sum\sum_{1\le i<j\le n} \int K_1^{(k)}\left(\frac{x - X_i}{h_1}\right) K_1^{(k)}\left(\frac{x - X_j}{h_1}\right) dx.$$

Under some moment and regularity conditions on the stationary process, assuming the differentiability of the kernel K_1, and choosing h_1 verifying $n^{-1/(4k+1)} \le h_1 \le 1$; Hall et al. (1995) proved that this plug-in method under dependence is consistent.

10.2.3 Bootstrap-Based Procedures

Since the introduction of the bootstrap method by Efron (1979), this technique has been widely used to approximate the sampling distribution of a statistic of interest (see Efron and Tibishirani 1993 for a deeper insight of the bootstrap method and its applications).

The essential idea to compute a bootstrap bandwidth selector is to obtain the bootstrap version of the mean integrated squared error (namely, $MISE$) and to find the smoothing parameter that minimizes this bootstrap version, given by

$$\begin{aligned} MISE^*(h) &= \mathbb{E}^*\left[\int \left(\hat{f}_h^*(x) - \hat{f}_g(x)\right)^2 dx\right] \\ &= B^*(h) + V^*(h), \end{aligned}$$

with

$$B^*(h) = \int \left[\mathbb{E}^*\left(\hat{f}_h^*(x)\right) - \hat{f}_g(x)\right]^2 dx, \text{ and}$$
$$V^*(h) = \int \text{Var}^*\left(\hat{f}_h^*(x)\right) dx,$$

where $\mathbb{E}^*$ denotes the expectation (Var*, the variance) with respect to the bootstrap sample $X_1^*, \ldots, X_n^*$, g is some pilot bandwidth, $\hat{f}_g$ is a kernel density estimation based on the sample $X_1, \ldots, X_n$, and $\hat{f}_h^*$ is the bootstrap version of the kernel density estimator with bandwidth h, based on the resample $X_1^*, \ldots, X_n^*$.

In the iid case, the bootstrap method has been used to produce bandwidth selectors (see, for instance, Cao 1993). The idea is basically to use the smoothed bootstrap proposed by Silverman and Young (1987) to approximate the $MISE$ of the kernel density estimator.

The main disadvantage of this procedure, however, is the necessity of Monte Carlo approximation whenever the bootstrap distribution of the bootstrap version of the statistic of interest cannot be explicitly computed. Nevertheless, as shown below, when dependence is considered, both smoothed stationary bootstrap (see Barbeito and Cao 2016) and smoothed moving blocks bootstrap (proposed subsequently) techniques, need no Monte Carlo in order to implement the bootstrap bandwidths.

10.2.3.1 Smoothed Stationary Bootstrap

This bootstrap resampling plan was proposed by Barbeito and Cao (2016). It is actually a smoothed version of the stationary bootstrap (see Politis and Romano 1994), that is, the bootstrap sample used in Eq. (10.2) below is drawn from the pilot density estimate $\hat{f}_g$. The pilot bandwidth g is proposed to be chosen as in the iid case (basic results involving this choice can be found in Cao 1993). The smoothed stationary bootstrap, SSB (see Cao 1999 for the unsmoothed case, SB), proceeds as follows:

1. Draw $X_1^{*(SB)}$ from F_n, the empirical distribution function of the sample.
2. Define $X_1^* = X_1^{*(SB)} + gU_1^*$, where U_1^* has been drawn with density K and independently from $X_1^{*(SB)}$.
3. Assume we have already drawn $X_1^*, \ldots, X_i^*$ (and, consequently, $X_1^{*(SB)}, \ldots, X_i^{*(SB)}$) and consider the index j, for which $X_i^{*(SB)} = X_j$. We define a binary auxiliary random variable I_{i+1}^*, such that $P^*\left(I_{i+1}^* = 1\right) = 1 - p$ and $P^*\left(I_{i+1}^* = 0\right) = p$. We assign $X_{i+1}^{*(SB)} = X_{(j \mod n)+1}$ whenever $I_{i+1}^* = 1$ and we use the empirical distribution function for $X_{i+1}^{*(SB)}|_{I_{i+1}^*=0}$, where mod stands for the modulus operator.
4. Once drawn $X_{i+1}^{*(SB)}$, we define $X_{i+1}^* = X_{i+1}^{*(SB)} + gU_{i+1}^*$, where, again, U_{i+1}^* has been drawn from the density K and independently from $X_{i+1}^{*(SB)}$.

In order to obtain the bandwidth parameter h^*_{SSB}, Monte Carlo is not needed because of the explicit formula for the smoothed stationary bootstrap version of the $MISE(h)$, given by:

$$\begin{aligned} MISE^*_{SSB}(h) &= n^{-2}\sum_{i,j=1}^{n}\left(K_g * K_g\right)\left(X_i - X_j\right) \\ &\quad - 2n^{-2}\sum_{i,j=1}^{n}\left(K_h * K_g * K_g\right)\left(X_i - X_j\right) \\ &\quad + \left[\frac{n-1}{n^3} - 2\frac{1-p-(1-p)^n}{pn^3}\right. \\ &\quad \left. + 2\frac{(n-1)(1-p)^{n+1} - n(1-p)^n + 1 - p}{p^2n^4}\right] \\ &\quad \times \sum_{i,j=1}^{n}\left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right]\left(X_i - X_j\right) \\ &\quad + n^{-1}h^{-1}R(K) \\ &\quad + 2n^{-3}\sum_{\ell=1}^{n-1}(n-\ell)(1-p)^{\ell}\sum_{k=1}^{n}\left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right] \\ &\qquad \left(X_k - X_{\lceil (k+\ell-1) \mod n\rceil+1}\right). \end{aligned} \tag{10.2}$$

The proof of the previous result can be found in Barbeito and Cao (2016), where an exact expression for $MISE(h)$ under dependence and stationarity is also obtained. The bootstrap smoothing parameter h^*_{SSB} turns out to be the one which minimizes in h the function in (10.2), that is:

$$h^*_{SSB} = h^{*SSB}_{MISE} = \arg\min_{h>0} MISE^*_{SSB}(h).$$

Similarly to the SSB, in a density estimation context it makes more sense to build a smoothed version of the moving blocks bootstrap by Künsch (1989) and Liu and Singh (1992). The method is presented in the next section, where a closed formula for its bootstrap version of $MISE$ is also stated.

10.3 Smoothed Moving Blocks Bootstrap

The smoothed moving blocks bootstrap, SMBB (see Cao (1999) for the unsmoothed case, MBB), proceeds as follows:

1. Fix the block length, $b \in \mathbb{N}$, and define $k = \min_{\ell\in\mathbb{N}} \ell \geq \frac{n}{b}$

2. Define:

$$B_{i,b} = (X_i, X_{i+1}, \ldots, X_{i+b-1})$$

3. Draw $\xi_1, \xi_2, \ldots, \xi_k$ with uniform discrete distribution on $\{B_1, B_2, \ldots, B_q\}$, with $q = n - b + 1$
4. Define $X_1^{*(MBB)}, \ldots, X_n^{*(MBB)}$ as the first n components of

$$(\xi_{1,1}, \xi_{1,2}, \ldots, \xi_{1,b}, \xi_{2,1}, \xi_{2,2} \ldots, \xi_{2,b}, \ldots, \xi_{k,1}, \xi_{k,2}, \ldots, \xi_{k,b})$$

5. Define $X_i^* = X_i^{*(MBB)} + gU_i^*$, where U_i^* has been drawn with density K and independently from $X_i^{*(MBB)}$, for all $i = 1, 2, \ldots, n$

This resampling plan depends on a parameter b, which is the block length, to be chosen by the user. The pilot bandwidth, g, also needs to be chosen. The following result presents an exact expression for the smoothed moving blocks bootstrap version of the $MISE(h)$.

Theorem 1 *If the kernel K is a symmetric density function, then the smoothed moving blocks bootstrap version of $MISE$ admits the following closed expression, considering n an integer multiple of b:*

1. If $b < n$,

$$\begin{aligned}
MISE^*_{SMBB}(h) = {} & \frac{R(K)}{nh} + \sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j \psi(X_i - X_j) \\
& - \frac{2}{n} \sum_{i=1}^{n} a_i \sum_{j=1}^{n} \left[\left(K_h * K_g \right) * K_g \right] (X_i - X_j) \\
& + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[K_g * K_g \right] (X_i - X_j) \\
& - \frac{b-1}{n(n-b+1)^2} \sum_{i=b-1}^{n-b+1} \sum_{j=b}^{n-b+2} \psi(X_i - X_j) \\
& - \frac{1}{nb(n-b+1)^2} \left[\sum_{i=1}^{b-1} \sum_{j=1}^{b-1} \min\{i, j\} \psi(X_i - X_j) \right. \\
& + \sum_{i=1}^{b-1} i \sum_{j=b}^{n-b+1} \psi(X_i - X_j) \\
& + \sum_{i=1}^{b-1} \sum_{j=n-b+2}^{n} \min\{(n-b+i-j+1), i\} \psi(X_i - X_j) \\
& \left. + \sum_{i=b}^{n-b+1} \sum_{j=1}^{b-1} j \psi(X_i - X_j) \right)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=n-b+2}^{n}\min\{(n-i+1),b\}\sum_{j=b}^{n-b+1}\psi(X_i-X_j)\\
&+\sum_{i=b}^{n-b+1}\sum_{j=n-b+2}^{n}\min\{(n-j+1),b\}\psi(X_i-X_j)\\
&+\sum_{i=n-b+2}^{n}\sum_{j=1}^{b-1}\min\{(n-b+j-i+1),j\}\psi(X_i-X_j)\\
&+b\sum_{i=b}^{n-b+1}\sum_{j=b}^{n-b+1}\psi(X_i-X_j)\\
&+\sum_{i=n-b+2}^{n}\sum_{j=n-b+2}^{n}(n+1-\max\{i,j\})\,\psi(X_i-X_j)\Bigg]\\
&+\frac{2}{nb(n-b+1)}\sum_{s=1}^{b-1}\sum_{j=1}^{n-s}(\min\{j,b-s\}\\
&-\max\{1,j+b-n\}+1)\psi(X_{j+s}-X_j)\\
&-\frac{2}{nb(n-b+1)^2}\Bigg[\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\Bigg[\sum_{i=k}^{b-2}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)\\
&+\sum_{i=n-b+2}^{n-b+k}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\\
&+\sum_{i=k}^{b-2}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)+\sum_{i=n-b+2}^{n-b+k}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)\Bigg]\\
&+\sum_{k=1}^{b-1}(b-k)\sum_{i=k}^{b-2}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\\
&+\sum_{\ell=2}^{b}(\ell-1)\sum_{i=b-1}^{n-b+1}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)\\
&+\sum_{\ell=2}^{b}(\ell-1)\sum_{i=b-1}^{n-b+1}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\\
&+\sum_{k=1}^{b-1}(b-k)\sum_{i=n-b+2}^{n-b+k}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\Bigg],
\end{aligned}
$$

where $\psi(u) = \left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right](u)$ *and:*

$$a_j = \frac{\min\{j, n-j+1, b\}}{b(n-b+1)}, \; j = 1, 2, \ldots, n. \tag{10.3}$$

2. *If* $b = n$,

$$\begin{aligned} MISE^*_{SMBB}(h) = \frac{R(K)}{nh} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(X_i - X_j) \\ -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(K_h * K_g\right) * K_g\right](X_i - X_j) \\ +\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[K_g * K_g\right](X_i - X_j) + \frac{\psi(0)}{n}. \end{aligned}$$

Theorem 1 is proven in the Appendix.

A bootstrap bandwidth selector, h^*_{SMBB}, can be defined as the minimizer, in h, of $MISE^*(h)$ given in Theorem 1.

$$h^*_{SMBB} = h^{*SMBB}_{MISE} = \arg\min_{h>0} MISE^*_{SMBB}(h).$$

It is worth mentioning that the exact expression for the $MISE^*_{SMBB}(h)$ is really useful since Monte Carlo approximation is not necessary to compute the bootstrap bandwidth selector.

Remark 1 The closed expression for $MISE^*_{SMBB}$ in Theorem 1 has been worked out in order to alleviate its computational cost. Alternative closed expressions for $MISE^*_{SMBB}$ could be easier and shorter to write but they would imply a larger computational cost when implemented.

10.4 Simulations

10.4.1 General Description of the Study

A simulation study is carried out to compare the practical behaviour of the bandwidth parameters described in Sect. 10.2, that is, h_{CV_l}, h_{PCV}, h_{SMCV}, h_{PI}, h^*_{SSB} and h^*_{SMBB}. We will consider seven different populations (six of them already used by Cao et al. 1993 and Barbeito and Cao 2016) so as to show the empirical results of every bandwidth in different situations:

Model 1: The sample is drawn from an $AR(2)$ model given by $X_t = -0.9X_{t-1} - 0.2X_{t-2} + a_t$, where a_t are iid random variables with common distribution $N(0, 1)$. The marginal distribution is $X_t \overset{d}{=} N(0, 0.42)$.

Model 2: A $MA(2)$ model given by $X_t = a_t - 0.9a_{t-1} + 0.2a_{t-2}$ is considered, where the innovations a_t are iid standard normal random variables. The marginal distribution is $X_t \overset{d}{=} N(0, 1.85)$.

Model 3: An $AR(1)$ model: $X_t = \phi X_{t-1} + (1 - \phi^2)^{1/2} a_t$. Here a_t are iid random variables with common standard normal distribution. The autocorrelation was set to the values $\phi = \pm 0.3, \pm 0.6, \pm 0.9$. The marginal distribution is a standard normal.

Model 4: The time series is generated from an $AR(1)$ model given by $X_t = \phi X_{t-1} + a_t$. In this case, the distribution of a_t has exponential structure:

$$\mathbb{P}(I_t = 1) = \phi, \mathbb{P}(I_t = 2) = 1 - \phi, \text{ with}$$
$$a_t|_{I_t=1} \overset{d}{=} 0 \text{ (constant)}, \; a_t|_{I_t=2} \overset{d}{=} \exp(1).$$

The values of ϕ chosen were $\phi = 0, 0.3, 0.6, 0.9$. The marginal distribution is $X_t \overset{d}{=} \exp(1)$.

Model 5: An $AR(1)$ model with a double-exponential structure: $X_t = \phi X_{t-1} + a_t$, which means that the innovations have to be drawn from the following distribution:

$$\mathbb{P}(I_t = 1) = \phi^2, \mathbb{P}(I_t = 2) = 1 - \phi^2, \text{ with}$$
$$a_t|_{I_t=1} \overset{d}{=} 0 \text{ (constant)}, \; a_t|_{I_t=2} \overset{d}{=} \text{Dexp}(1),$$

The values of ϕ used were $\phi = \pm 0.3, \pm 0.6, \pm 0.9$. The marginal distribution is $X_t \overset{d}{=} \text{Dexp}(1)$.

Model 6: A mixture of two normal densities, with probability $1/2$ each, associated to the model:

$$X_t = \begin{cases} X_t^{(1)} \text{ with probability } 1/2 \\ X_t^{(2)} \text{ with probability } 1/2 \end{cases},$$

where $X_t^{(j)} = (-1)^{j+1} + 0.5X_{t-1}^{(j)} + a_t^{(j)}$ with $j = 1, 2$, $\forall t \in \mathbb{Z}$, and $a_t^{(j)} \overset{d}{=} N(0, 0.6)$. The marginal distribution is a normal mixture $X_t \overset{d}{=} \frac{1}{2}N(2, 0.8) + \frac{1}{2}N(-2, 0.8)$.

Model 7: A mixture of three normal densities (Model 9 of Marron and Wand 1992) inducing dependence using a Markovian regime change:

$$X_t = \begin{cases} X_t^{(1)} \text{ with probability } 9/20 \\ X_t^{(2)} \text{ with probability } 9/20 \\ X_t^{(3)} \text{ with probability } 1/10 \end{cases},$$

where $X_t^{(j)} = 0.9X_{t-1}^{(j)} + a_t^{(j)}$ with $j = 1, 2, 3$, $\forall t \in \mathbb{Z}$, and $a_t^{(1)} \overset{d}{=} N(-0.12, 0.0684)$, $a_t^{(2)} \overset{d}{=} N(0.12, 0.0684)$, $a_t^{(3)} \overset{d}{=} N(0, 0.011875)$. The marginal distribution is a normal mixture $X_t \overset{d}{=} \frac{9}{20}N\left(-\frac{6}{5}, \frac{9}{25}\right) + \frac{9}{20}N\left(\frac{6}{5}, \frac{9}{25}\right) + \frac{1}{10}N\left(0, \frac{1}{16}\right)$.
The transition matrix used is given by:

$$T = \begin{pmatrix} 0.90790619 & 0.08152211 & 0.0105717 \\ 0.08152211 & 0.90790619 & 0.0105717 \\ 0.04757265 & 0.04757265 & 0.9048547 \end{pmatrix}.$$

This produces a first order autocorelation of $\phi = 0.6144$.

For every model, 1000 random samples of size $n = 100$ were drawn. The Gaussian kernel is used to compute the Parzen-Rosenblatt estimator. For the three cross-validation bandwidths the value of l was $l = 5$. The parameter p used in the SSB was $p = \frac{1}{2\sqrt{n}}$ and the parameter b used in the SMBB was $b = 2\sqrt{n}$, while the pilot bandwidth, g, has been chosen as in Cao (1993) in both cases. The pilot bandwidth used in the plug-in method was $h_1 = C\tilde{h}_1 n^{4/45}$, where $C = 1.12$ and $\tilde{h}_1$ is the bandwidth chosen as in Sheather and Jones (1991) for iid data. The bandwidth selectors $h^*_{SSB}, h^*_{SMBB}, h_{CV_l}$ and h_{SMCV} are the minimizers, in h, of four empirical functions. Since these minimizers do not have explicit expressions, a numerical method is used to approximate them. The algorithm proceeds as follows:

Step 1: Let us consider a set of 5 equally spaced values of h in the interval $[0.01, 10]$.
Step 2: For each method, a bandwidth h is chosen among the five given in the preceeding step, by minimizing the objective function ($MISE^*_{SSB}$, $MISE^*_{SMBB}$, CV_l or $SMCV$). We denote it by h_{OPT_1}.
Step 3: Among the set of 5 bandwidth parameters defined in Step 1, we consider the previous and the next one to h_{OPT_1}. If h_{OPT_1} is the smallest (largest) bandwidth in the grid, then h_{OPT_1} is used instead of the previous (next) value of h_{OPT_1} in the grid.
Step 4: A set of 5 equally spaced values of h is constructed within the interval whose endpoints are the two values selected in Step 3.
Step 5: Finally, Steps 2–4 are repeated 10 times, retaining the optimal bandwidth selector in the last stage.

It is worth mentioning that, to avoid oversmoothing of the SMCV procedure, h_{SMCV} is considered as the smallest h for which $SMCV(h)$ attains a local minimum, not its global one.

The six bandwidth selectors are compared in terms of how close they are to the optimal $MISE$ bandwidth and also in terms of the error committed when using each one of them. Thus, using the 1000 samples, the following expressions were approximated by simulation:

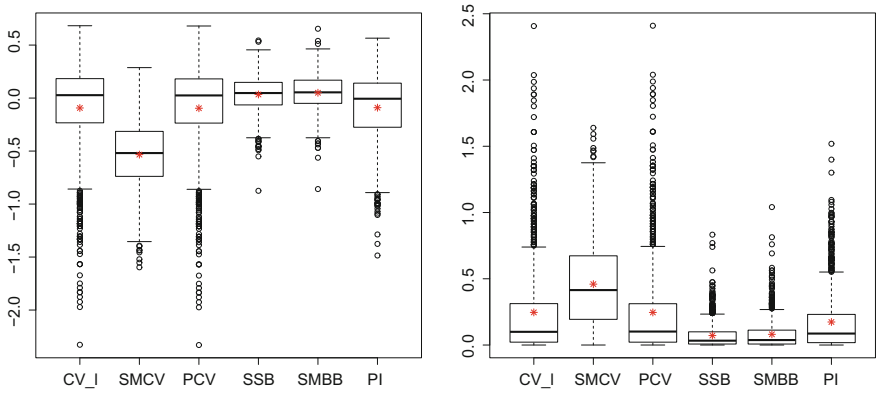
$$\log\left(\frac{\hat{h}}{h_{MISE}}\right) \text{ and} \tag{10.4}$$

$$\log\left(\frac{MISE(\hat{h})}{MISE(h_{MISE})}\right), \tag{10.5}$$

where $\hat{h} = h_{CV_l}, h_{PCV}, h_{SMCV}, h_{PI}, h^*_{SSB}, h^*_{SMBB}$, and h_{MISE} is the smoothing parameter which minimizes the error criterion, $MISE(h)$.

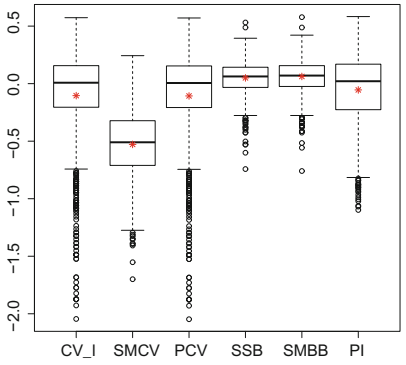
10.4.2 Discussion and Results

Figures 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, 10.7 and 10.8 show boxplots with the results obtained for expressions (10.4) and (10.5) approximated by simulation. The simulation results remarkably show that smoothing parameters h^*_{SSB} and h^*_{SMBB} display a similar performance, actually the best one, even for heavy dependence. For both h^*_{SSB} and h^*_{SMBB}, expression (10.4), shown in Figs. 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, 10.7 and 10.8 (left side) exhibit that the median of h^*_{SSB} and h^*_{SMBB} is approximately h_{MISE}. Moreover, h^*_{SSB} and h^*_{SMBB} present less variance than the three cross-validation smoothing parameters.

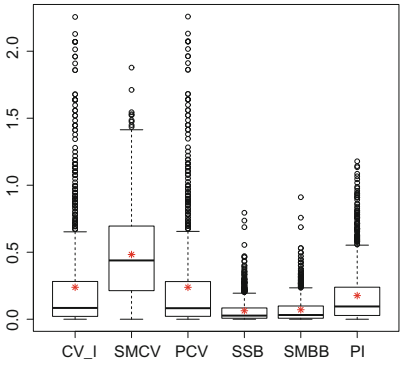


(a) Model 1 (b) Model 1

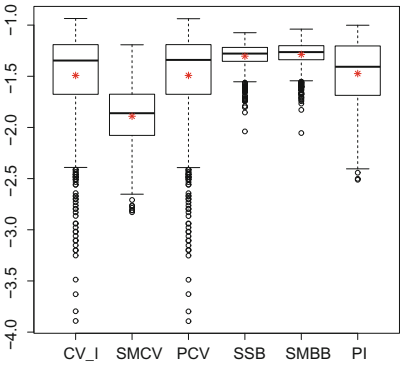
Fig. 10.1 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 1, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)



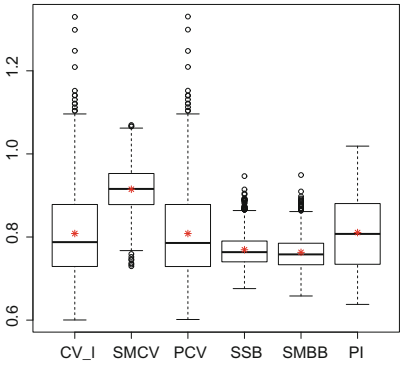
(a) Model 2



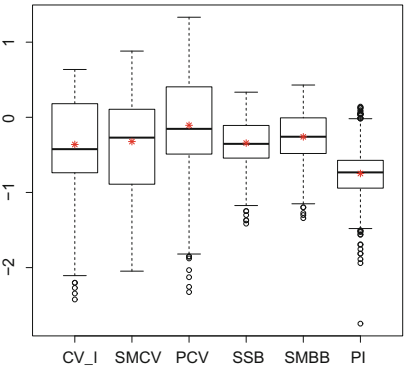
(b) Model 2



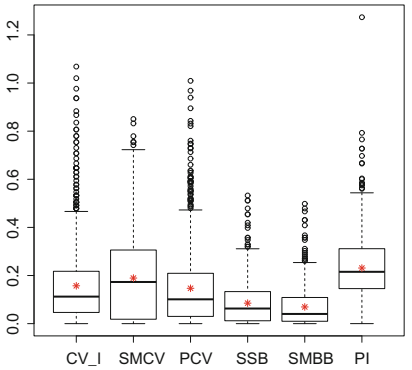
(c) Model 6



(d) Model 6

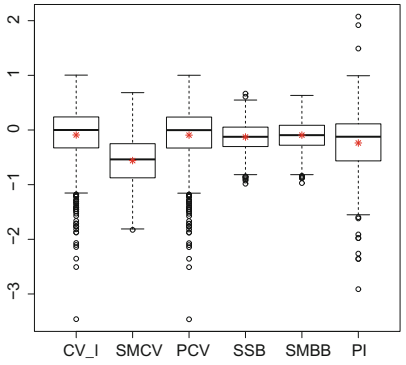


(e) Model 7

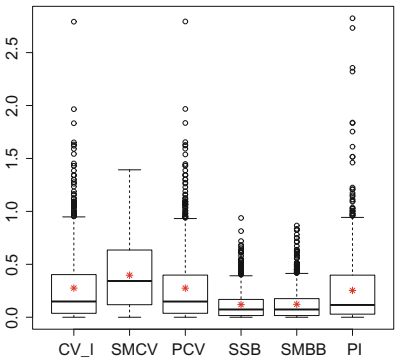


(f) Model 7

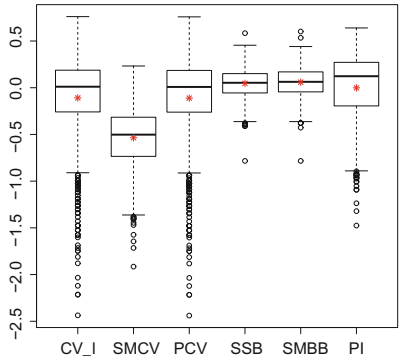
Fig. 10.2 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Models 2, 6 and 7, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)



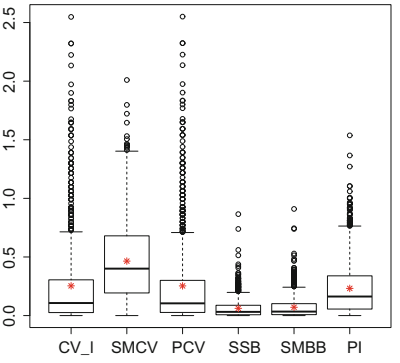
(a) Model 3, $\phi = -0.9$



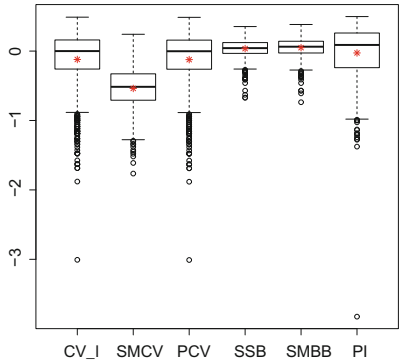
(b) Model 3, $\phi = -0.9$



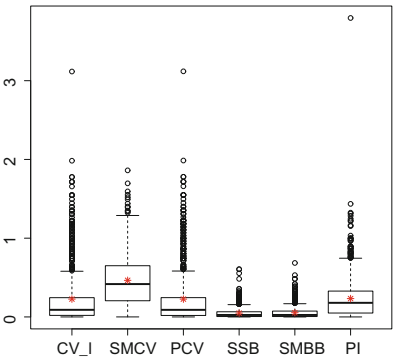
(c) Model 3, $\phi = -0.6$



(d) Model 3, $\phi = -0.6$

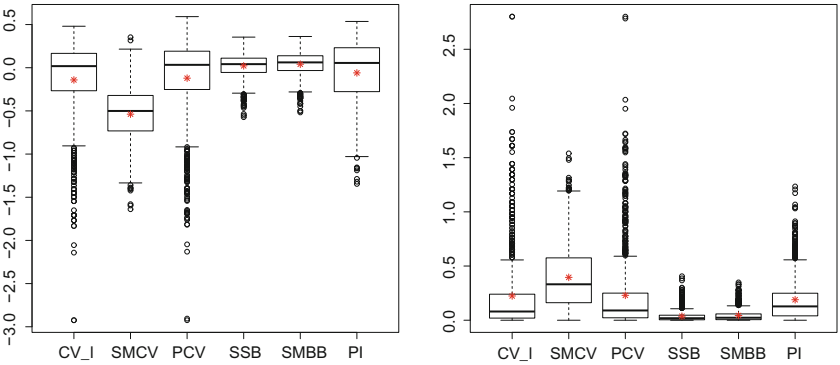


(e) Model 3, $\phi = -0.3$

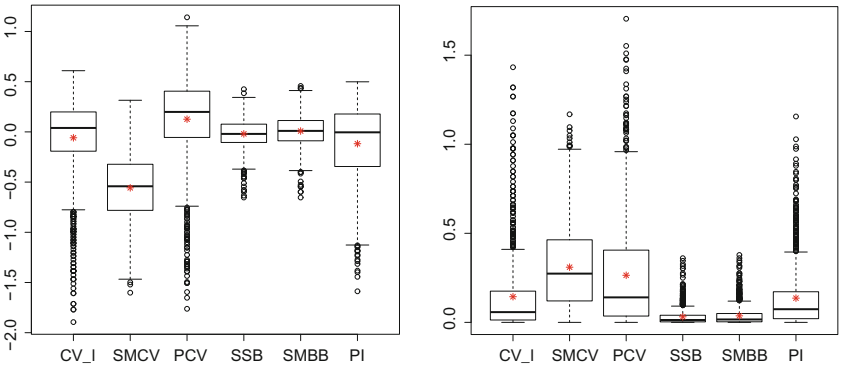


(f) Model 3, $\phi = -0.3$

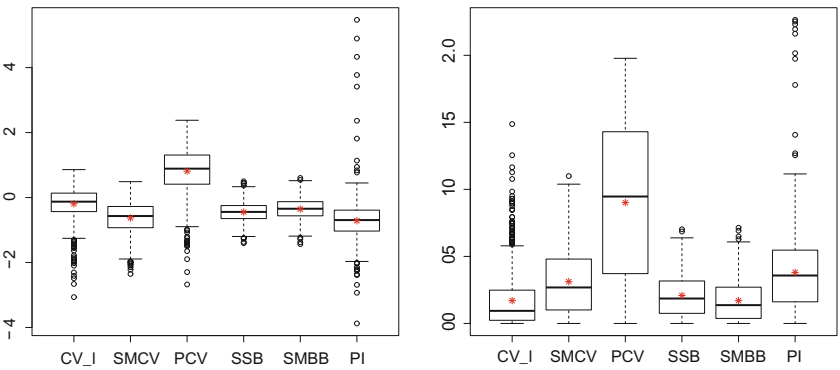
Fig. 10.3 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 3 with autocorrelation $\phi = -0.9, -0.6, -0.3$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)



(a) Model 3, $\phi = 0.3$ (b) Model 3, $\phi = 0.3$

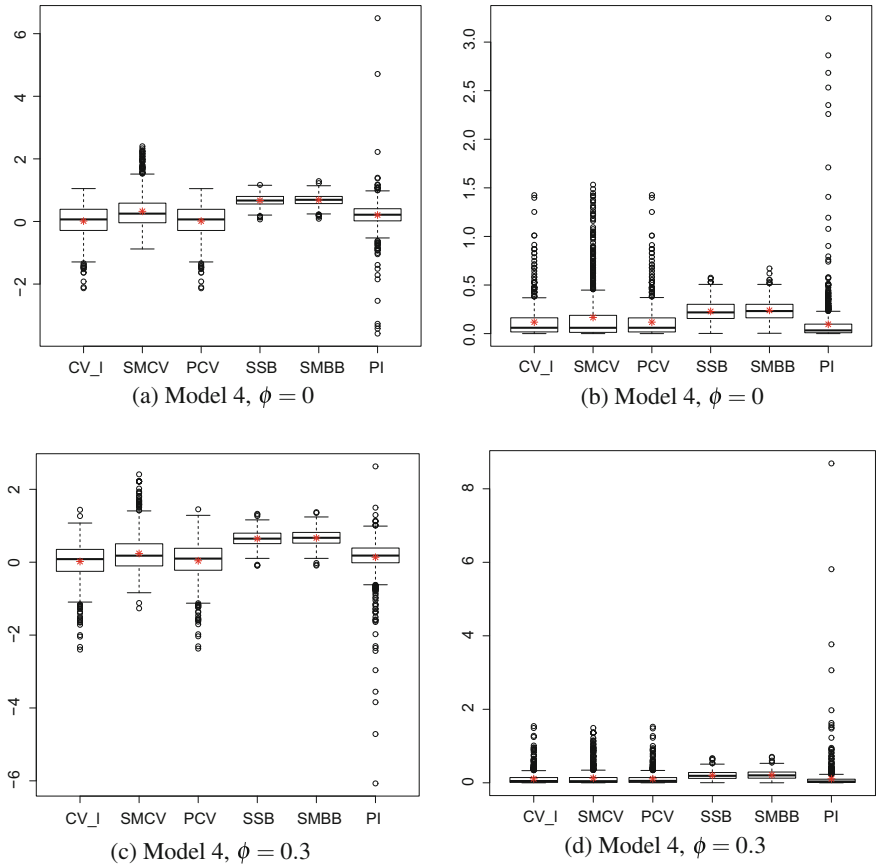


(c) Model 3, $\phi = 0.6$ (d) Model 3, $\phi = 0.6$



(e) Model 3, $\phi = 0.9$ (f) Model 3, $\phi = 0.9$

Fig. 10.4 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 3 with autocorrelation $\phi = 0.3, 0.6, 0.9$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)

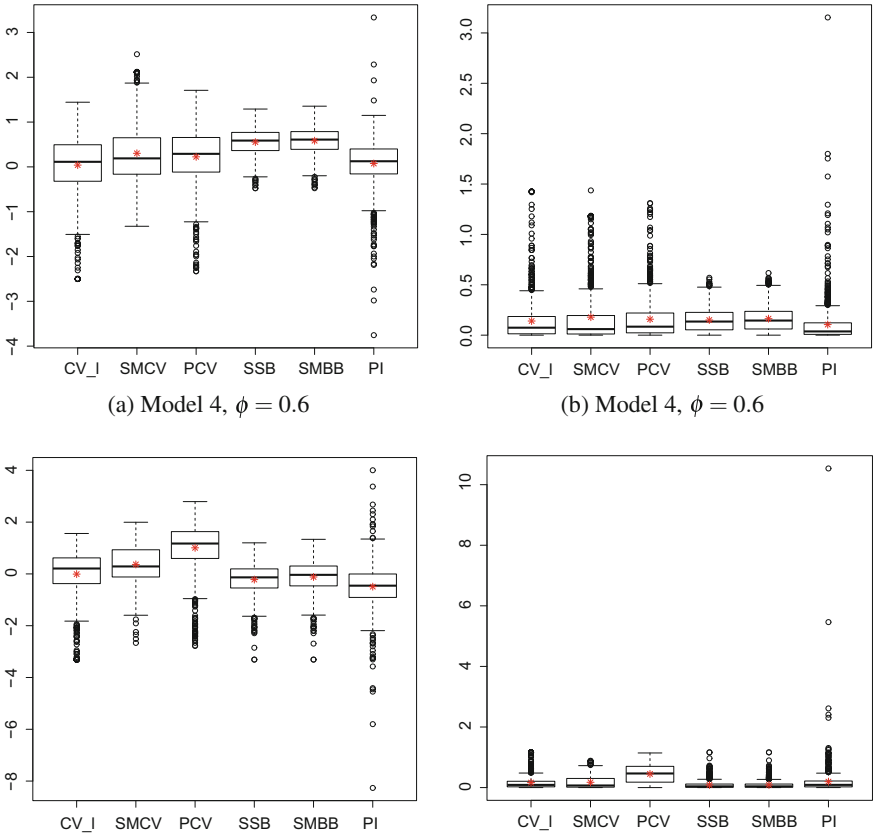


(a) Model 4, $\phi = 0$ (b) Model 4, $\phi = 0$

(c) Model 4, $\phi = 0.3$ (d) Model 4, $\phi = 0.3$

Fig. 10.5 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 4 with autocorrelation $\phi = 0, 0.3$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)

The three cross-validation bandwidths exhibit a worse behaviour than the bootstrap selectors. The bandwidth h_{CV_l} tends to severely undersmooth for some trials in almost every model. This is not satisfactorily corrected by h_{PCV}, which, paradoxically, sometimes shows a general tendency to oversmoothing (see Figs. 10.4e and 10.8e). Although h_{SMCV} typically corrects the extreme udersmoothing cases of h_{CV_l} and h_{PCV}, on the average it tends to give smaller values than the target h_{MISE}. The undersmoothing feature is also present in h_{PI}, which, in turn, for some models, presents a remarkable proportion of trials with severe oversmoothing.

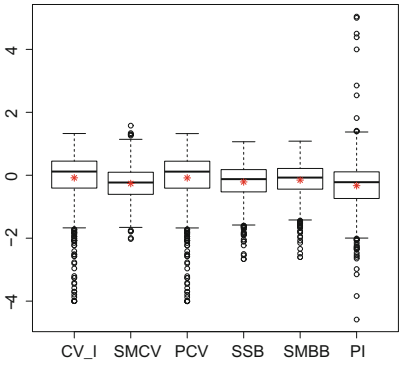


(a) Model 4, $\phi = 0.6$ (b) Model 4, $\phi = 0.6$

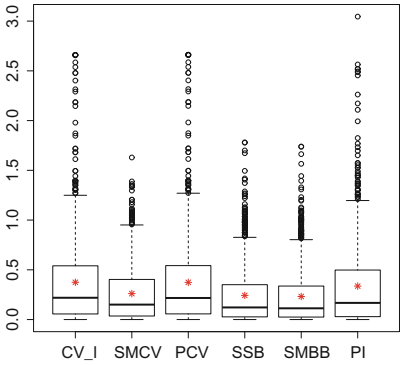
(c) Model 4, $\phi = 0.9$ (d) Model 4, $\phi = 0.9$

Fig. 10.6 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 4 with autocorrelation $\phi = 0.6, 0.9$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)

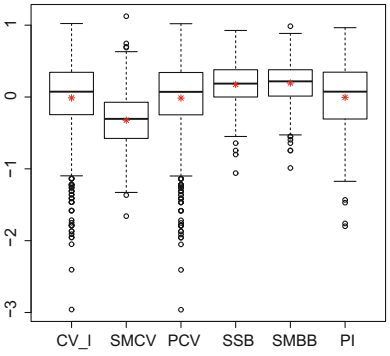
All in all, it is clear in Figs. 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, 10.7 and 10.8 (right side) that the two bootstrap-based bandwidth selectors present the best results in terms of $MISE$. It is also worth pointing out that h^*_{SMBB} actually performs better than its main competitor, h^*_{SSB}, when there exists heavy and positive correlation (specifically, $\phi = 0.9$), as can be noticed in Figs. 10.4f and 10.8f. Additionally, even for moderate autocorrelation, it can be easily checked by looking at Fig. 10.2f that, for Model 7, the empirical behaviour presented by h^*_{SMBB} is by far the best (in terms of $MISE$). However, Model 7 is in itself difficult to analyze in a nonparametric way, due to the fact that its underlying theoretical density is trimodal.



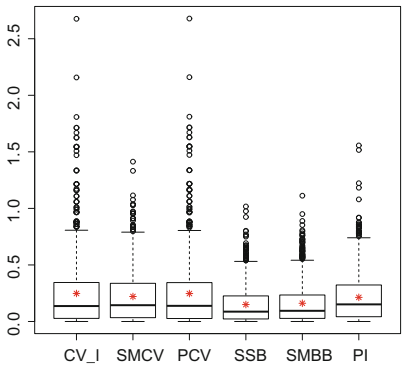
(a) Model 5, $\phi = -0.9$



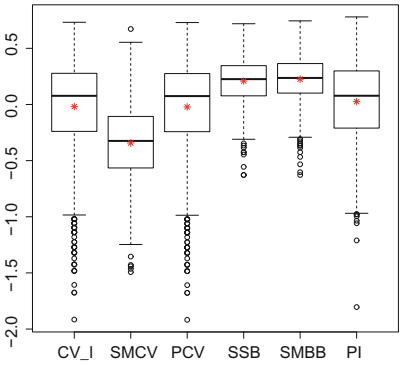
(b) Model 5, $\phi = -0.9$



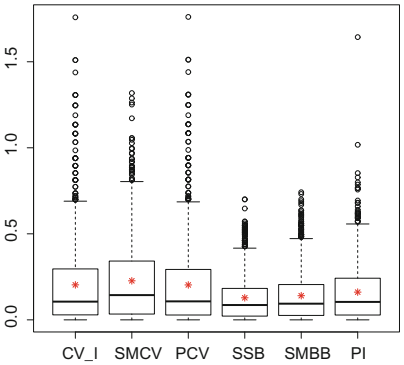
(c) Model 5, $\phi = -0.6$



(d) Model 5, $\phi = -0.6$

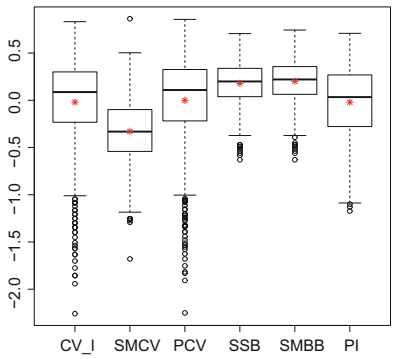


(e) Model 5, $\phi = -0.3$

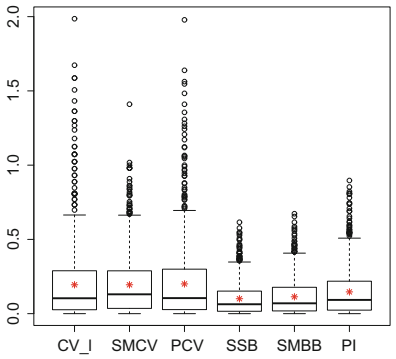


(f) Model 5, $\phi = -0.3$

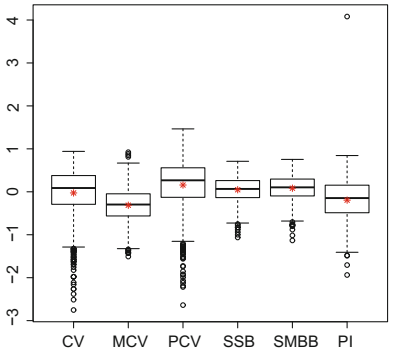
Fig. 10.7 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 5 with autocorrelation $\phi = -0.9, -0.6, -0.3$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)



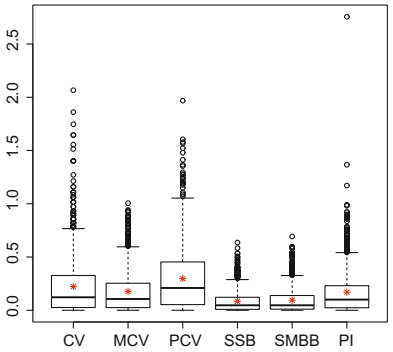
(a) Model 5, $\phi = 0.3$



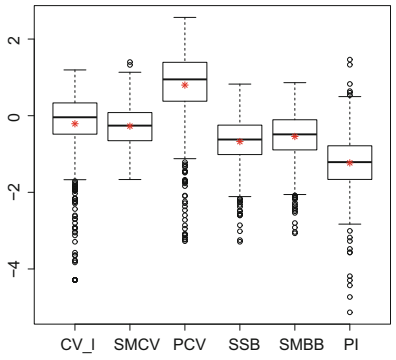
(b) Model 5, $\phi = 0.3$



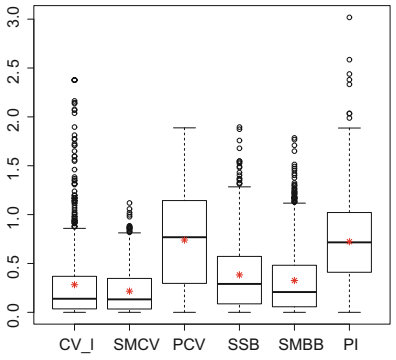
(c) Model 5, $\phi = 0.6$



(d) Model 5, $\phi = 0.6$



(e) Model 5, $\phi = 0.9$



(f) Model 5, $\phi = 0.9$

Fig. 10.8 Boxplots of $\log\left(\hat{h}/h_{MISE}\right)$ (*left side*) and $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (*right side*) for Model 5 with autocorrelation $\phi = 0.3, 0.6, 0.9$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box)

Acknowledgements The authors acknowledge partial support by MINECO grant MTM2014-52876-R and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF. They would also like to thank an anonymous referee for his/her comments that have helped to improve this chapter.

10.5 Appendix

Proof (*Proof of Theorem* 1) Let us take into account a random sample $(X_1, \ldots, X_n)$ which comes from a stationary process and the smoothed moving blocks bootstrap version of the kernel density estimator, $\hat{f}_h^*(x)$. The bootstrap version of the mean integrated squared error is given by:

$$MISE^*(h) = B^*(h) + V^*(h), \tag{10.6}$$

where

$$B^*(h) = \int \left[\mathbb{E}^*\left(\hat{f}_h^*(x)\right) - \hat{f}_g(x)\right]^2 dx, \text{ and}$$
$$V^*(h) = \int \mathrm{Var}^*\left(\hat{f}_h^*(x)\right) dx.$$

Now, straight forward calculations lead to

$$B^*(h) = \int \left[\mathbb{E}^*\left(\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i^*)\right) - \hat{f}_g(x)\right]^2 dx$$
$$= \int \left[\frac{1}{n}\sum_{i=1}^{n} \int K_h(x-y)\hat{f}_g^{(i)}(y)dy - \hat{f}_g(x)\right]^2 dx,$$

where

$$\hat{f}_g^{(i)}(y) = \frac{1}{n-b+1}\sum_{j=t_i}^{n-b+t_i} K_g(y - X_j),$$

considering $t_i = [(i-1)\mathrm{mod}\, b] + 1$.

Let us now assume that n is an integer multiple of b:

$$
\begin{aligned}
&\int \left[\frac{1}{n}\sum_{i=1}^{n}\int K_h(x-y)\hat{f}_g^{(i)}(y)dy - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{n}\sum_{i=1}^{b}\frac{n}{b}\left(K_h * \hat{f}_g^{(i)}\right)(x) - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{b}\sum_{i=1}^{b}\left(K_h * \hat{f}_g^{(i)}\right)(x) - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{b}\sum_{i=1}^{b}\left(\frac{1}{n-b+1}\sum_{j=t_i}^{n-b+t_i} K_h * K_g(\cdot - X_j)\right)(x) - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{b}\sum_{i=1}^{b}\left(\frac{1}{n-b+1}\sum_{j=t_i}^{n-b+t_i}\int K_h(x-y)K_g(y-X_j)dy\right) - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{b}\sum_{i=1}^{b}\left(\frac{1}{n-b+1}\sum_{j=t_i}^{n-b+t_i}\int K_h(x-u-X_j)K_g(u)du\right) - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{b}\sum_{i=1}^{b}\left(\frac{1}{n-b+1}\sum_{j=t_i}^{n-b+t_i} K_h * K_g(x-X_j)\right) - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\frac{1}{b(n-b+1)}\sum_{i=1}^{b}\sum_{j=t_i}^{n-b+t_i} K_h * K_g(x-X_j) - \hat{f}_g(x)\right]^2 dx.
\end{aligned}
$$

Furthermore, if $b < n$

$$
\begin{aligned}
&\frac{1}{b(n-b+1)}\sum_{i=1}^{b}\sum_{j=t_i}^{n-b+t_i} K_h * K_g(x-X_j) \\
&= \frac{1}{n-b+1}\sum_{j=b}^{n-b+1} K_h * K_g(x-X_j) + \frac{1}{b(n-b+1)}\sum_{j=1}^{b-1} j(K_h * K_g)(x-X_j)
\end{aligned}
$$

$$+\frac{1}{b(n-b+1)}\sum_{j=n-b+2}^{n}(n-j+1)(K_h * K_g)(x-X_j)$$
$$=\sum_{j=1}^{n}a_j(K_h * K_g)(x-X_j),$$

where a_j (10.3).
If $b=n$,

$$\frac{1}{b(n-b+1)}\sum_{i=1}^{b}\sum_{j=t_i}^{n-b+t_i}K_h * K_g(x-X_j)=\frac{1}{n}\sum_{j=1}^{n}K_h * K_g(x-X_j)$$
$$=\sum_{j=1}^{n}a_j\left(K_h * K_g\right)(x-X_j),$$

considering $a_j=\frac{1}{n}$, if $b=n$.
Hence, carrying on with the calculations of the integrated bootstrap bias (including several changes of variable and using the symmetry of K) results in:

$$B^*(h)=\int\left[\sum_{j=1}^{n}a_j\left(K_h * K_g\right)(x-X_j)-\hat{f}_g(x)\right]^2dx$$
$$=\int\left[\sum_{j=1}^{n}a_j\left(K_h * K_g\right)(x-X_j)-\frac{1}{n}\sum_{j=1}^{n}K_g(x-X_j)\right]^2dx$$
$$=\sum_{j=1}^{n}\sum_{k=1}^{n}\int\left[a_j\left(K_h * K_g\right)(x-X_j)-\frac{1}{n}K_g(x-X_j)\right]$$
$$\times\left[a_k\left(K_h * K_g\right)(x-X_k)-\frac{1}{n}K_g(x-X_k)\right]dx$$

$$
\begin{aligned}
&= \sum_{j=1}^{n}\sum_{k=1}^{n} \int \Big[a_j a_k \left(K_h * K_g\right)(x - X_j)\left(K_h * K_g\right)(x - X_k) \\
&\quad - \frac{2a_j}{n}\left(K_h * K_g\right)(x - X_j) K_g(x - X_k) + \frac{1}{n^2} K_g(x - X_j) K_g(x - X_k) \Big] dx \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \int \left[\left(K_h * K_g\right)(x - X_j)\left(K_h * K_g\right)(x - X_k)\right] dx \\
&\quad - \frac{2}{n}\sum_{j=1}^{n} a_j \sum_{k=1}^{n} \int \left[\left(K_h * K_g\right)(x - X_j) K_g(x - X_k)\right] dx \\
&\quad + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \int \left[K_g(x - X_j) K_g(x - X_k)\right] dx \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \int \left[\left(K_h * K_g\right)(-v)\left(K_h * K_g\right)(X_j - X_k - v)\right] dv \\
&\quad - \frac{2}{n}\sum_{j=1}^{n} a_j \sum_{k=1}^{n} \int \left[\left(K_h * K_g\right)(-v) K_g(X_j - X_k - v)\right] dv \\
&\quad + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \int \left[K_g(-v) K_g(X_j - X_k - v)\right] dv \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right](X_j - X_k) \\
&\quad - \frac{2}{n}\sum_{j=1}^{n} a_j \sum_{k=1}^{n} \left[\left(K_h * K_g\right) * K_g\right](X_j - X_k) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \left[K_g * K_g\right](X_j - X_k).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
B^*(h) &= \sum_{j=1}^{n} a_j \sum_{k=1}^{n} a_k \left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right](X_j - X_k) \\
&\quad - \frac{2}{n}\sum_{j=1}^{n} a_j \sum_{k=1}^{n} \left[\left(K_h * K_g\right) * K_g\right](X_j - X_k) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \left[K_g * K_g\right](X_j - X_k).
\end{aligned} \tag{10.7}
$$

We now focus on the integrated bootstrap variance, which needs a deeper insight:

$$
\begin{aligned}
V^*(h) &= \int Var^*\left(n^{-1}\sum_{i=1}^{n} K_h\left(x - X_i^*\right)\right)dx \\
&= n^{-2}\int \sum_{i=1}^{n} Var^*\left(K_h\left(x - X_i^*\right)\right)dx \\
&\quad + n^{-2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}\int Cov^*\left(K_h\left(x - X_i^*\right), K_h\left(x - X_j^*\right)\right)dx \\
&= n^{-2}\sum_{i=1}^{n}\int \mathbb{E}^*\left(K_h\left(x - X_i^*\right)^2\right)dx \\
&\quad - n^{-2}\sum_{i=1}^{n}\int \left[\mathbb{E}^*\left(K_h\left(x - X_i^*\right)\right)\right]^2 dx \\
&\quad + n^{-2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}\int Cov^*\left(K_h\left(x - X_i^*\right), K_h\left(x - X_j^*\right)\right)dx. \qquad (10.8)
\end{aligned}
$$

The first term in (10.8), after some changes of variable, is given by:

$$
\begin{aligned}
&n^{-2}\sum_{i=1}^{n}\int \mathbb{E}^*\left(K_h\left(x - X_i^*\right)^2\right)dx = n^{-2}\sum_{i=1}^{n}\int\left[\int K_h(x-y)^2 \hat{f}_g^{(i)}(y)dy\right]dx \\
&= n^{-2}\sum_{i=1}^{n}\int\left[\int K_h(x-y)^2\left[\frac{1}{n-b+1}\sum_{j=t_i}^{n-b+t_i} K_g(y - X_j)\right]dy\right]dx \\
&= \frac{1}{n^2(n-b+1)}\sum_{i=1}^{n}\sum_{j=t_i}^{n-b+t_i}\int K_g(y - X_j)\left[\int K_h(x-y)^2 dx\right]dy \\
&= \frac{1}{n^2(n-b+1)}\sum_{i=1}^{n}\sum_{j=t_i}^{n-b+t_i}\int K_g(y - X_j)\left[\frac{1}{h}\int K(z)^2 dz\right]dy \\
&= \frac{R(K)}{n^2(n-b+1)h}\sum_{i=1}^{n}\sum_{j=t_i}^{n-b+t_i}\int K_g(y - X_j)dy \\
&= \frac{R(K)}{n^2(n-b+1)h}\sum_{i=1}^{n}\sum_{j=t_i}^{n-b+t_i}\int K(u)\,du = \frac{R(K)}{nh}.
\end{aligned}
\qquad (10.9)
$$

Focusing now on the second term, including several changes of variable and using the symmetry of K:

$$
\begin{aligned}
&n^{-2}\sum_{i=1}^{n}\int\left[\mathbb{E}^*\left(K_h\left(x-X_i^*\right)\right)\right]^2dx = n^{-2}\sum_{i=1}^{n}\int\left[\int K_h(x-y)\hat{f}_g^{(i)}(y)dy\right]^2dx\\
&= n^{-1}b^{-1}\sum_{i=1}^{b}\int\left[\left(K_h * \hat{f}_g^{(i)}\right)(x)\right]^2dx\\
&= n^{-1}b^{-1}\sum_{i=1}^{b}\int\left[\sum_{j=t_i}^{n-b+t_i}\frac{1}{n-b+1}(K_h*K_g)(x-X_j)\right]\\
&\quad\times\left[\sum_{k=t_i}^{n-b+t_i}\frac{1}{n-b+1}(K_h*K_g)(x-X_k)\right]dx\\
&= \frac{1}{nb(n-b+1)^2}\sum_{i=1}^{b}\sum_{j=t_i}^{n-b+t_i}\sum_{k=t_i}^{n-b+t_i}\int(K_h*K_g)(x-X_j)(K_h*K_g)(x-X_k)dx\\
&= \frac{1}{nb(n-b+1)^2}\sum_{i=1}^{b}\sum_{j=t_i}^{n-b+t_i}\sum_{k=t_i}^{n-b+t_i}\int(K_h*K_g)(v)(K_h*K_g)(X_j-X_k-v)dv\\
&= \frac{1}{nb(n-b+1)^2}\sum_{i=1}^{b}\sum_{j=i}^{n-b+i}\sum_{k=i}^{n-b+i}\left[(K_h*K_g)*(K_h*K_g)\right](X_j-X_k).
\end{aligned}
$$

Let us consider the function ψ defined in Theorem 1. Whenever $b < n$, we have:

$$
\begin{aligned}
&n^{-2}\sum_{i=1}^{n}\int\left[\mathbb{E}^*\left(K_h(x-X_j^*)\right)\right]^2dx\\
&= \frac{1}{nb(n-b+1)^2}\sum_{i=1}^{b}\sum_{j=i}^{n-b+i}\sum_{k=i}^{n-b+i}\left[(K_h*K_g)*(K_h*K_g)\right](X_j-X_k)\\
&= \frac{1}{nb(n-b+1)^2}\sum_{i=1}^{b}\sum_{j=i}^{n-b+i}\sum_{k=i}^{n-b+i}\psi(X_j-X_k)\\
&= \frac{1}{nb(n-b+1)^2}\left[\sum_{i=1}^{b}\sum_{j=i}^{b-1}\sum_{k=i}^{b-1}\psi(X_j-X_k)\right.\\
&\quad\left.+\sum_{i=1}^{b}\sum_{j=i}^{b-1}\sum_{k=b}^{n-b+1}\psi(X_j-X_k)+\sum_{i=1}^{b}\sum_{j=i}^{b-1}\sum_{k=n-b+2}^{n-b+i}\psi(X_j-X_k)\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=1}^{b}\sum_{j=b}^{n-b+1}\sum_{k=i}^{b-1}\psi(X_j-X_k)+\sum_{i=1}^{b}\sum_{j=b}^{n-b+1}\sum_{k=b}^{n-b+1}\psi(X_j-X_k)\\
&+\sum_{i=1}^{b}\sum_{j=b}^{n-b+1}\sum_{k=n-b+2}^{n-b+i}\psi(X_j-X_k)+\sum_{i=1}^{b}\sum_{j=n-b+2}^{n-b+i}\sum_{k=i}^{b-1}\psi(X_j-X_k)\\
&+\sum_{i=1}^{b}\sum_{j=n-b+2}^{n-b+i}\sum_{k=b}^{n-b+1}\psi(X_j-X_k)+\sum_{i=1}^{b}\sum_{j=n-b+2}^{n-b+i}\sum_{k=n-b+2}^{n-b+i}\psi(X_j-X_k)\Bigg]\\
=&\frac{1}{nb(n-b+1)^2}\Bigg[\sum_{j=1}^{b-1}\sum_{k=1}^{b-1}\sum_{i=1}^{\min\{j,k\}}\psi(X_j-X_k)\\
&+\sum_{j=1}^{b-1}\sum_{k=b}^{n-b+1}\sum_{i=1}^{j}\psi(X_j-X_k)+\sum_{j=1}^{b-1}\sum_{k=n-b+2}^{n}\sum_{i=\max\{(k+b-n),1\}}^{j}\psi(X_j-X_k)\\
&+\sum_{j=b}^{n-b+1}\sum_{k=1}^{b-1}\sum_{i=1}^{k}\psi(X_j-X_k)+\sum_{j=b}^{n-b+1}\sum_{k=b}^{n-b+1}\sum_{i=1}^{b}\psi(X_j-X_k)\\
&+\sum_{j=b}^{n-b+1}\sum_{k=n-b+2}^{n}\sum_{i=\max\{(k-n+b),1\}}^{b}\psi(X_j-X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=1}^{b-1}\sum_{i=\max\{(j+b-n),1\}}^{k}\psi(X_j-X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=b}^{n-b+1}\sum_{i=\max\{(j-n+b),1\}}^{b}\psi(X_j-X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=n-b+2}^{n}\sum_{i=\max\{(j-n+b),(k-n+b)\}}^{b}\psi(X_j-X_k)\Bigg]\\
=&\frac{1}{nb(n-b+1)^2}\Bigg[\sum_{j=1}^{b-1}\sum_{k=1}^{b-1}\min\{j,k\}\psi(X_j-X_k)\\
&+\sum_{j=1}^{b-1}j\sum_{k=b}^{n-b+1}\psi(X_j-X_k)\\
&+\sum_{j=1}^{b-1}\sum_{k=n-b+2}^{n}\min\{(n-b+j-k+1),j\}\psi(X_j-X_k)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{j=b}^{n-b+1}\sum_{k=1}^{b-1} k\psi(X_j - X_k) + b\sum_{j=b}^{n-b+1}\sum_{k=b}^{n-b+1}\psi(X_j - X_k)\\
&+\sum_{j=b}^{n-b+1}\sum_{k=n-b+2}^{n}\min\{(n-k+1), b\}\psi(X_j - X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=1}^{b-1}\min\{(n-b+k-j+1), k\}\psi(X_j - X_k)\\
&+\sum_{j=n-b+2}^{n}\min\{(n-j+1), b\}\sum_{k=b}^{n-b+1}\psi(X_j - X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=n-b+2}^{n}\min\{(n-j+1), (n-k+1)\}\psi(X_j - X_k)\Bigg]\\
=\ &\frac{1}{nb(n-b+1)^2}\Bigg[\sum_{j=1}^{b-1}\sum_{k=1}^{b-1}\min\{j,k\}\psi(X_j - X_k)\\
&+\sum_{j=1}^{b-1} j\sum_{k=b}^{n-b+1}\psi(X_j - X_k)\\
&+\sum_{j=1}^{b-1}\sum_{k=n-b+2}^{n}\min\{(n-b+j-k+1), j\}\psi(X_j - X_k)\\
&+\sum_{j=b}^{n-b+1}\sum_{k=1}^{b-1} k\psi(X_j - X_k) + b\sum_{j=b}^{n-b+1}\sum_{k=b}^{n-b+1}\psi(X_j - X_k)\\
&+\sum_{j=b}^{n-b+1}\sum_{k=n-b+2}^{n}\min\{(n-k+1), b\}\psi(X_j - X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=1}^{b-1}\min\{(n-b+k-j+1), k\}\psi(X_j - X_k)\\
&+\sum_{j=n-b+2}^{n}\min\{(n-j+1), b\}\sum_{k=b}^{n-b+1}\psi(X_j - X_k)\\
&+\sum_{j=n-b+2}^{n}\sum_{k=n-b+2}^{n}(n+1-\max\{j,k\})\,\psi(X_j - X_k)\Bigg].
\end{aligned}
\tag{10.10}
$$

On the other hand, if $b = n$:

$$\frac{1}{nb(n-b+1)^2}\sum_{i=1}^{b}\sum_{j=i}^{n-b+i}\sum_{k=i}^{n-b+i}\psi\left(X_j - X_k\right) = \frac{1}{n^2}\sum_{i=1}^{n}\psi(X_i - X_i) = \frac{\psi(0)}{n}.$$

Finally, we investigate the covariance term further. It is now necessary to take into account the following notation, naming the n/b blocks as follows:

$$J_r = \{(r-1)b+1, (r-1)b+2, \ldots, rb\}, r = 1, 2, \ldots, n/b.$$

Thus, X_i^* and X_j^* turn out to be independent (in the bootstrap universe) whenever it does not exist $r \in \{1, 2, \ldots, n/b\}$ which satisfies $i, j \in J_r$. In that case, X_i^* and X_j^* do not belong to the same bootstrap block, implying:

$$Cov^*\left(K_h(x - X_i^*), K_h(x - X_j^*)\right) = 0.$$

On the other hand, if there exists $r \in \{1, 2, \ldots, n/b\}$ satisfying $i, j \in J_r$, then the bootstrap distribution of the pair (X_i^*, X_j^*) is exactly identical to that of the pair $(X_{t_i}^*, X_{t_j}^*)$, where $t_i = [(i-1)\text{mod } b] + 1$. Let us consider $r \in \{1, 2, \ldots, n/b\}$ satisfying $i, j \in J_r$, then X_i^* e X_j^* belong to the same bootstrap block. As a consequence,

$$Cov^*\left(K_h(x - X_i^*), K_h(x - X_j^*)\right) = Cov^*\left(K_h(x - X_{t_i}^*), K_h(x - X_{t_j}^*)\right).$$

Thus:

$$Cov^*\left(K_h(x - X_i^*), K_h(x - X_j^*)\right) = \begin{cases} Cov^*\left(K_h(x - X_{t_i}^*), K_h(x - X_{t_j}^*)\right), & \text{if } \exists r / i, j \in J_r \\ 0, & \text{otherwise} \end{cases}.$$

Notice that: $\mathbb{E}^*\left[K_h\left(x - X_i^*\right)\right] = \left(K_h * \hat{f}_g^{(i)}\right)$, and $\mathbb{E}^*\left[K_h\left(x - X_j^*\right)\right] = \left(K_h * \hat{f}_g^{(j)}\right)$. Now, consider $k, \ell \in \{1, 2, \ldots, b\}$ satisfying $k < \ell$. Carrying on with the calculations of the covariance term and using:

$$\mathbb{P}^*\left(\left(X_k^{*(d)}, X_\ell^{*(d)}\right) = \left(X_j, X_{j+\ell-k}\right)\right) = \frac{1}{n-b+1}, j = k, k+1, \ldots, n-b+k,$$

leads to:

$$\begin{aligned} &\frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} Cov^*\left(K_h\left(x - X_i^*\right), K_h\left(x - X_j^*\right)\right) \\ &= \frac{1}{n^2}\frac{n}{b}\sum_{\substack{k,\ell=1 \\ k\neq\ell}}^{b} Cov^*\left(K_h\left(x - X_k^*\right), K_h\left(x - X_\ell^*\right)\right) \end{aligned}$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} Cov^* \left(K_h \left(x - X_k^* \right), K_h \left(x - X_\ell^* \right) \right)$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\mathbb{E}^* \left(K_h \left(x - X_k^* \right) K_h \left(x - X_\ell^* \right) \right) - \mathbb{E}^* \left(K_h \left(x - X_k^* \right) \right) \mathbb{E}^* \left(K_h \left(x - X_\ell^* \right) \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\mathbb{E}^* \left[\mathbb{E}^* \left(K_h \left(x - X_k^* \right) K_h \left(x - X_\ell^* \right) \big|_{U_k^*, U_\ell^*} \right) \right] - \left(K_h * \hat{f}_g^{(k)} \right) \left(K_h * \hat{f}_g^{(\ell)} \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\mathbb{E}^* \left[\mathbb{E}^* \left(K_h \left(x - X_k^{*(d)} - g U_k^* \right) K_h \left(x - X_\ell^{*(d)} - g U_\ell^* \right) \Big|_{U_k^*, U_\ell^*} \right) \right] \right.$$
$$\left. - \left(K_h * \hat{f}_g^{(k)}(x) \right) \left(K_h * \hat{f}_g^{(\ell)}(x) \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} \mathbb{E}^* \left[K_h \left(x - X_j - g U_k^* \right) K_h \left(x - X_{j+\ell-k} - g U_\ell^* \right) \right] \right.$$
$$\left. - \left(K_h * \hat{f}_g^{(k)}(x) \right) \left(K_h * \hat{f}_g^{(\ell)}(x) \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} \int \int K_h(x - X_j - gu) K_h(x - X_{j+\ell-k} - gv) K(u) K(v) du dv \right.$$
$$\left. - \left(K_h * \hat{f}_g^{(k)}(x) \right) \left(K_h * \hat{f}_g^{(\ell)}(x) \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} \int \int K_h(x - X_j - s) K_h(x - X_{j+\ell-k} - t) K_g(s) K_g(t) ds dt \right.$$
$$\left. - \left(K_h * \hat{f}_g^{(k)}(x) \right) \left(K_h * \hat{f}_g^{(\ell)}(x) \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} (K_h * K_g)(x - X_j)(K_h * K_g)(x - X_{j+\ell-k}) \right.$$
$$\left. - \left(K_h * \hat{f}_g^{(k)}(x) \right) \left(K_h * \hat{f}_g^{(\ell)}(x) \right) \right]$$

$$= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} (K_h * K_g)(x - X_j)(K_h * K_g)(x - X_{j+\ell-k}) \right.$$
$$- \left(\frac{1}{n-b+1} \sum_{i=k}^{n-b+k} \int K_h(x - y) K_g(y - X_i) dy \right)$$
$$\left. \times \left(\frac{1}{n-b+1} \sum_{j=\ell}^{n-b+\ell} \int K_h(x - y) K_g(y - X_j) dy \right) \right]$$

$$
\begin{aligned}
&= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} (K_h * K_g)(x - X_j)(K_h * K_g)(x - X_{j+\ell-k}) \right. \\
&\quad - \left(\frac{1}{n-b+1} \sum_{i=k}^{n-b+k} \int K_h(x - X_i - u) K_g(u) du \right) \\
&\quad \times \left. \left(\frac{1}{n-b+1} \sum_{j=\ell}^{n-b+\ell} \int K_h(x - X_j - u) K_g(u) du \right) \right] \\
&= \frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} (K_h * K_g)(x - X_j)(K_h * K_g)(x - X_{j+\ell-k}) \right. \\
&\quad \left. - \frac{1}{(n-b+1)^2} \sum_{i=k}^{n-b+k} \sum_{j=\ell}^{n-b+\ell} (K_h * K_g)(x - X_i)(K_h * K_g)(x - X_j) \right].
\end{aligned}
$$

The integral with respect to x is now computed (using some changes of variable and the symmetry of the kernel K):

$$
\begin{aligned}
&\int \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} Cov^* \left(K_h \left(x - X_i^* \right), K_h \left(x - X_j^* \right) \right) dx \\
&= \int \left[\frac{2}{nb} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} (K_h * K_g)(x - X_j)(K_h * K_g)(x - X_{j+\ell-k}) \right. \right. \\
&\quad \left. \left. - \frac{1}{(n-b+1)^2} \sum_{i=k}^{n-b+k} \sum_{j=\ell}^{n-b+\ell} (K_h * K_g)(x - X_i)(K_h * K_g)(x - X_j) \right] \right] dx \\
&= \frac{2}{nb} \left[\sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \left[\frac{1}{n-b+1} \sum_{j=k}^{n-b+k} \int (K_h * K_g)(X_{j+\ell-k} - X_j - u)(K_h * K_g)(u) du \right. \right. \\
&\quad \left. \left. - \frac{1}{(n-b+1)^2} \sum_{i=k}^{n-b+k} \sum_{j=\ell}^{n-b+\ell} \int (K_h * K_g)(u)(K_h * K_g)(X_i - X_j - u) du \right] \right]
\end{aligned}
$$

$$= \frac{2}{nb(n-b+1)} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{j=k}^{n-b+k} \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+\ell-k} - X_j)$$
$$- \frac{2}{nb(n-b+1)^2} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=k}^{n-b+k} \sum_{j=\ell}^{n-b+\ell} \left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j).$$

Notice that, whenever $b < n$:

$$\sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{j=k}^{n-b+k} \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+\ell-k} - X_j)$$
$$= \sum_{k=1}^{b-1} \sum_{j=k}^{n-b+k} \sum_{\ell=k+1}^{b} \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+\ell-k} - X_j)$$
$$= \sum_{k=1}^{b-1} \sum_{j=k}^{n-b+k} \sum_{s=1}^{b-k} \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+s} - X_j)$$
$$= \sum_{s=1}^{b-1} \sum_{j=1}^{n-s} \sum_{k=\max\{1,\, j+b-n\}}^{\min\{j,\, b-s\}} \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+s} - X_j)$$
$$= \sum_{s=1}^{b-1} \sum_{j=1}^{n-s} (\min\{j, b-s\} - \max\{1,\, j+b-n\} + 1) \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+s} - X_j).$$

Now, using the function ψ, and considering $b < n$, we have:

$$n^{-2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \int Cov^* \left(K_h(x - X_i^*), K_h(x - X_j^*)\right) dx$$
$$= \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=k}^{n-b+k} \sum_{j=\ell}^{n-b+\ell} \psi(X_i - X_j)$$
$$= \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=k}^{b-2} \sum_{j=\ell}^{b-1} \psi(X_i - X_j) + \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=k}^{b-2} \sum_{j=b}^{n-b+2} \psi(X_i - X_j)$$
$$+ \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=k}^{b-2} \sum_{j=n-b+3}^{n-b+\ell} \psi(X_i - X_j)$$
$$+ \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=b-1}^{n-b+1} \sum_{j=n-b+3}^{n-b+\ell} \psi(X_i - X_j) + \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=n-b+2}^{n-b+k} \sum_{j=\ell}^{b-1} \psi(X_i - X_j)$$

$$
\begin{aligned}
&+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\\
&=\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=k}^{b-2}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)+\sum_{k=1}^{b-1}\sum_{\ell=k+1}^{b}\sum_{i=k}^{b-2}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\\
&\quad+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=k}^{b-2}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\\
&\quad+\sum_{\ell=2}^{b}\sum_{k=1}^{\ell-1}\sum_{i=b-1}^{n-b+1}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)+\sum_{i=b-1}^{n-b+1}\sum_{j=b}^{n-b+2}\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\psi(X_i-X_j)\\
&\quad+\sum_{\ell=2}^{b}\sum_{k=1}^{\ell-1}\sum_{i=b-1}^{n-b+1}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)\\
&\quad+\sum_{k=1}^{b-1}\sum_{\ell=k+1}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\\
&\quad+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\\
&=\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=k}^{b-2}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)+\sum_{k=1}^{b-1}(b-k)\sum_{i=k}^{b-2}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\\
&\quad+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=k}^{b-2}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\\
&\quad+\sum_{\ell=2}^{b}(\ell-1)\sum_{i=b-1}^{n-b+1}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)+\frac{b(b-1)}{2}\sum_{i=b-1}^{n-b+1}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\\
&\quad+\sum_{\ell=2}^{b}(\ell-1)\sum_{i=b-1}^{n-b+1}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)\\
&\quad+\sum_{k=1}^{b-1}(b-k)\sum_{i=n-b+2}^{n-b+k}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\\
&\quad+\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\sum_{i=n-b+2}^{n-b+k}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j). \qquad (10.11)
\end{aligned}
$$

On the other hand, if $b = n$ and using the symmetry of the kernel K, we obtain:

$$\frac{2}{nb(n-b+1)} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{j=k}^{n-b+k} \left[(K_h * K_g) * (K_h * K_g)\right](X_{j+\ell-k} - X_j)$$

$$-\frac{2}{nb(n-b+1)^2} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{b} \sum_{i=k}^{n-b+k} \sum_{j=\ell}^{n-b+\ell} \left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j)$$

$$= \frac{2}{n^2} \sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} \psi\,(X_\ell - X_k) - \frac{2}{n^2} \sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} \psi\,(X_k - X_\ell) = 0. \tag{10.12}$$

Using (10.9), (10.10) and (10.11) in (10.8), and this and (10.7) in (10.6) gives the statement of Theorem 1 for $b < n$. The case $b = n$ is even simpler using (10.12).

References

Barbeito I, Cao R (2016) Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data. Comput Statist & Data Anal 104:130–147

Bowman A (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71:353–360

Cao R (1993) Bootstrapping the mean integrated squared error. J Mult Anal 45:137–160

Cao R (1999) An overview of bootstrap methods for estimating and predicting in time series. Test 8:95–116

Cao R, Quintela del Río A, Vilar Fernández J (1993) Bandwidth selection in nonparametric density estimation under dependence. a simulation study. Com Statist 8:313–332

Cao R, Cuevas A, González Manteiga W (1994) A comparative-study of several smoothing methods in density-estimation. Comput Statist & Data Anal 17:153–176

Chow Y, Geman S, Wu L (1983) Consistent cross-validated density-estimation. Ann Statist 11:25–38

Cox D, Kim T (1997) A study on bandwidth selection in density estimation under dependence. J Mult Anal 62:190–203

Devroye L (1987) A Course in Density Estimation. Birkhauser, Boston

Efron B (1979) Bootstrap methods: Another look at the jackniffe. Ann Statist 7:1–26

Efron B, Tibishirani R (1993) An Introduction to the Bootstrap. Chapman and Hall, New York

Estévez-Pérez G, Quintela del Río A, Vieu P (2002) Convergence rate for cross-validatory bandwidth in kernel hazard estimation from dependent samples. J Statisti Plann Infer 104:1–30

Faraway J, Jhun M (1990) Bootstrap choice of bandwidth for density estimation. J Amer Statist Assoc 85:1119–1122

Feluch W, Koronacki J (1992) A note on modified cross-validation in density-estimation. Comput Statist & Data Anal 13:143–151

Hall P (1983) Large sample optimality of least-squares cross-validation in densityestimation. Ann Statist 11:1156–1174

Hall P (1990) Using the bootstrap to estimate mean squared error and select smoothing parameter in nonprametric problems. J Multiv Anal 32:177–203

Hall P, Marron J (1987a) Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density-estimation. Probab Theor Relat Fields 74:567–581
Hall P, Marron J (1987b) On the amount of noise inherent in bandwidth selection for a kernel density estimator. Ann Statist 15:163–181
Hall P, Marron J (1991) Local minima in cross-validation functions. J Roy Statist Soc Ser B 53:245–252
Hall P, Lahiri S, Truong Y (1995) On bandwidth choice for density estimation with dependent data. Ann Statist 23:2241–2263
Hart J, Vieu P (1990) Data-driven bandwidth choice for density estimation based on dependent data. Ann Statist 18:873–890
Jones M, Marron J, Park B (1991) A simple root-n bandwidth selector. Ann Statist 19:1919–1932
Jones M, Marron J, Sheather S (1996) A brief survey of bandwidth selection for density estimation. J Amer Statist Assoc 91:401–407
Kreiss J, Paparoditis E (2011) Bootstrap methods for dependent data: A review. J Korean Statist Soc 40:357–378
Künsch H (1989) The jackknife and the bootstrap for general stationary observations. Ann Statist 17:1217–1241
Léger C, Romano J (1990) Bootstrap choice of tuning parameters. Ann Inst Statist Math 42:709–735
Liu R, Singh K (1992) Moving blocks jackknife and bootstrap capture weak dependence. In Exploring the Limits of Bootstrap, eds R LePage and L Billard pp 225–248
Marron J (1985) An asymptotically efficient solution to the bandwidth problem of kernel density-estimation. Ann Statist 13:1011–1023
Marron J (1987) A comparison of cross-validation techniques in density estimation. Ann Statist 15:152-162 A review and new proposals for bandwidth selection in density estimation under dependence 31
Marron J (1992) Bootstrap bandwidth selection. In Exploring the Limits of Bootstrap, eds R LePage and L Billard pp 249–262
Marron J, Wand M (1992) Exact mean integrated squared error. Ann Statist 20(2):712–736
Park B, Marron J (1990) Comparison of data-driven bandwidth selectors. J Amer Statist Assoc 85:66–72
Parzen E (1962) Estimation of a probability density-function and mode. Ann Math Statist 33:1065–1076
Politis D, Romano J (1994) The stationary bootstrap. J Amer Statist Assoc 89:1303–1313
Rosenblatt M (1956) Estimation of a probability density-function and mode. Ann Math Statist 27:832–837
Rudemo M (1982) Empirical choice of histograms and kernel density estimators. Scand J Statist 9:65–78
Scott D, Terrell G (1987) Biased and unbiased cross-validation in density-estimation. J Amer Statist Assoc 82:1131–1146
Sheather S, Jones M (1991) A reliable data-based bandwidth selection method for kernel density estimation. J Roy Statist Soc Ser B 53:683–690
Silverman B, Young G (1987) The bootstrap: To smooth or not to smooth? Biometrika 74:469–479
Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman & Hall, London
Stone C (1984) An asymptotically optimal window selection rule for kernel density estimates. Ann Statist 12:1285–1297
Stute W (1992) Modified cross-validation in density-estimation. J Statisti Plann Infer 30:293–305
Taylor C (1989) Bootstrap choice of the smoothing parameter in kernel density estimation. Biometrika 76:705–712

Estimating the Error Distribution in a Single-Index Model

11

Hira L. Koul, Ursula U. Müller and Anton Schick

11.1 Introduction

We consider the single-index regression model in which the response variable Y is linked to a p-dimensional covariate vector X via the formula

$$Y = \varrho(\theta_0^\top X) + \varepsilon, \tag{11.1}$$

where ϱ is a smooth function, θ_0 is a p-dimensional unit vector, and the error variable ε is independent of the covariate X, has mean zero and a finite variance. In order to guarantee identifiability, we require that the matrix $E[XX^\top]$ is positive definite and that θ_0 belongs to Θ, the set of all p-dimensional unit vectors whose first coordinate is positive, see e.g., Cui et al. (2011). Furthermore, we assume that ε has a density f and that $\theta^\top X$ has a density g_θ for each θ in Θ.

The single-index regression model was introduced to overcome the curse of dimensionality. Numerous applications and theoretical results can be found in Stoker (1986); Li (1991); Ichimura (1993); Xia and Li (1999); Xia et al. (2002a); Xia et al. (2002b); Xia and Härdle (2006); Xia (2008), and references therein. The primary focus of these and related papers has been the estimation of the parameter θ_0 and of the link function ϱ. Stute and Zhu (2005) provide asymptotically distribution free maximin tests for fitting a single-index model to the regression function against a large class of local alternatives.

H.L. Koul (✉)
Department of Statistics and Probability, Michigan State University, East Lansing 48824, USA
e-mail: koul@stt.msu.edu

U.U. Müller
Department of Statistics, Texas A&M University, College Station 77843-3143, USA

A. Schick
Department of Mathematical Sciences, Binghamton University, Binghamton 13902-6000, USA

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_11

Here we are interested in the estimation of the error distribution function F based on independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). Our goal is to derive a first order uniform stochastic expansion for a suitably weighted residual empirical distribution function. Such a uniform expansion has been obtained in linear, partially linear and nonparametric regression models by Koul (1969, 1970, 2002); Akritas and Van Keilegom (2001); Müller et al. (2007, 2009a), and Neumeyer and Van Keilegom (2010). In the context of time series, expansions of this type have been obtained in Boldin (1982, 1990, 1998); Koul (1991, 2002); Müller et al. (2009b), and Neumeyer and Selk (2013). The existing literature does not cover the case of interest here.

If we denote the single-index $\theta_0^\top X$ by S, then we can write the regression model as a nonparametric regression model

$$Y = \varrho(S) + \varepsilon.$$

A common assumption for estimating ϱ in such nonparametric regression models is that the single covariate S has a density that is bounded and bounded away from zero on its compact support. We call distributions of this type *quasi-uniform*. Thus, if θ_0 were known and if S were quasi-uniform, we could estimate ϱ by classical nonparametric curve estimators, and the results of Müller, Schick and Wefelmeyer ("MSW"2007) would yield the desired expansion of the corresponding residual empirical process. However, the assumption that S is quasi-uniform is not reasonable in our case as the following two examples demonstrate.

Example 1 Suppose that X is uniformly distributed on the unit disk $D = \{x \in \mathbb{R}^2 : \|x\| \le 1\}$. In this case $\theta^\top X$ has density

$$g_\theta(s) = g(s) = \frac{2}{\pi}\sqrt{1 - s^2}\, \mathbf{1}[|s| < 1]$$

for all θ in Θ, and this density is not bounded away from zero on $[-1, 1]$.

Example 2 Suppose that X is uniformly distributed on the unit square $[0, 1] \times [0, 1]$. Let $\theta = (a, b)^\top$ with $0 < a \le 1/\sqrt{2}$ and $b = \sqrt{1 - a^2}$. Then the density of $\theta^\top X$ is given by

$$g_\theta(s) = \frac{1}{ab}\Big[\mathbf{1}[0 \le s \le b] \min(s, a) + \mathbf{1}[b < s < a + b](a + b - s)\Big].$$

This density is piecewise linear, and its support depends on a.

Let $\hat\theta$ be an estimator of θ_0 and set

$$\hat S_j = \hat\theta^\top X_j \quad \text{and} \quad \hat\delta_j = \mathbf{1}[\hat S_j \in \hat I], \quad j = 1, \ldots, n,$$

where $\hat{I}$ is the random interval $[\hat{l}, \hat{u}]$ whose endpoints are functions of the estimated indices $\hat{S}_1, \ldots, \hat{S}_n$ and the estimator $\hat{\theta}$, say

$$\hat{l} = \phi_{n,l}(\hat{S}_1, \ldots, \hat{S}_n, \hat{\theta}) \quad \text{and} \quad \hat{u} = \phi_{n,u}(\hat{S}_1, \ldots, \hat{S}_n, \hat{\theta}). \tag{11.2}$$

Choices of such random intervals are discussed in Remark 1 below.

We estimate the link function ϱ by a local quadratic smoother $\hat{\varrho}$ treating $\hat{S}_j$ as the regressor. Our estimator $\hat{\mathbb{F}}_n$ of the distribution function F is based on the residuals $Y_j - \hat{\varrho}(\hat{S}_j)$ for which $\hat{\delta}_j = 1$, i.e.,

$$\hat{\mathbb{F}}_n(t) = \frac{1}{N_n} \sum_{j=1}^{n} \hat{\delta}_j \mathbf{1}[Y_j - \hat{\varrho}(\hat{S}_j) \le t], \quad t \in \mathbb{R}, \tag{11.3}$$

with $N_n = \sum_{j=1}^{n} \hat{\delta}_j$.

Remark 1 Let us briefly comment on choices of $\hat{I}$. The goal is to have g_{θ_0} bounded away from zero on $\hat{I}$ with high probability. If g_θ were known, we could choose intervals $I(\theta)$ on which g_θ is bounded away from zero, and then take $\hat{I} = I(\hat{\theta})$. If g_θ is known up to parameters, say g_θ is a normal density with mean $\theta^\top \mu$ and variance $\theta^\top \Sigma \theta$ for some unknown vector μ and some unknown dispersion matrix Σ, then we could take $\hat{I} = [\hat{\nu} - c\hat{\sigma}, \hat{\nu} + c\hat{\sigma}]$, where $\hat{\nu}$ is the sample mean and $\hat{\sigma}$ the sample standard deviation of the estimated indices $\hat{S}_1, \ldots, \hat{S}_n$. Another choice for $\hat{I}$ is $[\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}]$ for $0 < \alpha_1 < \alpha_2 < 1$, where $\hat{q}_\alpha$ denotes the α-th sample quantile of the estimated indices. Using only values $\hat{S}_j$ that are densely distributed in an interval $\hat{I}$ around the mean or median, e.g., between the upper and lower five percent quantiles, seems to be a natural choice: it ensures that the local smoother $\hat{\varrho}$ has enough 'observations' available to estimate the link function reasonably well.

The remainder of the paper is organized as follows. In Sect. 11.2 we describe our main result, a first order uniform stochastic expansion of $\hat{\mathbb{F}}_n$, and discuss in detail the assumptions used. An application of the main result is described in Sect. 11.3 by constructing asymptotically distribution free tests for fitting an error distribution in model (11.1). Some properties of local quadratic smoothers are given in Sect. 11.4. In Sect. 11.5 we generalize results from MSW (2007) for nonparametric regression with quasi-uniform covariates to the case when quasi-uniformity cannot be assumed. Sects. 11.4 and 11.5 play a major role in the proof of our main result given in Sect. 11.6. Our approach is to regard the single-index model as a nonparametric regression model with estimated covariates $\hat{S}_j$. The randomness caused by the estimators $\hat{\theta}$ is handled using discretization and contiguity arguments, which are standard techniques in the construction of efficient estimators in semiparametric models.

11.2 Main Result

We begin by describing the local quadratic smoother. The value $\hat{\varrho}(s)$ of this estimator at $s \in \mathbb{R}$ equals the first component $\hat{\beta}_0$ of the minimizer $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ of

$$\frac{1}{nb_n}\sum_{j=1}^{n}\Big(Y_j - \beta_0 - \beta_1\frac{\hat{S}_j - s}{b_n} - \beta_2\Big(\frac{\hat{S}_j - s}{b_n}\Big)^2\Big)^2 K\Big(\frac{\hat{S}_j - s}{b_n}\Big),$$

where K is a symmetric density with compact support $[-1, 1]$ and b_n is a bandwidth, i.e., b_n is a sequence of positive numbers that converges to zero.

We prove our main result, a uniform stochastic expansion of $\hat{\mathbb{F}}_n$, under the following conditions. Let $I = [a, b]$ be a compact interval of $\mathbb{R}$.

(R1) The regression function ϱ is twice continuously differentiable and satisfies

$$E[|\varrho(\theta^\top X) - \varrho(\theta_0^\top X) - (\theta - \theta_0)^\top X\varrho'(\theta_0{}^\top X)|^2] = o(\|\theta - \theta_0\|^2),$$

as $\|\theta - \theta_0\| \to 0$.

(R2) The $p \times p$ matrix

$$M = E[(\varrho'(S))^2(X - E[X|S])(X - E[X|S])^\top]$$

has rank $p - 1$.

(T) The estimator $\hat{\theta}$ satisfies $n^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$ and is discretized.

(I) There are interior points $l_0 < u_0$ of I and functions $\bar{\phi}_{n,l}$ and $\bar{\phi}_{n,u}$ such that for all θ_n in Θ with $n^{1/2}(\theta_n - \theta_0)$ bounded we have

$$\phi_{n,l}(S_1, \ldots, S_n, \theta_n) = \bar{\phi}_{n,l}(\theta_n) + o_p(n^{-1/4}) = l_0 + o_p(1),$$
$$\phi_{n,u}(S_1, \ldots, S_n, \theta_n) = \bar{\phi}_{n,u}(\theta_n) + o_p(n^{-1/4}) = u_0 + o_p(1),$$

with $\phi_{n,l}$ and $\phi_{n,u}$ as in (11.2) and $S_j = \theta_0^\top X_j$.

(G1) The density g_{θ_0} is bounded and also bounded away from zero on I.

(G2) The map $\theta \mapsto \sqrt{g_\theta}$ is differentiable at θ_0 in L_2, i.e., there is a measurable function $\dot{g}_{\theta_0}$ from $\mathbb{R}$ into $\theta_0^\perp = \{v \in \mathbb{R}^p : v^\top\theta_0 = 0\}$ such that $\|\dot{g}_{\theta_0}\|$ is square-integrable and

$$\int\Big(\sqrt{g_\theta(s)} - \sqrt{g_{\theta_0}(s)} - (\theta - \theta_0)^\top\dot{g}_{\theta_0}(s)\Big)^2 ds = o(\|\theta - \theta_0\|^2)$$

holds as $\|\theta - \theta_0\| \to 0$.

(F1) The error variable has a finite third moment.

(F2) The error density f has finite Fisher information for location.

We shall now discuss these assumption. The first part of (R1) is used to derive appropriate properties of the local quadratic smoother of ϱ. The bias of this estimator is of order $o(b_n^2)$ which needs to be of order $o(n^{-1/2})$. The choice $b_n \sim n^{-1/4}$ used in Theorem 1 is the largest bandwidth satisfying this requirement. Larger bandwidth are allowed under additional smoothness assumptions on ϱ. For example, if the second derivative of ϱ is Hölder with exponent α, then we can take larger b_n subject to the constraint $b_n^{2+\alpha} = o(n^{-1/2})$. In particular, for $\alpha > 1/2$, the familiar choice $b_n \sim n^{-2/5}$ works. Instead of a local quadratic smoother, we could have worked with a local linear smoother. The bias of this estimator is of order $O(b_n^2)$. This would require a smaller bandwidth such as $b_n \sim (n \log n)^{-1/4}$ to guarantee that the bias is of order $o(n^{-1/2})$. A local linear smoother with this choice of bandwidth was used in MSW (2007).

The matrix M in condition (R2) cannot have full rank p as

$$\theta_0^\top M \theta_0 = E[(\varrho(S))^2(\theta_0^\top(X - E[X|S]))^2] = E[(\varrho(S))^2(S - E[S|S])^2] = 0.$$

Condition (R2) guarantees that $v^\top M v > 0$ for every unit vector v orthogonal to θ_0. This is needed to guarantee the existence of a root-n consistent estimator of θ_0, as required in condition (T).

The set $\theta_0^\perp$ appearing in (G2) is the tangent space of Θ at θ_0. The requirement in (G2) that $\dot{g}_{\theta_0}$ takes values in $\theta_0^\perp$ ensures that the derivative $\dot{g}_{\theta_0}$ is uniquely determined (up to almost everywhere equivalence). Without this assumption, the differentiability requirement would also hold with $\dot{g}_{\theta_0}$ replaced by $\dot{g}_{\theta_0} + h\theta_0$ for each square-integrable h. This follows from the fact that $(\theta - \theta_0)^\top \theta_0$ equals $-\|\theta - \theta_0\|^2/2$ for all θ in Θ. On the other hand, it suffices to verify the differentiability condition for some $\dot{g}_{\theta_0}$ that is not $\theta_0^\perp$-valued, because it then holds with $\dot{g}_{\theta_0}$ replaced by $(I_p - \theta_0\theta_0^\top)\dot{g}_{\theta_0}$, where I_p is the $p \times p$ identity matrix, and this replacement is $\theta_0^\perp$-valued in view of $\theta_0^\top(I_p - \theta_0\theta_0^\top) = 0$.

By the same token, we can replace in (R1) the derivative $X\varrho'(\theta_0^\top X)$ by the $\theta_0^\perp$-valued derivative $(I_p - \theta_0\theta_0^\top)X\varrho'(\theta_0^\top X)$. This and (F2) show that the score function for θ_0 is given by

$$\ell(\varepsilon)\varrho'(S)(I_p - \theta_0\theta_0^\top)X,$$

with $\ell = -f'/f$, the score function for location. The tangent space $\mathscr{T}$ for the nuisance parameter (ϱ, F, G), with G the distribution of X, consists of the function

$$\ell(\varepsilon)a(S) + c(\varepsilon) + b(X)$$

with $E[a^2(S)]$ finite, $E[b(X)] = 0$ and $E[b^2(X)]$ finite, $E[c(\varepsilon)] = E[\varepsilon c(\varepsilon)] = 0$ and $E[c^2(\varepsilon)]$ finite. The projection of the score function onto $\mathscr{T}^p$ is given by

$$\ell(\varepsilon)\varrho'(S)(I_p - \theta_0\theta_0^\top)E[X|S].$$

Thus the efficient score for estimating θ_0 is

$$\ell(\varepsilon)\varrho'(S)(I_p - \theta_0\theta_0^\top)(X - E[X|S])$$

and the efficient information matrix is

$$J_* = E[\ell^2(\varepsilon)](I_p - \theta_0\theta_0^\top)M(I_p - \theta_0\theta_0^\top) = E[\ell^2(\varepsilon)]M.$$

While the information matrix is not invertible, the map ϕ from ${\theta_0}^\perp$ to ${\theta_0}^\perp$ defined by it, i.e. $\phi(v) = J_* v$, $v \in {\theta_0}^\perp$, is invertible. Finally, the efficient influence function for estimation $F(t)$ is given by

$$\mathbf{1}[\varepsilon \le t] - F(t) + f(t)\varepsilon.$$

This can be deduced from the results in Müller and Schick (2017) and the form of the present tangent space.

For the construction of root-n consistent estimators of θ_0 we refer to Carroll et al. (1997); Wang et al. (2010), and Xia and Härdle (2006), who develop $n^{1/2}$-consistent estimators of the underlying Euclidean parameters in a class of partially linear single-index models. Cui et al. (2011) use a method of estimating functions to develop estimators of θ_0 that satisfy condition (T) for a large class of single-index model. Their estimator of θ_0 is found to have smaller or equal limiting variance than that of Carroll et al. (1997). See also the correction note by Li et al. (2011) pertaining to the reference Wang et al. (2010). The method of Hall and Yao (2005) provides yet another approach to obtain a root-n consistent estimator.

Condition (T) also requires that the root-n consistent estimator of θ_0 is discretized. Such an estimator can be obtained by discretizing any preliminary root-n consistent estimator $\hat\theta$ on grids with mesh width $n^{-1/2}$, e.g., by replacing it by the closest point on the grid, so the change is at most $n^{-1/2}$ and consistency is preserved. This trick simplifies the proofs since we can replace $\hat\theta_n$ by a *nonrandom* sequence $\theta_n = \theta_0 + O(n^{-1/2})$, see, e.g., Le Cam (1986) or van der Vaart (1998).

We use condition (G2) to establish that the distributions of $(\theta_n^\top X_1, \ldots, \theta_n^\top X_n)$ and $(\theta_0^\top X_1, \ldots, \theta_0^\top X_n)$ are mutually contiguous whenever $\theta_n = \theta_0 + O(n^{-1/2})$. This implies that (I) holds with each $S_j = \theta_0^\top X_j$ replaced by $\theta_n^\top X_j$. This and (T) then allow us to conclude that $\hat l$ is a consistent estimator of l_0, more precisely, we have

$$\hat l = \bar\phi_{n,l}(\hat\theta) + o_p(n^{-1/4}) = l_0 + o_p(1).$$

An analogous statement holds for $\hat u$. Similar arguments yield

$$\frac{N_n}{n} = \frac{1}{n}\sum_{j=1}^n \hat\delta_j = P(l_0 \le \theta_0^\top X \le u_0) + o_p(1). \tag{11.4}$$

Let $\hat q_\alpha$ denote the α-th sample quantile constructed from the estimated indices $\hat S_1, \ldots, \hat S_n$. Recall that the sample quantile based on independent observations from

a density is a root-n consistent estimator of the quantile whenever the density is positive and continuous at this quantile. Thus condition (I) is met by

$$\hat{I} = [\hat{l}, \hat{u}] = [\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}],$$

with $0 < \alpha_1 < \alpha_2 < 1$, if g_{θ_0} is continuous and positive on an open interval containing the α_1 and α_2-quantiles of g_{θ_0}. In particular, condition (I) holds for any such α_1 and α_2 if g_{θ_0} is continuous and the set $\{g_{\theta_0} > 0\}$ is an interval.

The moment assumption (F1) is used to derive properties of the local quadratic smoothers. The assumption (F2) together with (R1) is used to obtain contiguity. It also guarantees that the density f is Hölder with exponent 1/2, which meets one of the requirements in MSW (2007), namely, the density f to be Hölder with exponent greater than $1/3$.

We now state our main result, the uniform stochastic expansion of the estimator $\hat{\mathbb{F}}_n$ introduced in (11.3). This expansion is similar to the expansions obtained in MSW (2007, 2009a) in semiparametric and nonparametric regression models. The difference is the presence of weights $w(\theta_0^\top X_j)$, where

$$w(s) = \frac{\mathbf{1}[l_0 \le s \le u_0]}{P(l_0 \le \theta_0^\top X \le u_0)}, \quad s \in \mathbb{R}.$$

Theorem 1 *Suppose the model* (11.1) *and the conditions (R1), (R2), (T), (I), (G1), (G2), (F1) and (F2) hold. In addition, assume that the kernel K has a Hölder continuous second derivative, and the bandwidth b_n satisfies $b_n \sim n^{-1/4}$. Then we have the uniform stochastic expansion*

$$\sup_{t \in \mathbb{R}} \left| \hat{\mathbb{F}}_n(t) - F(t) - \mathbb{W}_n(t) \right| = o_p(n^{-1/2}) \tag{11.5}$$

with

$$\mathbb{W}_n(t) = \frac{1}{n} \sum_{j=1}^{n} w(\theta_0^\top X_j) \big[\mathbf{1}[\varepsilon_j \le t] - F(t) + f(t)\varepsilon_j \big], \quad t \in \mathbb{R}.$$

Remark 2 The above result shows that the influence function of the estimator $\hat{\mathbb{F}}_n(t)$ is

$$\phi_t(Y, X) = w(\theta_0^\top X) \big[\mathbf{1}[\varepsilon \le t] - F(t) + f(t)\varepsilon \big],$$

which is the efficient influence function for estimating $F(t)$ multiplied by $w(\theta_0^\top X)$. The asymptotic variance of our estimator thus equals the efficient variance multiplied by $E[w^2(\theta_0^\top X)]$. This factor equals $1/p_0$ with

$$p_0 = P(l_0 \le \theta_0^\top X \le u_0).$$

Thus our estimator is nearly efficient if p_0 is close to one.

11.3 An Application

We shall now discuss an application of (11.5) for deriving an asymptotically distribution free (ADF) test for fitting a known error distribution in the model (11.1). For this we introduce the process

$$\mathbb{Z}_n(t) = \frac{1}{n}\sum_{j=1}^{n}\big[\mathbf{1}[\varepsilon_j \le t] - F(t) + f(t)\varepsilon_j\big], \quad t \in \mathbb{R}.$$

Note that $n\mathrm{Cov}(\mathbb{Z}_n(s), \mathbb{Z}_n(t)) = C(s,t)$ and $n\mathrm{Cov}(\mathbb{W}_n(s), \mathbb{W}_n(t)) = (1/p_0)C(s,t)$ with

$$C(s,t) = \mathrm{Cov}(\mathbf{1}[\varepsilon \le s] + f(s)\varepsilon, \mathbf{1}[\varepsilon \le t] + f(t)\varepsilon), \quad s,t \in \mathbb{R}.$$

Recall, say from Koul (2002), that $n^{1/2}\mathbb{Z}_n$ converges weakly to a continuous Gaussian process $\mathbb{Z}$ with mean zero and covariance function C. Thus Theorem 1 implies

$$n^{1/2}(\hat{\mathbb{F}}_n - F) \to_D p_0^{-1/2}\mathbb{Z},$$

where $\to_D$ denotes weak convergence in the Skorokhod space $D[-\infty, \infty]$ and uniform metric. By (11.4), $\hat{p}_n = N_n/n$ is a consistent estimator of p_0 and we conclude

$$N_n^{1/2}(\hat{\mathbb{F}}_n - F) \to_D \mathbb{Z}. \tag{11.6}$$

An analog of Theorem 1 is obtained in MSW (2009a) for the ordinary nonparametric residual empirical process $\hat{F}_n$ in a class of nonparametric regression models. They established, under some conditions on the regression function and F, the expansion

$$n^{1/2}\sup_{t\in\mathbb{R}}|\hat{F}_n(t) - F(t) - \mathbb{Z}_n(t)| = o_p(1). \tag{11.7}$$

Let F_0 be a known distribution function having zero mean, a finite third moment and finite Fisher information for location. Consider the problem of testing $H_0 : F = F_0$ versus the alternative that H_0 is not true. In the context of nonparametric regression models, Khmaladze and Koul (2009) (KK) used the expansion (11.7) to show that under H_0 a certain transform of $\hat{F}_n$ converges weakly in $D[-\infty, \infty]$ and uniform metric to $B \circ F_0$, where B is standard Brownian motion on $[0, \infty)$. The results (11.5) and (11.6) used with $F = F_0$ enable one to conclude that the analog of this transform will also converge weakly, under H_0, to $B \circ F_0$. For the sake of completeness we describe this transform here.

Let f_0 be density of F_0 and f_0' be its a.e. derivative. Define

$$h(x) = \big(1, -f_0'(x)/f_0(x)\big)^\top, \quad \sigma^2(x) = \int_x^\infty \Big(\frac{f_0'(y)}{f_0(y)}\Big)^2 dF_0(y),$$

$$\Gamma_{F_0(x)} = \int_x^\infty h(x)h^\top(x)dF_0(x) = \begin{pmatrix} 1 - F_0(x) & f_0(x) \\ f_0(x) & \sigma^2(x) \end{pmatrix}, \quad x \in \mathbb{R}.$$

Let

$$K_n(t) = \int_{-\infty}^{t} h^\top(s)\Gamma_{F_0(s)}^{-1} \int_s^\infty h(z)\, d\hat{\mathbb{F}}_n(z)\, dF_0(s), \quad t \in \mathbb{R}.$$

The transformed process is

$$U_n(t) = n^{1/2}\big(\hat{\mathbb{F}}_n(t) - K_n(t)\big), \quad t \in \mathbb{R}.$$

See the discussion in KK for the existence of this transform. Arguing as in KK, one can show with the help of (11.5) and (11.6) that under appropriate conditions $\hat{p}_n^{1/2} U_n \to_D B \circ F_0$. As a consequence, under H_0,

$$D_n = \sup_{t\in\mathbb{R}} |\hat{p}_n^{1/2} U_n(t)| = \sup_{t\in\mathbb{R}} |N_n^{1/2}(\hat{\mathbb{F}}_n(t) - K_n(t))| \to_D \sup_{0\le s\le 1} |B(s)|,$$

and the test based on D_n is ADF for testing H_0 in the single-index model (11.1). Perhaps it is worth pointing out that this test differs from its analog used for fitting an error distribution in the one sample location model only in that the scale factor $n^{1/2}$ is replaced by $N_n^{1/2}$ and the ordinary residual empirical distribution function of the one sample location model is replaced by $\hat{\mathbb{F}}_n$.

11.4 Properties of Local Quadratic Smoothers

In this section we describe some large sample properties of a local quadratic smoother of the regression function r on a closed interval $I = [a, b]$ with $a < b$ for the nonparametric regression model

$$Y = r(Z) + \varepsilon,$$

where ε and Z are independent random variables, ε has mean zero and a finite third moment, Z has a *bounded* density g, and r is twice continuously differentiable. Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ denote independent copies of the pair (Y, Z) from the above regression model.

The local quadratic smoother $\hat{r}$ associated with a kernel K and a bandwidth c_n is defined as follows. The value $\hat{r}(z)$ of this estimator at z is given by the first component of the minimizer $\hat{\beta}(z) = (\hat{\beta}_0(z), \hat{\beta}_1(z), \hat{\beta}_2(z))^\top$ of

$$L(\beta) = \frac{1}{nc_n} \sum_{j=1}^{n} \Big(Y_j - \beta_0 - \beta_1 \frac{Z_j - z}{c_n} - \beta_2\Big(\frac{Z_j - z}{c_n}\Big)^2\Big)^2 K\Big(\frac{Z_j - z}{c_n}\Big).$$

We assume throughout that K is a symmetric density with support $[-1, 1]$ and make additional assumptions as needed.

In what follows we use the following notation. For a $k \times m$ matrix A, we let $\|A\|$ denote its Euclidean norm

$$\|A\| = \left(\sum_{i=1}^{k} \sum_{j=1}^{m} A_{ij}^2 \right)^{1/2}.$$

For a function M from the interval I to the set of $k \times m$ matrices we set

$$\|M\|_* = \sup_{x \in I} \|M(x)\|.$$

If this function is differentiable with derivative M', then we set

$$\|M\|_{1,\gamma} = \|M\|_* + \|M'\|_* + \sup_{x,y \in I, x<y} \frac{\|M'(x) - M'(y)\|}{|x-y|^\gamma}, \qquad 0 < \gamma < 1.$$

These norms apply to vectors ($m = 1$) and scalars ($k = m = 1$).

Set $\psi(x) = (1, x, x^2)^\top$ for $x \in \mathbb{R}$. Then the above criterion function becomes

$$L(\beta) = \frac{1}{nc_n} \sum_{j=1}^{n} \Big(Y_j - \beta^\top \psi\Big(\frac{Z_j - z}{c_n}\Big)\Big)^2 K\Big(\frac{Z_j - z}{c_n}\Big).$$

Routine calculations show that the minimizer $\hat{\beta}(z)$ of $L(\beta)$ solves the normal equations

$$\hat{W}(z)\hat{\beta}(z) = \hat{A}(z) + \hat{B}(z),$$

where

$$\hat{W}(z) = \frac{1}{nc_n} \sum_{j=1}^{n} \psi\Big(\frac{Z_j - z}{c_n}\Big)\psi^\top\Big(\frac{Z_j - z}{c_n}\Big) K\Big(\frac{Z_j - z}{c_n}\Big),$$

$$\hat{A}(z) = \frac{1}{nc_n} \sum_{j=1}^{n} \varepsilon_j \psi\Big(\frac{Z_j - z}{c_n}\Big) K\Big(\frac{Z_j - z}{c_n}\Big),$$

$$\hat{B}(z) = \frac{1}{nc_n} \sum_{j=1}^{n} r(Z_j) \psi\Big(\frac{Z_j - z}{c_n}\Big) K\Big(\frac{Z_j - z}{c_n}\Big).$$

Since K has support $[-1, 1]$ and r'' is uniformly continuous on compact sets, a Taylor expansion yields

$$\|\hat{B} - \hat{W}\dot{r}_{c_n}\|_* = \sup_{z \in I} |\hat{B}(z) - \hat{W}(z)\dot{r}_{c_n}(z)| = o(c_n^2)$$

with $\dot{r}_{c_n}(z) = (r'(z), c_n r'(z), c_n^2 r''(z)/2)^\top$. Direct calculations show that

$$\bar{W}(z) = E[\hat{W}(z)] = \int \psi(u)\psi^\top(u)K(u)g(z + c_n u)\,du.$$

In order to prove Theorem 2 below, which lists some important properties of the smoother, we need the following two lemmas. The first lemma is an immediate consequence of the definition of $\bar{W}$ and the fact that the matrix

$$\int_A \psi(u)\psi^\top(u)K(u)\,du$$

is positive definite for any subinterval A of $[-1, 1]$ of positive length. Recall that g is bounded.

Lemma 1 *Suppose g is also bounded away from zero on I. Then there is an α, $0 < \alpha < 1$, such that the eigenvalues of $\bar{W}(z)$ fall into the interval $[\alpha, 1/\alpha]$ for all z in I and all c_n satisfying $c_n \le l/2$, where l is the length of the interval I.*

The next lemma is a consequence of Corollary 4.2 in MSW (2007) with their δ equal to zero. Note that Z has a bounded density as required there. We also use the fact that ε has finite third moment.

Lemma 2 *Suppose w is an integrable and Hölder continuous function and $\log n/(nc_n)$ is bounded. Then the rate*

$$\sup_{z\in I}\Big|\frac{1}{nc_n}\sum_{j=1}^{n} w\Big(\frac{Z_j - z}{c_n}\Big) - E[w\Big(\frac{Z_j - z}{c_n}\Big)]\Big| = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big)$$

holds. Moreover, if $E|\varepsilon|^3 < \infty$ and $\log n/(c_n n^{1/3})$ is bounded, then the rate

$$\sup_{z\in I}\Big|\frac{1}{nc_n}\sum_{j=1}^{n} \varepsilon_j w\Big(\frac{Z_j - z}{c_n}\Big)\Big| = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big)$$

holds.

In view of Lemma 2, from now on we assume that $\log n/(c_n n^{1/3})$ is bounded. It then follows from Lemma 1 that

$$\|\bar{W}\|_* = O(1) \quad \text{and} \quad \|\bar{W}^{-1}\|_* = O(1).$$

Furthermore, Lemma 2 applied to the entries of the matrices implies that

$$\|\hat{W} - \bar{W}\|_* = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big)$$

and

$$\|\hat{A}\|_* = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big),$$

provided K is Hölder. It follows that the matrices $\hat{W}(z)$, $z \in I$, are invertible on the event $\{\|\hat{W} - \bar{W}\|_* < \alpha\}$, whose probability converges to one. On this event we have

$$\|\hat{W}\|_* = O(1), \qquad \|\hat{W}^{-1}\|_* = O(1),$$

and

$$\|\hat{W}^{-1} - \bar{W}^{-1}\|_* = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big).$$

Moreover, on this event we have the identity

$$\begin{aligned}&\hat{\beta}(z) - \dot{r}_{c_n}(z) - \bar{W}(z)^{-1}\hat{A}(z)\\ &\quad = (\hat{W}(z)^{-1} - \bar{W}(z)^{-1})\hat{A}(z) + \hat{W}(z)^{-1}(\hat{B}(z) - \hat{W}(z)\dot{r}_{c_n}(z)), \quad z \in I.\end{aligned}$$

Using the above properties we obtain the following result. Recall that we assumed that g is bounded, that ε has a finite third moment and that K is a symmetric density with support $[-1, 1]$.

Proposition 1 *Suppose g is also bounded away from zero on I and the kernel K is also Hölder. Then the uniform stochastic expansion*

$$\|\hat{\beta} - \dot{r}_{c_n} - \bar{W}^{-1}\hat{A}\|_* = O_p\Big(\frac{\log n}{nc}\Big) + o_p(c_n^2)$$

holds.

Thus, under the assumptions of the proposition and $c_n \sim n^{-1/4}$, we have the expansion

$$\sup_{z\in I} |\hat{r}(z) - r(z) - [1, 0, 0]\bar{W}(z)^{-1}\hat{A}(z)| = o_p(n^{-1/2}).$$

Next, we investigate the magnitude of the process

$$\hat{C}(z) = \bar{W}^{-1}(z)\hat{A}(z), \quad z \in I.$$

Proposition 2 *Suppose g is bounded away from zero on I and K has a Hölder continuous second derivative. Then the map $z \mapsto \hat{C}(z)$ is twice differentiable and the following rates hold.*

$$\begin{aligned}\|\hat{C}\|_* &= O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big),\\ \|c_n\hat{C}'\|_* &= O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big),\\ \|c_n^2\hat{C}''\|_* &= O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big).\end{aligned}$$

Proof Note that

$$\|\hat{C}\|_* \le \|\bar{W}^{-1}\|_* \|\hat{A}\|_* = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big).$$

By the properties of the kernel K, the function $z \mapsto \hat{C}(z)$ is twice continuously differentiable with the first derivative given by

$$\hat{C}'(z) = \bar{W}^{-1}(z)\hat{A}'(z) - \bar{W}^{-1}(z)\bar{W}'(z)\bar{W}^{-1}(z)\hat{A}(z)$$

and the second derivative given by

$$\begin{aligned}\hat{C}''(z) &= \bar{W}^{-1}(z)\hat{A}''(z) - 2\bar{W}^{-1}(z)\bar{W}'(z)\bar{W}^{-1}(z)\hat{A}'(z) \\ &\quad + 2\bar{W}^{-1}(z)\bar{W}'(z)\bar{W}^{-1}(z)\bar{W}'(z)\bar{W}^{-1}(z)\hat{A}(z) - \bar{W}^{-1}(z)\bar{W}''(z)\bar{W}^{-1}(z)\hat{A}(z).\end{aligned}$$

We write the matrix $[c_n\hat{A}'(z), c_n^2\hat{A}''(z)]$ as

$$\frac{1}{nc_n}\sum_{j=1}^{n} \varepsilon_j \Psi\Big(\frac{Z_j - z}{c_n}\Big)$$

with $\Psi = [(K\psi)', (K\psi)'']$. By the assumption on K, the entries of Ψ are integrable and Hölder. Thus we obtain from Lemma 2 that

$$\|c_n\hat{A}'\|_* + \|c_n^2\hat{A}''\|_* = O_p\Big(\Big(\frac{\log n}{nc_n}\Big)^{1/2}\Big).$$

Rewrite the matrix $[c_n\bar{W}'(z), c_n^2\bar{W}''(z)]$ as

$$\int g(u + c_n x)[V'(x), V''(x)]\,dx$$

with $V = K\psi\psi^\top$. From this we conclude that $\|c_n\bar{W}'\|_* + \|c_n^2\bar{W}''\|_* = O(1)$. Combining the above we obtain

$$\|c_n\hat{C}'\|_* \le \|\bar{W}^{-1}\|_* \|c_n\hat{A}'\|_* + \|\bar{W}^{-1}\|_*^2 \|c_n\bar{W}'\|_* \|\hat{A}\|_*$$

and

$$\begin{aligned}\|c_n^2\hat{C}''\|_* &\le \|\bar{W}^{-1}\|_* \|c_n^2\hat{A}''\|_* + 2\|\bar{W}^{-1}\|_*^2 \|c_n\bar{W}'\|_* \|c_n\hat{A}'\|_* \\ &\quad + 2\|\bar{W}^{-1}\|_*^3 \|c_n\bar{W}'\|_*^2 \|\hat{A}\|_* + \|\bar{W}^{-1}\|_*^2 \|c_n^2\bar{W}''\|_* \|\hat{A}\|_*.\end{aligned}$$

This immediately yields the desired rates. □

We use Proposition 2 to obtain rates on the Hölder norms $\|\hat{C}\|_{1,\gamma}$, $0 < \gamma < 1$. Since we can bound $\|\hat{C}'(s) - \hat{C}'(t)\||s - t|^{-\gamma}$ by $\|\hat{C}''\|_* c_n^{1-\gamma}$ for $0 < |s - t| \le c_n$ and by $2\|\hat{C}'\|_* c_n^{-\gamma}$ for $|t - s| > c_n$, we have the following result.

Proposition 3 *Suppose the assumptions of Proposition 2 hold and* $0 < \gamma < 1$. *Then the rate*

$$\|\hat{C}\|_{1,\gamma} = O_p\Big(\Big(\frac{\log n}{nc_n^{3+2\gamma}}\Big)^{1/2}\Big)$$

holds. In particular, for $c_n \sim n^{-1/4}$ *and* $\gamma < 1/2$, *one has*

$$\|\hat{C}\|_{1,\gamma} = o_p(1).$$

The next result summarizes properties of the local quadratic smoother $\hat{r}$ if the bandwidth is proportional to $n^{-1/4}$.

Theorem 2 *Suppose g is bounded away from zero on I, K has a Hölder continuous second derivative and the bandwidth satisfies* $c_n \sim n^{-1/4}$. *Then the following hold with* $\hat{c} = [1, 0, 0]\hat{C}$ *the first coordinate of* $\hat{C}$.

$$\sup_{z\in I} |\hat{r}(z) - r(z) - \hat{c}(z)| = o_p(n^{-1/2}), \tag{11.8}$$

$$\int_I \hat{c}^2(z)g(z)\,dz = O_p(n^{-3/4}), \tag{11.9}$$

and, for $0 < \gamma < 1/2$,

$$\sup_{z\in I} |\hat{c}(z)| + \sup_{z\in I} |\hat{c}'(z)| + \sup_{s,t\in I, s<t} \frac{|\hat{c}'(t) - \hat{c}'(s)|}{|t-s|^\gamma} = o_p(1). \tag{11.10}$$

Moreover, for any square-integrable functions $v, v_1, v_2, \ldots$ *satisfying*

$$\int_I (v_n(z) - v(z))^2\,dz = o(1),$$

we have the expansion

$$\int_I \hat{c}(z)v_n(z)g(z)\,dz = \frac{1}{n}\sum_{j=1}^n \varepsilon_j \mathbf{1}[Z_j \in I]v(Z_j) + o_p(n^{-1/2}). \tag{11.11}$$

Proof Claim (11.8) is a consequence of Proposition 1 and (11.10) of Proposition 3. Statement (11.9) follows from the bounds $\|\bar{W}^{-1}\|_* = O(1)$ and

$$nc_n E[\|\hat{A}(z)\|^2] \le E[\varepsilon^2]E\Big[\frac{3}{c_n}K^2\Big(\frac{Z-z}{c_n}\Big)\Big] \le 3E[\varepsilon^2]\int g(z + c_n u)K^2(u)\,du$$

and the boundedness of g. Here we used $\|\psi K\|^2 \le 3K^2$.

In order to prove (11.11) we set

$$\tilde{v}_n(z) = \mathbf{1}[z \in I] v_n(z) g(z) [1, 0, 0] \bar{W}^{-1}(z),$$

and

$$V_n(z) = \int \tilde{v}_n(z - c_n u) \psi(u) K(u)\, du \quad \text{and} \quad V(z) = \mathbf{1}[z \in I] v(z).$$

Then rewrite the left-hand side of (11.11) as

$$\frac{1}{n c_n} \sum_{j=1}^{n} \varepsilon_j \int \tilde{v}_n(z) \psi\Big(\frac{Z_j - z}{c_n}\Big) K\Big(\frac{Z_j - z}{c_n}\Big)\, dz = \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j V_n(Z_j).$$

Let $\Psi = \int \psi(u) \psi^\top(u) K(u)\, du$. Since the density g is bounded, it is square-integrable. This and the translation continuity in L_2 yield the convergence

$$\int \Big\| \int g(z + c_n u) \psi(u) \psi^\top(u) K(u)\, du - g(z) \Psi \Big\|^2 dz \to 0.$$

Since g is bounded away from zero on I, we conclude from this that the map $z \mapsto \mathbf{1}[z \in I] \bar{W}^{-1}(z)$ converges to the map $z \mapsto \mathbf{1}[z \in I](g(z)\Psi)^{-1}$ in Lebesgue measure. An application of Lebesgue's dominated convergence theorem now yields that $\tilde{v}_n$ converges in L_2 to $\tilde{v}$, where

$$\tilde{v}(z) = \mathbf{1}[z \in I] v(z) [1, 0, 0] \Psi^{-1}.$$

Using this, the identity $\tilde{v}(z) \int \psi(u) K(u)\, du = \tilde{v}(z) \Psi [1, 0, 0]^\top = \mathbf{1}[z \in I] v(z) = V(z)$, and the translation continuity in L_2, we derive

$$\begin{aligned}
\Delta_n &= \int (V_n(z) - V(z))^2\, dz \\
&= \int \Big| \int \tilde{v}_n(z - c_n u) \psi(u) K(u)\, du - \mathbf{1}[z \in I] v(z) \Big|^2 dz \\
&\le 2 \int \|\tilde{v}_n(z) - \tilde{v}(z)\|^2\, dz \int \|\psi(u) K(u)\|^2\, du \\
&\quad + 2 \int \Big| \int (\tilde{v}(z - c_n u) - \tilde{v}(z)) \psi(u) K(u)\, du \Big|^2 dz = o(1).
\end{aligned}$$

From the above we conclude that n times the second moment of the difference

$$\int_I \hat{c}(z) v_n(z) g(z)\, dz - \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j v(Z_j) = \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j (V_n(Z_j) - V(Z_j))$$

equals $E[\varepsilon^2] E[(V_n(Z) - V(Z))^2]$, which is bounded by a constant times Δ_n. This implies the desired (11.11). □

Remark 3 Let $\hat{v}$ be an estimator of some square-integrable function v_0. Suppose there is a sequence of square-integrable functions v_n such that

$$\int_I (\hat{v}(z) - v_n(z))^2 \, dz = o(n^{-1/4}) \quad \text{and} \quad \int_I (v_n(z) - v_0(z))^2 \, dz = o(1).$$

Then under the assumption of the previous theorem the expansion

$$\int_I \hat{c}(z)\hat{v}(z)g(z) \, dz = \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j \mathbf{1}[Z_j \in I] v_0(Z_j) + o_p(n^{-1/2})$$

holds. This follows from (11.9), (11.11), the inequality

$$\Big| \int_I \hat{c}(z)(\hat{v}(z) - v_n(z))g(z) \, dz \Big|^2 \le \int_I \hat{c}^2(z)g(z) \, dz \int_I (\hat{v}(z) - v_n(z))^2 g(z) \, dz$$

and the fact that g is bounded.

11.5 Estimating the Error Distribution in Nonparametric Regression

In this section we modify results from MSW (2007) to the case when the regressor is not quasi-uniform. We begin by extending their Theorems 2.1 and 2.2.

Let ε be a random variable with distribution function F, and let Z be a k-dimensional random vector with distribution Q, independent of ε. Let D be a non-negative function in $L_2(Q)$, and $\mathscr{D}$ be a set of measurable functions a such that $|a| \le D$ and $0 \in \mathscr{D}$. Let $\mathscr{V}$ be a class of measurable functions from $\mathbb{R}^k$ into $[0, 1]$. We now give conditions on the classes $\mathscr{D}$ and $\mathscr{V}$ that imply that the class

$$\mathscr{H} = \{h_{a,v,t} : a \in \mathscr{D}, v \in \mathscr{V}, t \in \mathbb{R}\}$$

is $F \otimes Q$-Donsker, where

$$h_{a,v,t}(\varepsilon, Z) = v(Z)\mathbf{1}[\varepsilon - a(Z) \le t], \quad a \in \mathscr{D}, v \in \mathscr{V}, t \in \mathbb{R}.$$

For this we endow $\mathscr{D}$ with the $L_1(Q)$-pseudo-norm. By an η-bracket for $(\mathscr{D}, L_1(Q))$ we mean a set $[\underline{a}, \overline{a}] = [a \in \mathscr{D} : \underline{a} \le a \le \overline{a}]$ where $\underline{a}$ and $\overline{a}$ belong to $L_1(Q)$ and satisfy $\int |\underline{a} - \overline{a}| \, dQ \le \eta$. Recall that the *bracketing number* $N_{[\,]}(\eta, \mathscr{D}, L_1(Q))$ is the smallest integer m for which there are m η-brackets $[\underline{a}_1, \overline{a}_1], \ldots, [\underline{a}_m, \overline{a}_m]$ which cover $\mathscr{D}$ in the sense that the union of the brackets contains $\mathscr{D}$.

Proposition 4 *Suppose that $\mathscr{V}$ is Q-Donsker. Assume that F has a finite second moment and a bounded density and that the bracketing numbers satisfy*

$$\int_0^1 \sqrt{\log N_{[\,]}(\eta^2, \mathscr{D}, L_1(Q))}\, d\eta < \infty. \tag{11.12}$$

Then $\mathscr{H}$ is $F \otimes Q$-Donsker.

Proof Let ϕ be the projection map from $\mathbb{R} \times \mathbb{R}^k$ into $\mathbb{R}^k$ so that $\phi(\varepsilon, Z) = Z$. Since $\mathscr{V}$ is Q-Donsker, the class $\tilde{\mathscr{V}} = \{v \circ \phi : v \in \mathscr{V}\}$ is $F \otimes Q$-Donsker. Let $\mathscr{H}_1 = \{h_{a,1,t} : a \in \mathscr{D}, t \in \mathbb{R}\}$ with $h_{a,1,t}(\varepsilon, Z) = \mathbf{1}[\varepsilon - a(Z) \le t]$. It follows from Theorem 2.1 of MSW (2007) that the class $\mathscr{H}_1$ is $F \otimes Q$ is Donsker. Since $\tilde{\mathscr{V}}$ and $\mathscr{H}_1$ are uniformly bounded (by 1) $F \otimes Q$-Donsker classes, their pairwise product $\tilde{\mathscr{V}} \cdot \mathscr{H}_1 = \{\tilde{v}h : \tilde{v} \in \tilde{\mathscr{V}}, h \in \mathscr{H}_1\}$ forms a $F \otimes Q$-Donsker class by Example 2.10.8 in van der Vaart and Wellner (1996). This is the desired result as $\mathscr{H}$ equals $\tilde{\mathscr{V}} \cdot \mathscr{H}_1$. □

Now consider a regression model

$$Y = r(Z) + \varepsilon$$

and independent copies (Y_j, Z_j) of (Y, Z). For an estimator $\hat{r}$ of r define the residuals $\hat{\varepsilon}_j = Y_j - \hat{r}(Z_j)$. Define the processes

$$\hat{W}(t, v) = \frac{1}{n}\sum_{j=1}^n v(Z_j)\mathbf{1}[\hat{\varepsilon}_j \le t], \quad W(t, v) = \frac{1}{n}\sum_{j=1}^n v(Z_j)\mathbf{1}[\varepsilon_j \le t], \quad t \in \mathbb{R},\ v \in \mathscr{V}.$$

Proposition 5 *Let $\mathscr{D}$ and $\mathscr{V}$ be as in Proposition 4. Let $\mathscr{V}$ have envelope $\mathbf{1}_I$ for some compact convex set I with nonempty interior. Let F have a finite second moment and a density f that is Hölder with exponent $\xi \in (0, 1]$. Additionally, assume that there is an $\hat{a}$ such that*

$$P(\hat{a} \in \mathscr{D}) \to 1, \tag{11.13}$$

$$\int \mathbf{1}_I |\hat{a}|^{1+\xi}\, dQ = o_p(n^{-1/2}), \tag{11.14}$$

$$\sup_{z \in I} |\hat{r}(z) - r(z) - \hat{a}(z)| = o_p(n^{-1/2}). \tag{11.15}$$

Then the uniform expansion

$$\sup_{t \in \mathbb{R}, v \in \mathscr{V}} \left| \hat{W}(t, v) - W(t, v) - f(t) \int \hat{a}\, v\, dQ \right| = o_p(n^{-1/2})$$

holds.

Proof Without loss of generality we may assume $\hat{a}$ is $\mathscr{D}$-valued; otherwise replace $\hat{a}$ by $\hat{a}\mathbf{1}[\hat{a} \in \mathscr{D}]$. Let

$$\tilde{W}(t,v) = \frac{1}{n}\sum_{j=1}^{n} v(Z_j)\mathbf{1}[\varepsilon_j - \hat{a}(Z_j) \le t] \quad \text{and} \quad W_a(t,v) = \int F(t + a(z))v(z)\,dQ(z).$$

Then we can write

$$\hat{W}(t,v) - W(t,v) - f(t)\int \hat{a}\,v\,dQ = T_1(t,v) + T_2(t,v) + T_3(t,v),$$

where

$$\begin{aligned} T_1(t,v) &= \hat{W}(t,v) - \tilde{W}(t,v), \\ T_2(t,v) &= \tilde{W}(t,v) - W_{\hat{a}}(t,v) - W(t,v) + W_0(t,v), \\ T_3(t,v) &= W_{\hat{a}}(t,v) - W_0(t,v) - f(t)\int \hat{a}\,v\,dQ. \end{aligned}$$

Since f is Hölder, say with constant Λ, we obtain that

$$\begin{aligned} |T_3(t,v)| &\le \int \mathbf{1}_I |F(t + \hat{a}(z)) - F(t) - f(t)\hat{a}(z)|\,dQ(z) \\ &\le \Lambda \int \mathbf{1}_I(z)|\hat{a}|^{1+\xi}\,dQ = o_p(n^{-1/2}). \end{aligned}$$

To deal with T_1 and T_2, we introduce the empirical process

$$\begin{aligned} \nu_n(a,v,t) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\big\{v(Z_j)\mathbf{1}[\varepsilon_j - a(Z_j) \le t] - W_a(t,v)\big\} \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\big(h_{a,v,t}(\varepsilon_j, Z_j) - E[h_{a,v,t}(\varepsilon, Z)]\big), \quad a \in \mathscr{D}, v \in \mathscr{V}, t \in \mathbb{R}, \end{aligned}$$

associated with the Donsker class $\mathscr{H}$. Then we have the identity

$$n^{1/2}T_2(t,v) = \nu_n(\hat{a},v,t) - \nu_n(0,v,t)$$

and the bound

$$\begin{aligned} |n^{1/2}T_1(t,v)| &\le n^{1/2}(\tilde{W}(t + R_n, v) - \tilde{W}(t - R_n, v)) \\ &\le |\nu_n(\hat{a}, t + R_n, v) - \nu_n(\hat{a}, t - R_n, v)| \\ &\quad + n^{1/2}(W_{\hat{a}}(t + R_n, v) - W_{\hat{a}}(t - R_n, v)), \end{aligned}$$

where R_n denotes the left-hand side of (11.15). Since f is Hölder, f is bounded and F is Lipschitz with Lipschitz constant $\|f\|_\infty$. Thus we obtain

$$n^{1/2}(W_{\hat{a}}(t + R_n, v) - W_{\hat{a}}(t - R_n, v)) \le 2\|f\|_\infty n^{1/2}R_n = o_p(1). \tag{11.16}$$

Moreover, for $s, t \in \mathbb{R}$ and $a, b \in \mathscr{D}$, we have the bound

$$E[(h_{a,v,s}(\varepsilon, Z) - h_{b,v,t}(\varepsilon, Z))^2] \le E[v^2(Z)|F(s + a(Z)) - F(t + b(Z))|] \le \|f\|_\infty \big(|s - t| + E[|a(Z) - b(Z)|]\big).$$

In view of this and the stochastic equi-continuity of the empirical process, for every $\eta > 0$ there is a $\delta > 0$ such that, with P^* denoting outer measure,

$$\sup_n P^*\Big(\sup_{t\in\mathbb{R}, a\in\mathscr{D}, v\in\mathscr{V}, \int |a|\, dQ<\delta} |\nu_n(a, v, t) - \nu_n(0, v, t)| > \eta \Big) < \eta,$$

$$\sup_n P^*\Big(\sup_{a\in\mathscr{D}, v\in\mathscr{V}, s,t\in\mathbb{R}, |s-t|<\delta} |\nu_n(a, v, s) - \nu_n(a, v, t)| > \eta \Big) < \eta.$$

The first of these statements and (11.14) imply

$$\sup_{t\in\mathbb{R}, v\in\mathscr{V}} |T_2(t, v)| = o_p(n^{-1/2}),$$

while the second, (11.15) and (11.16) imply

$$\sup_{t\in\mathbb{R}, v\in\mathscr{V}} |T_1(t, v)| = o_p(n^{-1/2}).$$

This completes the proof. □

Now fix a v_0 in $\mathscr{V}$ and let $\hat{v}$ denote an estimator of v_0. Suppose this estimator satisfies

$$P(\hat{v} \in \mathscr{V}) \to 1 \tag{11.17}$$

and

$$\int (\hat{v}(z)) - v_0(z))^2\, dQ(z) = o_p(1). \tag{11.18}$$

It follows that

$$\hat{v}_* = \frac{1}{n}\sum_{j=1}^n \hat{v}(Z_j) = \int v_0\, dQ + o_p(1). \tag{11.19}$$

Moreover, under the assumptions of Proposition 5, the uniform expansion

$$\sup_{t\in\mathbb{R}} \Big| \hat{W}(t, \hat{v}) - W(t, \hat{v}) - f(t) \int \hat{a}\, \hat{v}\, dQ \Big| = o_p(n^{-1/2})$$

holds. We write $W(t, \hat{v}) = \hat{v}_* F(t) + U(t, \hat{v})$, where

$$U(t, v) = \frac{1}{n}\sum_{j=1}^n v(Z_j)(\mathbf{1}[\varepsilon_j \le t] - F(t)).$$

Note that the functions $(y, z) \mapsto v(z)(\mathbf{1}[\varepsilon \le t] - F(t))$ with $v \in \mathscr{V}$ and $t \in \mathbb{R}$ form an $F \otimes Q$-Donsker class. Thus we find

$$\sup_{t\in\mathbb{R}} \left| W(t, \hat{v}) - \hat{v}_* F(t) - U(t, v_0) \right| = o_p(n^{-1/2}).$$

Combining the above yields the uniform expansion

$$\sup_{t\in\mathbb{R}} \left| \hat{W}(t, \hat{v}) - \hat{v}_* F(t) - U(t, v_0) - f(t) \int \hat{a}\, \hat{v}\, dQ \right| = o_p(n^{-1/2}).$$

This finding is summarized in the following theorem.

Proposition 6 *Suppose the assumptions of Proposition 5 are met and $\hat{v}$ is an estimator which satisfies (11.17) and (11.18) for some $v_0 \in \mathscr{V}$ with $\bar{v}_0 = \int v_0\, dQ$ positive. Then the uniform expansion*

$$\sup_{t\in\mathbb{R}} \left| \hat{W}(t, \hat{v})/\hat{v}_* - F(t) - U(t, v_0)/\bar{v}_0 - f(t) \int \hat{a}\hat{v}\, dQ/\hat{v}_* \right| = o_p(n^{-1/2})$$

holds with $\hat{v}_*$ as in (11.19).

Now assume that Z has dimension 1 with a density g that is bounded and bounded away on the interval $I = [a, b]$ with $-\infty < a < b < \infty$. We take $\mathscr{D}$ to be the set of all functions h that vanish off I and satisfy

$$\|h\|_{1,1/4} = \sup_{z\in I} |h(z)| + \sup_{z\in I} |h'(z)| + \sup_{a\le s<t\le b} \frac{|h'(s) - h(t)|}{|t - s|^{1/4}} \le 1.$$

Here we have to understand h' as the derivative of the restriction of h to I so that $h'(a)$ is the right-hand derivative of h at a and $h'(b)$ is the left-hand derivative at b. It follows from Theorem 2.7.1 in van der Vaart and Wellner (1996) that the entropy condition (11.12) holds as $\log N_{[\,]}(\eta^2, \mathscr{D}, L_1(Q))$ is bounded by $C(b - a)(1/\eta)^{8/5}$, for some positive constant C. It follows from the results in the previous section that a local quadratic smoother with bandwidth $c_n \sim n^{-1/4}$ and appropriate kernel K satisfies the conditions (11.13)–(11.15) with $\hat{a} = \hat{c}$ and $\xi > 1/3$. Let $\mathscr{V}$ be the set of indicator functions of intervals $[l, u]$ with $a \le l < u \le b$. This is clearly a Donsker class. Now take

$$\hat{v} = \mathbf{1}_{[\hat{l},\hat{u}]}, \quad v_n = \mathbf{1}_{[l_n,u_n]} \quad \text{and} \quad v_0 = \mathbf{1}_{[l_0,u_0]}$$

with $l_0 < u_0$ interior points of I. Note that

$$\int (\mathbf{1}_{[s,t]}(z) - \mathbf{1}_{[l,u]}(z))^2\, dz \le |s - l| + |t - u|.$$

We have the following result for the regression problem with a one-dimensional Z.

Theorem 3 *Suppose Z has a bounded density that is bounded away from zero on the interval $I = [a, b]$, ε has mean zero, a finite third moment and a density f that is Hölder with exponent greater than $1/3$, the kernel K has a Hölder continuous second derivative, the bandwidth satisfies $c_n \sim n^{-1/4}$, the lower endpoints of the above intervals satisfy $\hat{l} = l_n + o_p(n^{-1/4})$ and $l_n \to l_0$ and the upper endpoints satisfy $\hat{u} = u_n + o_p(n^{-1/4})$ and $u_n \to u_0$. Then the estimator*

$$\hat{F}(t) = \frac{1}{\hat{N}} \sum_{j=1}^{n} \mathbf{1}[\hat{l} \le Z_j \le \hat{u}] \mathbf{1}[\hat{\varepsilon}_j \le t], \quad t \in \mathbb{R},$$

with $\hat{N} = \sum_{j=1}^{n} \mathbf{1}[\hat{l} \le Z_j \le \hat{u}]$, satisfies the uniform expansion

$$\sup_{t \in \mathbb{R}} \left| \hat{F}(t) - F(t) - \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{1}[l_0 \le Z_j \le u_0]}{P(l_0 \le Z \le u_0)} \Big[\mathbf{1}[\varepsilon_j \le t] - F(t)) + f(t)\varepsilon_j \Big] \right| = o_p(n^{-1/2}).$$

11.6 Proof of Theorem 1

A key technical tool for proving Theorem 1 will be the use of two contiguity results. For the sake of self-containment, we shall briefly review the notion of contiguity of Le Cam (1960) and give the needed contiguity results that will be used in the proof; see also Le Cam (1986) and Hájek and Šidák (1967).

Let $(\Omega_n, \mathscr{A}_n, \{P_n, Q_n\})$ be a sequence of binary experiments. Then Q_n is *contiguous* to P_n if for every sequence A_n, $A_n \in \mathscr{A}_n$, $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$. We say P_n and Q_n are *mutually contiguous* if Q_n is contiguous to P_n and P_n is contiguous to Q_n.

We now state a sufficient condition for contiguity of product measures. For this we assume that $(\Omega, \mathscr{A}, \mu)$ is a measure space and $\{\Gamma_\theta : \theta \in \Theta\}$ is a family of probability measures dominated by μ. Denote by γ_θ a density of Γ_θ with respect to μ. Suppose there is a measurable function $\dot{\gamma}_{\theta_0}$ from Ω into ${\theta_0}^{\perp}$ such that $\|\dot{\gamma}_{\theta_0}\|$ belongs to $L_2(\mu)$ and

$$\int (\gamma_\theta^{1/2} - \gamma_{\theta_0}^{1/2} - (\theta - \theta_0)^\top \dot{\gamma}_{\theta_0})^2 \, d\mu = o(\|\theta - \theta_0\|^2) \tag{11.20}$$

holds. Then the product measures $\Gamma_{\theta_n}^n$ and $\Gamma_{\theta_0}^n$ are mutually contiguous whenever $n^{1/2}(\theta_n - \theta_0)$ is bounded. See, e.g., van der Vaart (1998).

We shall use this result first with

$$\gamma_\theta(x, y) = \gamma_{1,\theta}(x, y) = f(y - \varrho(\theta^\top x)), \quad x \in \mathbb{R}^p, y \in \mathbb{R},$$

and $\mu = G \otimes \lambda$ where G is the distribution of X and λ is the Lebesgue measure. It follows from (R1), (R2) and (F2) that (11.20) holds with

$$\dot{\gamma}_{\theta_0}(x, y) = \frac{-f'(y - \varrho(\theta_0{}^\top x))}{2f^{1/2}(y - \varrho(\theta_0{}^\top x))} \varrho'(\theta_0{}^\top x)(I_p - \theta_0\theta_0^\top)x.$$

Then we shall apply the result with

$$\gamma_\theta(x, y) = \gamma_{2,\theta}(x, y) = f(y)g_\theta(x), \quad x \in \mathbb{R}, y \in \mathbb{R},$$

and $\mu = \lambda \otimes \lambda$. It follows from (G2) that (11.20) holds with $\dot{\gamma}_{\theta_0}(x, y) = f^{1/2}(y)\dot{g}_{\theta_0}(x)$.

By the properties of $\hat{\theta}$ specified in (T), it suffices to prove the result with $\hat{\theta}$ replaced by non-stochastic sequences θ_n such that $n^{1/2}(\theta_n - \theta_0)$ is bounded. This is a standard argument used in the construction of efficient estimators in semiparametric models, see, e.g., Schick (1986) and references therein.

Now fix such a sequence θ_n and set

$$S_{n,j} = \theta_n^\top X_j, \quad \delta_{n,j} = \mathbf{1}[S_{n,j} \in I(\theta_n)] \quad \text{and} \quad \varepsilon_{n,j} = Y_j - \varrho(S_{n,j})$$

for $j = 1, \ldots, n$. Let $\tilde{\varrho}$ denote the local linear smoother associated with minimizing

$$\frac{1}{nb_n} \sum_{j=1}^n \Big(Y_j - \beta_0 - \beta_1 \frac{S_{n,j} - s}{b_n} - \beta_2 \Big(\frac{S_{n,j} - s}{b_n}\Big)^2\Big)^2 K\Big(\frac{S_{n,j} - s}{b_n}\Big).$$

Moreover, we introduce

$$\tilde{F}(t) = \frac{\sum_{j=1}^n \mathbf{1}[\tilde{l}_n \le S_{nj} \le \tilde{u}_n]\mathbf{1}[Y_j - \tilde{\varrho}(S_{nj}) \le t]}{\sum_{j=1}^n \mathbf{1}[\tilde{l}_n \le S_{nj} \le u_n]},$$

with $\tilde{l}_n = \phi_{n,l}(S_{n,1}, \ldots, S_{n,n}, \theta_n)$ and $\tilde{u}_n = \phi_{n,u}(S_{n,1}, \ldots, S_{n,n}, \theta_n)$ and set

$$\tilde{\mathbb{W}}_n(t) = \frac{1}{n} \sum_{j=1}^n w(S_{n,j})\big[\mathbf{1}[\varepsilon_{n,j} \le t] - F(t)) + f(t)\varepsilon_{n,j}\big], \quad t \in \mathbb{R}.$$

We achieve our goal by verifying the uniform stochastic expansions

$$\sup_{t \in \mathbb{R}} \Big|\tilde{F}(t) - F(t) - \tilde{\mathbb{W}}_n(t)\Big| = o_p(n^{-1/2}) \tag{11.21}$$

and

$$\sup_{t \in \mathbb{R}} |\tilde{\mathbb{W}}_n(t) - \mathbb{W}_n(t)| = o_p(n^{-1/2}). \tag{11.22}$$

To stress dependence on the parameter θ_0 we now write P_θ for the underlying probability measure when $\theta_0 = \theta$ and write $P_{n,\theta}$ for the joint distribution of the data

$X_1, Y_1, \ldots, X_n, Y_n$ under P_θ, for each $\theta \in \Theta$. It follows from the above that the sequences of distributions P_{n,θ_n} and P_{n,θ_0} are mutually contiguous. Thus it suffices to prove (11.21) under the measure P_{θ_n}. Under the measure P_{θ_n}, we have

$$Y_j = \varrho(S_{n,j}) + \varepsilon_{n,j}, \quad j = 1, \ldots, n,$$

and derive that that the left-hand side of (11.21) is a function of the random vectors $(\varepsilon_{n,1}, S_{n,1})^\top, \ldots, (\varepsilon_{n,n}, S_{n,n})^\top$. Under P_{θ_n} these variables are independent with common density γ_{2,θ_n}. By another contiguity argument it thus suffices to prove (11.21) under the assumption that the random vectors $(\varepsilon_{n,1}, S_{n,1})^\top, \ldots, (\varepsilon_{n,n}, S_{n,n})^\top$ are independent with density γ_{2,θ_0}. The desired (11.21) then follows from Theorem 3.

Note that the distribution of the process defined by the first average in (11.22) under the measure P_{θ_n} equals the distribution of the process defined by the second average in (11.22) under P_{θ_0}. Thus, by contiguity, the difference of these two processes is tight under P_{θ_0}. It suffices to prove (11.22) without the supremum, but for all t in $\mathbb{R}$. Fix such a t. We are left to verify

$$\frac{1}{n}\sum_{j=1}^{n} h_{\theta_n}(X_j, Y_j) - \frac{1}{n}\sum_{j=1}^{n} h_{\theta_0}(X_j, Y_j) = o_p(n^{-1/2}) \tag{11.23}$$

with

$$h_\theta(X, Y) = w(\theta^\top X)\Big\{\mathbf{1}[Y - \varrho(\theta^\top X) \le t] - F(t) + f(t)(Y - \varrho(\theta^\top X))\Big\}.$$

Using the translation continuity in L_2, we verify

$$\iint \left|h_\theta(x, y)\sqrt{\gamma_{1,\theta}(x, y)} - h_{\theta_0}(x, y)\sqrt{\gamma_{1,\theta_0}(x, y)}\right|^2 dG(x)dy \to 0$$

as $\theta \to \theta_0$. With $\ell = -f'/f$ the score function for location, we verify

$$\begin{aligned} D_{\theta_0} &= -\iint h_{\theta_0}(x, y)\ell(y - \varrho({\theta_0}^\top x))\varrho'({\theta_0}^\top x)(I_p - \theta_0\theta_0^\top)x\gamma_{1,\theta_0}(x, y)\, dG(x)dy \\ &= -E\big[(\mathbf{1}[\varepsilon \le t] - F(t) + f(t)\varepsilon)\ell(\varepsilon)\big]\, E\big[w(\theta_0^\top X)\varrho'(\theta_0^\top X)(I_p - \theta_0\theta_0^\top)X\big] = 0, \end{aligned}$$

because the first expectation in the product equals $-f(t) - 0 + f(t) = 0$. Since the densities $\gamma_{1,\theta}$ are Hellinger differentiable at θ_0 with Hellinger derivative

$$\kappa_{\theta_0}(x, y) = \ell(y - \varrho({\theta_0}^\top x))\varrho'({\theta_0}^\top x)(I_p - \theta_0\theta_0^\top)x,$$

as shown above, the claim (11.23) follows from Theorem 2.3 in Schick (2001), which extends to the present parameter set Θ. His result is stated for open subsets of $\mathbb{R}^p$.

Dedication

The authors congratulate Winfried Stute on his 70th birthday and wish him uncountable years of good health and research productivity.

H.L.K.: Working with Winfried on several research projects has been one of my truly rewarding experiences. He is a great scholar and a great friend.

U.U.M.: In appreciation of encouragement and support on several occasions, generously given and gratefully received.

Acknowledgements Research of Hira L. Koul was in part supported by the NSF-DMS grant 1205271.

References

Akritas, M.G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.*, **28**, 549–567.

Boldin, M.V. (1982). An estimate of the distribution of the noise in an autoregressive scheme. *Teor. Veroyatnost. i Primenen.*, **27(4)**, 805–810.

Boldin, M.V. (1990). On testing hypotheses in the sliding average scheme by the Kolmogorov Smirnov and ω^2 tests. *Theory of Probability and Its Applications*, **34(4)**, 699–704.

Boldin, M.V. (1998). On residual empirical distribution functions in ARCH models with applications to testing and estimation. *Mitt. Math. Sem. Giessen*, No. **235**, 40–66.

Carroll, R.J., Fan, J., Gijbels, I., and Wand, M.P. (1997). Generalized Partially Linear Single-Index Models. *J. Amer. Statist. Assoc.*, **92**, 477–489.

Cui, X., Härdle, W. and Zhu, L. (2011). Generalized single-index models: The EFM approach. *Ann. Statist.*, **39**, 1658–1688.

Hájek, J. and Sidák, Z. (1967). *Theory of rank tests*. Academic Press, New York-London; Academia Publishing House of the Czechoslovak Academy of Sciences, Prague.

Hall, P.J. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Ann. Statist.*, **33**, 1404–1421.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, **58**, 71–120.

Khmaladze E.V. and Koul H.L. (2009). Goodness-of-fit problem for errors in nonparametric regression: Distribution free approach. *Ann. Statist.*, **37**, 3165–3185.

Koul, H.L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.*, **40**, 1950–1979

Koul, H.L. (1970). Some convergence theorems for ranks and weighted empirical cumulatives. *Ann. Math. Statist.*, **41**, 1768–1773.

Koul, H.L. (1991). A weak convergence result useful in robust autoregression. *J. Statist. Plann. Inference*, **29**, 291–308.

Koul, H.L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models.* Lecture Notes in Statistics, **166**. Springer-Verlag, New York.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics*, **3**, 37–98.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.

Li, K.C. (1991). Sliced Inverse Regression for Dimension Reduction. *J. Amer. Statist. Assoc.*, **86**, 316–342.

Li, T.T., Yang, H., Wang, J.L., Xue, L.G. and Zhu, L. (2011). Correction on Estimation for a partial linear single-index model. *Ann. Statist.*, **39**, 3441–3443.

Müller, U.U., Schick, A. (2017). Efficiency transfer for regression models with responses missing at random. *Bernoulli*, vol. 23, 2693–2719.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2007). Estimating the error distribution function in semiparametric regression. *Statist. Decisions*, **25**, 1–18.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2009a). Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statist. Probab. Lett.*, **79**, 957–964.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2009b). Estimating the innovation distribution in nonparametric autoregression. *Probab. Theory Related Fields*, **144**, 53–77.

Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. of Mult. Analysis,* **101(5)**, 1067–1078.

Neumeyer, N. and Selk, L. (2013). A note on non-parametric testing for gaussian innovations in ararch models. *J. of Time Series Analysis*, **34**, 362–367.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, **14**, 1139–1151.

Schick, A. (2001). On asymptotic differentiability of averages. *Statist. Probab. Letters*, **51**, 15–23.

Stoker, T.M. (1986). Consistent estimation of scaled coefficients. *Econometrics*, **54**, 1461–1481.

Stute, W. and Zhu, L. (2005). Nonparametric checks for single-index models. *Ann. Statist.*, **33**, 1048–1083.

van der Vaart, A.W. (1998). *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes*, Springer.

Xia, Y. (2008). A multiple-index model and dimension reduction. *J. Amer. Statist. Assoc.*, **103**, 1631–1640.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.*, **97**, 1162–1184.

Xia, Y. and Li, W.K. (1999). On the estimation and testing of functional-coefficient linear models. *Statist. Sinica*, **9**, 735–758.

Xia, Y., Tong, H. and Li, W.K. (2002a). Single-index volatility models and estimation. *Statist. Sinica*, **12(3)**, 785–799.

Xia, Y., Tong, H., Li, W.K. and Zhu, L. (2002b). An adaptive estimation of dimension reduction space. *J.R. Stat. Soc. B*, **64**, 363–410.

Wang, J.L., Xue, L., Zhu, L. and Chong, Y.S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.*, **38**, 246–274

Bounds and Approximations for Distributions of Weighted Kolmogorov-Smirnov Tests

12

Nino Kordzakhia and Alexander Novikov

12.1 Introduction

This study is motivated by applications of nonparametric testing to statistical genomic data treatment, in particular Gene Set Enrichment Analysis (GSEA), see Mootha et al. (2003) and Subramanian et al. (2005). Here we consider the so-called "weighted Kolmogorov-Smirnov" goodness-of-fit tests which have been extensively studied by Charmpi and Ycart (2015) in the context of GSEA. Further the one-sided and the two-sided weighted Kolmogorov-Smirnov tests will be referred to as wKS-1 and wKS-2, respectively.

In GSEA sample sizes are typically very large, thus limit distributions of test statistics can be used for analyzing data. The limit distributions of wKS-1 and wKS-2 test statistics can be defined as the distributions of the following random variables

$$D_g^+ = \max_{0\le t\le 1}(X_t), \quad D_g = \max_{0\le t\le 1}|X_t|, \tag{12.1}$$

where $X = \{X_t, 0 \le t \le 1\}$ is a continuous centered Gaussian process with the covariance function

$$R_X(t,s) = \min(t,s) - ts + g(t)g(s), \tag{12.2}$$

$g = \{g(t), 0 \le t \le 1\}$ is a continuous function such that $g(0) = g(1) = 0$.

N. Kordzakhia (✉)
Macquarie University, Balaclava Road, North Ryde, NSW 2109, Australia
e-mail: nino.kordzakhia@mq.edu.au

A. Novikov
Department of Mathematical Sciences, The University of Technology, PO Box 123, Broadway, Sydney, NSW 2007, Australia
e-mail: Alex.Novikov@uts.edu.au

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_12

The following family of functions g is of special interest in genomic applications

$$g(t) = (t^{\alpha} - t), \ \frac{1}{2} < \alpha < 1 \ . \tag{12.3}$$

In particular, $\alpha = 2/3$ corresponds to the case where gene expression ranks are tested against a given gene set, see Charmpi and Ycart (2015) or Kordzakhia et al. (2016) for more details.
Recall that for the two-sided case with $g = 0$, Kolmogorov (1933) found that

$$K(x) := P\{D_0 > x\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}\,\mathrm{e}^{-2k^2x^2}. \tag{12.4}$$

Kolmogorov also noted in Kolmogorov (1933) that for smaller values of $x > 0$ the approximation

$$1 - K(x) \approx \frac{\sqrt{2\pi}}{x}\mathrm{e}^{-\frac{\pi^2}{8x^2}}$$

holds. This can be deduced from another series expansion for $K(x)$ (see e.g. Feller (1971), Durbin (1973))

$$K(x) = 1 - \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}\mathrm{e}^{-\frac{(2k-1)^2\pi^2}{8x^2}}. \tag{12.5}$$

For the one-sided case with $g = 0$, Smirnov (1939) proved that

$$S(x) := P\{D_0^+ > x\} = \mathrm{e}^{-2x^2}. \tag{12.6}$$

The formula (12.6) first appeared in a personal letter from A. Kolmogorov to P. Aleksandrov written in 1931, see Shiryaev (2003), p. 436.

The covariance function

$$R_B(t,s) = \min(t,s) - ts$$

corresponds to the standard Brownian bridge $B = \{B_t, 0 \le t \le 1\}$ with values $B_0 = B_1 = 0$. It is well known that the process B admits the following two representations:

$$B = \{B_t, 0 \le t < 1\} \stackrel{d}{=} \{W_t - tW_1, 0 \le t < 1\} \stackrel{d}{=} \{(1-t)W_{t/(1-t)}, 0 \le t < 1\}, \tag{12.7}$$

where $W = \{W_t, t \ge 0\}$ is a standard Brownian motion defined on a filtered probability space $(\Omega, F, \{F_t\}, P)$ and "$\stackrel{d}{=}$" stands for "equality in distribution". By direct

calculation it can be easily checked that the Gaussian process X with covariance function (12.2) can be represented in the following form

$$X \stackrel{d}{=} \{B_t - g(t)\xi,\ 0 \le t \le 1\}, \tag{12.8}$$

where ξ follows the standard normal distribution and is independent of B.
It must be noted that the goodness-of-fit testing of models with unknown parameters (see e.g. Durbin (1957), Kulinskaya (1995), Tyurin (1984), del Barrio (2007)) typically involves limiting processes $Y = \{Y_t\}$ with covariance functions of the form

$$R_Y(t, s) = \min(t, s) - ts - g(t)g(s),$$

where $\int_0^1 (\frac{\mathrm{d}}{\mathrm{d}s} g(s))^2 \,\mathrm{d}s = 1$.
The covariance function $R_Y(t, s)$ looks similar to (12.2), however, the process Y has a more complicated representation:

$$Y \stackrel{d}{=} \{W_t - tW_1 - g(t) \int_0^1 g'(s)\,\mathrm{d}W_s,\ 0 \le t \le 1\}.$$

The present study is restricted to processes $\{X_t\}$ with covariance function $R_X(t, s)$ from (12.2).

In the literature, several approaches for approximating probabilities $P\{D_g^+ > x\}$ and $P\{D_g > x\}$ have been developed. Such probabilities can be seen as functionals of a diffusion process. Historically, it was Kolmogorov (1933) who used his own results on empirical processes and backward *Partial Differential Equations* (PDE) from the theory of diffusion processes for finding the analytical formula (12.4). Since then, the theory of empirical processes has been intensively developed in combination with PDE and other approaches (e.g. Anderson and Darling (1952), Gaenssler and Stute (1979), Tyurin (1984)) for solving related problems in the theory of goodness-of-fit tests. Under the PDE approach, for finding $P\{D_g^+ > x\}$ and $P\{D_g > x\}$ one needs to solve a parabolic time-varying PDE with some boundary conditions, for discussion also see p. 220 in Gaenssler and Stute (1979). A numerical solution of such PDEs can be obtained using finite-difference schemes although it requires a special software and rather complicated analysis of numerical errors.

The *martingale* approach has been used by Khmaladze (1981) and Stute and Anh (2012). Under this approach, an integral transformation is applied to an underlying empirical process to reduce the problem to the known distributions of $\max_{0\le t\le 1}(W_t)$ and $\max_{0\le t\le 1}|W_t|$. Finding such a transformation for our problem is a separate task which is worth pursuing.

Among others, we mention the *asymptotic* approach which is based on approximations to $P\{D_g^+ > x\}$for large x via the theory of extremes of Gaussian processes: see Durbin (1957), Parker (2013), Piterbarg (1996). Note that due to their asymptotic nature, such approximations are not accurate for smaller x (larger p-values) which are also of interest in GSEA.

The problem of finding accurate and fast approximations for one-sided weighted Kolmogorov-Smirnov test (wKS-1) has been discussed in Kordzakhia et al. (2016) using the reduction to boundary crossing probabilities (BCP) for the standard Brownian motion W. In Sect. 12.2 we demonstrate that BCP approach can be applied to both $P\{D_g^+ > x\}$ and $P\{D_g > x\}$ by approximating nonlinear boundaries with n-knot linear splines enabling us to use recurrent integrations. In Sect. 12.3, for wKS-1 and wKS-2 we provide general bounds for errors of approximations in terms of a specific distance defined between g and the approximating function g_n.

For the case wKS-1 in Kordzakhia et al. (2016) it has been shown via numerical experiments that the BCP approximations with 1-knot linear splines have relative errors smaller than 5% uniformly for all $\alpha \in (1/2, 1)$ in (12.3). In Sect. 12.4, we illustrate that for the important case wKS-1 with $g(t) = t^{2/3} - t$ the mid-point approximation based on *upper and lower bounds* has relative errors less than 1% (see Fig. 12.2). The refinement has been achieved with the use 2-knot linear splines while remaining computationally efficient. The advantage of using upper and lower bounds consists in the fact that one can control accuracy of calculations avoiding intensive analysis of numerical errors which could be computationally costly. A study comparing the PDE, BCP and MC approaches for $P\{D_g > x\}$ with various weighting functions g will be discussed elsewhere.

12.2 Reduction to BCP for a Brownian Motion

12.2.1 Notation and General Facts

In this section we provide the results of BCP approach for both wKS-1 and wKS-2 cases. For completeness of the exposition we provide the results for wKS-1 from Kordzakhia et al. (2016) along with some new aspects which were not discussed in Kordzakhia et al. (2016).

Using (12.8) and the second representation in (12.7) along with the change of time

$$\frac{t}{1-t} = s, \quad t = \frac{s}{1+s}, \tag{12.9}$$

the random variables D_g^+ and D_g from (12.1) can be represented as follows:

$$D_g^+ \stackrel{d}{=} \max_{0 \le s < \infty} \frac{W_s - G(s)\xi}{s+1}, \quad D_g \stackrel{d}{=} \max_{0 \le s < \infty} \frac{|W_s - G(s)\xi|}{s+1}, \tag{12.10}$$

where

$$G = \{G(s) := (s+1)g(\frac{s}{s+1}),\ s \in [0, \infty)\}. \tag{12.11}$$

In particular, for $g(s)$ from (12.3) we have

$$G(s) = (s+1)^{1-\alpha} s^{\alpha} - s\ , \quad 1/2 < \alpha < 1,$$

which is a non-negative bounded function.

Further we will assume that G is a monotone continuous function such that[1]

$$\inf_{0 \le s < \infty} G(s) = 0\ , \quad \sup_{0 \le s < \infty} G(s) := M < \infty. \tag{12.12}$$

Let $A = \{A(s),\ s \in [0, \infty)\}$ be a continuous function and define

$$S(x, y, A) := P\{\max_{0 \le s < \infty} \frac{W_s - A(s)y}{(s+1)} > x\}\ , \tag{12.13}$$

$$K(x, y, A) := P\{\max_{0 \le s < \infty} \frac{|W_s - A(s)y|}{s+1} > x\}\ . \tag{12.14}$$

Since ξ is independent of B in (12.8) we obtain for the case $A = G$

$$P\{D_g^+ > x\} = \int_{-\infty}^{\infty} \phi(y) S(x, y, G)\, \mathrm{d}y,\ P\{D_g > x\} = 2\int_0^{\infty} \phi(y) K(x, y, G)\, \mathrm{d}y, \tag{12.15}$$

where the function G is defined above in (12.11), $\phi(y) = \mathrm{e}^{-\frac{y^2}{2}}/\sqrt{2\pi}$ is the density function of the standard normal distribution and in the last integral representation for $P\{D_g > x\}$ we have used the symmetry property $K(x, y, A) = K(x, -y, A)$. Now, to approximate $P\{D_g > x\}$ and $P\{D_g^+ > x\}$ we can substitute an approximating function g_n instead of g and use the representation (12.15) with the corresponding function

$$G_n := \{G_n(s) := (s+1) g_n(\frac{s}{s+1})\ ,\ s \ge 0\}.$$

In fact, we found that it is more convenient to discuss at first a n-knot piecewise linear function G_n as an approximation to G and then define the corresponding approximating function

$$g_n := \{g_n(t) := (1-t) G_n(\frac{t}{1-t}),\ 0 \le t < 1\}\ ,$$

which is also a n-knot piecewise linear function. It is also important to note that under the assumption that G is a nonnegative continuously differentiable function,

[1]This assumption is used just for convenience, the exposition can be adapted for a general case as well.

we can always construct sequences of piecewise linear functions $G_n^L(s)$ and $G_n^U(s)$ such that for all $s \in (0, \infty)$

$$G_n^L(s) \le G(s) \le G_n^U(s),$$
$$\lim_{n\to\infty} G_n^L(s) = \lim_{n\to\infty} G_n^U(s) = G(s).$$

Hence, for $y \ge 0$, as $n \to \infty$

$$S(y, x, G) \le P\{\sup_{0\le s<\infty} \frac{W_s - G_n^L(s)y}{(s+1)} > x\} = S(y, x, G_n^L) \to S(y, x, G) ,$$

and for $y < 0$, as $n \to \infty$

$$S(y, x, G) \ge P\{\sup_{0\le s<\infty} \frac{W_s - G_n^U(s)y}{s+1} > x\} = S(y, x, G_n^U) \to S(y, x, G) .$$

Therefore

$$P\{D_g^+ > x\} \le UB(x) := \int_0^\infty \phi(y)[S(x, y, G_n^L) + S(x, -y, G_n^U)]\,\mathrm{d}y \to P\{D_g^+ > x\} , \tag{12.16}$$

$$P\{D_g^+ > x\} \ge LB(x) := \int_0^\infty \phi(y)[S(x, y, G_n^U) + S(x, -y, G_n^L)]\,\mathrm{d}y \to P\{D_g^+ > x\} . \tag{12.17}$$

Analogously, one can obtain the following upper and lower bounds for $P\{D_g > x\}$:

$$P\{D_g > x\} \le 2\int_0^\infty \phi(y)(1 - P\{-x(1+s) + G_n^U(s)y \le W_s \le G_n^L(s)y + x(1+s), s \ge 0\})dy$$
$$\to P\{D_g > x\},$$

$$P\{D_g > x\} \ge 2\int_0^\infty \phi(y)(1 - P\{-x(1+s) + G_n^L(s)y \le W_s \le G_n^U(s)y + x(1+s), s \ge 0\})dy$$
$$\to P\{D_g > x\}.$$

12.2.2 Recurrent Numerical Integration for BCP Approximations

Let $f(s)$ and $q(s)$ be boundaries for W on the interval $[0, T]$. Let

$$p(i, f, q|W_{t_i}, W_{t_{i+1}}) := P\{f(s) < W_s < q(s), t_i \le s \le t_{i+1}|W_{t_i}, W_{t_{i+1}}\}. \tag{12.18}$$

The following result, first proved in Novikov et al. (1999), can be used for finding BCP for one-sided and two-sided boundaries via recurrent integrations when the function $p(i, f, q|x, y)$ is known.

Proposition 1 (Novikov et al. 1999) *Let* $f(s)$*and* $q(s)$ *be deterministic continuous functions,*

$$t_0 = 0 < t_1 < \cdots < t_n = T.$$

Then

$$P\{f(s) < W_s < q(s), 0 \le s \le T\} = E\prod_{i=0}^{n-1} p(i, f, q|W_{t_i}, W_{t_{i+1}}).$$

In the particular case of one-sided boundaries where $f = -\infty$and $q(s)$ is a linear boundary on the interval $s \in [t_i, t_{i+1}]$, we have

$$p(i, -\infty, q|W_{t_i}, W_{t_{i+1}}) = 1 - \exp\{-2(q(t_i) - W_{t_i})^+(q(t_{i+1}) - W_{t_{i+1}})^+/(t_{i+1} - t_i)\}. \quad (12.19)$$

The latter can be found in Wang and Pötzelberger (1997). The case of two-sided linear boundaries was discussed in Novikov et al. (1999) and then it was also studied with the use of the Monte Carlo method in Pötzelberger and Wang (2001) and Pötzelberger (2012).

For completeness of the exposition we outline here the proof of Proposition 1 which was presented in Novikov et al. (1999).

Proof Let

$$Z_t^{(i)} := W_t - E[W_t|(W_{t_i}, W_{t_{i+1}})] = W_t - W_{t_i} - \frac{t - t_i}{t_{i+1} - t_i}(W_{t_{i+1}} - W_{t_i}),\ t \in [t_i, t_{i+1}]\ .$$

Then the Gaussian processes $Z_t^{(i)}$ and $E[W_t|(W_{t_i}, W_{t_{i+1}})]$ are independent for all $t \in [t_i, t_{i+1}]$. Note that $Z_t^{(i)}, i = 1, ..., n$ are independent Brownian bridges, the covariance function of $Z_t^{(i)}$ is

$$\mathrm{cov}(Z_t^{(i)}, Z_s^{(i)}) = \min(t - t_i, s - s_i) - (t - t_i)(s - s_i)/(t_{i+1} - t_i)\ ,\ s, t \in [t_i, t_{i+1}].$$

Using the change of time (12.9) and the Bachelier formula (see Doob 1949) with the substitutions

$$a = (q(t_i) - W_{t_i})^+\ ,\ b = (q(t_{i+1}) - W_{t_{i+1}})^+/(t_{i+1} - t_i), \quad (12.20)$$

we obtain (12.19).

An explicit representation for $p(i, f, q|W_{t_i}, W_{t_{i+1}})$ with linear functions f and q can be found in Doob (1949). Here we reproduce the result of Escriba (1987), as it is presented in computationally more convenient form due to nonnegativity of the summands in the series (12.22).

For $a > 0,\ c > 0,\ b \ge d$

$$P\{-c - ds < W_s < a + bs, 0 \le s < \infty\} = 1 - H(a, b, c, d)\ , \quad (12.21)$$

where

$$H(a,b,c,d) = Q(a,b,c,d) + Q(c,d,a,b),\ \ Q(a,b,c,d) := \sum_{k=1}^{\infty} q_k(a,b,c,d)\ , \tag{12.22}$$

$$\begin{aligned} q_k(a,b,c,d) := \exp\{&2(b+d)(-c)(k-1)^2 - \\ &2(b+d)a(k-1)k - ab + b(2(-c-a)(k-1) - a)\} \\ &\times(1 - \exp\{-2c((b+d)(2k-1) + b)\}). \end{aligned}$$

Make in (12.22) the change (12.20) for the upper boundary q and correspondingly, the following change for the lower boundary f:

$$c = (f(t_i) - W_{t_i})^+, d = (f(t_{i+1}) - W_{t_{i+1}})^+/(t_{i+1} - t_i)$$

an explicit representation of $p(i,\ f, q|W_{t_i},\ W_{t_{i+1}})$ as an infinite series can be obtained. Note in literature one can find other representations for $p(i,\ f, q|W_{t_i},\ W_{t_{i+1}})$ with linear functions f and q, see e.g. Anderson (1960) and Hall (1997).

12.2.3 Computing $S(x,\ y,\ G_n)$ and $K(x,\ y,\ G_n)$ with n-knot Linear Splines

Here we consider the case where instead of $G(s)$ in (12.10) a piecewise linear continuous function $G_n(s)$ is used. Let

$$\{t_i : t_0 = 0 < t_1 < \cdots < t_n = T\}.$$

Suppose that $G_n(s)$ is truncated by a constant at point $t_n = T < \infty$ i.e.

$$G_n(s) \equiv G_n(T)\ ,\ s \geq T\ .$$

Proposition 2 (Kordzakhia et al. (2016))

$$\begin{aligned} S(x,y,G_n) = 1 - E\prod_{i=0}^{n-1}(1 - \exp\{-2(q(t_i) - W_{t_i})^+(q(t_{i+1}) - W_{t_{i+1}})^+/(t_{i+1} - t_i)\}) \\ (1 - \mathrm{e}^{-2(G_n(T)y - W_T + (1+T)x)^+ x})\ . \end{aligned}$$

Remark 1 The computation of $S(y,x,G_n)$ can be reduced to n-fold integration of the factorized expression which contains the transition density of the Brownian motion W. Hence, we need to make n recurrent integrations to find $S(y,x,G_n)$ and it only remains to make one additional integration over y to find $P\{D^+_{g_n} > x\}$.

Remark 2 To find approximations based on the expressions (12.22) for wKS-2 one needs to make a proper truncation of the infinite series in (12.22); we shall discuss this procedure elsewhere in details. Here we just note that for large x, a rather accurate upper bound can be obtained via the obvious inequality

$$P\{D_g > x\} \le P\{D_g^+ > x\} + P\{D_{-g}^+ > x\} = 2P\{D_g^+ > x\}, \tag{12.23}$$

where $P\{D_g^+ > x\}$ can be approximated using (12.15) in Proposition 2.3 and (12.16).

The following simple lower bound

$$P\{D_g > x\} \ge K(x), \tag{12.24}$$

is a result of the Anderson inequality, see Anderson (1955). By direct calculations based on the second representation for $K(x)$ from (12.5) one can check that $1 - K(x) < 10^{-4}$ for $x \le 0.33$, hence, we have $P\{D_g > x\} \approx 1$ for the range $0 \le x \le 0.33$ with a high accuracy for *any function* g.

12.3 Accuracy of BCP Approximations

Here we present general upper bounds for distances $|S(x, y, A_1) - S(x, y, A_2)|$ and $|K(x, y, A_1) - K(x, y, A_2)|$. Such estimates lead to the theoretical justification of numerical consistency of the BCP approach using n-knot linear splines and recurrent integration formula from Proposition 2 with $n \to \infty$. In fact, for wKS-1 a comparison with MC simulations demonstrates that an accuracy of order 1% is achieved with the use of 2-knot linear splines, see Sect. 12.4.

Theorem 1 *Let $A_1(s)$ and $A_2(s)$ be continuously differentiable functions such that $A_1(0) = A_2(0)$ and*

$$\Delta(A_1, A_2) := \int_0^\infty (\frac{\mathrm{d}}{\mathrm{d}s}(A_1(s) - A_2(s)))^2 \,\mathrm{d}s < \infty\,. \tag{12.25}$$

Then

$$\begin{aligned} &|S(x, y, A_1) - S(x, y, A_2)| \\ &\le 2erf(|y|\sqrt{\Delta(A_1, A_2)/8}) \le |y|\sqrt{\frac{2\Delta(A_1, A_2)}{\pi}}\,, \end{aligned} \tag{12.26}$$

and

$$|K(x, y, A_1) - K(x, y, A_2)| \tag{12.27}$$
$$\leq 2erf(|y|\sqrt{\Delta(A_1, A_2)/8}) \leq |y|\sqrt{\frac{2\Delta(A_1, A_2)}{\pi}},$$

where $erf(x) := \int_0^x \phi(\sqrt{2}u)du/\sqrt{2}$.

To simplify notations further we set

$$\Delta(A_1, A_2) := \Delta_{12}, \quad \zeta_{12} := \int_0^\infty (\frac{\mathrm{d}}{\mathrm{d}s}(A_1(s) - A_2(s)))\,\mathrm{d}W_s\,. \tag{12.28}$$

The proof of Theorem 1 relies on the following lemma.

Lemma 1 *Let* $A_i(s), i = 1, 2$ *be continuous differentiable functions,* $A_i(0) = 0$. *Then for any* $x > 0$ *and* y

$$S(x, y, A_1) = EI\{\max_{0\leq s<\infty} \frac{W_s - A_2(s)y}{s+1} > x\}\,\mathrm{e}^{-y\zeta_{12}-y^2\Delta_{12}/2}, \tag{12.29}$$

and

$$K(x, y, A_1) = EI\{\max_{0\leq s<\infty} \frac{|W_s - A_2(s)y|}{s+1} > x\}\,\mathrm{e}^{-y\zeta_{12}-y^2\Delta_{12}/2}. \tag{12.30}$$

Further we show validity of (12.30) by adapting the proof of (12.29) given in Kordzakhia et al. (2016).

Proof of Lemma 1. We define the Girsanov measure transformation on $(\Omega, F, \{F_t\}, P)$ as follows

$$\tilde{P}(A) = EI\{A\}\,\mathrm{e}^{-y\zeta_{12}-y^2\Delta_{12}/2},\ A \in F.$$

By Girsanov's theorem (see Liptser and Shiryaev (2001)) the process W_t has drift $y(A_2(t) - A_1(t))$ with respect to the probability space $(\Omega, F, \{F_t\}, \tilde{P})$. This implies

$$EI\{\max_{0\leq s<\infty} \frac{|W_s - A_2(s)y|}{s+1} > x\}\,\mathrm{e}^{-y\zeta_{12}-y^2\Delta_{12}/2}$$

$$= \tilde{E}I\{\max_{0\leq s<\infty} \frac{|\tilde{W}_s + y(A_2(s) - A_1(s)) - A_2(s)y|}{s+1} > x\}$$
$$= \tilde{E}I\{\max_{0\leq s<\infty} \frac{|\tilde{W}_s - yA_1(s)|}{s+1} > x\} = EI\{\max_{0\leq s<\infty} \frac{|W_s - yA_1(s)|}{s+1} > x\},$$

where $\tilde{E}(.)$ is the symbol of expectation and $\tilde{W}_s$ is a standard Brownian motion with respect to the measure $\tilde{P}(A)$. Thus (12.30) is proved.

Consequently, the proof of Theorem 1 is a straightforward adaptation of that of Theorem 1 from Kordzakhia et al. (2016).

Remark 3 Using Theorem 1, the numerical consistency of the BCP approximations via piecewise linear functions can be deduced. Indeed, let $G_n(s)$ be a piecewise linear continuous function such that

$$G_n(s_i) = G(s_i),\ s_i = iT/n,\ i = 0, \ldots, n,$$
$$G_n(s) = G(T), s \geq T,\ T < \infty\ ,$$

and consider the following approximations

$$\hat{P}_n\{D_g^+ > x\} = \int_{-\infty}^{\infty} \phi(y) S(x, y, G_n) \mathrm{d}y$$

and

$$\hat{P}_n\{D > x\} = 2\int_0^{\infty} \phi(y) K(x, y, G_n)\, \mathrm{d}y,$$

obtained by substituting G_n instead of G in (12.15). Set

$$\Delta_n := \int_0^T (\frac{\mathrm{d}}{\mathrm{d}s}(G_n(s) - G(s)))^2\, \mathrm{d}s,\ \delta_T := \int_T^{\infty} (\frac{\mathrm{d}}{\mathrm{d}s} G(s)))^2\, \mathrm{d}s\ . \tag{12.31}$$

By Theorem 1 we have

$$|S(x, y, G) - S(x, y, G_n)| \leq 2\mathrm{erf}(|y|\sqrt{(\Delta_n + \delta_T)/8})$$

and

$$|K(x, y, G) - K(x, y, G_n)| \leq 2\mathrm{erf}(|y|\sqrt{(\Delta_n + \delta_T)/8}).$$

Hence, for $S(.)$ as well as $K(.)$, we have

$$|P\{D_g^+ > x\} - P\{D_{g_n}^+ > x\}| \leq$$
$$\int_{-\infty}^{\infty} \phi(y)|S(x, y, G) - S(x, y, G_n)|\, \mathrm{d}y \leq$$
$$\int_{-\infty}^{\infty} \phi(y) 2\mathrm{erf}(|y|\sqrt{(\Delta_n + \delta_T)/8})\, \mathrm{d}y =$$
$$\frac{4}{\pi} \arctan(\sqrt{(\Delta_n + \delta_T)}/2) \leq \frac{2\sqrt{(\Delta_n + \delta_T)}}{\pi}.$$

If $G(s)$ has a continuous second derivative then one can easily check that

$$\Delta_n \leq 2 \max_{s \leq T} |G''(s)| T^3/n^2 = \text{const} T^3/n^2 \ .$$

The quantity δ_T can be made arbitrary small. By choosing T and large enough n, $\sqrt{\Delta_n + \delta_T} < \varepsilon$ for any $\varepsilon > 0$. This implies the numerical consistency of the BCP approach for $\hat{P}_n\{D_g^+ > x\}$ and $\hat{P}_n\{D_g > x\}$ as $n \to \infty$.

12.4 Numerical Results

Here we provide the numerical results obtained for $P\{D_g^+ > x\}$ for the most important case $\alpha = 2/3$ in (12.3). In this case

$$G(s) = (s+1)^{1/3} s^{2/3} - s,$$

and the conditions (12.12) and (12.25) hold.

Figure 12.1 illustrates a choice of 2-knot upper and lower bounds for $G(s)$. The chosen nodes minimize the distance $\Delta(G, G_2)$ defined in (12.25). Results of numerical calculation are shown in Table 12.1 for the following functions: the upper bound $UB(x)$ (see (12.16)), the lower bound $LB(x)$ (see (12.17)), the Durbin's asymptotic approximation (Durbin (1985))

$$DurbA(x) := 0.94088\, \mathrm{e}^{-1.87055 x^2},$$

obtained from Kordzakhia et al. (2016), and the mid-point approximation:

$$Mid(x) := (LB(x) + UB(x))/2.$$

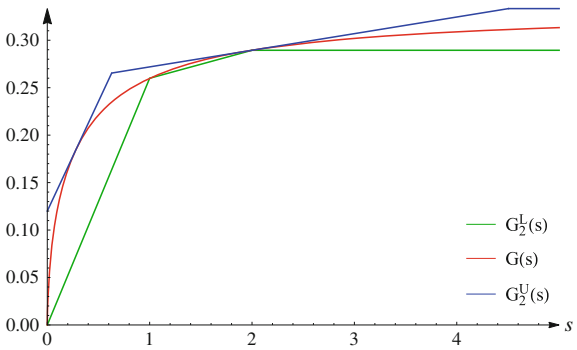


Fig. 12.1 2-knot *upper* and *lower* bounds for $G = G(s)$

Table 12.1 Numerical results for approximations to $P\{D_g^+ > x\}$, $g(t) = t^{2/3} - t$

X	LB	UB	Mid	DurbA	MC
0.1	0.968012	0.983650	0.975831	0.923446	0.972634
0.2	0.902706	0.931352	0.917029	0.873053	0.910118
0.3	0.811256	0.847397	0.829327	0.795101	0.821336
0.4	0.702729	0.741782	0.722256	0.697520	0.714913
0.5	0.586754	0.625236	0.605995	0.589445	0.600191
0.6	0.472232	0.507648	0.489940	0.479825	0.485830
0.7	0.366321	0.397126	0.381724	0.376249	0.379100
0.8	0.273871	0.299367	0.286619	0.284197	0.285115
0.9	0.197324	0.217484	0.207404	0.206784	0.206706
1.0	0.137001	0.152276	0.144639	0.144933	0.144388
1.1	0.091655	0.102759	0.097207	0.097852	0.097190
1.2	0.059082	0.066837	0.062959	0.063639	0.063017
1.3	0.036693	0.041901	0.039297	0.039869	0.039383
1.4	0.021956	0.025319	0.023637	0.024060	0.023713
1.5	0.012656	0.014747	0.013702	0.013987	0.013758
1.6	0.007028	0.008279	0.007654	0.007832	0.007692
1.7	0.003760	0.004480	0.004120	0.004225	0.004138
1.8	0.001938	0.002336	0.002137	0.002195	0.002144
1.9	0.000962	0.001174	0.001068	0.001099	0.001074
2.0	0.000460	0.000569	0.000515	0.000530	0.000516

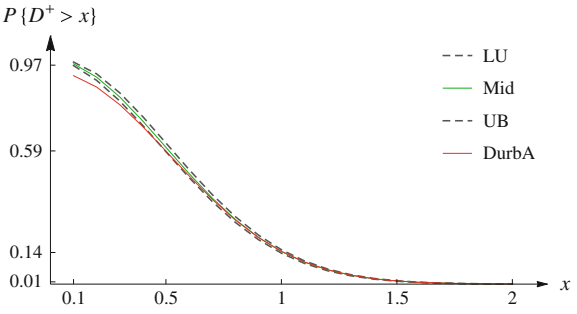


Fig. 12.2 *Lower* (LB) and *upper* (UB) bounds, *mid-point* (Mid) and Durbin's (DurbA) approximations to $P\{D_g^+ > x\}$ from Table 12.1

In Table 12.1 Monte Carlo simulation results are included for the discretised maximum:

$$\hat{D}_{g,n}^+ := \max_{0 \le j \le n} (X_{j/n}) \ ,$$

based on 10^7 trajectories and $n = 2 \times 10^5$ points uniformly located on $[0,1]$.

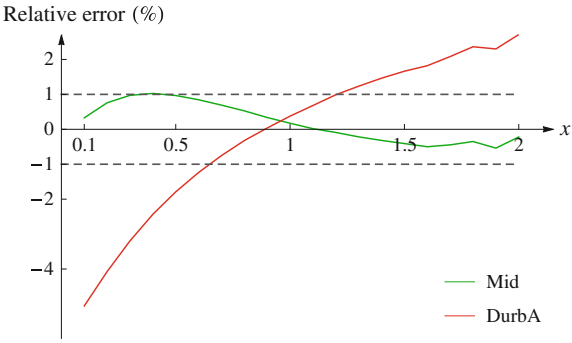


Fig. 12.3 Relative errors of *Mid* and *DurbA* approximations versus *MC*

The numerical results for mid-point and Durbin's approximations to $P\{D_g^+ > x\}$ provided in Table 12.1 for $g(t) = t^{2/3} - t,\ 0 \le t \le 1$, are plotted in Fig. 12.2 along with its upper and lower bound approximations.

Figure 12.3 illustrates the relative errors (%) of $Mid(x)$ and $DurbA(x)$ approximations evaluated with respect to Monte Carlo (MC) results shown in Table 12.1. The absolute value of the relative errors of the mid-point approximation $Mid(x)$ does not exceed 1% for all $x \in (.1, 2)$.

Acknowledgements It is our pleasure to thank our colleagues Albert Shiryaev and Dan Wu for their helpful discussions. We also thank Lin Yee Hin for boosting the computational capacity of our Monte Carlo study. We would like to extend our sincere gratitude to Bernard Ycart for his valuable comments that led to significant improvement of the paper. The work of A. Novikov was supported by the Russian Science Foundation under Grant 14-21-00162. The work of N. Kordzakhia was supported by the Australian Research Council Grant DP150102758.

References

Anderson, T. W., Darling, D. A. (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Statist. 23(2), 193–212.

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. Proc. Amer. Math. Soc. 6, 170-176.

Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. Ann. Math. Statist. 31, 165-197.

Charmpi K., Ycart B. (2015) Weighted Kolmogorov-Smirnov testing: an alternative for Gene Set Enrichment Analysis. Statist. Appl. in Genetics and Molecular Biology, 14(3), 279–295.

del Barrio E. (2007) Empirical and quantile processes in the asymptotic theory of goodness-of-fit tests. In: del Barrio E., Deheuvels P., van de Geer S. (eds.) Lectures on empirical processes: theory and statistical applications, EMS series of lectures in Mathematics, European Mathematical Society, Zurich, 1–92.

Doob, J. L. (1949) Heuristic approach to the Kolmogorov-Smirnov theorems. Ann. Math. Statist. 20(3), 393–403.

Durbin, J. (1973) Distribution Theory for Tests Based on the Sample Distribution Function. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics vol. 9, SIAM, Philadelphia, PA.

Durbin, J. (1975) Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. Biometrika 62(1), 5–22.

Durbin, J. (1985) The first-passage density of a continuous Gaussian process to a general boundary. J. Appl. Probab. 22(1), 99–122.

Escribá, L. B. (1987) A stopped Brownian motion formula with two sloping line boundaries. Ann. Probab. 15(4), 1524–26.

Feller, W. (1971) An introduction to probability theory and its applications. Vol. II. Second edition John Wiley & Sons, Inc., New York-London-Sydney.

Gaenssler, P. and Stute W. (1979) Empirical Processes: A Survey of Results for Independent and Identically Distributed Random Variables. Annals of Probability 7(2), 193-243.

Hall, W. J. (1997). The distribution of Brownian motion on linear stopping boundaries. Sequential Analysis 4, 345-352.

Khmaladze È. V. (1981) A martingale approach in the theory of goodness-of-fit tests. (In Russian) Teor. Veroyatnost. i Primenen. 26(2), 246–265.

Kolmogorov A. (1933) Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari 4, 83–91.

Kordzakhia N., Novikov A., Ycart B. (2016) Approximations for weighted Kolmogorov-Smirnov distributions via boundary crossing probabilities. Statistics and Computing, online since 15 September, doi:10.1007/s11222-016-9701-y.

Kulinskaya, E. (1995). Coefficients of the asymptotic distribution of the Kolmogorov-Smirnov statistic when parameters are estimated. J. Nonparametr. Statist. 5(1), 43–60.

Liptser, R. S., Shiryaev, A. N. (2001) Statistics of random processes. I. General Theory. Springer-Verlag, Berlin.

Mootha V. K., Lindgren C. M., Eriksson K. F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., Ridderstråle M., Laurila E., Houstis N., Daly M. J., Patterson N., Mesirov J. P., Golub T. R., Tamayo P., Spiegelman B., Lander E. S., Hirschhorn J. N., Altshuler D., Groop L. C. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34(3), 267–273.

Novikov A., Frishling V., Kordzakhia N. (1999) Approximations of boundary crossing probabilities for a Brownian motion. J. Appl. Probab. 36(4), 1019–1030.

Parker T. (2013) A comparison of alternative approaches to supremum-norm goodness of fit tests with estimated parameters. Econometric theory 29(5), 969–1008.

Piterbarg V. (1996) Asymptotic Methods in the Theory of Gaussian Processes and Fields, vol. 148. Translations of Mathematical Monographs. American Mathematical Society.

Pötzelberger, K. (2012) Improving the Monte Carlo estimation of boundary crossing probabilities by control variables. Monte Carlo Methods Appl. 18, no. 4, 353–377.

Pötzelberger K., Wang L. (2001) Boundary crossing probability for Brownian motion. J. Appl. Probab. 38(1), 152–164.

Shiryaev A. (2003) Kolmogorov, Book 2: Selecta from the correspondence between A. N. Kolmogorov and P. S. Aleksandrov. Moscow.

Smirnov N. (1939) Sur les écarts de la courbe de distribution empirique. Rec. Math. [Mat. Sbornik] N.S., 6(48), 3–26.

Stute W., Anh T. L. (2012) Principal component analysis of martingale residuals. J. Indian Statist. Assoc. 50(1-2), 263–276.

Subramanian A., Tamayo P., Mootha V. K., Mukherjee S., Ebert B. L., Gillette M. A., Paulovich A., Pomeroy S. L., Golub T. R., Lander E. S., Mesirov J. P. Gene Set Enrichment Analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA (PNAS) 102(43), 15545–550.

Tyurin Y. N. (1984) The limit distribution of the Kolmogorov-Smirnov statistics for a composite hypothesis. (In Russian) Izv. Akad. Nauk SSSR Ser. Mat. 48(6), 1314–43.

Wang L., Pötzelberger K. (1997), Boundary crossing probability for Brownian motion and general boundaries. J. Appl. Probab. 34, no. 1, 54–65.

Nonparametric Stopping Rules for Detecting Small Changes in Location and Scale Families

13

P.K. Bhattacharya and Hong Zhou

AMS 2000 Subject Classification: Primary 60G40 · 62G99 · 62L99 · Secondary 62G20 · 60F17 · 60G48

13.1 Introduction and Summary

Let $X_1, \ldots, X_k, \ldots$ be sequentially observed independent random variables whose Lebesgue density changes either within a location family

$$C_L(f) = \{f(x, \theta) = f(x - \theta), \theta \in R\},$$

or within a scale family

$$C_S(f) = \{f(x, \theta) = e^{-\theta} f\left(xe^{-\theta}\right), \theta \in R\},$$

from $f(x, \theta_0)$ to $f(x, \theta_0 + \Delta)$ after an unknown (possibly infinite) change-point τ. Here $\Delta > 0$ is a specified amount of change to be detected quickly and with a low rate of false alarm, while the initial θ_0 may or may not be known. The literature on this problem mostly deals with known f. The aim of this paper is to construct nonparametric stopping rules when f is unknown, which would compare favorably with their parametric counterparts based on a possibly misspecified form of density. This will be attempted in an asymptotic setting, considering contiguous changes in the following two models. We denote the distribution of $\{X_i\}$ by $P_{f,\infty}$ if the density remains $f(\cdot)$ throughout the sequence (i.e., $\tau = \infty$), and by $P_{f,\tau,\Delta}$ or $P_{1f,\tau,\Delta}$ if the

P.K. Bhattacharya (✉) · H. Zhou
University of California, Davis, USA
e-mail: pkbhattacharya@ucdavis.edu

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_13

density changes from $f(\cdot)$ to $f(\cdot - \Delta)$ in $C_L(f)$ or from $f(\cdot)$ to $e^{-\Delta} f(\cdot/e^{\Delta})$ in $C_S(f)$ after X_τ with $\tau < \infty$.

- Model 1. The distribution of $\{X_i\}$ is either $P_{f,\infty}$ or $P_{f,\tau,\Delta}$ in $C_L(f)$, where τ is unknown, $\Delta > 0$ is specified and f is symmetric about 0 but otherwise unknown.
- Model 2. The distribution of $\{X_i\}$ is either $P_{f,\infty}$ or $P_{f,\gamma,\Delta}$ in $C_L(f)$, or $P_{1f,\tau,\Delta}$ in $C_S(f)$, where τ is unknown, $\Delta > 0$ is specified and f is arbitrary and unknown.

In both models we shall consider contiguous changes. Indeed, our evaluation of a stopping rule will be determined by its asymptotic behavior over the first n observations under $P_n = P_{f,\infty}$, and under either $Q_n = P_{f,\tau_n,\Delta_n}$ in $C_L(f)$ or $Q_{1n} = P_{1f,\tau_n,\Delta_n}$ in $C_S(f)$ as $n \to \infty$, where $\tau_n = [n\lambda]$ for some $0 < \lambda < 1$ and $\Delta_n = \delta n^{-1/2}$ with a specified $\delta > 0$.

The stopping rules constructed in this paper are rank analogues of the Page-CUSUM procedure and its generalization. When f and θ_0 are both known, the CUSUM procedure due to Page (1954) calculates the log likelihood ratio (LLR) statistic to test $H_k : \tau \geq k$ against the alternative $H_k' : \tau < k$ based on $X_1, \ldots, X_k$ for each $k \geq 1$ and stops as soon as the LLR exceeds a prescribed boundary c . The LLR for H_k vs H_k' is

$$T_k(f, \Delta) = \max_{0 \leq j < k} T_{jk}(f, \Delta), \ k \geq 1, \tag{13.1a}$$

where

$$T_{jk}(f, \Delta) = \sum_{j+1 \leq i \leq k} \log[f(X_i, \theta_0 + \Delta)/f(X_i, \theta_0)]. \tag{13.1b}$$

The Page-CUSUM stopping rule is thus defined as

$$N_P(c, f, \Delta) = \min\{k : T_k(f, \Delta) \geq c\}, \tag{13.1c}$$

where the constant c, known as the decision boundary, is chosen so as to control the false alarm rate by keeping $E\,[N_P(c, f, \Delta)]$ or $P\,[N_P(c, f, \Delta) \leq n]$ for given n at a specified value when there is no change. This rule has been shown to be optimal in a minimax sense by Lorden (1971) and Moustakides (1986). However, it may perform poorly if f is misspecified.

Pursuing the likelihood approach when f is known and θ_0 is unknown, the Page-CUSUM procedure generalizes by replacing the Neyman-Pearson LLR, $T_{jk}(f, \Delta)$ for $H_k : \tau \geq k$ against the alternative $H_j^* : \tau = j$ with θ_0 known by the corresponding Wilks' Λ, using maximum likelihood estimators $\hat{\theta}_k$ and $\tilde{\theta}_{jk}$ of the unknown θ_0 based on $X_1, \ldots, X_k$ under H_k and H_j^* respectively. Thus at the kth stage sampling, we calculate

$$W_k(f, \Delta) = \max_{1 \leq j \leq k-1} W_{jk}(f, \Delta), \ k \geq 2, \tag{13.2a}$$

where

$$W_{jk}(f,\Delta)=\sum_{1\le i\le j}\log f(X_i,\tilde{\theta}_{jk})+\sum_{j+1\le i\le k}\log f(X_i,\tilde{\theta}_{jk}+\Delta)-\sum_{1\le i\le k}\log f(X_i,\hat{\theta}_k). \tag{13.2b}$$

For uniformity of notations, we extend the definitions of W_k and W_{jk} in (13.2a) and (13.2b) by letting

$$W_{0k}(f,\Delta)=0 \text{ and } W_k(f,\Delta)=\max_{0\le j<k}W_{jk}(f,\Delta) \text{ for } k\ge 1.$$

The generalized CUSUM stopping rule is thus defined as

$$N_G(c,f,\Delta)=\min\{k\ge 1: W_k(f,\Delta)\ge c\}. \tag{13.2c}$$

In practice, the above rules are often implemented with assumed density g (possibly different from the true f) in 1(a,b,c) or 2(a,b,c), which we shall call the "working density". For a family of densities $\{g(x,\theta),\theta\in R\}$, with $g(x,0)=g(x)$, write score function and the Fisher information as

$$\psi(x;g)=\frac{\partial\log g(x,\theta)}{\partial\theta}\,|_{\theta=0}\ ,\ I(g)=\int\psi^2(x;g)g(x)dx, \tag{13.3}$$

and in particular

$$\psi_L(x;g):=\psi(x;g)=-g'(x)/g(x),\ I_L(g):=I(g) \text{ in } C_L(g), \tag{13.3a}$$

$$\psi_S(x;g):=\psi(x;g)=-1-xg'(x)/g(x),\ I_S(g):=I(g) \text{ in } C_S(g). \tag{13.3b}$$

Then under $P_{g,\infty}$ and for $j=[ns]<k=[nt]$, $0\le s<t\le 1$ and $\Delta=\delta n^{-1/2}$, the building blocks T_{jk} of N_P and W_{jk} of N_G have the following approximations as $n\to\infty$, subject to some Cramér-type regularity conditions (see Bhattacharya and Zhou (1996)):

$$T_{jk}(g,\Delta)=\Delta\hat{T}_{jk}(g,\Delta)+o_P(1),\ W_{jk}(g,\Delta)=\Delta\hat{W}_{jk}(g,\Delta)+o_P(1),$$

where

$$\hat{T}_{jk}(g,\Delta)=\sum_{j+1\le i\le k}\psi(X_i;g)-(1/2)(k-j)\Delta I(g) \tag{13.4}$$

and

$$\hat{W}_{jk}(g,\Delta)=\sum_{j+1\le i\le k}\psi(X_i;g)-\frac{k-j}{k}\sum_{1\le i\le k}\psi(X_i;g)-\frac{j}{2k}(k-j)\Delta I(g) \tag{13.5}$$

with $\psi=\psi_L$, $I=I_L$ in $C_L(g)$ and $\psi=\psi_S$, $I=I_S$ in $C_S(g)$.

Now let $s(X_i) = sign(X_i)$, $R_{k:i}$ = the rank of X_i among $(X_1, \ldots, X_k)$, $R^+_{k:i}$ = the rank of $|X_i|$ among $(|X_1|, \ldots, |X_k|)$, and consider the σ-fields

$$\mathcal{F}_k = \sigma\{(s(X_1), \ldots, s(X_k), (R^+_{k:1}, \ldots, R^+_{k:k})\} \text{ and } \mathcal{G}_k = \sigma\{R_{k:1}, \ldots, R_{k:k}\}.$$

To construct nonparametric analogues of N_P and N_G, it seems reasonable to use

$$\bar{T}_{jk}(g, \Delta) = E_{g,\infty}\left[\hat{T}_{jk}(g, \Delta)|\mathcal{F}_k\right] \text{ and } \bar{W}_{jk} = E_{g,\infty}\left[\hat{W}_{jk}(g, \Delta)|\mathcal{G}_k\right]$$

as building blocks. This leads to the signed rank CUSUM stopping rule

$$N^+_R(c, g, \Delta) = \min\left\{k \geq 1 : \max_{0\leq j\leq k-1} \bar{T}_{jk}(g, \Delta) \geq c\right\}$$

to detect changes in Model 1 and the rank CUSUM stopping rule

$$N_R(c, g, \Delta) = \min\left\{k \geq 1 : \max_{0\leq j\leq k-1} \bar{W}_{jk}(g, \Delta) \geq c\right\}$$

to detect changes in Model 2. Explicit formulas for $\bar{T}_{jk}(g, \Delta)$ and $\bar{W}_{jk}(g, \Delta)$ are given in Sect. 13.2.2 in terms of notations introduced in Sect. 13.2.1.

In Sect. 13.3 we present our main results concerning the asymptotics of $N^+_R(c, g, \Delta_n)$ and $N_R(c, g, \Delta_n)$ over the first n observations under $P_n = P_{f,\infty}$ (no change) and $Q_n = P_{f,\tau_n,\Delta_n}$ (or $Q_{1n} = P_{1f,\tau_n,\Delta_n}$) with $\tau_n = [n\lambda]$ for some $0 < \lambda < 1$ and $\Delta_n = \delta n^{-1/2}$, $\delta > 0$. These asymptotics are described in Theorems 1 and 2 in terms of weak convergence properties of suitably normalized versions of doubly-indexed processes $\left\{\bar{T}_{jk}(g, \Delta), 0 \leq j < k \leq n\right\}$ and $\left\{\bar{W}_{jk}(g, \Delta), 0 \leq j < k \leq n\right\}$. Comparing these results with the weak limits of $\{T_{jk}(g, \Delta)\}$ and $\{W_{jk}(g, \Delta)\}$ obtained by Bhattacharya and Zhou (1996), we see that the nonparametric rules have the same asymptotic behaviors as their parametric counterparts under P_n as well as Q_n (or Q_{1n}) if $g = f$. This parallels the Chernoff and Savage (1958) result in fixed sample hypothesis testing. Moreover, due to the distribution-free property of ranks in the null case, false alarm rates of N^+_R and N_R are unaffected by misspecification (i.e., when $g \neq f$), while these rates for N_P and N_G may be quite adversely affected when $g \neq f$. The weak convergence results also show how the drift terms which set in when a change occurs, and drive the underlying stochastic processes towards the decision boundary, slow down under model misspecification for all procedures. We prove the two theorems by establishing the convergence of finite-dimensional distributions in Sect. 13.4 and tightness in Sect. 13.5 of normalized versions of $\{\bar{T}_{jk}(g, \Delta)\}$ and $\{\bar{W}_{jk}(g, \Delta)\}$. Convergence of finite-dimensional distributions is established by the usual Hájek-projection technique. However, the proof of tightness requires derivation of some fluctuation inequalities for doubly-indexed rank sums, using intricate martingale properties of these processes which are believed to be new in the literature (Lemma 8).

Asymptotic properties of $N_R^+(c, g, \Delta)$ were stated in Bhattacharya and Zhou (1994) without proof together with a simulation study comparing N_R^+ with N_P.

13.2 Construction of Nonparametric Rules to Detect Change

13.2.1 Preliminaries

Throughout the paper we assume that the true density f and the working density g satisfy the following condition.

Condition C. (a) In $C_L(f)$ and $C_L(g)$, f and g are absolutely continuous, f' and g' are integrable, and the Fisher informations $I_L(f)$, $I_L(g)$ are positive and finite. (b) In $C_S(f)$ and $C_S(g)$, f and g are absolutely continuous, xf' and xg' are integrable, and the Fisher informations $I_S(f)$, $I_S(g)$ are positive and finite.

The following notations are standard in the theory of ranks (see Hájek and Šidák (1967)).

Let $U_{k:1} < \cdots < U_{k:k}$ denote the order statistics in a random sample $(U_1, \ldots, U_k)$ from Uniform(0,1) and let F, G denote the distribution functions of f, g respectively. Under $P_{g,\infty}$, the score functions given by 13.3(a,b) are

$$\psi_L(X_i; g) \stackrel{\mathcal{D}}{=} -\frac{g'oG^{-1}}{goG^{-1}}(U_i) := \phi(U_i; g) \text{ in } C_L(g), \tag{13.6}$$

$$\psi_S(X_i; g) \stackrel{\mathcal{D}}{=} -1 - G^{-1}(U_i)\frac{g'oG^{-1}}{goG^{-1}}(U_i) := \phi_1(U_i; g) \text{ in } C_S(g). \tag{13.7}$$

Moreover, with a symmetric g, $\psi_L(X_i; g) = s(X_i)\psi_L(|X_i|; g)$, and

$$\psi_L(|X_i|; g) \stackrel{\mathcal{D}}{=} -\frac{g'oG^{-1}}{goG^{-1}}(1/2 + U_i/2) := \phi^+(U_i; g). \tag{13.8}$$

For a square-integrable function ϕ on $[0, 1]$ with $\int_0^1 \phi(u)du = 0$ and $\int_0^1 \phi^2(u)du = ||\phi||^2$, let

$$a_k(i, \phi) = E[\phi(U_{k:i})],\ A_k^2 = k^{-1}\sum_{1\le i\le k} a_k^2(i, \phi). \tag{13.9}$$

Then $\sum_{1\le i\le k} a_k(i, \phi) = 0$ and $\lim_{k\to\infty} A_k^2 = ||\phi||^2$. Under Condition C, the functions $\phi(\cdot; g)$, $\phi_1(\cdot; g)$ and $\phi^+(\cdot; g)$ defined by (13.6), (13.7) and (13.8) are square-integrable,

$$\int_0^1 \phi(u; g)du = \int_0^1 \phi_1(u; g)du = 0, \tag{13.10}$$
$$||\phi(\cdot; g)||^2 = ||\phi^+(\cdot; g)||^2 = I_L(g),\ ||\phi_1(\cdot; g)||^2 = I_S(g),$$

and specializing (13.9) to these functions, we define $a_k(i, g)$, $a_{1k}(i, g)$, and $a_k^+(i, g)$ as means of $\phi(U_{k:i}; g)$, $\phi_1(U_{k:i}; g)$ and $\phi^+(U_{k:i}; g)$ respectively. Clearly, (13.6), (13.7), (13.8) and (13.10) also hold for f and we define $a_k(i, f)$, $a_{1k}(i, f)$ and $a_k^+(i, f)$ analogously.

13.2.2 The Stopping Rules N_R^+ and N_R

In the following two lemmas, we obtain explicit formulas for the building blocks $\bar{T}_{jk}$ of N_R^+ and $\bar{W}_{jk}$ of N_R. These formulas allow us to do the computations for the stopping rule N_R^+ in terms of $s(X_i)$ and $R_{k:i}^+$ and for N_R in terms of $R_{k:i}$.

Lemma 1 *In $C_L(g)$, with a symmetric g satisfying Condition C,*

$$\bar{T}_{jk}(g, \Delta) = \sum_{j+1 \le i \le k} s(X_i) a_k^+(R_{k:i}^+, g) - (1/2)(k - j)\Delta I_L(g).$$

Proof For a symmetric g, the signs $s(X_i)$, the ranks $R_{k:i}^+$ and the order statistics $|X_{k:i}|$ are mutually independent under $P_{g,\infty}$. Now use (13.4) and (13.8) and the definition of a_k^+. □

Lemma 2 *For arbitrary g satisfying Condition C,*

$$\bar{W}_{jk}(g, \Delta) = \begin{cases} \sum_{j+1 \le i \le k} a_k(R_{k:i}, g) - \frac{j}{2k}(k - j)\Delta I_L(g), & \text{in } C_L(g), \\ \sum_{j+1 \le i \le k} a_{1k}(R_{k:i}, g) - \frac{j}{2k}(k - j)\Delta I_S(g), & \text{in } C_S(g). \end{cases}$$

Proof For arbitrary g ,the ranks $R_{k:i}$ and the order statistics $X_{k:i}$ are independent under $P_{g,\infty}$. Now use (13.5), (13.6) and (13.7) and the definition of a_k in $C_L(g)$ (and of a_{1k} in $C_S(g)$) and because $\sum_{1 \le i \le k} a_k(R_{k:i}, g) = 0$ (similarly for a_{1k}). □

13.3 Main Results

13.3.1 Weak Convergence of $\{\bar{T}_{jk}\}$ and $\{\bar{W}_{jk}\}$

For $\Delta_n = \delta n^{-1/2}$, $j = [ns]$, $k = [nt]$, $0 \le s < t \le 1$, dividing by $\{nI_L(g)\}^{1/2}$ or $\{nI_S(g)\}^{1/2}$ as appropriate, and letting

$$\xi_n(s,t;g) = n^{-1/2} \sum_{[ns]+1 \le i \le [nt]} s(X_i) a^+_{[nt]} \left(R^+_{[nt]:i}, g \right) / ||\phi^+(\cdot; g)||, \tag{13.11a}$$

$$\eta_n(s,t;g) = n^{-1/2} \sum_{[ns]+1 \le i \le [nt]} a_{[nt]} \left(R_{[nt]:i}, g \right) / ||\phi(\cdot; g)||, \tag{13.11b}$$

$$\eta_{1n}(s,t;g) = n^{-1/2} \sum_{[ns]+1 \le i \le [nt]} a_{1,[nt]} \left(R_{[nt]:i}, g \right) / ||\phi_1(\cdot; g)||, \tag{13.11c}$$

the normalized versions of $\{\bar{T}_{jk}\}$ and $\{\bar{W}_{jk}\}$ in $C_L(g)$ and $C_S(g)$ are respectively defined as

$$Y^+_n(s,t;g,\delta) = \xi_n(s,t;g) - (1/2)(t-s)\delta I_L(g)^{1/2} + o(1), \tag{13.12a}$$

$$Y_n(s,t;g,\delta) = \eta_n(s,t;g) - (1/2)st^{-1}(t-s)\delta I_L(g)^{1/2} + o(1), \tag{13.12b}$$

$$Y_{1n}(s,t;g,\delta) = \eta_{1n}(s,t;g) - (1/2)st^{-1}(t-s)\delta I_S(g)^{1/2} + o(1), \tag{13.12c}$$

on $(s,t) \in \Gamma = \{(s,t) : 0 \le s \le t \le 1\}$, where ξ_n, η_n, η_{1n} and Y^+_n, Y_n, Y_{1n} are all 0 for $s = t$. The $o(1)$ terms in 13.12(a,b,c) are uniform, because they represent the discrepancy due to treating ns, nt as integers, which we shall do all through.

We now describe the weak convergence properties of these processes in the following two theorems. In what follows, $\{B(t) : t \ge 0\}$ is a standard Brownian motion.

Theorem 1 *Suppose that f and g are symmetric densities, satisfying Condition C in $C_L(f)$ and $C_L(g)$. Then $\left\{Y^+_n(s,t;g,\delta), (s,t) \in \Gamma\right\}$ converges weakly to*

(a) $\left\{Y^+(s,t;g,\delta) = B(t) - B(s) - (1/2)\delta I_L(g)^{1/2}(t-s),\ (s,t) \in \Gamma\right\}$ *under $P_n = P_{f,\infty}$, and to*

(b) $\left\{Y^+(s,t;g,\delta) + \delta\rho^+(f,g)I_L(f)^{1/2}\alpha(s,t),\ (s,t) \in \Gamma\right\}$ *under $Q_n = P_{f,\tau_n,\Delta_n}$, where $\rho^+(f,g)$ is correlation between $\phi^+(\cdot; f)$ and $\phi^+(\cdot; g)$ and $\alpha(s,t) = (t-\lambda)^+ - (s-t)^+$.*

Theorem 2 *Suppose f, g are densities, satisfying Condition C in $C_L(f)$ and $C_L(g)$ or in $C_S(f)$ and $C_S(g)$. Then*

I. $\{Y_n(s,t;g,\delta), (s,t) \in \Gamma\}$ *converges weakly to*

(a) $\left\{Y(s,t;g,\delta) = st^{-1}B(t) - B(s) - (1/2)\delta I_L(g)^{1/2}st^{-1}(t-s), (s,t) \in \Gamma\right\}$ *under $P_n = P_{f,\infty}$, and to*

(b) $\{Y(s,t;g,\delta)+\delta\rho(f,g)I_L(f)^{1/2}\beta(s,t),\ (s,t)\in\Gamma\}$ *under* $Q_n = P_{f,\tau_n,\Delta_n}$, *where* $\rho(f,g)$ *is correlation between* $\phi(\cdot;f)$ *and* $\phi(\cdot;g)$ *and* $\beta(s,t) = st^{-1}(t-\lambda)^+ - (s-t)^+$.

II. *The weak limits of* $\{Y_{1n}(s,t;g,\delta),(s,t)\in\Gamma\}$ *under* P_n *and* $Q_{1n} = P_{1f,\tau_n,\Delta_n}$ *are obtained by substituting* $I_S(g)$, $I_S(f)$ *and* $\rho_1(f,g)$ =*correlation between* $\phi_1(\cdot;f)$ *and* $\phi_1(\cdot;g)$ *for* $I_L(g)$, $I_L(f)$ *and* $\rho(f,g)$ *respectively in the limits of* $\{Y_n(s,t;g,\delta)\}$ *stated in* $I(a,b)$.

By simple algebraic rearrangement it can be seen that the stopping time $N_n = N_R^+(c,g,\Delta)$ in model 1 with (i) $j=[ns]$, $k=[nt]$, (ii) $\Delta=\Delta_n=\delta n^{-1/2}$, (iii) $c=c_n=d\{nI_L(g)\}^{1/2}$ can be written as:

$$N_n = \min\{1\le[nt]\le n : \sup_{0\le s\le t} Y_n^+(s,t;g,\delta)\ge d\}.$$

Hence

$$P[N_n\le[nx]] = P[\sup_{0\le t\le x}\sup_{0\le s\le t} Y_n^+(s,t;g,\delta)\ge d]$$

for all $0\le x\le 1$. By Theorem 1(a) for $\lambda=1$ and Theorem 1(b) for $\lambda<1$, and the Continuous Mapping Theorem, it follows that in the limit as $n\to\infty$, this probability is obtained by replacing Y_n^+ by Y^+ in the above expression if d is a continuity point of the limit distribution. The limiting distribution of Y_n obtained in Theorem 2 serves the same purpose.

The above comment regarding the actual use of these theorems is added at the suggestion of a referee who also pointed out some typographic errors. We extend our thanks to the referee.

13.3.2 Comparison Between the Parametric and Nonparametric Stopping Rules

Similar to the normalization of $\{\bar{T}_{jk}\}$ and $\{\bar{W}_{jk}\}$ for $(s,t)\in\Gamma$, define

$$Z_n^0(s,t;g,\delta) = T_{[ns][nt]}\left(g,\delta n^{-1/2}\right)/\{nI_L(g)^{1/2}\}$$

for a symmetric g, and

$$Z_n(s,t;g,\delta) = W_{[ns][nt]}\left(g,\delta n^{-1/2}\right)/\{nI_L(g)^{1/2}\},$$
$$Z_{1n}(s,t;g,\delta) = W_{[ns][nt]}\left(g,\delta n^{-1/2}\right)/\{nI_S(g)^{1/2}\},$$

in $C_L(g)$ and $C_S(g)$ respectively for an arbitrary g. The weak limits of $\{Z_n^0\}$, $\{Z_n\}$ and $\{Z_{1n}\}$ which describe the asymptotic behaviors of N_P and N_G have been derived by Bhattacharya and Zhou (1996). Comparing their results with those given in Theorems

1 and 2, we see that for $g = f$ (i.e., when the working density is the same as the true density), the weak limits of $\{Z_n^0\}$, $\{Z_n\}$ and $\{Z_{1n}\}$ are the same as the weak limits of $\{Y_n^+\}$, $\{Y_n\}$ and $\{Y_{1n}\}$ respectively, under P_n as well as Q_n (or Q_{1n}). This parallels the results of Chernoff and Savage (1958) in fixed sample hypothesis testing.

Next consider the case of $g \neq f$. Here, under P_n, the weak limits of $\{Y_n^+\}$, $\{Y_n\}$ and $\{Y_{1n}\}$ remain the same as in the case of $g = f$ except for $I_L(f)^{1/2}$ (or $I_S(f)^{1/2}$) being replaced by $I_L(g)^{1/2}$ (or $I_S(g)^{1/2}$), but the weak limits of $\{Z_n^0\}$, $\{Z_n\}$ and $\{Z_{1n}\}$ depend on the unknown f. Thus the false alarm rates of N_R^+ and N_R are distribution-free, but the false alarm rates of N_P and N_G are vulnerable to model misspecification.

Under contiguous change, the deterministic components in part (b) of Theorem 1 and 2, which drive the respective processes towards the decision boundary, get weakened by the factors $\rho^+(f, g)$ in Model 1 and $\rho(f, g)$ (or $\rho_1(f, g)$) in the location (or scale) problem in Model 2 if $g \neq f$. Although these factors are less than 1, they may still attain reasonable levels unless g is drastically different from f. For a measure of distance between f and g which is relevant in the present context, see Hájek and Šidák (1967), page 22. Note that

(i) $\rho^+(f, g) > 0$ if f and g are symmetric and unimodal,

(ii) $\rho(f, g) > 0$ if f and g are strongly unimodal (i.e., f'/f, g'/g are non-increasing),

(iii) $\rho_1(f, g) > 0$ if $xf'(x)/f(x)$, $xg'(x)/g(x)$ are non-increasing.

The linear drifts N_R^+ and N_R in Theorems 1(b) and 2(b), are improvements upon the adhoc stopping rules based on cumulative sums of sequential ranks (or their scores) considered by Bhattacharya and Frierson (1981) and others, which are driven towards the decision boundary by logarithmic drifts after change.

The behaviors of $\{Z_n^0\}$, $\{Z_n\}$ and $\{Z_{1n}\}$ under contiguous change when $g \neq f$, are quite complicated. However, here also the linear drift terms which set in after change, slow down due to misspecification of model by factors which are correlations between certain scores. See Bhattacharya and Zhou (1996) for details.

13.3.3 Overview of Proofs of Theorems 1 and 2

Let p_n, q_n and q_{1n} denote the joint densities of $(X_1, \ldots, X_n)$ under P_n, Q_n and Q_{1n} respectively. Consider the likelihood ratios $L_n = q_n/p_n, L_{1n} = q_{1n}/p_n$. Then $\log L_n$ and $\log L_{1n}$ are the same as $T_{[n\lambda],n}(f, \delta n^{-1/2})$ in $C_L(f)$ and $C_S(f)$ respectively. Using (13.4), we then see that under P_n

$$\log L_n = \delta n^{-1/2} \sum_{[n\lambda]+1 \leq i \leq n} \psi_L(X_i; f) - (1/2)(1-\lambda)\delta^2 I_L(f) + o_P(1), \tag{13.13}$$

$$\log L_{1n} = \delta n^{-1/2} \sum_{[n\lambda]+1 \leq i \leq n} \psi_S(X_i; f) - (1/2)(1-\lambda)\delta^2 I_S(f) + o_P(1). \tag{13.14}$$

Thus under P_n, $\log L_n \xrightarrow{\mathcal{L}} N\left(-\sigma_L^2/2, \sigma_L^2\right)$ where $\sigma_L^2 = (1-\lambda)\delta^2 I_L(f)$ and $\log L_{1n} \xrightarrow{\mathcal{L}} N\left(-\sigma_S^2/2, \sigma_S^2\right)$ where $\sigma_S^2 = (1-\lambda)\delta^2 I_S(f)$. Consequently, both $\{Q_n\}$ and $\{Q_{1n}\}$ are contiguous to $\{P_n\}$. In view of this, the proofs of Theorems 1 and 2 will proceed as follows.

We shall work with $a_k(i) = E\left[\phi(U_{k:i})\right]$ for arbitrary square-integrable function ϕ on $[0, 1]$, and the process

$$\left\{\xi_n(s,t) = n^{-1/2} \sum_{[ns]+1 \le i \le [nt]} s(X_i) a_{[nt]}\left(R^+_{[nt]:i}\right)/||\phi||,\ (s,t) \in \Gamma\right\} \tag{13.15}$$

and

$$\left\{\eta_n(s,t) = n^{-1/2} \sum_{[ns]+1 \le i \le [nt]} a_{[nt]}\left(R_{[nt]:i}\right)/||\phi||,\ (s,t) \in \Gamma,\right\} \tag{13.16}$$

with the additional requirement $\int_0^1 \phi(u)du = 0$ in (13.16).

Approximating $\{\xi_n(s,t)\}, \{\eta_n(s,t)\}$ by normalized sums of iid random variables, finite-dimensional limit laws of $\{\xi_n\}, \{\eta_n\}$ will be derived first under P_n, and then under Q_n (or Q_{1n}) using (13.13) and (13.14) and LeCam's Third Lemma. This will establish the following two theorems in Sect. 13.4.

Theorem 1 *The finite-dimensional distributions of* $\{\xi_n(s,t)\}$ *converge to those of*
(a) $\{B(t) - B(s)\}$ *under* P_n, *and*
(b) $\left\{B(t) - B(s) + \delta||\phi||^{-1}\langle\phi, \phi^+(\cdot; f\rangle\alpha(s,t)\right\}$ *under* Q_n, *where* $\alpha(s,t) = (t-\lambda)^+ - (s-\lambda)^+$.

Theorem 2 *The finite-dimensional distributions of* $\{\eta_n(s,t)\}$ *converge to those of*
(a) $\{st^{-1}B(t) - B(s)\}$ *under* P_n, *and*
(b) $\left\{st^{-1}B(t) - B(s) + \delta||\phi||^{-1}\langle\phi, \phi(\cdot; f\rangle\beta(s,t)\right\}$ *under* Q_n *and the same expression with* $\phi(\cdot; f)$ *replaced by* $\phi_1(\cdot; f)$ *under* Q_{1n}, *where* $\beta(s,t) = st^{-1}(t-\lambda)^+ - (s-\lambda)^+$.

In Sect. 13.5, we shall establish tightness of $\{\xi_n(s,t)\}, \{\eta_n(s,t)\}$ under P_n, and then tightness under Q_n and Q_{1n} will follow by contiguity, thus yielding

Theorem 3 $\{\xi_n(s,t)\}$ *is tight under* P_n *and* Q_n.

Theorem 4 $\{\eta_n(s,t)\}$ *is tight under* P_n, Q_n *and* Q_{1n}.

Specializing Theorems 1 and 3 to $a_k(i) = E\left[\phi^+(U_{k:i}; g)\right]$, Theorem 1 will follow and specializing Theorems 2 and 4 to $a_k(i) = E\left[\phi(U_{k:i}; g)\right]$ and $a_k(i) =$

$E\,[\phi_1(U_{k:i};\,g)]$, Theorem 2 will follow. Therefore, the main task is to prove Theorems 1, 2, 3 and 4 in the general setting of $a_k(i) = E\,[\phi(U_{k:i})]$ with arbitrary square-integrable ϕ satisfying the additional requirement of $\int_0^1 \phi(u)du = 0$ in Theorems 2 and 4. This will be done in the rest of the paper.

13.4 Convergence of Finite-Dimensional Distributions—Proofs of Theorems 1 and 2

13.4.1 Finite-Dimensional Distributions of $\{\xi_n(s,\,t)\}$

Let $S_{jk}^{+} = \sum_{j+1\le i\le k} s(X_i)a_k\left(R_{k:i}^{+}\right)$ for $0 \le j \le k-1$ and $S_{kk}^{+} = 0$. Then $\xi_n(s,t) = S_{ns,nt}^{+}/(n^{1/2}||\phi||)$. Now let

$$\mathcal{F}_{jk} = \sigma\{(s(X_{j+1}),\ldots,s(X_k),(R_{k:j+1}^{+},\ldots,R_{k:k}^{+})\},\ 0 \le j \le k-1,$$

and let $\mathcal{F}_{kk} = \{\emptyset,\,\Omega\}$ denote the trivial σ-field. In particular,

$$\mathcal{F}_{0k} = \sigma\{(s(X_1),\ldots,s(X_k),(R_{k:1}^{+},\ldots,R_{k:k}^{+})\} = \mathcal{F}_k$$

in the context of Lemma 1. Note that S_{jk}^{+} is $\mathcal{F}_{jk}$-measurable and $\mathcal{F}_{j-1,k} \supset \mathcal{F}_{jk} \subset \mathcal{F}_{j,k+1}$. Now let $F_{+}(x) = 2F(x) - 1$ denote the distribution function of $|X_1|$ having a symmetric density f, and define

$$T_{jk}^{*} = \sum_{j+1\le i\le k} s(X_i)\phi\left(F^{+}(|X_i|)\right),\ 0 \le j \le k-1 \text{ and } T_{kk}^{*} = 0.$$

Then we have the following Lemma.

Lemma 3 *Under $P_n = P_{f,\infty}$ with symmetric f, the following hold:*

(a) $E\left[T_{jk}^{*}|\mathcal{F}_{0k}\right] = E\left[T_{jk}^{*}|\mathcal{F}_{jk}\right] = S_{jk}^{+}$.

(b) $E\left[T_{jk}^{*}\right] = E\left[S_{jk}^{+}\right] = 0,\ E\left[T_{jk}^{*2}\right] = (k-j)||\phi||^2,$

$E\left[S_{jk}^{+2}\right] = (k-j)A_k^2,$

$E\left[\left(T_{jk}^{*} - S_{jk}^{+}\right)^2\right] = (k-j)\left(||\phi||^2 - A_k^2\right),$

where $A_k^2 = k^{-1}\sum_{1\le i\le k} a_k(i)^2 = k^{-1}\sum_{1\le i\le k}\{E\,[\phi(U_{k:i})]\}^2$.

(c) $E\left[S^+_{j,k+1}|\mathcal{F}_{0k}\right] = E\left[S^+_{j,k+1}|\mathcal{F}_{jk}\right] = S^+_{jk} = E\left[S^+_{j-1,k}|\mathcal{F}_{jk}\right]$, *i.e.,*

$\left\{\left(S^+_{jk}, \mathcal{F}_{0k}\right),\ k \geq j\right\}$ *is a martingale for a fixed* j *and*

$\left\{\left(S^+_{jk}, \mathcal{F}_{jk}\right),\ 0 \leq j \leq k\right\}$ *is a reverse martingale for fixed* k.

(d) For $a \leq c \leq d \leq b$, $E\left[\left(S^+_{ab} - S^+_{cd}\right)^2\right] = (b-a)A_b^2 - (d-c)A_d^2$.

Proof of Lemma 3 Argue as in Lemma 1 to see that $E\left[T^*_{jk}|\mathcal{F}_{0k}\right] = S^+_{jk}$. Since S^+_{jk} is $\mathcal{F}_{jk}$-measurable and $\mathcal{F}_{jk} \subset \mathcal{F}_{0k}$ we now have

$$E\left[T^*_{jk}|\mathcal{F}_{jk}\right] = E\left[E\left[T^*_{jk}|\mathcal{F}_{0k}\right]|\mathcal{F}_{jk}\right] = E\left[S^+_{jk}|\mathcal{F}_{jk}\right] = S^+_{jk},$$

proving (a).

The first three parts of (b) follow because $s(X_i)$ is independent of $|X_i|$, has mean 0 and variance 1 and $A_k^2 = k^{-1}\sum_{1\leq i\leq k} a_k(i)^2$. Now use the proven parts of the lemma to prove the rest of (b) by conditional expectation argument.

Next, for fixed j, write $T^*_{j,k+1} = T^*_{jk} + s(X_{k+1})\phi\left(F_+\left(|X_{k+1}|\right)\right)$ and note that $s(X_{k+1})$ and $|X_{k+1}|$ are mutually independent and independent of $\mathcal{F}_{0k}$, S^+_{jk} is $\mathcal{F}_{jk}$-measurable and $\mathcal{F}_{jk} \subset \mathcal{F}_{0k} \subset \mathcal{F}_{0,k+1}$ for all $k \geq j$. Thus

$$\begin{aligned}
E\left[S^+_{j,k+1}|\mathcal{F}_{0k}\right] &= E\left[E\left[T^*_{j,k+1}|\mathcal{F}_{0,k+1}\right]|\mathcal{F}_{0k}\right] = E\left[T^*_{j,k+1}|\mathcal{F}_{0k}\right] \\
&= E\left[T^*_{jk}|\mathcal{F}_{0k}\right] + E\left[s(X_{k+1})\phi\left(F_+\left(|X_{k+1}|\right)\right)|\mathcal{F}_{0k}\right] = S^+_{jk} + 0 = S^+_{jk}, \\
E\left[S^+_{j,k+1}|\mathcal{F}_{jk}\right] &= E\left[E\left[S^*_{j,k+1}|\mathcal{F}_{0k}\right]|\mathcal{F}_{jk}\right] = E\left[S^*_{jk}|\mathcal{F}_{jk}\right] = S^+_{jk},
\end{aligned}$$

while for fixed k, writing $S^+_{j-1,k} = S^+_{jk} + s(X_j)a_k\left(R^+_{k:j}\right)$, the remaining part of (c) follows, because $s(X_j)$ is independent of $R^+_{k:j}$ and $\mathcal{F}_{jk}$.

Finally, let $a \leq c \leq d \leq b$. If $c = d$, then $S^+_{cd} = 0$, so (d) follows from (b). If $c < d$, then $\mathcal{F}_{cd} \subset \mathcal{F}_{ad}$, so by (c),

$$E\left[S^+_{ab}|\mathcal{F}_{cd}\right] = E\left[E\left[S^+_{ab}|\mathcal{F}_{ad}\right]|\mathcal{F}_{cd}\right] = E\left[S^+_{ad}|\mathcal{F}_{cd}\right] = S^+_{cd}.$$

Hence

$$\begin{aligned}
E\left[\left(S^+_{ab} - S^+_{cd}\right)^2\right] &= EE\left[\left\{S^+_{ab} - E\left(S^+_{ab}|\mathcal{F}_{cd}\right)\right\}^2|\mathcal{F}_{cd}\right] \\
&= E\left[E\left(S^{+2}_{ab}|\mathcal{F}_{cd}\right) - E^2\left(S^+_{ab}|\mathcal{F}_{cd}\right)\right] \\
&= E\left[S^{+2}_{ab}\right] - E\left[S^{+2}_{cd}\right] = (b-a)A_b^2 - (d-c)A_d^2
\end{aligned}$$

by (b), proving (d). □

Proof of Theorem 1 First suppose that P_n holds with symmetric f and let

$$V_i = s(X_i)\phi\left(F_+(|X_i|)\right)/||\phi||,\ W_i = \delta\psi_L(X_i; f).$$

Then $\{(V_i, W_i)\}$ is an iid sequence with $E[V_i] = E[W_i] = 0$ and

$$Var[V_i] = 1,\ Var[W_i] = \delta^2 I_L(f),\ Cov[V_i, W_i] = \delta||\phi||^{-1}\langle\phi, \phi^+(\cdot; f)\rangle.$$

By (13.13),

$$\log L_n = n^{-1/2}\sum_{n\lambda+1\le i\le n} W_i - (1/2)(1-\lambda)\delta^2 I_L(f) + o_P(1),$$

and if we let

$$\xi_n^*(s,t) = n^{-1/2}T^*_{ns,nt}/||\phi|| = n^{-1/2}\sum_{ns+1\le i\le nt} V_i$$

for $(s,t)\in\Gamma$, then

$$\xi_n(s,t) = n^{-1/2}S^+_{ns,nt}/||\phi|| = \xi_n^*(s,t) + o_P(1),$$

because of Lemma 3(b) and by virtue of $\lim_{k\to\infty} A_k^2 = ||\phi||^2$,

$$\begin{aligned}\lim_{n\to\infty} E\left[\left\{\xi_n(s,t) - \xi_n^*(s,t)\right\}^2\right] &= \lim_{n\to\infty}\left(n||\phi||^2\right)^{-1} E\left[\left(S^+_{ns,nt} - T^*_{ns,nt}\right)^2\right]\\ &= (t-s)\lim_{n\to\infty}\left(1 - A^2_{nt}/||\phi||^2\right) = 0.\end{aligned}$$

Thus under P_n, for arbitrary $(s_l, t_l)\in\Gamma$, $1\le l\le r$, where $\Gamma = \{(s,t) : 0\le s\le t\le 1\}$,

$$\begin{aligned}&(\log L_n, \xi_n(s_1,t_1), \ldots, \xi_n(s_r,t_r))\\ &= \left(-(1/2)(1-\lambda)\delta^2 I_L(f), 0, \ldots, 0\right) + n^{-1/2}\left(\sum_{n\lambda+1\le i\le n} W_i, \sum_{ns_1+1\le i\le nt_1} V_i, \ldots, \sum_{ns_r+1\le i\le nt_r} V_i\right) + o_P(1)\\ &\xrightarrow{\mathcal{L}} \left(-(1/2)(1-\lambda)\delta^2 I_L(f), 0, \ldots, 0\right) + (\zeta_0, \zeta_1, \ldots, \zeta_r)\end{aligned}$$

as $n\to\infty$, where $(\zeta_0, \zeta_1, \ldots, \zeta_r)$ is a Gaussian random vector with

(i) ζ_0 distributed as $N\left(0, (1-\lambda)\delta^2 I_L(f)\right)$,

(ii) $(\zeta_1, \ldots, \zeta_r)$ distributed as $(B(t_1) - B(s_1), \ldots, B(t_r) - B(s_r))$,

(iii) $Cov[\zeta_0, \zeta_l] = \left\{(t_l-\lambda)^+ - (s_l-\lambda)^+\right\} Cov[W_1, V_1] = \alpha(s_l, t_l)\delta||\phi||^{-1}\langle\phi, \phi^+(\cdot; f)\rangle$.

Thus the finite-dimensional distributions of $\{\xi_n(s,t)\}$ converge in law to those of $\{B(t) - B(s)\}$ under P_n, and by LeCam's Third Lemma, to those of $\left\{B(t) - B(s) + \alpha(s,t)\delta||\phi||^{-1}\langle\phi, \phi^+(\cdot; f)\rangle\right\}$ under Q_n. □

In the above proof, we have only used Lemma 3(b). Parts (c) and (d) of Lemma 3 will be used to control the fluctuations of $\left\{S^+_{jk}\right\}$ for the purpose of proving tightness claimed in Theorem 3.

13.4.2 Finite-Dimensional Distributions of $\{\eta_n(s,t)\}$

Proceeding as in Sect. 13.4.1, let $S_{jk} = \sum_{j+1\le i\le k} a_k(R_{k:i})$ for $0 \le j \le k-1$ and $S_{kk} = 0$. Then $\eta_n(s,t) = n^{-1/2} S_{ns,nt}/||\phi||$ can be approximated by $\eta_n^*(s,t) = n^{-1/2} W^*_{ns,nt}/||\phi||$, where

$$W^*_{jk} = jk^{-1} \sum_{1\le i\le k} \phi(F(X_i)) - \sum_{1\le i\le j} \phi(F(X_i)),\ 1 \le j \le k-1 \text{ and } W^*_{kk} = 0.$$

Slight modifications in the proof of Lemma 3(a,b) lead to

$$E_{f,\infty}\left[\left(W^*_{jk} - S_{jk}\right)^2\right] = j(k-j)\left\{||\phi||^2/k - A_k^2/(k-1)\right\}, \tag{13.17}$$

as will be seen as a special case of Lemma 4(c) with $\mathcal{G}_k = \sigma\{R_{k:1}, \ldots, R_{k:k}\}$. Hence

$$\lim_{n\to\infty} E_{f,\infty}\left[\left\{\eta_n(s,t) - \eta_n^*(s,t)\right\}^2\right] = st^{-1}(t-s) \lim_{n\to\infty}\left(1 - A_{nt}^2/||\phi||^2\right) = 0.$$

Thus $\eta_n(s,t) = \eta_n^*(s,t) + o_P(1)$ under P_n.

Proof of Theorem 2 Argue under P_n in $C_L(f)$ and let

$$V_i = \phi(F(X_i))/||\phi||,\ W_i = \delta\psi_L(X_i; f).$$

Then everything is as in the proof of Theorem 1, the only exception being

$$Cov[V_i, W_i] = \delta||\phi||^{-1}\langle\phi, \phi(\cdot; f)\rangle,$$

$$\eta_n^*(s,t) = n^{-1/2}\left[st^{-1}\sum_{1\le i\le nt} V_i - \sum_{1\le i\le ns} V_i\right],$$

and for for arbitrary $(s_l, t_l) \in \Gamma$, $1 \le l \le r$, as in the proof of Theorem 1,

$$\begin{aligned}
&(\log L_n, \eta_n(s_1,t_1), \ldots, \eta_n(s_r,t_r))\\
&= \left(-(1/2)(1-\lambda)\delta^2 I_L(f), 0, \ldots, 0\right)\\
&+ n^{-1/2}\left(\sum_{n\lambda+1\le i\le n} W_i, s_1t_1^{-1}\sum_{1\le i\le nt_1} V_i - \sum_{1\le i\le ns_1} V_i, \ldots, s_rt_r^{-1}\sum_{1\le i\le nt_r} V_i - \sum_{1\le i\le ns_r} V_i\right)\\
&\xrightarrow{\mathcal{L}} \left(-(1/2)(1-\lambda)\delta^2 I_L(f), 0, \ldots, 0\right) + (\zeta_0, \zeta_1, \ldots, \zeta_r)
\end{aligned}$$

as $n \to \infty$, where $(\zeta_0, \zeta_1, \ldots, \zeta_r)$ is a Gaussian random vector with

(i) ζ_0 distributed as $N\left(0, (1-\lambda)\delta^2 I_L(f)\right)$,

(ii) $(\zeta_1, \ldots, \zeta_r)$ distributed as $\left(s_1t_1^{-1}B(t_1) - B(s_1), \ldots, s_rt_r^{-1}B(t_r) - B(s_r)\right)$,

(iii) $Cov[\zeta_0, \zeta_l] = \left\{ s_l t_l^{-1}(t_l - \lambda)^+ - (s_l - \lambda)^+ \right\} Cov[W_1, V_1] = \beta(s_l, t_l)\delta$ $||\phi||^{-1}\langle \phi, \phi(\cdot; f)\rangle$.

The convergence of finite-dimensional distributions of $\eta_n(s,t)$ under P_n and under Q_m now follows exactly as in the proof of Theorem 1 In the scale family, take $W_i = \delta\psi_S(X_i; f)$ and replace $I_L(f)$ by $I_S(f)$ and $\phi(\cdot; f)$ by $\phi_1(\cdot; f)$. Otherwise, use the same proof. □

13.4.3 Generalization of the Process $\{S_{jk}\}$

Analogue of Lemma 3(b) holds for $\left(S_{jk}, W^*_{jk}\right)$ replacing $\left(S^+_{jk}, T^*_{jk}\right)$ and this enabled us to prove Theorem 2, but there are no simple analogues of Lemma 3(c,d) for $\left\{S_{jk}\right\}$ due the absence of the $s(X_i)$ terms. To prove tightness of $\{\eta_n\}$ we shall rely on a more complicated martingale structure to control the fluctuations of $\left\{S_{jk}\right\}$. To this end, we introduce a more general process, viz.,

$$S_{q:jk} = \sum_{j+1 \le i \le k} a_q(R_{q:i}),\ 0 \le j < k \le q \le n,\ S_{q:kk} = 0, \tag{13.18}$$

and let

$$W^*_{q:jk} = \sum_{j+1 \le i \le k} \phi\left(F(X_i)\right) - (k-j)q^{-1} \sum_{1 \le i \le q} \phi\left(F(X_i)\right),\ 0 \le j < k \le q, \tag{13.19}$$

with $W^*_{q:kk} = 0$.

As before, $\mathcal{G}_q = \sigma\left\{R_{q:1}, \ldots, R_{q:k}\right\}$. Note that for $q = k$, $S_{k:jk} = S_{jk}$, $W^*_{k:jk} = W^*_{jk}$. Then for $(s,t) \in \Gamma$,

$$\eta_n(s,t) = n^{-1/2} S_{nt:ns,nt}/||\phi|| = n^{-1/2} S_{ns,nt}/||\phi||, \tag{13.20}$$

which will be approximated by

$$\eta^*_n(s,t) = n^{-1/2} W^*_{nt:ns,nt}/||\phi|| = n^{-1/2} W^*_{ns,nt}/||\phi||. \tag{13.21}$$

The following two lemmas will be needed to prove tightness of $\{\eta_n(s,t)\}$ in the proof of Theorem 4 which is omitted.

Lemma 4 *Under P_n, the following hold:*

(a) $E\left[W^*_{q:jk} | \mathcal{G}_q\right] = S_{q:jk}$.

(b) $E\left[W^*_{q:jk}\right] = E\left[S_{q:jk}\right] = 0, \qquad E\left[W^{*2}_{q:jk}\right] = q^{-1}(k-j)(q-k+j)||\phi||^2,$
$E\left[S^2_{q:jk}\right] = (q-1)^{-1}(k-j)(q-k+j)A^2_q$.

(c) $E\left[\left(W^*_{q:jk} - S_{q:jk}\right)^2\right] = (k-j)(q-k+j)\left[q^{-1}||\phi||^2 - (q-1)^{-1}A_q^2\right]$.

(d) $E\left[S_{k+1:d(k+1)}|\mathcal{G}_k\right] = S_{k:dk}$, $E\left[S_{k+1:cd}|\mathcal{G}_k\right] = S_{k:cd}$ *and* $E\left[S_{k+1:cd} - S_{d:cd}|\mathcal{G}_k\right] = S_{k:cd} - S_{d:cd}$ *for* $c \le d \le k$.
Consequently, $\{(|S_{k:dk}|, \mathcal{G}_k), k \ge d\}$ *and* $\{(|S_{k:cd} - S_{d:cd}|, \mathcal{G}_k), k \ge d \ge c\}$ *are nonnegative submartingales starting at 0.*

(e) Fix $a \le c \le b$ *and let* $T_k = \sup_{a \le j \le c} |S_{k:jc}|$ *for* $a \le c \le k \le b$. *Then* $\{(T_k, \mathcal{G}_k), c \le k \le b\}$ *is a submartingale.*

(f) Fix $a \le b$ *and let* $T'_k = \sup_{a \le j \le k} |S_{k:jk}|$ *for* $a \le k \le b$. *Then* $\left\{\left(T'_k, \mathcal{G}_k\right), a \le k \le b\right\}$ *is a submartingale.*

Proof of Lemma 4 For $1 \le i \le q$ we have

$$E\left[\phi(F(X_i))|\mathcal{G}_q\right] = a_q(R_{q:i}) \text{ and } \sum_{1 \le i \le q} a_q(R_{q:i}) = qE\left[\phi(U_1)\right] = 0.$$

Also by definition,

$$E\left[\phi^2(F(X_i))\right] = ||\phi||^2 \text{ and } q^{-1}\sum_{1 \le i \le q} a_q^2(i) = A_q^2.$$

Part (a) and all but the last item of part (b) follow immediately from these facts, and

$$E\left[S^2_{q:jk}\right] = E\left[\left\{\sum_{j+1 \le i \le k} a_q(R_{q:i})\right\}^2\right] = (q-1)^{-1}(k-j)(q-k+j)A_q^2$$

by routine simplification. This completes the proof of part (b).

Using (a), part (d) of the lemma follows by routine conditional expectation arguments and because $E\left[|S_{k+1:d(k+1)}||\mathcal{G}_k\right] \ge |E\left[S_{k+1:d(k+1)}|\mathcal{G}_k\right]| = |S_{k:dk}|$. Parts (e) and (f) follow in the same manner using

$$E\left[\max_{a \le j \le k+1} |S_{k+1:j(k+1)}||\mathcal{G}_k\right] \ge \max_{a \le j \le k+1} \left|E\left[S_{k+1:j(k+1)}|\mathcal{G}_k\right]\right| = \max_{a \le j \le k} |S_{k:jk}|.$$

□

Lemma 5 *For* $s \le t_1 \le t_2$, *the following hold, as* $n \to \infty$:

(a) $n^{-1/2}S_{nt_2:ns,nt_1} \xrightarrow{\mathcal{L}} N\left(0, t_2^{-1}(t_1 - s)(t_2 - t_1 + s)||\phi||^2\right)$,

(b) $n^{-1/2}\left[S_{nt_2:ns,nt_1} - S_{nt_1:ns,nt_1}\right] \xrightarrow{\mathcal{L}} N\left(0, (t_1^{-1} - t_2^{-1})(t_1 - s)^2||\phi||^2\right)$.

Proof By Lemma 4(c), $n^{-1}E\left[\left(S_{nt_2:ns,nt_1} - W^*_{nt_2:ns,nt_1}\right)^2\right] \to 0$, and

$$n^{-1/2}W^*_{nt_2:ns,nt_1} \xrightarrow{\mathcal{L}} ||\phi||\left[\{B(t_1) - B(s)\} - t_2^{-1}(t_1 - s)B(t_2)\right],$$

proving (a). The proof of (b) is similar. □

13.5 Tightness-Proofs of Theorems 1 and 3

We shall verify the tightness criterion in Theorem 2 of Wichura (1969).

Remark 1 Consider the spaces D_2 on $[0, 1]^2$ and H_2 on $\Gamma = \{(s, t) : 0 \le s \le t \le 1\}$ of functions which are continuous from above with limits from below, endowed with the topology of uniform convergence. Let $\mathcal{A}$ and $\mathcal{B}$ denote respectively the σ-fields on D_2 and H_2 generated by coordinate mappings. For weak convergence of probability measures on $(D_2, \mathcal{A})$, Wichura (1969) has given conditions which consist of convergence of fdd's and a fluctuation inequality for tightness. On the other hand, the processes Y_n^+, Y_n and Y_{1n} are in H_2, so their weak convergence properties must be examined on $(H_2, \mathcal{B})$. However, if we extend each $y : \Gamma \to R$ to $\bar{y} : D_2 = [0, 1]^2 \to R$ by letting $\bar{y}(s, t) = 0$ on $[0, 1]^2 \backslash \Gamma$, and observe that Y_n^+, Y_n and Y_{1n} are 0 on the diagonal line, then it follows that the extensions $\bar{Y}_n^+$, $\bar{Y}_n$ and $\bar{Y}_{1n}$ are 0 outside Γ, so it is enough to demonstrate the weak convergence of fdd's and validate the fluctuation inequality on Γ to establish weak convergence of $\bar{Y}_n^+$, $\bar{Y}_n$ and $\bar{Y}_{1n}$ on $(D_2, \mathcal{A})$. □

Let

$$B(\delta) = \left\{\left((s, t), (s', t')\right) \in \Gamma \times \Gamma : |s - s'| < \delta, |t - t'| < \delta\right\}.$$

A sequence $\{X_n(s, t), (s, t) \in \Gamma\}$ satisfies this tightness criterion if for arbitrary $\varepsilon, \varepsilon' > 0$, there exist $\delta_0 \in (0, 1)$ and $n(\delta)$ for each $\delta > 0$, such that for $0 < \delta \le \delta_0$ and $n \ge n(\delta)$,

$$P\left[\sup_{B(\delta)} |X_n(s, t) - X_n(s', t')| > \varepsilon\right] < \varepsilon'. \tag{13.22}$$

We shall verify this for $\{\xi_n\}$ under $P_{f,\infty}$ with symmetric f and for $\{\eta_n\}$ under $P_{f,\infty}$ with arbitrary f. The rest will follow by contiguity.

For notational simplicity, treat $1/\delta$ and $n\delta$ as integers and for $1 \le l \le m \le 1/\delta - 1$, let

$$R_{lm}(\delta) = \{(s, t) \in \Gamma : (l - 1)\delta \le s \le (l + 1)\delta, (m - 1)\delta \le t \le (m + 1)\delta\}$$

and

$$P_{lm}(\delta; X_n) = P\left[\sup_{(s,t),(s',t')\in R_{lm}(\delta)} |X_n(s,t) - X_n(s',t')| > \varepsilon\right].$$

Since $((s,t),(s',t')) \in B(\delta)$ implies that both (s,t), (s',t') belong to the same $R_{lm}(\delta)$ for some $l \le m$,

$$\begin{aligned} P\left[\sup_{B(\delta)} |X_n(s,t) - X_n(s',t')| > \varepsilon\right] &\le \sum_{1\le l\le \delta^{-1}-1} \sum_{l\le m\le \delta^{-1}-1} P_{lm}(\delta; X_n) \\ &= \sum_{1\le l\le \delta^{-1}-3} \sum_{l+2\le m\le \delta^{-1}-1} P_{lm}(\delta; X_n) \\ &+ \sum_{1\le l\le \delta^{-1}-2} \sum_{l\le m\le l+1} P_{lm}(\delta; X_n) + P_{\delta^{-1}-1,\delta^{-1}-1}(\delta; X_n), \end{aligned} \tag{13.23}$$

so we need to get appropriate upper bounds for $P_m(\delta; \xi_n)$ and $P_{lm}(\delta; \eta_n)$ for all these (l,m). Of these, the $(\delta^{-1}-3)(\delta^{-1}-2)/2$ terms for $m \ge l+2$ in the first term will be treated differently from the $(2\delta^{-1}-3)$ terms for $m = l, l+1$.

We now use the martingale (and submartingale) properties of $\{S^+_{jk}\}$ proved in Lemma 3(c) and of $\{S_{q:jk}\}$ proved in Lemma 4(d,e,f) to bound $P_{lm}(\delta; \xi_n)$ and $P_{lm}(\delta; \eta_n)$ under P_n to verify (13.22).

13.5.1 Fluctuations of $\{S^+_{jk}\}$ and Proof of Theorem 3

In the following Lemma, we state two maximal inequalities for nonnegative submartingales.

Lemma 6 *Suppose that* $\{(X_n, \mathcal{F}_n), n \ge 1\}$ *is a nonnegative submartingale. Then:*

(a) $P\left[\max_{1\le k\le n} X_k \ge 2t\right] \le P[X_1 \ge t] + t^{-1}E^{1/2}\left[X_n^2\right] P^{1/2}[X_n \ge t]$ *for* $t > 0$.

(b) $E\left[\max_{1\le k\le n} X_k^2\right] \le 4E\left[X_n^2\right]$.

Remark 2 Part (b) of the Lemma is from Doob (1953), p. 317 and part (a) is proved in Bhattacharya (2005). If $X_1 = 0$, then the first term on the right-hand side of (a) is 0. Lemma 6(a) is an extension of the following inequality for martingales. □

Lemma 6(a1) *Suppose that* $\{(X_n, \mathcal{F}_n), n \ge 1\}$ *is a martingale with* $X_1 = 0$. *Then*

$$P\left[\max_{1\le k\le n} |X_k| \ge 2t\right] \le t^{-1}E^{1/2}\left[X_n^2\right] P^{1/2}[|X_n| \ge t] \text{ for } t > 0.$$

See Hall and Heyde (1980) for Lemma 6(a1), which is a weaker version of Brown (1971) inequality

$$P\left[\max_{1\le k\le n}|X_k|\ge 2t\right]\le P[X_n\ge t]+t^{-1}E\left[(|X_n|-2t)\,I(|X_n|\ge 2t)\right]\ \text{for } t>0.$$

From Lemma 6 we now obtain the following.

Lemma 7 *Suppose that P_n holds. Then*

(a) $P\left[\max_{a\le j<k\le b}|S^+_{jk}|\ge 4t\right]\le t^{-3/2}E^{3/4}\left[S^{+2}_{ab}\right]P^{1/4}\left[\left[|S^+_{ab}|\right]\ge t\right]$ *for* $a<b$,

(b) for $a\le c\le d\le b$

$$\begin{aligned}P\left[\max_{a\le j\le c\le d\le k\le b}|S^+_{jk}-S^+_{cd}|\ge 4t\right]&\le t^{-1}E^{1/2}\left[|S^+_{ad}-S^+_{cd}|^2\right]P^{1/2}\left[|S^+_{ad}-S^+_{cd}|\ge t\right]\\&\quad+t^{-1}E^{1/2}\left[|S^+_{ab}-S^+_{cd}|^2\right]P^{1/2}\left[|S^+_{ab}-S^+_{cd}|\ge t\right]\\&\quad+t^{-3/2}E^{3/4}\left[|S^+_{ab}-S^+_{cd}|^2\right]P^{1/4}\left[|S^+_{ad}-S^+_{cd}|\ge t\right].\end{aligned}$$

Proof of Lemma 7 For $a\le k\le b$, let $T_k=\max_{a\le j<k}|S_{jk}|$ and $T_a=0$. By Lemma 3(c), $\{(T_k,\mathcal{F}_{0k}),k\ge a\}$ is a submartingale and $\left\{\left(|S^+_{jb}|,\mathcal{F}_{jb}\right),a\le j\le b\right\}$ is a reverse submartingale, both nonnegative and both starting at 0. Now note that $\max_{a\le j<k\le b}|S^+_{jk}|=\max_{a\le k\le b}T_k$ and use Lemma 6 on $\{T_k\}$ and $\left\{|S^+_{jb}|\right\}$. Thus

$$P\left[\max_{a\le j<k\le b}|S^+_{jk}|\ge 4t\right]\le t^{-3/2}E^{3/4}\left[|S^+_{ab}|^2\right]P^{1/4}\left[|S^+_{ab}|\ge t\right],$$

proving (a). Next note that S^+_{cd} is $\mathcal{F}_{jk}$-measurable and use Lemma 3(c) to obtain

$$E\left[S^+_{j,k+1}-S^+_{cd}|\mathcal{F}_{0k}\right]=S^+_{jk}-S^+_{cd}=E\left[S^+_{j-1,k}-S^+_{cd}|\mathcal{F}_{jk}\right].\tag{13.24}$$

By the first equality in (13.24), $\left\{\left(S^+_{jk}-S^+_{cd},\mathcal{F}_{0k}\right),k\ge d\right\}$ is a martingale for fixed j. As in part (a), it now follows that if we let $T'_k=\max_{a\le j\le c}|S^+_{jk}-S^+_{cd}|$ for $d\le k\le b$, then $\left\{\left(T'_k,\mathcal{F}_{0k}\right),k\ge d\right\}$ is a nonnegative submartingale starting at T'_d. Hence by Lemma 6(a),

$$\begin{aligned}&P\left[\max_{a\le j\le c\le d\le k\le b}|S^+_{jk}-S^+_{cd}|\ge 4t\right]\\&=P\left[\max_{d\le k\le b}T'_k\ge 4t\right]\\&\le P\left[T'_d\ge 2t\right]+(2t)^{-1}E^{1/2}\left[T'^2_b\right]P^{1/2}\left[T'^2_b\ge 2t\right]\\&=P\left[\max_{a\le j\le c}|S^+_{jd}-S^+_{cd}|\ge 2t\right]+(2t)^{-1}E^{1/2}\left[\max_{a\le j\le c}|S^+_{jb}-S^+_{cd}|^2\right]P^{1/2}\left[\max_{a\le j\le c}|S^+_{jb}-S^+_{cd}|\ge 2t\right].\end{aligned}$$

Now use the second inequality in (13.24) for $k = d$ and $k = b$ to see that $\left\{\left(S^+_{jd} - S^+_{cd}, \mathcal{F}_{jd}\right), a \le j \le c\right\}$ is a reverse martingale starting at 0 and $\left\{\left(|S^+_{jb} - S^+_{cd}|, \mathcal{F}_{jb}\right), a \le j \le c\right\}$ is a reverse martingale starting at 0 and $\left\{\left(|S^+_{jb} - S^+_{cd}|, \mathcal{F}_{jb}\right), a \le j \le c\right\}$ is a nonnegative reverse submatringale starting at $|S^+_{cb} - S^+_{cd}|$. This makes Lemma 6(a,a1,b) applicable to all terms in the last expression, leading to the bound claimed in (b). □

This lemma now leads to the following, which provides the tool to establish the tightness of $\{\xi_n(s,t)\}$.

Lemma 8 *Suppose that P_n holds, and let $\delta < 1$.*
(a) Let $(a,b) = ((l-1)n\delta, (l+2)n\delta)$. Then

$$\lim_{n\to\infty} P\left[\max_{a\le j<k\le b} |S^+_{jk}| > n^{1/2}||\phi||\varepsilon\right] \le c(\varepsilon)\delta^{1/2}\left\{1 - \Phi\left(a(\varepsilon)\delta^{-1/2}\right)\right\}^{1/4}.$$

(b) For $m \ge l+2$, let $(a,b) = ((l-1)n\delta, (m+1)n\delta)$ and $(c,d) = ((l+1)n\delta, (m-1)n\delta)$. Then

$$\lim_{n\to\infty} P\left[\max_{a\le j\le c\le d\le k\le b} |S^+_{jk} - S^+_{cd}| > n^{1/2}||\phi||\varepsilon\right] \le c(\varepsilon)\delta^{1/2}\left\{1 - \Phi\left(a(\varepsilon)\delta^{-1/2}\right)\right\}^{1/4}.$$

Here Φ is the standard normal distribution function and $c(\varepsilon)$, $a(\varepsilon)$ are generic constants depending only on $\varepsilon > 0$.

Proof In Lemma 7, take $t = n^{1/2}||\phi||\varepsilon/4$. Since $\xi_n(s,t) \xrightarrow{\mathcal{L}} B(t) - B(s)$ by Theorem 1 and $\lim_{k\to\infty} A_k^2 = ||\phi||^2$, the lemma follows. □

Proof of Theorem 3 We shall show that under P_n, the bound in (13.23) for $\{\xi_n(s,t)\}$ tends to 0 as $n \to \infty$ and then invoke continguity.

Of the $(\delta^{-1}-3)(\delta^{-1}-2)/2$ terms in the first sum, consider the (l,m)-th term and let

$$(a_{lm}, b_{lm}) = ((l-1)n\delta, (m+1)n\delta),\ (c_{lm}, d_{lm}) = ((l+1)n\delta, (m-1)n\delta).$$

By the triangle inequality and Lemma 8(b), we have

$$\begin{aligned} P_{lm}(\delta, \xi_n) &\le 2P\left[\sup_{(s,t)\in R_{lm}(\delta)} |\xi_n(s,t) - \xi_n\left(n^{-1}c_{lm}, n^{-1}d_{lm}\right)| > \varepsilon/2\right] \\ &= 2P\left[\sup_{a_{lm}\le j<c_{lm}\le d_{lm}<k\le b_{lm}} |S^+_{jk} - S_{c_{lm},d_{lm}}| > n^{1/2}||\phi||\varepsilon/2\right] \\ &\le c(\varepsilon)\delta^{1/2}\left\{1 - \Phi\left(a(\varepsilon)\delta^{-1/2}\right)\right\}^{1/4}, \end{aligned}$$

as $n \to \infty$.

Next for $m = l$ and $m = l + 1$, $(s, t) \in R_{lm}(\delta)$ implies $a_l \le ns \le nt \le b_l$, where $(a_l, b_l) = ((l-1)n\delta, (l+2)n\delta)$. Hence

$$\sup_{(s,t),(s',t') \in R_{lm}(\delta)} |\xi_n(s,t) - \xi_n(s',t')| \le 2 \sup_{(s,t) \in R_{lm}(\delta)} |\xi_n(s,t)|$$
$$= 2n^{-1/2}||\phi||^{-1} \sup_{(s,t) \in R_{lm}(\delta)} |S^+_{ns,nt}| \le 2n^{-1/2}||\phi||^{-1} \max_{a_l \le j < k \le b_l} |S^+_{jk}|.$$

Using this and Lemma 8(a) on each of the remaining $(2\delta^{-1} - 3)$ terms in (13.23), we have

$$P_{lm}(\delta, \xi_n) \le P\left[\max_{a_l \le j < k \le b_l} |S^+_{jk}| > n^{1/2}||\phi||\varepsilon/2\right] \le c(\varepsilon)\delta^{1/2}\left\{1 - \Phi\left(a(\varepsilon)\delta^{-1/2}\right)\right\}^{1/4},$$

as $n \to \infty$.

Putting all these together in the bound in (13.23),

$$\lim_{n\to\infty} P\left[\sup_{B(\delta)} |\xi_n(s,t) - \xi_n(s',t')| > \varepsilon\right] \le (\delta^{-2} - \delta^{-1})c(\varepsilon)\delta^{1/2}\left\{1 - \Phi\left(a(\varepsilon)\delta^{-1/2}\right)\right\}^{1/4}.$$

Examining the fluctuations of $\{S_{jk}\}$ and proving Theorem 4 involves a lot more technicalities, which we omit. □

References

Bhattacharya, P.K., Frierson, D., Jr. (1981) A nonparametric control chart for detecting small disorders. *Ann. Statist.* 9, 544-554.

Bhattachrya, P.K., Zhou, Hong (1994) A rank CUSUM procedure for detecting small changes in a symmetric distribution. *Change-Point Problems, IMS Lecture Notes-Monograph Series* 23, 57-65.

Bhattachrya, P.K., Zhou, Hong (1996) A generalized CUSUM procedure for sequential detection of change-point in a parametric family when the initial distribution is unknown. *Sequential. Anal.* 15, 311-325.

Bhattachrya, P.K. (2005) A maximal inequality for nonnegative submartingales. *Statist. Probab. Letters* 72, 11-12.

Brown, B.M. (1971) Martingale central limit theorems. *Ann. Math. Statist.* 42, 59-66.

Chernoff, H., Savage, I.R. (1958) Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist*. 29, 972-994.

Doob, J.L. (1953) *Stochastic Processes*. Wiley.

Hájek, J., Šidák, Z. (1967) *Theory of Rank Tests*. Academic Press.

Hall, P., Heyde, C.C. (1980) *Martingale Limit Theory and Applications*. Academic Press.

Lorden, G. (1971) Procedures for reacting to a change in distributions. *Ann. Math. Statist.* 42, 1897-1908.

Moustakides, G.V. (1986) Optimal stopping times for detecting changes in distributions. *Ann. Statist.* 14, 1379-1387.

Page, E.S. (1954) Continuous inspection schemes. *Biometrika* 41, 100-115.

Wichura, M.J. (1969) Inequalities with applications to the weak convergence of random processes with multi-dimensional time parameters. *Ann. Math. Statist.* 40, 681-687

Change Point Detection with Multivariate Observations Based on Characteristic Functions

14

Zdeněk Hlávka, Marie Hušková and Simos G. Meintanis

14.1 Introduction

When observing a certain random quantity over a given time period, the assumption of time-invariance of the underlying stochastic structure is often made as a benchmark assumption. Although invariance may well hold for a short period of time, it is not a truly realistic assumption when observations are collected over a long horizon. On the contrary, it is expected that institutional changes cause structural breaks in the stochastic properties of certain variables, particularly in the macroeconomic and financial world. Hence, change detection procedures are of undeniable interest. For univariate independent observations there are numerous procedures for change-point detection. These are nicely reviewed in Horváth and Rice (2014). On the other hand, corresponding methods for multivariate and/or dependent observations have not yet been considered as much, and an on-going effort has recently begun in order to extend the existing procedures towards such situations.

Z. Hlávka (✉) · M. Hušková
Faculty of Mathematics and Physics, Department of Statistics, Charles University, Prague, Czech Republic
e-mail: hlavka@karlin.mff.cuni.cz

M. Hušková
e-mail: huskova@karlin.mff.cuni.cz

S.G. Meintanis
Department of Economics, National and Kapodistrian University of Athens, Athens, Greece
e-mail: simosmei@econ.uoa.gr

S.G. Meintanis
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_14

In this paper, we propose change-point detectors for multivariate independent observations, as well as corresponding methods involving observations which are driven by vector autoregressive (VAR) models. The methods make use of characteristic functions (CFs). Apart from other favorable features which will be mentioned along the paper regarding CF-based procedures, an extra reason for using CFs is that with CFs vector observations are linearly projected onto the real line and the resulting statistics may be written in convenient closed-form expressions. This feature of simplicity is particularly important when dealing with multivariate data and it is not always true if one employs classical procedures based on the empirical distribution function. Papers which are closest to the current work, either in terms of the problems being considered, and/or in terms of methodology employed are Hlávka et al. (2012, 2016, 2017); Kirch et al. (2015); Lee et al. (2009) and Selk and Neumeyer (2013).

The remainder of the paper is as follows. In Sect. 14.2, we formulate the null hypothesis and introduce the corresponding criteria with independent observations while in Sect. 14.3 we do the same for VAR observations. In Sect. 14.4, we study the large sample behavior of the new methods. Section 14.5 is devoted to computational aspects and the implementation of the procedures on the basis of suitable resampling techniques. The results of a Monte Carlo study for the finite-sample properties of the methods are presented in Sect. 14.6, along with some empirical applications.

14.2 Change Detection Under Independence

Let $\{\boldsymbol{X}_t,\ t = 1, 2, \ldots, T\}$ be a sequence of independent vectors of dimension d ($d \geq 1$), with corresponding distribution function (DF) denoted by F_t, $1 \leq t \leq T$. Then the classical change-point detection problem is formulated in the following hypotheses:

$$\mathscr{H}_0 : F_t \equiv F_0 \text{ for all } t = 1, \ldots, T, \quad \text{vs.} \quad \mathscr{H}_1 : F_t \equiv F_0,\ t \leq t_0;\ F_t \equiv F^0,\ t > t_0, \tag{14.1}$$

where the DFs F_0 and F^0 ($F_0 \neq F^0$) are considered unknown.

Our approach will be based on the fact that the null hypothesis $\mathscr{H}_0$ in (14.1) is tantamount to accepting the hypothesis

$$\varphi_t \equiv \varphi_0 \text{ for all } t = 1, \ldots, T \tag{14.2}$$

and vice versa, where $\varphi_t(u) := \mathbb{E}(e^{i\boldsymbol{u}'\boldsymbol{X}_t})$ stands for the characteristic function (CF) of $\boldsymbol{X}_t$. Based on this fact, Hušková and Meintanis (2006) develop detectors for the same problem with univariate observations, while Matteson and James (2014) consider multivariate data. In both cases, the resulting procedures were found to have nice asymptotic properties and to be competitive to other methods in finite samples.

The proposed detector involves the quantity

$$\delta_t(\boldsymbol{u}) = \left|\phi_t(\boldsymbol{u}) - \phi^t(\boldsymbol{u})\right|^2, \tag{14.3}$$

where

$$\phi_t(u) = \frac{1}{t}\sum_{\tau=1}^{t} e^{i\boldsymbol{u}'\boldsymbol{X}_\tau}, \quad \phi^t(u) = \frac{1}{T-t}\sum_{\tau=t+1}^{T} e^{i\boldsymbol{u}'\boldsymbol{X}_\tau}, \tag{14.4}$$

are the empirical CFs computed from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$ and $\boldsymbol{X}_{t+1}, \ldots, \boldsymbol{X}_T$, $t = 1, \ldots, T$, respectively. Clearly under $\mathscr{H}_0$ (resp. $\mathscr{H}_1$) in (14.1), $\delta_t(\boldsymbol{u})$ is expected to be "small"(resp. "large"), and this should hold uniformly in the argument $\boldsymbol{u}$. In fact, if the change point t_0 were known then a two-sample test statistic such as those suggested by Hušková and Meintanis (2008) will be appropriate. However, in the present setting t_0 is considered unknown and therefore some extra weighting scheme is needed which will allow detection of early as well as late changes. Both options are made possible through a proper choice of the parameter γ below. Based on these considerations, we propose to reject the null hypothesis $\mathscr{H}_0$ for large values of the detector

$$Q_{T,w}(\gamma) = \max_{1\le t<T}\left(\frac{t(T-t)}{T^2}\right)^{2+\gamma} T\, D_{t,w}, \tag{14.5}$$

where

$$D_{t,w} = \int_{\mathbb{R}^d} \delta_t(\boldsymbol{u}) w(\boldsymbol{u}) d\boldsymbol{u}. \tag{14.6}$$

Here $\gamma \in (-1, 1]$ is a tuning constant and $w(\boldsymbol{u})$ denotes a weight function the choice of which will be discussed later.

14.3 Change Detection in VAR Models

For fixed $p > 0$, assume that we observe $\boldsymbol{X}_t$, $t = 1, \ldots, T$, coming from the VAR(p) model

$$\boldsymbol{X}_t = \sum_{j=1}^{p} \boldsymbol{A}_j \boldsymbol{X}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{14.7}$$

where $\{\boldsymbol{\varepsilon}_t\}$ is a sequence of $(d \times 1)$ i.i.d. random vectors (termed innovations) satisfying $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$, $\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s') = 0$, $t \neq s$. The $(d \times d)$ square matrices $\{\boldsymbol{A}_j\}_{j=1}^p$ contain the unknown coefficients, and we assume the usual stability condition $\det(\mathbb{I}_d - \sum_{j=1}^p \boldsymbol{A}_j z^j) \neq 0$, $|z| \le 1$, with $\mathbb{I}_d$ denoting the identity matrix of dimension $(d \times d)$.

Equation (14.7) expresses the typical VAR model of fixed order, whereby the basic ingredients are assumed to be time-invariant. Nevertheless we may consider several

kinds of departures from (14.7). The most popular ones are changes in the parameters $\boldsymbol{A}_j$, the possibility of additive or innovation outliers, as well as breaks in the correlation structure of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ or time-varying volatility; see, among others, Lanne et al. (2010), Herwartz and Lütkepohl (2014), and Lütkepohl (2012). A further aspect, which has not been studied as much, is a possible break due to a change in the shape of the conditional distribution of the innovations. Specifically the innovations $\{\boldsymbol{\varepsilon}_t\}$ are typically assumed to be normally distributed which corresponds to the classical Gaussian VAR model. However, from the time of Mandelbrot (1963) and Fama (1965) there is strong evidence that the distribution of economic data, particularly in the financial world, could be heavy-tailed and possibly asymmetric, which makes the normality assumption unrealistic. In this connection and although in univariate autoregressions non-Gaussian innovations have been considered by several authors (see for instance, Hannan and Kanter 1977, Brockwell and Davis 1992, Davis 1996, Tiku et al. 2000, and Andrews et al. 2009), the corresponding literature with multivariate models is rather poor; exceptions include Siegfried (2002) and Lanne and Lütkepohl (2010) who consider non-Gaussian VAR models.

Earlier work on change point detectors in the context of VAR models includes Bai et al. (1998), Bai (2000), Ng and Vogelsang (2002), Qu and Perron (2007), Dvořák and Prášková (2013), Dvořák (2015, 2016). Here we consider change detectors in VAR models, but our approach deviates from earlier approaches in two basic features: (i) Although we are interested in all kinds of breaks mentioned above, including parameter breaks, our procedures are targeted not on the estimates of the parameters as it is typically the case, but on the resulting residuals. This approach, which is also followed by Hlávka et al. (2012, 2016) and Kirch et al. (2015), enables us to capture arbitrary changes in model (14.7) since any such change will be immediately reflected in the behavior of the resulting residuals. (ii) We employ the empirical CF as our main tool for the reasons already mentioned above.

Motivated by this discussion, we consider the detection problem in (14.1) for model (14.7) where F_t denotes the distribution of $\boldsymbol{\varepsilon}_t$, $t \geq 1$. Given that innovations are unobserved, our test statistic will be based on corresponding residuals

$$\widehat{\boldsymbol{\varepsilon}}_t = \boldsymbol{X}_t - \sum_{j=1}^{p} \widehat{\boldsymbol{A}}_j \boldsymbol{X}_{t-j}, \tag{14.8}$$

where $\widehat{\boldsymbol{A}}_j$, $j = 1, \ldots, p$, are $\sqrt{T}$ consistent estimators of $\boldsymbol{A}_j$, $j = 1, \ldots, p$, resulting from some standard method of estimation such the as the method of OLS, the QMLE or Yule–Walker type estimation (see for instance Hamilton (1994) or Lütkepohl (2005)). For the purpose of estimation, we moreover suppose that a set of starting values $\boldsymbol{X}_{1-p}, \ldots, \boldsymbol{X}_0$, exists. Then the suggested criterion based on $\widehat{Q}_{T,w}(\gamma)$ is given by (14.5) but with $D_{t,w}$ replaced by

$$\widehat{D}_{t,w} = \int_{\mathbb{R}^d} \widehat{\delta}_t(\boldsymbol{u}) w(\boldsymbol{u}) d\boldsymbol{u}, \tag{14.9}$$

where $\widehat{\delta}_t(\boldsymbol{u}) := \left|\widehat{\phi}_t(\boldsymbol{u}) - \widehat{\phi}^t(\boldsymbol{u})\right|^2$ incorporates the empirical CFs

$$\widehat{\phi}_t(u) = \frac{1}{t}\sum_{\tau=1}^{t} e^{i\boldsymbol{u}'\widehat{\boldsymbol{\varepsilon}}_\tau}, \quad \widehat{\phi}^t(u) = \frac{1}{T-t}\sum_{\tau=t+1}^{T} e^{i\boldsymbol{u}'\widehat{\boldsymbol{\varepsilon}}_\tau}, \tag{14.10}$$

computed from $\widehat{\boldsymbol{\varepsilon}}_1, \ldots, \widehat{\boldsymbol{\varepsilon}}_t$ and $\widehat{\boldsymbol{\varepsilon}}_{t+1}, \ldots, \widehat{\boldsymbol{\varepsilon}}_T$, $t = 1, \ldots, T$, respectively.

The test statistic here is similar to that in the independent situation. However the $\boldsymbol{X}_t$'s are being replaced by the residuals $\widehat{\boldsymbol{\varepsilon}}_t$'s defined in (14.8), which brings forward the need for additional assumptions; details are postponed to subsection 14.4.2.

14.4 Asymptotics

14.4.1 Independent Setup

Here we present the results on the limit behavior of $Q_{T,w}(\gamma)$ both under the null as well as under a class of alternatives.

Theorem 1 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be a sequence of independent identically d-dimensional random vectors with finite second moment, let $\gamma \in (-1, 1]$ and let $w(\cdot)$ be a non-negative measurable weight function defined on $\mathbb{R}^d$ such that*

$$w(\boldsymbol{u}) = w(-\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathbb{R}^d, \quad 0 < \int_{\mathbb{R}^d} \|\boldsymbol{u}\|^2 w(\boldsymbol{u}) d\boldsymbol{u} < \infty. \tag{14.11}$$

Then, as $T \to \infty$,

$$Q_{T,w}(\gamma) \xrightarrow{d} \sup_{s\in(0,1)} (s(1-s))^\gamma \int_{\mathbb{R}^d} \big(V(\boldsymbol{u}, s) - sV(\boldsymbol{u}, 1)\big)^2 w(\boldsymbol{u}) d\boldsymbol{u}, \tag{14.12}$$

where $\{V(\boldsymbol{u}, s);\ \boldsymbol{u} \in \mathbb{R}^d,\ s \in (0, 1)\}$ is a Gaussian process with zero mean and covariance structure

$$\operatorname{cov}(V(\boldsymbol{u}_1, s_1), V(\boldsymbol{u}_2, s_2)) = \min(s_1, s_2) C(\boldsymbol{u}_1, \boldsymbol{u}_2),$$

where $C(\boldsymbol{u}_1, \boldsymbol{u}_2) = \operatorname{cov}\big(\cos(\boldsymbol{u}_1'\boldsymbol{X}_1) + \sin(\boldsymbol{u}_1'\boldsymbol{X}_1), \cos(\boldsymbol{u}_2'\boldsymbol{X}_1) + \sin(\boldsymbol{u}_2'\boldsymbol{X}_1)\big)$.

Proof It is postponed to the Appendix. □

Remark 1 The one-dimensional situation is treated in the paper of Hušková and Meintanis (2006), however there the limit distribution is formulated in a different, but equivalent, way. In any case, the explicit form of the distribution of the quantity on the r.h.s. of (14.12) is unknown, and moreover it depends on unknown quantities.

Possible solutions to the approximation of this distribution are (i) to estimate these quantities and then simulate the limit distribution by Monte Carlo or (ii) to apply a proper version of resampling.

Remark 2 The proof of Theorem 1 in the Appendix still holds, and hence the same test criterion can be used, even when the observations are dependent, e.g., if they are α–mixing. In fact, it can be further extended to testing of no change in the joint distribution of the vector $(\boldsymbol{X}_t, \ldots, \boldsymbol{X}_{t+q})'$, for given $q \geq 1$. This is quite straightforward but we do not pursue this here any further as it is technically more complex.

Next, we focus on the behavior of $Q_{T,w}(\gamma)$ under alternatives.
Denote the CFs before and after the change by φ_0 and φ^0, respectively, and let

$$B_0(\boldsymbol{u}) = E\Big(\cos(\boldsymbol{u}'\boldsymbol{X}_t) + \sin(\boldsymbol{u}'\boldsymbol{X}_t)\Big), \quad 1 \leq t \leq t_0,$$
$$B^0(\boldsymbol{u}) = E\Big(\cos(\boldsymbol{u}'\boldsymbol{X}_t) + \sin(\boldsymbol{u}'\boldsymbol{X}_t)\Big), \quad t_0 + 1 \leq t \leq T.$$

Theorem 2 *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be independent d-dimensional random vectors and let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{t_0}$ and $\boldsymbol{X}_{t_0+1}, \ldots, \boldsymbol{X}_T$ have CF φ_0 and φ^0, respectively. Let assumption (14.11) on the weight function $w(\cdot)$ be satisfied and assume that $t_0 = \lfloor T s_0 \rfloor$, for some $s_0 \in (0, 1)$. Then, as $T \to \infty$,*

$$\frac{(s(1-s))^2}{T} D_{\lfloor Ts \rfloor, w} \xrightarrow{P} (\min(s, s_0)(1 - \max(s, s_0)))^2 \int_{\mathbb{R}^d} \big(B_0(\boldsymbol{u}) - B^0(\boldsymbol{u})\big)^2 w(\boldsymbol{u}) d\boldsymbol{u} \quad (14.13)$$

for $s \in (0, 1)$.

From (14.13) it follows that $T \int_{\mathbb{R}^d} \big(B_0(\boldsymbol{u}) - B^0(\boldsymbol{u})\big)^2 w(\boldsymbol{u}) d\boldsymbol{u} \to +\infty$ implies that $Q_{T,w}(\gamma) \xrightarrow{P} +\infty$ and, hence, the test is consistent. In fact, it may be shown that the test is consistent even for some local alternatives.

The assertion of Theorem 2 motivates us in accordance with change-point procedures in simple models to define estimators of the change point t_0 as

$$\widehat{t_0} = \underset{1 \leq t < T}{\arg\max}\, D_{t,w}(t(T - t))^2.$$

Some weak consistency of this estimator can be shown quite straightforwardly.

14.4.2 Dependent setup

Here we formulate Theorems 3 and 4 in an analogous manner to Theorems 1 and 2, respectively. As already noted, additional and/or more stringent assumptions are needed.

Theorem 3 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be a sequence of d-dimensional random vectors following model (14.7) including the assumptions below it and let $\widehat{\boldsymbol{A}}_j,\ j = 1, \ldots, p$, be estimators of $\boldsymbol{A}_j,\ j = 1, \ldots, p$, satisfying*

$$\sqrt{T}(\widehat{\boldsymbol{A}}_j - \boldsymbol{A}_j) = O_P(1), \quad j = 1, \ldots, p. \tag{14.14}$$

Let $\gamma \in (-1, 1]$ and let $w(\cdot)$ be a nonnegative measurable weight function defined on $\mathbb{R}^d$ such that

$$w(\boldsymbol{u}) = w(-\boldsymbol{u}), \quad \forall \boldsymbol{t} \in \mathbb{R}^d, \quad 0 < \int_{\mathbb{R}^d} \|\boldsymbol{u}\|^4 w(\boldsymbol{u}) d\boldsymbol{u} < \infty. \tag{14.15}$$

Then, as $T \to \infty$,

$$\widehat{Q}_{T,w}(\gamma) \xrightarrow{d} \sup_{s\in(0,1)} (s(1-s))^{\gamma} \int_{\mathbb{R}^d} \big(\widehat{V}(\boldsymbol{u}, s) - s\widehat{V}(\boldsymbol{u}, 1)\big)^2 w(\boldsymbol{u}) d\boldsymbol{u}, \tag{14.16}$$

where $\{\widehat{V}(\boldsymbol{u}, s);\ \boldsymbol{u} \in \mathbb{R}^d,\ s \in (0, 1)\}$ is a Gaussian process with zero mean and covariance structure

$$\operatorname{cov}\big(\widehat{V}(\boldsymbol{u}_1, s_1), \widehat{V}(\boldsymbol{u}_2, s_2)\big) = \min(s_1, s_2)\widehat{C}(\boldsymbol{u}_1, \boldsymbol{u}_2),$$

where

$$\widehat{C}(\boldsymbol{u}_1, \boldsymbol{u}_2) = \operatorname{cov}\big(\cos(\boldsymbol{u}_1'\boldsymbol{\varepsilon}_1) + \sin(\boldsymbol{u}_1'\boldsymbol{\varepsilon}_1), \cos(\boldsymbol{u}_2'\boldsymbol{\varepsilon}_1) + \sin(\boldsymbol{u}_2'\boldsymbol{\varepsilon}_1)\big).$$

Proof It is postponed to the Appendix. □

Next, we shortly consider the alternatives. To this end, consider the situation of a change-in-distribution of innovations

$$\boldsymbol{X}_t = \sum_{j=1}^{p} \boldsymbol{A}_j \boldsymbol{X}_{t-j} + \boldsymbol{\varepsilon}_t, \quad 1 \le t \le t_0, \tag{14.17}$$

$$\boldsymbol{X}_t = \sum_{j=1}^{p} \boldsymbol{A}_j \boldsymbol{X}_{t-j} + \boldsymbol{\varepsilon}_{t,T}, \quad t_0 < t \le T, \tag{14.18}$$

where the change point is such that $t_0 = \lfloor T s_0 \rfloor$ for some $s_0 \in (0, 1)$, $\{\boldsymbol{\varepsilon}_t\}$ and $\{\boldsymbol{\varepsilon}_{t,T}\}$ are independent sequences of $(d \times 1)$ i.i.d. random vectors with zero means, finite variances, and with CFs φ_0 and φ^0, respectively, and where the assumptions for the matrices $\{\boldsymbol{A}_j\}_{j=1}^p$ given below (14.7) continue to hold.

Theorem 4 *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ satisfy the above assumptions. Let assumption (14.15) on the weight function $w(\cdot)$ be satisfied. Then, as $T \to \infty$,*

$$\frac{(s(1-s))^2}{T}\widehat{D}_{\lfloor Ts \rfloor, w} \xrightarrow{P} (\min(s, s_0)(1 - \max(s, s_0)))^2 \int_{\mathbb{R}^d} \left(B_{0,\varepsilon}(\boldsymbol{u}) - B^{0,\varepsilon}(\boldsymbol{u})\right)^2 w(\boldsymbol{u}) d\boldsymbol{u}, \tag{14.19}$$

for $s \in (0, 1)$, where

$$B_{0,\varepsilon}(\boldsymbol{u}) = E\Big(\cos(\boldsymbol{u}'\boldsymbol{\varepsilon}_t) + \sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_t)\Big), \quad 1 \le t \le t_0$$
$$B^{0,\varepsilon}(\boldsymbol{u}) = E\Big(\cos(\boldsymbol{u}'\boldsymbol{\varepsilon}_{t,T}) + \sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_{t,T})\Big), \quad t_0 + 1 \le t \le T.$$

Moreover, as soon as $T \int_{\mathbb{R}^d} \left(B_{0,\varepsilon}(\boldsymbol{u}) - B^{0,\varepsilon}(\boldsymbol{u})\right)^2 w(\boldsymbol{u}) d\boldsymbol{u} \to +\infty$, the test based on $\widehat{Q}_{T,w}(\gamma)$ is consistent.

Remark 3 Most of the remarks and comments in Sect. 14.4.1 hold true here also.

14.5 Computations and Resampling Procedures

14.5.1 Computations

In what follows, we discuss only the test statistic (14.6) of Sect. 14.2 but analogous computations apply to the criterion (14.9) in Sect. 14.3. As already mentioned, our procedures enjoy the advantage of computational simplicity. To see this, we first proceed from (14.3) by using simple algebra and the trigonometric identity $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$ to get

$$\delta_t(\boldsymbol{u}) = \frac{1}{t^2}\sum_{\tau,s=1}^{t} \cos(\boldsymbol{u}'\boldsymbol{X}_{\tau,s}) + \frac{1}{(T-t)^2}\sum_{\tau,s=t+1}^{T} \cos(\boldsymbol{u}'\boldsymbol{X}_{\tau,s}) - \frac{2}{t(T-t)}\sum_{\tau=1}^{t}\sum_{s=t+1}^{T} \cos(\boldsymbol{u}'\boldsymbol{X}_{\tau,s}), \tag{14.20}$$

where $\boldsymbol{X}_{\tau,s} = \boldsymbol{X}_\tau - \boldsymbol{X}_s$. Then, by making use of the previous equation in (14.6), we conclude that the test statistic can be written as

$$D_{t,W} = \frac{1}{t^2}\sum_{\tau,s=1}^{t} I_w(\boldsymbol{X}_{\tau,s}) + \frac{1}{(T-t)^2}\sum_{\tau,s=t+1}^{T} I_w(\boldsymbol{X}_{\tau,s}) - \frac{2}{t(T-t)}\sum_{\tau=1}^{t}\sum_{s=t+1}^{T} I_w(\boldsymbol{X}_{\tau,s}), \tag{14.21}$$

where

$$I_w(\boldsymbol{x}) = \int_{\mathbb{R}^d} \cos(\boldsymbol{u}'\boldsymbol{x}) w(\boldsymbol{u}) d\boldsymbol{u}. \tag{14.22}$$

The weight function $w(\cdot)$ in (14.22) may be chosen in a way that avoids numerical integration which is problematic in higher dimension. To this end, we follow Henze and Wagner (1997) and adopt the weight function $w(\boldsymbol{u}) = e^{-a\|\boldsymbol{u}\|^2}$, $a > 0$, which leads to

$$I_w(\boldsymbol{x}) = \left(\frac{\pi}{a}\right)^{d/2} e^{-\|\boldsymbol{x}\|^2/4a}, \tag{14.23}$$

where $\|z\| = \sqrt{\sum_{m=1}^d z_m^2}$ denotes the Euclidian norm of an arbitrary vector z of dimension d. Alternative choices for $w(\cdot)$ are also possible but we will not pursue this issue further here.

14.5.2 Resampling Procedures

As already shown in Sect. 14.4, the null distribution of the proposed test statistic depends, among other things, on the underlying stochastic properties of the random variables involved which, however, are assumed unknown in the present setting. In order to deal with these issues, we apply appropriate resampling procedures for computing critical points and actually carrying out the tests. We present below such procedures for all the detection problems considered.

14.5.3 Resampling for Independent Data

Let $Q = Q(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$ be a test statistic which depends on a sample of size T of observations $\boldsymbol{X}_t$, $1 \le t \le T$. We will apply the permutation procedure whereby we randomly generate a permutation $b = \{b_1, \ldots, b_T\}$ of $\{1, \ldots, T\}$, and compute the test statistic $Q_b = Q(\boldsymbol{X}_{b_1}, \ldots, \boldsymbol{X}_{b_T})$. The procedure is repeated a number of times $b = 1, \ldots, \mathrm{B}$, and the critical point of the test of size α is determined as the corresponding $(1-\alpha)$ quantile $Q_{((1-\alpha)\mathrm{B})}$ of the values Q_b, $b = 1, \ldots, \mathrm{B}$. The null hypothesis is then rejected if $Q > Q_{((1-\alpha)\mathrm{B})}$.

14.5.4 Non-parametric Bootstrap for the VAR Model

First, we estimate the model (14.7) based on the observations $\boldsymbol{X}_t$, $t = 1, \ldots, T$, and initial values $\boldsymbol{X}_{1-p}, \ldots, \boldsymbol{X}_0$, and obtain the residuals $\widehat{\boldsymbol{\varepsilon}}_t$, $t = 1, \ldots, T$, and the corresponding value of the criterion $Q := Q(\widehat{\boldsymbol{\varepsilon}}_1, \ldots, \widehat{\boldsymbol{\varepsilon}}_T)$. Let $\overline{\widehat{\boldsymbol{\varepsilon}}} = T^{-1} \sum_{t=1}^T \widehat{\boldsymbol{\varepsilon}}_t$ be the residual sample mean and write $\widetilde{\boldsymbol{\varepsilon}}_t = \widehat{\boldsymbol{\varepsilon}}_t - \overline{\widehat{\boldsymbol{\varepsilon}}}$ for the centered residuals. Obtain

$\{\boldsymbol{\varepsilon}_1^*, \ldots, \boldsymbol{\varepsilon}_T^*\}$, by sampling from the empirical distribution of $\{\widetilde{\boldsymbol{\varepsilon}}_1, \ldots, \widetilde{\boldsymbol{\varepsilon}}_T\}$. Compute the value of the criterion $Q^* = Q(\boldsymbol{\varepsilon}_1^*, \ldots, \boldsymbol{\varepsilon}_T^*)$ and repeat this step a number of times B. This gives rise to the bootstrap statistics Q_b^*, $b = 1, \ldots, \mathrm{B}$, and then we calculate the critical point in the same manner as in the case of the permutation procedure above.

14.6 Simulations and Real Data

The setup of the simulation study has been inspired by Dvořák (2015): We simulate observations from a two-dimensional VAR(1) model (14.7), where

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

with the parameter σ controlling the scale and the parameter ρ the correlation. Apart from these parameters, we consider several distributions of the random error terms:

1. multivariate normal (N),
2. multivariate t_{df} with df degrees of freedom,
3. multivariate χ^2_{df} with df degrees of freedom.

All distributions are standardized, i.e., $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$ and $\mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. The multivariate normal and t_{df} distributions were simulated using R library `mvtnorm` (Genz et al. 2014; Genz and Bretz 2009). The multivariate χ^2_{df} distribution was simulated according to Minhajuddin et al. (2004).

Throughout this simulation study, we investigate the behavior of the test statistic (14.21) with the weight function $w(\boldsymbol{u}) = e^{-a\|\boldsymbol{u}\|^2}$, $a > 0$. The VAR coefficients are estimated using OLS method.

14.6.1 Empirical Level

In Table 14.1, we display the empirical level obtained from 1000 computer simulations with 2000 bootstrap replications for different values of the parameters γ and a. We consider two sample sizes ($T = 200$ or 400) and five distributions of the random errors. The scale and correlation parameters were set to $\sigma = 1$ and $\rho = 0.2$.

Looking at Table 14.1, it seems that the empirical level lies very close to the true (nominal) level α considered in the simulation study (0.01, 0.05, and 0.10).

Table 14.1 Empirical level (in %) for five error distributions

	a	α	$T = 200$					$T = 400$				
			N	t_3	t_4	χ^2_2	χ^2_4	N	t_3	t_4	χ^2_2	χ^2_4
$\gamma = -0.5$	1	0.01	0.9	0.8	0.6	1.1	0.9	1.2	0.7	1.4	1.3	1.3
		0.05	3.3	3.9	4.9	4.9	4.2	4.9	4.5	5.4	4.1	5.1
		0.10	7.9	9.9	10.3	10.2	9.6	9.8	9.9	11.0	8.2	11.0
	2	0.01	1.0	0.5	0.3	0.9	1.0	0.8	1.3	1.5	1.5	0.6
		0.05	4.6	3.7	5.4	5.1	5.3	4.3	4.8	5.4	4.5	4.1
		0.10	9.8	8.1	10.8	9.9	9.8	11.0	9.8	10.2	10.0	7.9
	3	0.01	1.0	1.3	0.6	1.1	1.0	1.0	1.1	0.5	1.1	1.4
		0.05	5.0	4.2	4.4	5.0	3.8	5.6	4.7	4.3	5.6	4.9
		0.10	10.4	8.2	9.2	9.8	9.3	9.1	9.6	11.2	10.9	9.9
$\gamma = 0$	1	0.01	1.2	0.7	0.8	0.9	1.3	0.9	1.2	0.6	1.0	0.8
		0.05	4.6	4.7	3.6	3.9	6.4	5.2	4.3	3.6	5.5	3.2
		0.10	10.3	8.8	8.5	8.3	11.6	10.8	7.5	7.8	11.7	7.5
	2	0.01	1.6	1.1	0.5	1.0	1.1	1.2	0.9	1.6	1.0	1.2
		0.05	6.4	4.4	5.0	4.7	4.9	5.2	4.5	4.9	4.4	4.5
		0.10	11.9	7.9	9.0	8.6	10.0	9.3	8.8	8.7	8.4	10.2
	3	0.01	1.0	0.9	1.5	1.0	1.0	1.0	0.9	0.9	1.3	0.9
		0.05	5.8	5.1	5.1	4.2	4.6	5.0	5.0	5.1	6.2	5.2
		0.10	10.8	8.9	10.8	8.7	10.4	9.9	11.1	9.8	10.6	11.2
$\gamma = 0.5$	1	0.01	0.9	0.7	0.5	1.0	1.0	1.5	1.4	1.1	0.7	1.0
		0.05	5.2	3.8	5.5	4.3	5.2	5.2	4.8	5.3	4.8	4.9
		0.10	10.0	8.4	10.8	9.3	9.9	10.4	9.5	10.3	9.5	9.1
	2	0.01	0.6	0.6	1.2	1.1	0.9	0.8	0.7	1.1	1.4	0.8
		0.05	4.7	4.4	5.5	5.3	5.2	3.7	3.6	4.0	5.3	5.0
		0.10	11.0	9.2	11.4	10.0	10.7	8.6	8.8	9.9	8.9	11.0
	3	0.01	0.8	0.6	0.8	1.3	0.8	0.9	0.7	1.5	1.2	1.3
		0.05	3.8	4.3	5.2	5.9	5.0	4.6	4.3	6.0	5.7	4.0
		0.10	8.3	8.9	10.9	10.8	9.8	8.6	9.4	10.6	12.0	8.2

14.6.2 Empirical Power

In this section, we investigate the power of the change–point test with respect to changes in the error distribution.

We assume that the distribution before the change–point $t_0 = \tau_0 T$ is bivariate normal with the variance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ defined in the previous section with $\sigma_1 = 1$ and $\rho_1 = 0.2$ and we consider the following types of change:

Table 14.2 Empirical power (in %) for several types of change in the error distribution with changepoint $t_0 = \tau_0 T$. The symbol $\star$ denotes 100%, $a = 2$

	τ_0	γ	$\sigma_1 \to \sigma_2$			$\rho_1 \to \rho_2$			$N \to t_4$			$N \to \chi^2_4$		
			-0.5	0.0	0.5	-0.5	0.0	0.5	-0.5	0.0	0.5	-0.5	0.0	0.5
$T = 200$	0.1		37.0	20.6	19.4	4.7	5.9	4.5	5.7	4.4	4.7	6.8	4.8	5.1
	0.2		98.7	97.4	91.6	4.9	5.2	7.3	6.8	6.3	5.3	10.0	9.5	9.7
	0.5		$\star$	99.9	$\star$	9.2	7.9	8.7	8.9	7.2	8.6	19.0	20.0	20.5
	0.8		99.4	98.0	95.0	6.4	5.9	4.9	5.9	6.6	5.7	11.3	9.2	8.1
$T = 400$	0.1		96.7	64.5	43.4	5.6	5.7	5.0	6.6	5.5	6.8	8.9	6.4	7.4
	0.2		$\star$	$\star$	$\star$	6.3	6.0	5.1	8.7	8.4	6.8	18.4	14.1	13.0
	0.5		$\star$	$\star$	$\star$	11.4	12.7	13.4	13.1	12.9	16.9	36.8	37.5	40.2
	0.8		$\star$	$\star$	$\star$	7.7	5.1	6.9	6.7	7.6	6.9	15.2	14.7	11.4
$T = 600$	0.1		$\star$	97.1	74.0	6.5	5.3	6.1	5.8	4.9	5.5	11.5	8.6	6.7
	0.2		$\star$	$\star$	$\star$	8.2	8.3	6.2	11.4	10.0	7.7	24.4	21.5	19.7
	0.5		$\star$	$\star$	$\star$	15.5	19.6	20.3	18.6	21.4	21.7	50.2	57.0	56.6
	0.8		$\star$	$\star$	$\star$	8.8	7.6	7.1	7.7	7.9	7.9	21.8	20.1	18.5

1. change in scale (the parameter $\sigma_1 = 1$ changes to $\sigma_2 = 2$),
2. change in correlation (the parameter $\rho_1 = 0.2$ changes to $\rho_2 = 0.6$),
3. change in distribution (normal distribution changes to t_4 or χ^2_4).

The results of the simulation study are summarized in Table 14.2. It seems that the test has good power against changes in the variance of the random errors. The empirical power against other types of alternatives is much lower. With $T = 600$ observations, the test rejects the null hypothesis of no change with probability 20% for the change in the correlation of random errors and for the change from Normal to t_4 distribution. The probability of detecting the change from Normal to χ^2_4 distribution with the same number of observations is approximately 50%.

Concerning the choice of the parameter γ, it seems that $\gamma = 0.5$ works somewhat better for changes occurring in the center of the time series ($\tau_0 = 0.5$) and $\gamma = -0.5$ works somewhat better especially for changes occurring earlier. In our opinion, the value $\gamma = 0.0$ provides a reasonable compromise.

14.6.3 Real Data Analysis

We apply the proposed test on the bivariate time series consisting of monthly log returns of IBM and S&P500 from January 1926 until December 1999 (Tsay 2010). This data set has been already investigated in Dvořák (2015, Sect. 3.6), who considered VAR(5) model and identified a change in its parameters in December 1932.

Table 14.3 p–values for monthly IBM and S&P500 log returns for seven decades

Decade	1930s	1940s	1950s	1960s	1970s	1980s	1990s
p-value	0.2380	0.1595	0.8590	0.4740	0.2430	0.4245	0.0185

Looking at the time series ($T = 888$) and applying the proposed test with parameters $a = 2$ and $\gamma = 0$, we also reject the null hypothesis of no change (p-value $= 0.0045$).

In order to investigate the changes in more detail, we test the existence of a change-point in the error distribution of the VAR(5) model separately in each decade. Interestingly, the p-values summarized in Table 14.3 suggest that significant changes in the error distribution of the VAR model can be detected only in 1990s.

Acknowledgements The research of Simos Meintanis was partially supported by grant number 11699 of the Special Account for Research Grants (EΛΚΕ) of the National and Kapodistrian University of Athens. The research of Marie Hušková and Zdeněk Hlávka was partially supported by grant GAČR 15-09663S and AP research network grant Nr. P7/06 of the Belgian government (Belgian Science Policy).

14.7 Appendix

Proof (Theorem 1) Notice that little algebra leads to an equivalent expression for $D_{t,w}$:

$$T \cdot D_{t,w} = T \int_{\mathbb{R}^d} \Big(\frac{1}{t} \sum_{j=1}^{t} Z(\boldsymbol{u}, \boldsymbol{X}_j) - \frac{1}{T-t} \sum_{j=t+1}^{T} Z(\boldsymbol{u}, \boldsymbol{X}_j) \Big)^2 w(\boldsymbol{u}) d\boldsymbol{u} \quad (14.24)$$

$$= \Big(\frac{T^2}{t(T-t)} \Big)^2 \int_{\mathbb{R}^d} \Big(V_T(\boldsymbol{u}, t) - \frac{t}{T} V_T(\boldsymbol{u}, T) \Big)^2 w(\boldsymbol{u}) d\boldsymbol{u}, \quad (14.25)$$

where

$$Z(\boldsymbol{u}, \boldsymbol{X}_j) = \cos(\boldsymbol{u}' \boldsymbol{X}_j) + \sin(\boldsymbol{u}' \boldsymbol{X}_j), \quad \boldsymbol{u} \in \mathbb{R}^d, \quad j = 1, \ldots, T, \quad (14.26)$$

$$V_T(\boldsymbol{u}, t) = \frac{1}{\sqrt{T}} \sum_{j=1}^{t} Z(\boldsymbol{u}, \boldsymbol{X}_j), \quad \boldsymbol{u} \in \mathbb{R}^d, \quad t = 1, \ldots, T. \quad (14.27)$$

Notice that, under the null hypothesis, the process $\{V_T(\boldsymbol{u}, t),\ \boldsymbol{u} \in \mathbb{R}^d,\ t = 1, \ldots, T\}$ has the expectation and the covariance structure

$$EV_T(\boldsymbol{u}, t) = \frac{t}{\sqrt{T}} EZ(\boldsymbol{u}, \boldsymbol{X}_1),$$
$$\operatorname{cov}\Big(V_T(\boldsymbol{u}_1, t_1), V_T(\boldsymbol{u}_2, t_2)\Big) = \frac{\min(t_1, t_2)}{T} \operatorname{cov}\Big(Z(\boldsymbol{u}_1, \boldsymbol{X}_1), Z(\boldsymbol{u}_2, \boldsymbol{X}_1)\Big).$$

Now the proof follows the lines of that of Theorem 4.1 in Hlávka et al. (2017), therefore we will be brief.

At first, it should be proved for any $s \in (0, 1)$ fixed

$$\int_{\mathbb{R}^d} \Big(V_T(\boldsymbol{u}, \lfloor Ts \rfloor) - sV_T(\boldsymbol{u}, T)\Big)^2 w(\boldsymbol{u}) d\boldsymbol{u} \xrightarrow{d} \int_{\mathbb{R}^d} \Big(V(\boldsymbol{u}, s) - sV(\boldsymbol{u}, 1)\Big)^2 w(\boldsymbol{u}) d\boldsymbol{u}, \quad (14.28)$$

where the process $\{V(\boldsymbol{u}, s), \boldsymbol{u} \in \mathbb{R}^d,\ s \in (0, 1)\}$ is defined in Theorem 1. Towards this, notice that under our assumptions $V_T(\boldsymbol{u}, \lfloor Ts \rfloor) - EV_T(\boldsymbol{u}, \lfloor Ts \rfloor)$ has asymptotically a normal distribution with zero mean and variance $s \cdot var(Z(\boldsymbol{u}, \boldsymbol{X}_1))$. Moreover, standard arguments give for any $s \in (0, 1)$ and any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$

$$E|V_T(\boldsymbol{u}_1, \lfloor Ts \rfloor) - V_T(\boldsymbol{u}_2, \lfloor Ts \rfloor)| \le C\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^\beta$$

for some $\beta > 0$ and $C > 0$ and also

$$E\Big(\frac{T^2}{t(T-t)}\Big)^2 \int_{\mathbb{R}^d} \Big(V_T(\boldsymbol{u}, t) - \frac{t}{T} V_T(\boldsymbol{u}, T)\Big)^2 w(\boldsymbol{u}) d\boldsymbol{u} = \int_{\mathbb{R}^d} var(Z(\boldsymbol{u}, \boldsymbol{X}_1)) w(\boldsymbol{u}) d\boldsymbol{u} < \infty.$$

Then the convergence (14.28) follows from Theorem 22 in Ibgagimov and Has'minskii (1981).

Additionally, to derive properties of the process

$$Y_T(s) = \sqrt{\int_{\mathbb{R}^d} \Big(V_T(\boldsymbol{u}, \lfloor Ts \rfloor) - sV_T(\boldsymbol{u}, T)\Big)^2 w(\boldsymbol{u}) d\boldsymbol{u}}, \quad s \in (0, 1)$$

it suffices to follow the proof of Theorem 4.1 b) in Hlávka et al. (2017), therefore we omit it. □

Proof (Theorem 2) Following the considerations in the proof of Theorem 1, we easily receive that

$$\sup_{s \in (0,1)} (s(1-s))^\gamma \int_{\mathbb{R}^d} \Big(V_T(\boldsymbol{u}, \lfloor Ts \rfloor) - sV_T(\boldsymbol{u}, T)$$
$$- E\Big(V_T(\boldsymbol{u}, \lfloor Ts \rfloor) - sV_T(\boldsymbol{u}, T)\Big)\Big)^2 w(\boldsymbol{u}) d\boldsymbol{u} = O_P(1)$$

and

$$\frac{1}{T}\int_{\mathbb{R}^d}\Big(E\Big(V_T(\boldsymbol{u},\lfloor Ts\rfloor)-sV_T(\boldsymbol{u},T)\Big)\Big)^2 w(\boldsymbol{u})d\boldsymbol{u}$$
$$=\min(s,s_0)(1-\max(s,s_0))\int_{\mathbb{R}^d}\Big(B_0(\boldsymbol{u})-B^0(\boldsymbol{u})\Big)^2 w(\boldsymbol{u})d\boldsymbol{u}.$$

The assertion of Theorem 2 easily follows. □

Proof (Theorem 3) It follows the same line as that of Theorem 1 but the situation is slightly more complicated due to the presence of a nuisance parameter. We have

$$T\widehat{D}_{t,w}=T\int_{\mathbb{R}^d}\Big(\frac{1}{t}\sum_{j=1}^{t}\widehat{Z}(\boldsymbol{u},\widehat{\boldsymbol{\varepsilon}}_j)-\frac{1}{T-t}\sum_{j=t+1}^{T}\widehat{Z}(\boldsymbol{u},\widehat{\boldsymbol{\varepsilon}}_j)\Big)^2 w(\boldsymbol{u})d\boldsymbol{u}$$
$$=\Big(\frac{T^2}{t(T-t)}\Big)^2\int_{\mathbb{R}^d}\Big(\widehat{V}_T(\boldsymbol{u},t)-\frac{t}{T}\widehat{V}_T(\boldsymbol{u},T)\Big)^2 w(\boldsymbol{u})d\boldsymbol{u},$$

where

$$\widehat{Z}(\boldsymbol{u},\widehat{\boldsymbol{\varepsilon}}_j)=\cos(\boldsymbol{u}'\widehat{\boldsymbol{\varepsilon}}_j)+\sin(\boldsymbol{u}'\widehat{\boldsymbol{\varepsilon}}_j),\quad \boldsymbol{u}\in\mathbb{R}^d,\quad j=1,\ldots,T,$$
$$\widehat{V}_T(\boldsymbol{u},t)=\frac{1}{\sqrt{T}}\sum_{j=1}^{t}\widehat{Z}(\boldsymbol{u},\widehat{\boldsymbol{\varepsilon}}_j).$$

Since the residuals $\widehat{\boldsymbol{\varepsilon}}_j,\ j=1,\ldots,T$ are dependent, we have to do additional steps. Particularly, we apply the Taylor expansion

$$\cos(\boldsymbol{u}'\widehat{\boldsymbol{\varepsilon}}_t)=\cos(\boldsymbol{u}'\boldsymbol{\varepsilon}_t)-\boldsymbol{u}'(\widehat{\boldsymbol{\varepsilon}}_t-\boldsymbol{\varepsilon}_t)\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_t)+R_T(\boldsymbol{u},t),$$

where the first term on the r.h.s. is influential while the others are not. Straightforward calculations give

$$|R_T(\boldsymbol{u},t)|\le C\|\boldsymbol{u}\|^2\cdot\frac{1}{T}\sum_{j=1}^{T}\|(\widehat{\boldsymbol{\varepsilon}}_j-\boldsymbol{\varepsilon}_j)\|^2$$

and therefore

$$\frac{1}{T}\int\Big(\sum_{t=1}^{T}|R_T(\boldsymbol{u},t)|\Big)^2 w(\boldsymbol{u})d\boldsymbol{u}=o_P(1).$$

Next, we have to study

$$L_T(\boldsymbol{u},t)=\frac{1}{\sqrt{T}}\sum_{j\le t}(\widehat{\boldsymbol{\varepsilon}}_j-\boldsymbol{\varepsilon}_j)\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)-\frac{t}{T}\frac{1}{\sqrt{T}}\sum_{j\le T}(\widehat{\boldsymbol{\varepsilon}}_j-\boldsymbol{\varepsilon}_j)\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j).$$

Noticing that $\widehat{\boldsymbol{\varepsilon}}_j-\boldsymbol{\varepsilon}_j=(\boldsymbol{A}-\widehat{\boldsymbol{A}}_T)\boldsymbol{X}_{j-1},\ j=1,\dots,T$, we have

$$L_T(\boldsymbol{u},t)=\sqrt{T}(\widehat{\boldsymbol{A}}-\boldsymbol{A})\frac{1}{T}\Big(\sum_{j\le t}\boldsymbol{X}_{j-1}\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)-\frac{t}{T}\sum_{j\le T}\boldsymbol{X}_{j-1}\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)\Big).$$

By assumptions (14.14) and using the same arguments as in the proof of Theorem 1:

$$\int\frac{1}{T^2}\Big(\sum_{j\le t}\boldsymbol{X}_{j-1}\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)-\frac{t}{T}\sum_{j\le T}\boldsymbol{X}_{j-1}\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)\Big)^2w(\boldsymbol{t})d\boldsymbol{t}=o_P(T^{-\xi})$$

for some $\xi>0$. Here we utilize properties of

$$\sum_{j\le t}\Big(\boldsymbol{X}_{j-1}\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)-E\big(\boldsymbol{X}_{j-1}\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_j)\big)\Big),\quad t>p,$$

that are for each fixed $\boldsymbol{u}$ partial sums of martingale differences. Particularly, the asymptotic normality holds true under the considered assumptions. As a consequence, we get that the respective term is not influential.

Finally, we get after some standard steps, that the limit distribution of $\widehat{Q}_{w,T}(\gamma)$ is the same as $Q_{w,T}(\gamma)$ with the Gaussian process $\{\widehat{V}(\boldsymbol{u},s),\ \boldsymbol{u}\in\mathbb{R}_d, s\in(0,1)\}$ with the expectation

$$E\widehat{V}(\boldsymbol{u},s)=\sqrt{T}sE\big(\cos(\boldsymbol{u}'\boldsymbol{\varepsilon}_1)+\sin(\boldsymbol{u}'\boldsymbol{\varepsilon}_1)\big)$$

and the covariance structure

$$\begin{aligned}&\operatorname{cov}\big(\widehat{\mathrm{V}}(\boldsymbol{u}_1,\mathrm{s}_1),\widehat{\mathrm{V}}(\boldsymbol{u}_2,\mathrm{s}_2)\big)\\&=\min(s_1,s_2)\operatorname{cov}\big(\cos(\boldsymbol{u}_1'\boldsymbol{\varepsilon}_1)+\sin(\boldsymbol{u}_1'\boldsymbol{\varepsilon}_1),\cos(\boldsymbol{u}_2'\boldsymbol{\varepsilon}_2)+\sin(\boldsymbol{u}_2'\boldsymbol{\varepsilon}_2)\big)\end{aligned}$$

instead of $V(\boldsymbol{u},s),\ \boldsymbol{u}\in\mathbb{R}_d, s\in(0,1)$. □

Proof (Theorem 4) Since it is assumed that the change occurs only in the distribution of the error term, but not in $\boldsymbol{A}_j,\ j=1,\dots,p$, we realize going through the proof of Theorem 3 that the limit distribution of

$$\max_{1<t<T}\Big(\frac{t(T-t)}{T^2}\Big)^{2+\gamma}\int_{\mathbb{R}^d}\Big(\frac{1}{t}\sum_{j=1}^{t}\widehat{Z}(\boldsymbol{u},\widehat{\boldsymbol{\varepsilon}}_j)-\frac{1}{T-t}\sum_{j=t+1}^{T}\widehat{Z}(\boldsymbol{u},\widehat{\boldsymbol{\varepsilon}}_j)\Big)^2w(\boldsymbol{u})d\boldsymbol{u}$$

is the same as if $\widehat{\boldsymbol{\varepsilon}}_j,\ j = 1, \ldots, T$, are replaced by $\boldsymbol{\varepsilon}_j,\ j = 1, \ldots, T$. Then it remains to study

$$\Big(\frac{t(T-t)}{T^2}\Big)^{2+\gamma} \int_{\mathbb{R}^d} \Big(\frac{1}{t}\sum_{j=1}^{t} \widehat{Z}(\boldsymbol{u}, \boldsymbol{\varepsilon}_j) - \frac{1}{T-t}\sum_{j=t+1}^{T} \widehat{Z}(\boldsymbol{u}, \boldsymbol{\varepsilon}_j)\Big)^2 w(\boldsymbol{u}) d\boldsymbol{u}.$$

The proof can be finished as in the proof of Theorem 2. The details are omitted. □

References

Andrews, B., Calder, M., Davis, R.A.: Maximum likelihood estimation for α–stable autoregressive processes. Ann. Statist. **37**, 1946–1982 (2009)

Bai, J.: Vector autoregressions with structural changes in regression coefficients and in variance–covariance matrices. Ann. Econom. Financ. **1**, 303–339 (2000)

Bai, J., Lumsdaine, R.L., Stock, J.H.: Testing for and dating common breaks in multivariate time series. Rev. Econom. Stud. **65**, 395–432 (1998)

Brockwell, P.J., Davis, R.A.: Estimating the noise parameters from observations of a linear process with stable innovations. J. Statist. Plann. Inferen. **33**, 175–186 (1992)

Davis, R.A.: Gauss–Newton and M–estimation for ARMA processes with infinite variance. Stoch. Process. Applic. **63**, 75–95 (1996)

Dvořák. M.: Stability in autoregressive time series models. PhD. Thesis, Charles University in Prague, Czech Republic (2015)

Dvořák, M.: Darling–Erdős type test for change detection in stationary VAR models. Commun. Statist. – Theor. Meth. (2016) doi:10.1080/03610926.2014.995828

Dvořák, M., Prášková, Z.: On testing changes in autoregressive parameters of a VAR model. Commun. Statist. – Theor. Meth. **42**, 1208–1226 (2013)

Fama, E.: The behavior of stock market prices. J. Business **38**, 34–105 (1965)

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T.: mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-0 (2014) http://CRAN.R-project.org/package=mvtnorm

Genz, A., Bretz, F.: Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195. Springer-Verlag, Heidelberg (2009)

Hamilton, J.D.: Time Series Analysis. Princeton University Press, New Jersey (1994)

Hannan, E.J., Kanter, M.: Autoregressive processes with infinite variance. J. Appl. Probab. **14**, 411–415 (1977)

Henze, N., Wagner, T.: A new approach to the BHEP tests for multivariate normality. J. Multivar. Anal. **62**, 1–23 (1997)

Herwartz, H., Lütkepohl, H.: Structural vector autoregressions with Markov Markov switching: Combining conventional with statistical identification of shocks. J. Econometr. **183**, 104–116 (2014)

Hlávka, Z., Hušková, M., Kirch, C., Meintanis, S.G.: Monitoring changes in the error distribution of autoregressive models based on Fourier methods. TEST **21**, 605–634 (2012)

Hlávka, Z., Hušková, M., Kirch, C., Meintanis, S.G.: Bootstrap procedures for online monitoring of changes in autoregressive models. Commun. Statist. – Simul. Comput. **45**, 2471–2490 (2016)

Hlávka, Z., Hušková, M., Kirch, C., Meintanis, S.G.: Fourier-type tests involving martingale difference processes. Econometr. Rev. **36**, 468–492 (2017)

Horváth, L., Rice, G.: Extensions of some classical methods in change point analysis. Test **23**, 219–255 (2014)
Hušková, M., Meintanis, S.G.: Change point analysis based on empirical characteristic functions. Metrika **63**, 145–168 (2006)
Hušková, M., Meintanis, S.G.: Tests for the multivariate k–sample problem based on the empirical characteristic function. J. Nonparametr. Statist. **20**, 263–277 (2008)
Ibgagimov, I., Has'minskii, R.: Statistical Estimation: Asymptotic Theory. Springer Verlag, New York (1981)
Kirch, C., Muhsal, B., Ombao, H.: Detection of changes in multivariate time series with aplication to EEG data. J. Amer. Statist. Assoc. **110**, 1197–1216 (2015)
Lanne, M. and Lütkepohl, H.: Structural vector autoregressions with nonnormal residuals. J. Buss. Econom. Statist. **28**, 159–168 (2010)
Lanne, M., Lütkepohl, H., Maciejowska, K.: Structural vector autoregressions with Markov switching. J. Econom. Dynam. Contr. **34**, 121–131 (2010)
Lee, S., Lee, Y., Na, O.: Monitoring distributional changes in autoregressive models. Commun. Stastist. – Theor. Meth. **38**, 2969–2982 (2009)
Lütkepohl, H.: New Introduction to Multiple Time Series. Springer–Verlag, Berlin (2005)
Lütkepohl, H.: Identifying structural vector autoregressions via changes in volatility. Universität Berlin, Berlin (2012)
Mandelbrot, B.: The variation of certain speculative prices. J. Business **36**, 394–419 (1963)
Matteson, D.S., James, N.A.: A non–parametric approach for multiple change point analysis of multivariate data. J. Amer. Statist. Assoc. **109**, 334–345 (2014)
Minhajuddin, A.T., Harris, I.R., Schucany, W.R.: Simulating multivariate distributions with specific correlations. J. Statist. Comput. Simulation **74**, 599–607 (2004)
Ng, S., Vogelsang, T.J.: Analysis of vector autoregressions in the presence of shifts in the mean. Econometr. Rev. **21**, 353–381 (2002)
Qu, Z., Perron, P.: Estimating and testing structural changes in multivariate regressions. Econometrica **75**, 459–502 (2007)
Selk, L., Neumeyer, N.: Testing for a change of the innovation distribution in nonparametric autoregression: The sequential empirical process approach. Scand. J. Statist. **40**, 770–788 (2013).
Siegfried, N.A.: An information-theoretic extension to structural VAR modelling. Hamburg University, Dept. Economics, Working Paper 02–03 (2002)
Tiku, M.L., Wong, W. – K., Vaughan, D.C., Bian, G.: Time series models in non–normal situations: Symmetric innovations. J. Time Ser. Anal. **21**, 571–596 (2000)
Tsay, R.: Analysis of Financial Time Series. Wiley, New Jersey (2010)

Kader—An R Package for Nonparametric Kernel Adjusted Density Estimation and Regression

15

Gerrit Eichner

15.1 Introduction

Kernel methods in nonparametric density estimation and in nonparametric regression estimation have been a core topic in theoretical statistical research and in practical applications likewise. Standard references for the theory of nonparametric density estimation are, e.g., Silverman (1986) and Wand and Jones (1995), while Härdle (1990) and again Wand and Jones (1995) provide a review of nonparametric regression estimation. With respect to the practical applications various methods have been implemented in almost every professional statistical software package. This is, in particular, the case for the open-source programming language and environment for statistical computing R, R Core Team (2016), both in its so-called base distribution and in readily available add-on packages.

In the following we shall very briefly recall the goals and very basics of nonparametric kernel density estimation and of nonparametric kernel regression estimation. References to some R packages which provide implementations of respective methods will be given. We then introduce the fully adaptive kernel methods of Srihera and Stute (2011) for density estimation, the robustified approach by Eichner and Stute (2013), and the extension to nonparametric regression estimation by Eichner and Stute (2012). (The extension of the pointwise method for density estimation to an L_2-approach as presented by Eichner and Stute (2015) is beyond the scope of this paper and the current version of package `kader`, but shall be included in a future version of the package.) It should be emphasized that this and the next section intend to just recall and summarize quantities and results which are most relevant for the

G. Eichner (✉)
Mathematical Institute, Justus-Liebig-University Giessen, Arndtstr. 2,
35392 Giessen, Germany
e-mail: gerrit.eichner@math.uni-giessen.de

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_15

implementation of the methods while all proofs will be omitted and, instead, the reader referred to the cited publications.

15.1.1 Nonparametric Density Estimation

The goal is to recover an unknown density $f = F'$ (on the real line) from a random sample $X_1, \ldots, X_n$ of independent replicates of $X \sim F$ with an absolutely continuous distribution function F. A classical approach due to Rosenblatt (1956) and Parzen (1962) proposes to estimate $f(x)$ through

$$f_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \quad x \in \mathbb{R}, \tag{15.1}$$

with bandwidth (or window size) $h > 0$ and a kernel K, which typically satisfies $K \geq 0$ and $\int K(u)du = 1$, i.e., is itself a density.

If, in addition, K is symmetric about zero, i.e., $K(-u) = K(u)$ for all u, or at least satisfies $\int uK(u)du = 0$, and if f is twice continuously differentiable in a neighborhood of x then asymptotic expansions of bias and variance, and hence of the mean squared error (MSE) are well-known for bandwidths $h = h_n$ such that $h_n \to 0$ with $nh_n \to \infty$ while $n \to \infty$. The asymptotically optimal choice of h minimizing the leading term of the MSE is also known, at least in theory (see, e.g., Silverman (1986) or Wand and Jones (1995)). However, in practice it is unknown since it depends on f and f''. Although it also depends on K, the choice of K has little effect on MSE (see again, e.g., Silverman (1986)). Consequently, considerable efforts have been made to get close to the optimal bandwidth. Among those efforts are, e.g., iterative approaches (in which, first, a preliminary bandwidth is chosen to estimate quantities that enter the optimal bandwidth and, second, the optimal bandwidth is used with unknown quantities replaced by their estimates), cross-validation strategies, or using a parametric family of centered densities with a scale parameter σ to compute the optimal bandwidth, and then apply the optimal bandwidth with an estimated σ. (The first method is not fully satisfactory since the choice of a preliminary bandwidth is subjective. Feluch and Koronacki (1992) criticize the second approach and Silverman (1986) the third in certain situations; see Srihera and Stute (2011) and Eichner and Stute (2013) where very brief accounts of those popular methods and of their criticism are given.)

There are numerous functions in various R packages which provide univariate (or multivariate) kernel density estimation. A non-exhaustive list of examples (e.g., found on www.RSeek.org with the search term "kernel density estimation") contains the function `density` in R's base distribution, the R packages `KernSmooth` (Wand 2015), `sm` (Bowman and Azzalini 2014), `np` (Hayfield and Racine 2008), `feature` (Duong and Wand 2015), `ks` (Duong 2016), and `kedd` (Guidoum (2015), also offering functions for kernel density derivative estimation).

15.1.2 Nonparametric Regression Estimation

Here, the goal is to nonparametrically estimate an unknown regression function $x \mapsto m(x) := \mathbb{E}[Y|X = x]$, $x \in \mathbb{R}$, from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent replicates of $(X, Y) \in \mathbb{R}^2$. Nadaraya (1964) and Watson (1964) proposed the kernel estimator

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}, \quad x \in \mathbb{R}, \tag{15.2}$$

where $h = h_n$ and K are a positive bandwidth and a typically non-negative kernel, respectively, and where $h \to 0$ at a particular rate if $n \to \infty$, while K has to satisfy certain conditions like being integrable and having short tails. See Härdle (1990) and Wand and Jones (1995) for a review. In those references the optimal choice of the bandwidth h minimizing the leading term of the MSE (and also of the MISE) under smoothness conditions on m is also discussed. The methods in practice, when crucial quantities are unknown, are analogous to the ones mentioned in the previous section on kernel density estimation.

An example for a function and two examples for packages which provide functionality for nonparametric kernel regression estimation are function `ksmooth` in R's base distribution and the R packages `KernSmooth` (Wand 2015) and `lokerns` (Herrmann 2016). This listing is certainly also not exhaustive.

15.2 New Kernel Adaptive Methods

Srihera and Stute (2011) proposed a new, fully adaptive approach of pointwise kernel density estimation which modifies the third method for optimal bandwidth selection mentioned above in Sect. 15.1. Eichner and Stute (2013) robustified it by basing it on ranks, and Eichner and Stute (2015) adapted the latter to an L_2-approach. Eichner and Stute (2012) extended the pointwise approach to nonparametric regression estimation. In this section we will only very briefly summarize the methods and the results relevant for their implementation while omitting all proofs and referring the reader to the cited publications. Instead, we will present important quantities in representations which are suitable to simplify their implementation in R or to reduce their computational complexity, or, in the ideal case, achieve both.

It is evident that mathematical elegance or brevity of analytical formulae not necessarily coincide with usefulness or effectiveness when it comes to their concrete computational numerical evaluation. It was not the intention to achieve the fastest and most efficient implementation that guided us in designing the current version of the presented package. Instead, it was the idea to provide a pure, modular R-implementation that is easy to modify, utilizes R's computational strengths in matrix

calculus, and hence is passably efficient under this premises. For future versions of the package enhancements in respect to computational efficiency and speed shall therefore be the focus.

15.2.1 Kernel Adjusted Density Estimation

Srihera and Stute (2011) suggested to "update" the kernel K in a data-adaptive way by replacing it by a location-scale family associated with the classical Parzen-Rosenblatt estimator in (15.1). More precisely, an initial kernel K is used to first construct f_n and then to replace K in (15.1) by $\sigma f_n(\sigma \cdot + \theta)$ with a location-parameter $\theta \in \mathbb{R}$ and a scale-parameter $\sigma > 0$ yielding

$$\frac{\sigma}{nh} \sum_{i=1}^{n} f_n \left(\sigma \frac{x - X_i}{h} + \theta \right) = \frac{\sigma}{n^2 h^2} \sum_{1 \le i,j \le n} K \left(\frac{\sigma}{h} \cdot \frac{x - X_i}{h} + \frac{\theta - X_j}{h} \right)$$

The actual estimator was set to

$$\tilde{f}_n(x) \equiv \tilde{f}_n(x; h, \theta, \sigma) := \frac{\sigma}{n(n-1)h^2} \sum_{1 \le i \ne j \le n} K \left(\frac{\sigma}{h} \cdot \frac{x - X_i}{h} + \frac{\theta - X_j}{h} \right) \quad (15.3)$$

and does exclude the summands for $i = j$ in the double-sum to reduce possible bias (see Eq. (1.6) in Srihera and Stute (2011)). Then (under, e.g., K being a symmetric density with compact support, f being twice continously differentiable in a neigborhood of x, and $\mathbb{E}[X^2] < \infty$), the asymptotically leading terms of bias and variance were obtained. It turned out that choosing $\theta = \mathbb{E}[X]$ allowed an interesting asymptotic representation of the asymptotic MSE:

$$\begin{aligned} \mathrm{MSE}\, \tilde{f}_n(x) &\equiv \left(\mathrm{Bias}\, \tilde{f}_n(x) \right)^2 + \mathrm{Var}\, \tilde{f}_n(x) \\ &= \frac{1}{4} \left(f''(x) \right)^2 \frac{h^4}{\sigma^4} \mathrm{Var}^2 X + \frac{\sigma f(x) \int f^2(u) du}{nh} \end{aligned} \quad (15.4)$$

as $h \to 0$ and $n \to \infty$ such that $nh \to \infty$ (see Theorem 1.1 and Eq. (2.1) in Srihera and Stute (2011)). It follows from (15.4) that, besides the fact that the quality of the estimator $\tilde{f}_n(x)$ is affected by both local and global properties of the true density f, it merely depends on the *ratio* of h and σ. In fact, the MSE in (15.4) is minimal if

$$\frac{h}{\sigma} = \left(\frac{1}{n} \right)^{1/5} \left(\frac{f(x) \int f^2(u) du}{[f''(x) \mathrm{Var} X]^2} \right)^{1/5}$$

Consequently, $h = n^{-1/5}$ is feasible without loss of generality so that the bias is asymptotically neglible and the MSE is a function only of σ. Recall that θ has been set to the (typically) unknown $\mathbb{E}[X]$. However, the asymptotic representation of the MSE

in (15.4) is also true for consistent estimators $\hat{\theta}$ of θ, so that in a real data situation, where θ is unknown (but contained in a compact set) and hence needs to be replaced by an estimator, $\hat{\theta} = n^{-1}\sum_{i=1}^{n} X_i$ can (and typically will) be chosen. This holds as well for estimators $\hat{\sigma}$ with $\hat{\sigma} \to \sigma$ if σ is bounded away from 0 and K' is continuous (Theorem 1.3 in Srihera and Stute (2011)). Distributional convergence of $\tilde{f}_n(x)$ and of $\tilde{f}_n(x; n^{-1/5}, \hat{\theta}, \hat{\sigma}_x)$ was obtained under mild conditions (see Theorems 1.2 and 1.3 in Srihera and Stute (2011)).

The proofs in Srihera and Stute (2011) further revealed that minimization of an estimator $\widehat{\mathrm{MSE}}_x(\sigma)$ of $\mathrm{MSE}\,\tilde{f}_n(x)$ in σ does not have to be based on its representation in (15.4), but instead can be based on estimators for explicit expressions of $\mathrm{Bias}\,\tilde{f}_n(x)$ and $\mathrm{Var}\,\tilde{f}_n(x)$, respectively, without the need for preliminary estimates of $f(x)$ and $f''(x)$. Consequently, with F_n denoting the empirical distribution function of the sample, the bias estimator was set to

$$\begin{aligned}\widehat{\mathrm{Bias}}_x(\sigma) &:= \iint K(u) f_n\left(x + h\frac{\hat{\theta} - y - hu}{\sigma}\right) F_n(dy)du - f_n(x) \\ &= \int K(u)\frac{1}{n}\sum_{i=1}^{n} f_n\left(x + h\frac{\hat{\theta} - X_i - hu}{\sigma}\right)du - f_n(x) \qquad (15.5)\end{aligned}$$

with $\hat{\theta} := n^{-1}\sum_{i=1}^{n} X_i$ as estimator of $\theta \equiv \mathbb{E}[X]$ and f_n the "initial" estimator from (15.1) with $h = n^{-1/5}$. The variance of $\tilde{f}_n(x)$ in turn was first approximated by the variance of the Hájek-projection $\tilde{f}_n^{(0)}(x)$ of $\tilde{f}_n(x)$, i.e., by

$$\mathrm{Var}\,\tilde{f}_n^{(0)}(x) = \sigma^2 h^{-4} n^{-1}\mathrm{Var}(Z_i(x))$$

with the iid quantities

$$Z_i(x) \equiv \int K\left(\frac{\sigma}{h}\cdot\frac{x - X_i}{h} + \frac{\theta - y}{h}\right)F(dy) + \int K\left(\frac{\sigma}{h}\cdot\frac{x - z}{h} + \frac{\theta - X_i}{h}\right)F(dz) \qquad (15.6)$$

Then, the estimator $\widehat{\mathrm{Var}}_x(\sigma)$ of $\mathrm{Var}(Z_i(x))$ was set to be the sample variance of estimators $\hat{Z}_i$ of $Z_i(x)$ for $1 \le i \le n$ (obtained by replacing F in (15.6) with its empirical analogue F_n and θ with $\hat{\theta}$ as above) where

$$\hat{Z}_i := \frac{1}{n}\sum_{j=1}^{n}\left[K\left(\frac{\sigma}{h}\cdot\frac{x - X_i}{h} + \frac{\hat{\theta} - X_j}{h}\right) + K\left(\frac{\sigma}{h}\cdot\frac{x - X_j}{h} + \frac{\hat{\theta} - X_i}{h}\right)\right] \qquad (15.7)$$

Finally, the choice for σ was taken to be a minimizer $\hat{\sigma}_x$ of (see (2.3) in Srihera and Stute (2011))

$$\widehat{\mathrm{MSE}}_x(\sigma) := \left(\widehat{\mathrm{Bias}}_x(\sigma)\right)^2 + \sigma^2 h^{-4} n^{-1}\widehat{\mathrm{Var}}_x(\sigma) \qquad (15.8)$$

which is apparently possible without referring to higher order derivatives of the unknown density f. Typically, $\sigma \mapsto \widehat{\mathrm{MSE}}_x(\sigma)$ is virtually convex for small σ, so that $\hat{\sigma}_x$ is usually uniquely determined.

15.2.2 Rank Transformations in Kernel Density Estimation

Eichner and Stute (2013) robustified the approach of Srihera and Stute (2011) in so far as they constructed an estimator of f with similar features as the above $\tilde{f}_n$, but which does not require moments of the X's. For that robust version of $\tilde{f}_n$ the (typically centered) $\theta - X_j$ in (15.3) was replaced by a suitable transformation J of the rank of X_j to obtain the kernel estimator

$$\hat{f}_n(x) \equiv \hat{f}_n(x; h, J, \sigma) := \frac{\sigma}{n(n-1)h^2} \sum_{1 \le i \ne j \le n} K\left(\frac{\sigma}{h} \cdot \frac{x - X_i}{h} - \frac{J(F_n(X_j))}{h}\right) \tag{15.9}$$

(which is Eq. (4) in Eichner and Stute (2013)). The transformation J, defined on the unit interval, needs to be strictly increasing and continuously differentiable with $\int_0^1 J(u)du = 0$ to ensure (among other things) that $\mathbb{E}[J(F(X_j))] = 0$ and that the bias is of order h^2 (and not h). Note: No location parameter θ needs to be estimated.

From Theorem 2.1 in Eichner and Stute (2013) it follows (under f being twice continuously differentiable in a neighborhood of x, $\int K(w)dw = 1$, $\int wK(w)dw = 0$, $\int w^2 K(w)dw < \infty$, K being thrice differentiable with bounded K''', and the above-mentioned conditions on J) that the asymptotic bias and variance and hence the asymptotic MSE again depend on the *ratio* of h and σ. Omitting negligible terms it was obtained that

$$\begin{aligned} \mathrm{MSE}\,\hat{f}_n(x) &\equiv \left(\mathrm{Bias}\,\hat{f}_n(x)\right)^2 + \mathrm{Var}\,\hat{f}_n(x) \\ &= \left(\frac{h}{\sigma}\right)^4 \frac{[f''(x)]^2}{4} \left[\int_0^1 J^2(u)du\right]^2 + \frac{\sigma}{h}\frac{f(x)}{n} \int_0^1 \frac{1}{J'(du)} du \end{aligned} \tag{15.10}$$

as $h \to 0$ and $n \to \infty$ such that $nh \to \infty$. This is minimal if

$$\frac{h}{\sigma} = \left(\frac{1}{n}\right)^{1/5} \left(\frac{f(x) \int_0^1 \frac{1}{J'(u)} du}{\left[f''(x) \int_0^1 J^2(u)du\right]^2}\right)^{1/5} \tag{15.11}$$

In particular, one may (as in the non-robustified version of the previous section) choose $h = n^{-1/5}$ without loss of generality so that again the bias is neglible and the MSE is a function of σ only.

Analogously to the strategy in Srihera and Stute (2011) described in the previous section, minimization of an estimator of $\mathrm{MSE}\hat{f}_n(x)$ in σ was not based on its representation in (15.10), but instead on estimators for explicit expressions of $\mathrm{Bias}\hat{f}_n(x)$ and $\mathrm{Var}\hat{f}_n(x)$, respectively: The bias estimator is

$$\widehat{\mathrm{Bias}}_x(\sigma) := \int K(u)\frac{1}{n}\sum_{i=1}^{n} f_n\left(x + h\frac{-J(F_n(X_i)) - hu}{\sigma}\right)du - f_n(x) \tag{15.12}$$

where f_n is the classical Parzen-Rosenblatt kernel estimator with $h = n^{-1/5}$, and the variance estimator $\widehat{\mathrm{Var}}_x(\sigma)$ is the sample variance of

$$\hat{Z}_i := \frac{1}{n}\sum_{j=1}^{n}\left[K\left(\frac{\sigma}{h}\cdot\frac{x - X_i}{h} - \frac{J(F_n(X_j))}{h}\right) + K\left(\frac{\sigma}{h}\cdot\frac{x - X_j}{h} - \frac{J(F_n(X_i))}{h}\right)\right] \tag{15.13}$$

for $1 \le i \le n$. Consequently, $\hat{\sigma}_x$ is obtained as a minimizer of an expression equivalent to that in (15.8). (For all that, see p. 431 in Eichner and Stute (2013).) The remark after (15.8) regarding convexity of $\sigma \mapsto \widehat{\mathrm{MSE}}_x(\sigma)$ applies here as well.

Note the structural similarities of (15.12) and (15.13) to (15.5) and (15.7), respectively, which can be utilized for a unified implementation of both methods. Note further that $F_n(X_i)$ is the rank of X_i so that summation over the order statistics in (15.12) and (15.13) simplifies $J(F_n(X_i))$ to $J(F_n(X_{i:n})) = J(i/n)$.

15.2.2.1 The Optimal Rank Transformation

The choice of J was determined by the goal to minimize the leading term of the $\mathrm{MSE}\hat{f}_n(x)$ in (15.10) for the optimal ratio h/σ given in (15.11). It turned out that the optimal J is a minimizer of

$$J \to \int_0^1 \frac{1}{J'(u)}du \quad \text{subject to } J' > 0,\ \int_0^1 J(u)du = 0 \text{ and } \int_0^1 J^2(u)du = 1 \tag{15.14}$$

(see Eqs. (8) and (9) in Eichner and Stute (2013)).

An explicit solution was obtained by solving an isoperimetric problem in variational calculus in a class of functions parameterized by a real-valued $c > 0$. This boiled down to solving an Euler-Lagrange differential equation (see Eq. (10) in Eichner and Stute (2013)) within that parameterized function class, which, in turn, led to solving the following cubic equation (see Lemma 3.1 in Eichner and Stute (2013)) within that class:

$$y(u)^3 + 3p_c y(u) + 2q_c = 0, \quad 0 \le u \le 1, \tag{15.15}$$

where, provided that $c \neq \sqrt{3}$,

$$p_c := \frac{1}{5} \cdot \frac{3c^2 - 5}{3 - c^2} \cdot c^2 \quad \text{and} \quad q_c \equiv q_c(u) := \frac{2}{5} \cdot \frac{c^5}{3 - c^2} \cdot (1 - 2u)$$

For $c = \sqrt{3}$ the cubic equation is not necessary because the mentioned differential equation and its solution simplify (see below). Finally, the "admissible" solutions of the cubic equation (15.15) had to be identified in dependence of $c \neq \sqrt{3}$, in the sense that they are the ones which are real-valued, minimize the "target integral" in (15.14) while satisfying the conditions therein and the symmetry condition $y(1) = c = -y(0)$, as well as solve the differential equation. The following results were obtained (see p. 434 in Eichner and Stute (2013)):

- For $c < \sqrt{5/3}$ and for $c > \sqrt{5}$ the real-valued solutions of (15.15) are not admissible because minimization of the target integral in (15.14) is not warranted.
- For $\sqrt{5/3} \leq c < \sqrt{3}$ the only real-valued solution of (15.15) is

$$y(u) = J_1(u; c) := \sqrt[3]{-q_c(u) + \sqrt{q_c^2(u) + p_c^3}} + \sqrt[3]{-q_c(u) - \sqrt{q_c^2(u) + p_c^3}} \tag{15.16}$$

  Note: y tends to the function $u \mapsto \sqrt{3}(2u - 1)$ if $c \nearrow \sqrt{3}$.
- For $c = \sqrt{3}$: $y(u) = \sqrt{3}(2u - 1)$.
- For $\sqrt{3} < c \leq \sqrt{5}$ the solution is

$$y(u) = J_2(u; c) := 2\sqrt{-p_c} \cdot \sin\left\{\frac{1}{3} \arcsin\left\{\frac{q_c(u)}{(-p_c)^{3/2}}\right\}\right\} \tag{15.17}$$

All the above forms of y for $\sqrt{5/3} \leq c \leq \sqrt{5}$ are admissible, but in fact only $c = \sqrt{5}$ minimizes the target integral (see Theorem 3.3 in Eichner and Stute (2013)). Hence, $u \mapsto J(u) := J_2(u; \sqrt{5})$ is the sought-after optimal rank transformation.

15.2.3 Kernel Adjusted Nonparametric Regression

Eichner and Stute (2012) suggested to replace the kernel in the Nadaraya-Watson estimator (15.2) analogously to the approach in Sect. 15.2.1 by a location-scale family associated with the classical kernel density estimator f_n in (15.1) using an initial kernel K (preferably nonnegative and with unbounded support). They thus obtained as estimator the weighted sum

$$\hat{m}_n(x) \equiv \hat{m}_n(x; h, \theta, \sigma) := \sum_{i=1}^{n} W_{ni}(x) \cdot Y_i \tag{15.18}$$

with weights

$$W_{ni}(x) \equiv W_{ni}(x; h, \theta, \sigma) := \frac{f_n\left(\sigma \frac{x - X_i}{h} + \theta\right)}{\sum_{k=1}^{n} f_n\left(\sigma \frac{x - X_k}{h} + \theta\right)}$$
$$= \frac{\sum_{j=1}^{n} K\left(\frac{\sigma}{h} \cdot \frac{x - X_i}{h} + \frac{\theta - X_j}{h}\right)}{\sum_{1 \le k, j \le n} K\left(\frac{\sigma}{h} \cdot \frac{x - X_k}{h} + \frac{\theta - X_j}{h}\right)} \quad (15.19)$$

which are non-negative and sum to 1 for all $x \in \mathbb{R}$ (see proof of Theorem 1 on p. 2542 in Eichner and Stute (2012)).

With $\theta = \mathbb{E}[X]$, Theorem 1 in Eichner and Stute (2012) implies (under $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}[X^2] < \infty$ with X having a density f that satisfies $f(x) > 0$ and is continuously differentiable in a neighborhood of x, m being twice continuously differentiable in a neighborhood of x, and K being a symmetric probability density with $\int |w^3| K(w) dw < \infty$) that the asymptotic MSE (omitting negligible terms) satisfies

$$\begin{aligned} \text{MSE}\, \hat{m}_n(x) &\equiv \left(\text{Bias}\, \hat{m}_n(x)\right)^2 + \text{Var}\, \hat{m}_n(x) \\ &= \frac{1}{4}\left(\frac{2f'(x)m'(x) + f(x)m''(x)}{f(x)}\right)^2 \frac{h^4}{\sigma^4}\, \text{Var}^2 X \\ &\quad + \frac{\sigma \int f^2(u) du}{nhf(x)}\, \text{Var}(Y|X = x) \end{aligned} \quad (15.20)$$

as $h \to 0$ and $n \to \infty$ such that $nh \to \infty$ (see also Eq. (2.2) in Eichner and Stute (2012)). Apparently, the asymptotic MSE depends (as in kernel adjusted density estimation) on the *ratio* of h and σ. This MSE is minimal if

$$\frac{h}{\sigma} = \left(\frac{1}{n}\right)^{1/5} \left(\frac{f(x) \int f^2(u) du\, \text{Var}(Y|X = x)}{[2f'(x)m'(x) + f(x)m''(x)]^2\, \text{Var}^2 X}\right)^{1/5}$$

In particular, one may choose also here $h = n^{-1/5}$ without loss of generality so that the bias is asymptotically neglible and the MSE is a function of σ alone. Recall that θ has been set to the (typically) unknown $\mathbb{E}[X]$. Fortunately, the asymptotic representation in (15.20) holds also for $\sqrt{n}$-consistent estimators $\hat{\theta}$ of θ, so that in a real data situation where θ is unknown (but realistically assumed to be contained in a compact set), the estimator $\hat{\theta} = n^{-1} \sum_{i=1}^{n} X_i$ can (and typically will) replace it.

Minimization of an estimator $\widehat{\text{MSE}}_x(\sigma)$ of MSE $\hat{m}_n(x)$ in σ, however, is not based on the analytic form of its representation in (15.20). Moreover, estimators for explicit expressions of Bias $\hat{m}_n(x)$ and Var $\hat{m}_n(x)$, respectively, are utilized without the need

for preliminary estimates of the unknown local and global quantities appearing in (15.20). In particular, the bias estimator was set to

$$\widehat{\mathrm{Bias}}_x(\sigma) := \sum_{i=1}^{n} W_{ni}(x)\{m_n(X_i) - m_n(x)\} \tag{15.21}$$

and the variance estimator to

$$\widehat{\mathrm{Var}}_x(\sigma) := \sum_{i=1}^{n} W_{ni}^2(x)\{Y_i - m_n(X_i)\}^2, \tag{15.22}$$

where in both definitions m_n is the classical Nadaraya-Watson estimate from (15.2) with $h = n^{-1/5}$. Note that σ only enters the weights W_{ni}. Note also that in Eichner and Stute (2012) it is recommended to not replace m_n by $\hat{m}_n$ here. Note further that this estimation procedure is completely different from the one followed in kernel adjusted density estimation of Srihera and Stute (2011) summarized in Sect. 15.2.1. (See p. 2540 in Eichner and Stute (2012) for the derivation and the reasoning behind it.)

Finally, the choice for σ was taken to be a minimizer $\hat{\sigma}_x$ of

$$\widehat{\mathrm{MSE}}_x(\sigma) := \left(\widehat{\mathrm{Bias}}_x(\sigma)\right)^2 + \widehat{\mathrm{Var}}_x(\sigma)$$

which is apparently possible without referring to the unknown local and global quantities in (15.20). According to empirical evidence, the remark after (15.8) regarding convexity of $\sigma \mapsto \widehat{\mathrm{MSE}}_x(\sigma)$ appears to apply here only in a weaker form.

15.3 Implementations

Caveat: For the following it is assumed that the reader is familiar with R.

Package `kader` in its latest version should be readily available for download and installation from "The Comprehensive R Archive Network" (CRAN) at https://CRAN.R-project.org/ in the "Packages" section, or directly from https://CRAN.R-project.org/package=kader. After successful installation, attaching the package to R's search path with `library(kader)` should enable one to reproduce the following examples.

15.3.1 R Function `kade` for Kernel Adjusted Density Estimation

The main function to compute the kernel adjusted density estimators of Sects. 15.2.1 and 15.2.2 with an MSE-estimator-minimizing σ for a given data set of real values is `kade`. To illustrate its mode of operation we shall apply it to a univariate data set

that is available in R. The data are $n = 272$ durations of eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. They are contained in the component `eruptions` of the data frame `faithful`. (Details may be found in `faithful`'s help page.) Fig. 15.1 shows just for illustrative purposes a "classical" kernel density estimate obtained with the R-function `density` which comes with R's base distribution in its package `stats`. It computes a kernel density estimate with a selectable kernel (the gaussian kernel by default) and a global bandwidth that is (here) choosen by a method which has been suggested by Sheather and Jones (1991) and is also recommended by Venables and Ripley (2002). More information is available on `density`'s and `bw.SJ`'s help pages. In addition, the raw data are indicated by a rug plot on the x-axis. Figure 15.1 was created by

```
> plot(density(faithful$eruptions, bw = "SJ"))
> rug(faithful$eruptions, col = "blue")
```

Before proceeding to concrete examples of use of `kade` we give an (incomplete) overview over its arguments:

- `x`: Vector of location(s) at which the density estimate is to be computed.
- `data`: Vector $(X_1, \ldots, X_n)$ of the data from which the estimate is to be computed.
- `kernel`: A character string naming the kernel to be used for the adjusted estimator. This must partially match (currently) one of `"gaussian"`, `"rectangular"`, or `"epanechnikov"`, with default `"gaussian"`. It may be abbreviated to a unique prefix.
- `method`: A character string naming the method to be used for the adaptive estimator. This must partially match one of `"both"`, `"ranktrafo"`, or `"nonrobust"`, with default `"both"`, and may be abbreviated to a unique prefix.

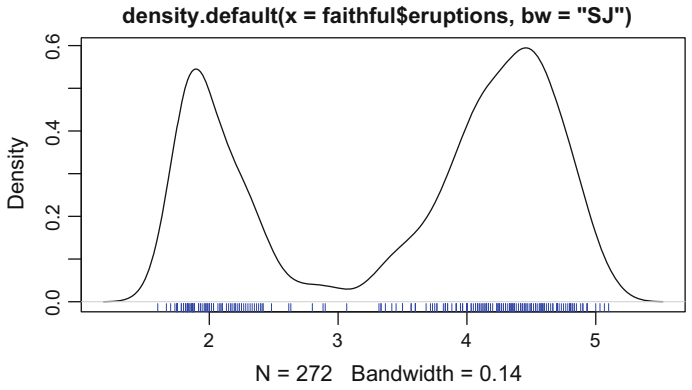


Fig. 15.1 "Classical" kernel density estimate for the Old Faithful eruptions data as produced by R's built-in function `density` using the gaussian kernel with a provided bandwidth selection procedure, augmented with a rug plot of the data

- `Sigma`: Vector of value(s) of the scale parameter σ. If of length 1 no adaptation is performed, but just the estimator of (15.3) or (15.9) for the given value of σ computed. Otherwise, `Sigma`'s value is considered as the grid over which the optimization of the adaptive method will be performed. Defaults to `seq(0.01, 10, length = 51)`.
- `h`: Numeric scalar for bandwidth h. Defaults to `NULL` which leads to automatically setting $h := n^{-1/5}$.
- `theta`: Numeric scalar for the value of the location parameter θ. It defaults to `NULL` which leads to setting it to the arithmetic mean of the data $X_1, \ldots, X_n$.
- `ranktrafo`: Function used for the rank transformation. Defaults to the optimal rank transformation in (15.17) with $c = \sqrt{5}$ (which is implemented in a function named `J2` with its default `cc = sqrt(5)`; see Sect. 15.3.1.3).
- `plot`: Should graphical output be produced? Defaults to `FALSE` which means no plotting. If set to `TRUE` a plot of the estimators of the squared bias, the variance, and of the MSE as functions of σ (for the values in `Sigma`) is produced *for each* value in `x`. If provided with a character string (which can, of course, also contain a file path) the output is directed to automatically numbered (from 1 to the length of `x`) pdf-files whose names and location in your file system are determined by the given character string.
- `parlist`: A list of graphical parameters (which is passed to a function named `adaptive_fnhat` coordinating the adaptive procedure and doing some of the plotting). It affects only the pdf-files which are created if the aforementioned `plot` is set to `TRUE`. Default: `NULL`.

15.3.1.1 Non-robust Kernel Adjusted Density Estimation

To compute the kernel adjusted density estimator of (15.3) using the gaussian kernel, we have to set `method = "nonrobust"` and would have to set `kernel = "gaussian"`, if the latter wasn't the default. So, we can simply omit the `kernel`-argument in the argument list of the call:

```
> X <- faithful$eruptions
> est1 <- kade(x = pretty(X, n = 50), data = X,
+   method = "nonrobust", Sigma = seq(0.01, 10, length = 21),
+   plot = "Plots/Est1_MSEx",     #'Plots' must already exist!
+   parlist = list(mar = c(2.5, 2, 2, 0.5), tcl = -0.3,
+                  mgp = c(1.3, 0.5, 0), cex = 1.1))
```

The argument `x` of `kade` expects the point(s) at which the density estimator is to be evaluated. Here, we let R determine an equidistant grid of roughly 50 "pretty" points which covers the range of the data (see `?pretty` for details). The sample data are passed to `kade` through its argument `data`. A vector of values for the scale parameter σ can be provided to `Sigma`, but doesn't have to because its default is a vector with a grid of equidistant values (currently `seq(0.01, 10, length = 51)`). Here we have chosen a coarser grid to reduce the computational workload since the optimization of the adaptive method will be performed on the provided grid

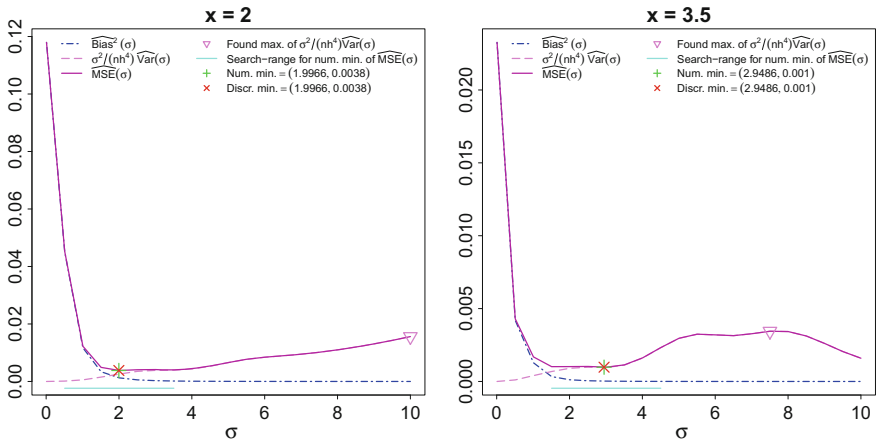


Fig. 15.2 Typical estimators of squared bias, variance, and MSE as functions of σ in kernel adjusted density estimation for the Old Faithful eruptions data, each for the single x which is reported in the respective plot title

as follows: First, a minimizer of $\widehat{\mathrm{MSE}}$ is searched on that grid, and then it is refined by a numerical minimization in its neighborhood. (More details are provided in the remarks on page 306–307.) Since `plot` receives a character string one (colored) plot is produced *for each* value in the vector that is assigned to `x` presenting the estimators of the "technical quantities" squared bias, variance and MSE as functions of σ for the values in `Sigma`. (Here, the plots are placed in a subdirectory "Plots" which has to exist already in R's current working directory.) Two selected examples of this plot are shown in Fig. 15.2 and indicate the mentioned convexity of $\widehat{\mathrm{MSE}}$ for small σ. (Argument `parlist` is here provided with settings to reduce the amount of "white space" around the graph.)

`kade`'s return-value is a data frame with one row for each element of `x`; here we only show the first three. But note also that, since `kade` usually produces numerous status messages while it is working, the presented output is anyway not complete:

```
> head(est1, 3)
```

```
     x         y sigma.adap  msehat.min discr.min.smaller sig.range.adj
1 1.60 0.1920315   1.253887 0.001658418             FALSE             0
2 1.65 0.2132562   1.547541 0.002023333             FALSE             0
3 1.70 0.2414115   1.962614 0.002261519             FALSE             0
```

Column `x` contains each x-value at which the density estimator was computed, `y` the corresponding values of the density estimator, `sigma.adap` the respective values of $\hat{\sigma}_x$, and `msehat.min` the pertaining values of $\widehat{\mathrm{MSE}}(\hat{\sigma}_x)$. (The remaining two columns report technical information: if the minimizer found by grid search is smaller than the refined one found by numerical minimization (`discr.min.smaller`), and the number of times that the search range for $\hat{\sigma}_x$ had to be extended beyond the initial σ-grid during the minimization process (`sig.range.adj`).)

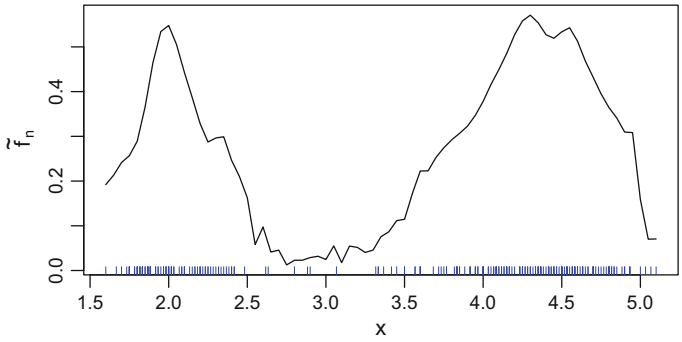


Fig. 15.3 Kernel adjusted density estimation (pointwise adapted) for the Old Faithful eruptions data on an x-grid of 71 equidistant points, overlaid with a rug plot of the data

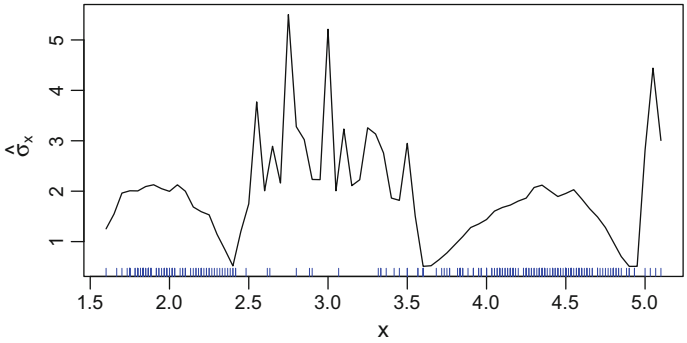


Fig. 15.4 Kernel adjusted density estimation: $\hat{\sigma}_x$ versus x for the same x-grid and data as in Fig. 15.3, overlaid with a rug plot of the data

With a sufficiently fine grid of x-values, and after having the necessary patience regarding the required computing time, a plot of the graph of the (pointwise (!) adapted) estimated density is immediately created from `kade`'s result (see Fig. 15.3). Similarly, a plot of $\hat{\sigma}_x$ vs. x is as easily produced, e.g., to gain some empirical insight into how the minimizer $\hat{\sigma}_x$ "depends" on x and the data distribution (see Fig. 15.4). The following two code snippets lead to the two figures.

```
> with(est1, plot(x, y, type = "l",
+                 ylab = expression(tilde(f)[n])))   # Fig. 3
> rug(faithful$eruptions, col = "blue")   # Rug plot of data

> with(est1, plot(x, sigma.adap, type = "l",
+                 ylab = expression(hat(sigma)[x]))) # Fig. 4
> rug(faithful$eruptions, col = "blue")   # Rug plot of data
```

15.3.1.2 Kernel Adjusted Density Estimation Using Rank Transformations

To compute the robustified kernel adjusted density estimator of (15.9) using rank transformations while otherwise using the same settings as in the previous section, i.e., the gaussian kernel, etc., we just have to change the `method`-argument to `method = "ranktrafo"`. It is also possible to provide a user-defined function for the rank transformation to the argument `ranktrafo`. Its default is the MSE-optimal rank transformation implemented in `J2`. Section 15.3.1.3 provides more details on this. (Only the call to `kade`, but nothing of its numerical output is shown here since it is completely analogous to that in the previous section.) For the sake of comparison Fig. 15.5 shows the plots of estimators of the squared bias, variance and MSE as functions of σ for the values in `Sigma` corresponding to the plots in Fig. 15.2. They also indicate the mentioned convexity of $\widehat{\mathrm{MSE}}$ for small σ (even more convincing than for the non-robust method).

```
> X <- faithful$eruptions
> est2 <- kade(x = pretty(X, n = 50), data = X,
+   method = "ranktrafo", Sigma = seq(0.01, 10, length = 21),
+   plot = "Plots/Est2_MSEx",
+   parlist = list(mar = c(2.5, 2, 2, 0.5), tcl = -0.3,
+                  mgp = c(1.3, 0.5, 0), cex = 1.1))
```

The plot of the graph of the (pointwise (!) adapted) robustly estimated density using the MSE-optimal rank transformation is seen in Fig. 15.6, and that of $\hat{\sigma}_x$ versus x is presented in Fig. 15.7. (The code that produced those two graphs is completely analogous to the respective code near the end of Sect. 15.3.1.1 and thus not shown.) Note the increased smoothness of both the density estimator and of $x \mapsto \hat{\sigma}_x$ in comparison to Figs. 15.3 and 15.4 of the non-robust method.

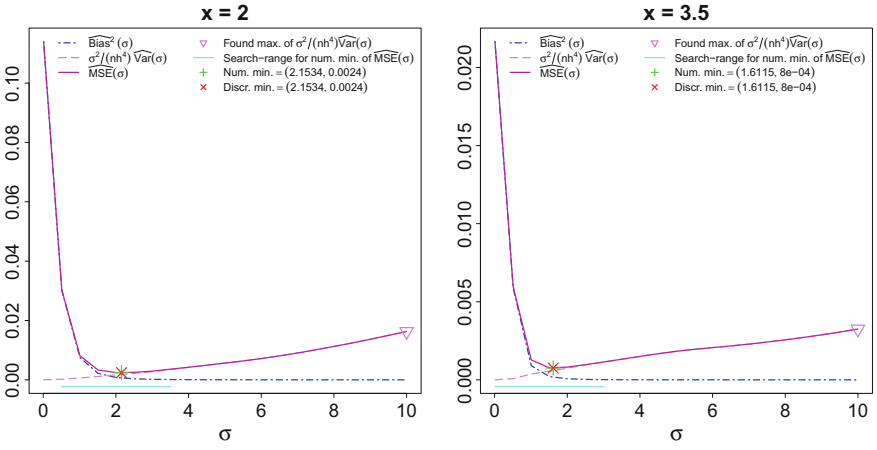


Fig. 15.5 Typical estimators of squared bias, variance, and MSE as functions of σ using the MSE-optimal rank transformation in robust kernel adjusted density estimation for the Old Faithful eruptions data, each for the single x reported in the plot title

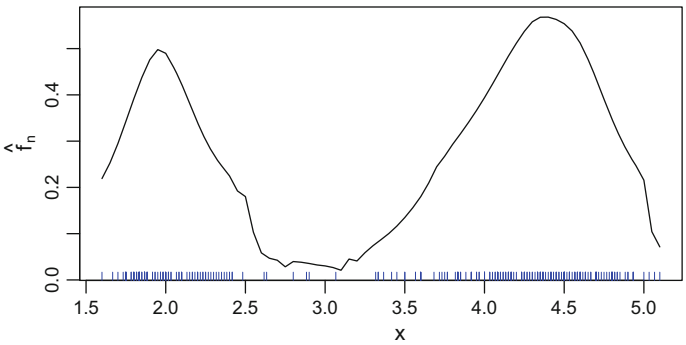


Fig. 15.6 Kernel adjusted density estimation (pointwise adapted) using the MSE-optimal rank transformation for the Old Faithful eruptions data on an x-grid of 71 equidistant points, overlaid with a rug plot of the data

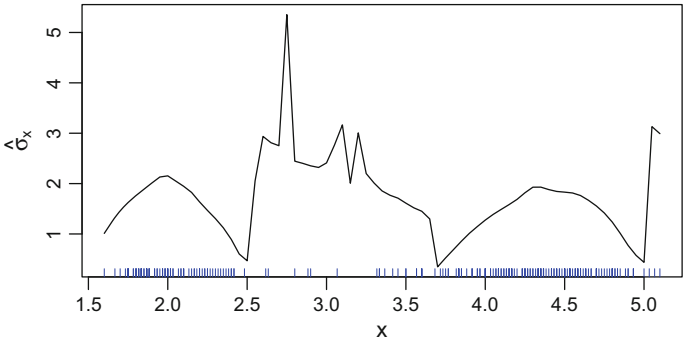


Fig. 15.7 Kernel adjusted density estimation using the optimal rank transformation: $\hat{\sigma}_x$ versus x for the same x-grid and data as in Fig. 15.6, overlaid with a rug plot of the data

Remarks concerning the implementations of both estimation methods:

- The help pages of the introduced R-functions provide further information, and in particular the examples there illustrate additional variants of using them.
- There are custom-made functions `fnhat_SS2011` and `fnhat_ES2013` which prepare the evaluation of (15.3) and (15.9), respectively, for a set of x-values, a data vector $(X_1, \ldots, X_n)$, a kernel K with bandwidth h, a single scale parameter value σ, and either a location parameter θ or a rank transformation J: `fnhat_ES2013` actually mainly precomputes the vector $(-J(1/n), \ldots, -J(n/n))$ for the rank-transformation method while `fnhat_SS2011` does the respective for $(\hat{\theta} - X_1, \ldots, \hat{\theta} - X_n)$ for the non-robust method. They then both call the function `compute_fnhat` which does what its name already insinuates.

The examples section on the help pages of those functions demonstrate their use extensively. Note, however, that the graphs which are created there are computed for a single fixed σ, i.e., without adaption! They are produced for demonstration purposes only.

- Most of the number-crunching parts of the implementation are without loops, but instead use matrix-vector-calculus. (This is usually pretty fast, but can, in turn, be quite memory-intensive already for moderate sample sizes of $n \geq 200$ in combination with even not-so-large x-grids.) Exceptions from the matrix-calculus are the handling of the elements in `x` and (for each `x`-element) the discrete miminizer search along the σ-grid. Both processes iterate through the respective vectors `Sigma` and `x`.
- The search for the miminizer of the estimated MSE (like in (15.8)) proceeds by first searching a minimizer of $\widehat{\mathrm{MSE}}$ on the σ-grid provided in `Sigma`, and then refining the found discrete minimizer by a numerical minimization in its neighborhood. This takes currently the most time, both in the discrete grid-search and when using R's numerical optimization routine `optimize`. This is due to the fact that the integral in the bias-estimator in (15.5) and (15.12) is computed numerically by means of R's `integrate`. A function named `bias_AND_scaledvar` coordinates and triggers this repeatedly so that it could be termed *the* "workhorse" function here, but actually `integrate` does most of the (so-to-say internal) computational work together with a function `kfn_vectorized`. The latter realizes a vectorized (in u) evaluation of the integrand $n^{-1} \sum_{i=1}^{n} K(u) \cdot f_n(x - h/\sigma \cdot (\hat{\theta} - X_i - hu))$ in (15.5) and of an analogous version of (15.12). This integration is currently the main source of computing effort and time, and one of the targets of future improvements.
- Empirical evidence suggests that the σ-grid should start near zero. By default this is currently the case with the smallest σ-value being 0.01.
- The numerous status messages of `kade` can generally be suppressed if a call of `suppressMessages` is "wrapped around" the call of `kade`.

15.3.1.3 The Optimal Rank Transformation

All functions from Sect. 15.2.2.1 are implemented, notably:

- p_c and $q_c(u)$ of the cubic equation (15.15) as `pc` and `qc`, respectively.
- $J_1(u; c)$ of (15.16) and $J_2(u; c)$ of (15.17) as `J1` and `J2`, respectively. Note that the optimal rank transformation is hence implemented as `J2`, because its default is to use $c = \sqrt{5}$.
- The function `J_admissible` provides the complete family of admissible function forms of y mentioned at the end of Sect. 15.2.2.1.

Thus, it is simple to take an "empirical" look at the shapes and the behavior of the admissible functions. Figure 15.8, as an example, is the result of the following code:

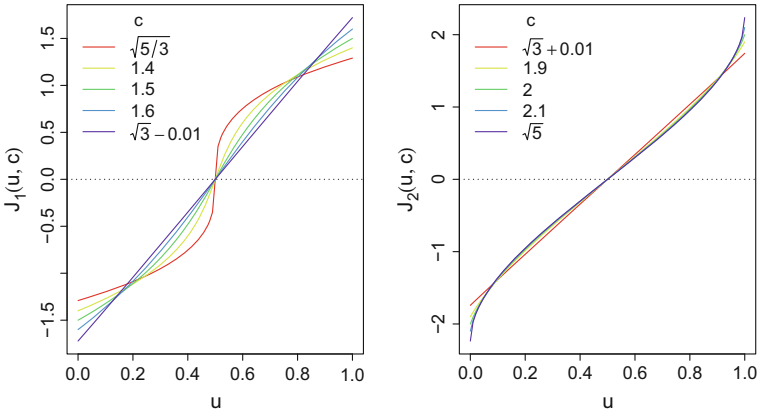


Fig. 15.8 Shapes of admissible forms of the rank transformation for selected values of c. Recall: For $c = \sqrt{3}$ the admissible form is linear, namely $u \mapsto \sqrt{3}(2u - 1)$, but this is not shown here

```
> par(mfrow = c(1, 2), mar = c(3, 3, 0.5, 0.5), mgp = c(1.5, 0.5, 0),
+   tcl = -0.3)   # Reduce white space around plots.
> u <- seq(0, 1, by = 0.01)
>  # Family of functions for c-grid in [\sqrt(5/3), \sqrt(3)); left plot.
>  # (expression() used only to have mathematical notation in legend.)
>  #-------------------------------------------------------------------------
> c0 <- expression(sqrt(5/3));         c1 <- expression(sqrt(3) - 0.01)
> cgrid <- seq(1.4, 1.6, by = 0.1);   cvals <- c(eval(c0), cgrid, eval(c1))
> Y <- sapply(cvals, function(cc, u) J1(u, cc = cc), u = u)
> cols <- rainbow(ncol(Y), end = 3/4) # Just to let the graphs look fancy.
> matplot(u, Y, type = "l", lty = "solid", col = cols,
+   ylab = expression(J[1](u, c)));   abline(h = 0, lty ="dotted")
> legend("topleft", title = "c", legend = c(c0, cgrid, c1), lty = 1,
+   col = cols, bty = "n")

>  # Family of functions for c-grid in (\sqrt(3), \sqrt(5)]; right plot.
>  #-------------------------------------------------------------------------
> c0 <- expression(sqrt(3) + 0.01);   c1 <- expression(sqrt(5))
> cgrid <- seq(1.9, 2.1, by = 0.1);   cvals <- c(eval(c0), cgrid, eval(c1))
> Y <- sapply(cvals, function(cc, u) J2(u, cc = cc), u = u)
> cols <- rainbow(ncol(Y), end = 3/4)
> matplot(u, Y, type = "l", lty = "solid", col = cols,
+   ylab = expression(J[2](u, c)));   abline(h = 0, lty = "dotted")
> legend("topleft", title = "c", legend = c(c0, cgrid, c1), lty = 1,
+   col = cols, bty = "n")
```

Note: `J1` accepts $c \notin [\sqrt{5/3}, \sqrt{3})$ and issues then only a warning; likewise, `J2` accepts $c \notin (\sqrt{3}, \sqrt{5}]$ and issues also only a warning!

15.3.2 R Function `kare` for Kernel Adjusted Regression Estimation

`kare` is the main function to compute the kernel adjusted regression estimator in (15.18) with an MSE-estimator-minimizing σ for a given data set of univariate regressor values $X_1, \ldots, X_n$ and the pertaining univariate response values $Y_1, \ldots, Y_n$. To illustrate its use we shall apply it to a data set available in R that comprises $n = 116$ (non-missing) observations of temperature and ozone concentration. They are contained in the respective components `Ozone` and `Temp` of the data frame `airquality`. (For details, see `?airquality`.) Fig. 15.9 shows the scatter plot of the ozone versus the temperature values, for the sake of illustration overlaid with two curves obtained with the "classical" kernel regression estimator of Nadaraya and Watson in (15.2) using two different bandwidths. The R-function `ksmooth` which comes with R's base distribution in its package `stats` was used to produce this result. It computes the Nadaraya-Watson estimate (here) with the "normal", i.e., gaussian, kernel, a given bandwidth and, by default, on a grid of at least 100 equidistant points over the range of the regressor variable. Details are available on `ksmooth`'s help page. Figure 15.9 was created by

```
> with(na.omit(airquality[c("Temp","Ozone")]), {
+   plot(Temp, Ozone)
+   lines(ksmooth(Temp, Ozone,"normal", bandwidth = 2), col = "red")
+   lines(ksmooth(Temp, Ozone,"normal", bandwidth = 5), col = "green")
+ })
```

Before applying `kare` to the example data we list its main arguments:

- `x.points`: Vector of location(s) at which the regression estimate is to be computed.
- `data`: Data frame or list with components named `x` and `y`, where `x` contains a numeric vector of the regressor values $X_1, \ldots, X_n$ and `y` a numeric vector of the

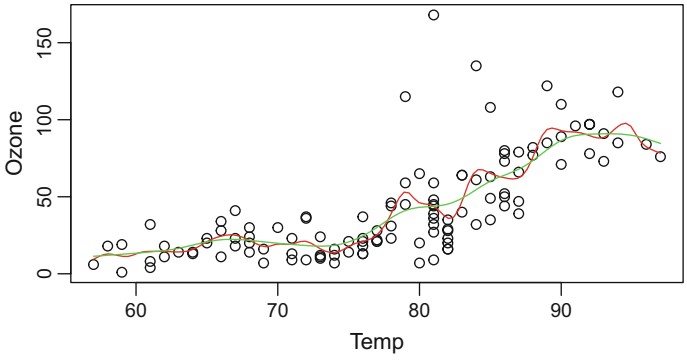


Fig. 15.9 "Classical" Nadaraya-Watson kernel regression estimate for the ozone data as produced by R's built-in function `ksmooth` using the gaussian kernel with two different bandwidths

pertaining response values $Y_1, \ldots, Y_n$ of the data for which the estimate is to be computed.

- `kernel`: A character string naming the kernel to be used for the adjusted estimator. This must partially match (currently) one of `"gaussian"`, `"rectangular"`, or `"epanechnikov"`, with default `"gaussian"`. It may be abbreviated to a unique prefix.
- `Sigma`: Vector of value(s) of the scale parameter σ. If of length 1 no adaptation is performed, but just the estimator in (15.18) for the given value of σ computed. Otherwise, `Sigma`'s contents is considered as the grid over which the optimization of the adaptive method will be performed. Defaults to `seq(0.01, 10, length = 51)`.
- `h`: Numeric scalar for bandwidth h. Defaults to `NULL` which leads to automatically setting $h := n^{-1/5}$.
- `theta`: Numeric scalar for the value of the location parameter θ. It defaults to `NULL` which leads to setting it to the arithmetic mean of the design values $X_1, \ldots, X_n$.

The following code snippet prepares the input for `kare`'s argument `data` by creating a data frame with suitably named components and removing its rows with missing values. The, it calls `kare` with a single value for its argument `x.points` at which the regression estimator is to be evaluated. The sample data are passed through its argument `data`. A vector of values for the scale parameter σ can be provided to `Sigma`, but doesn't have to because its default is a vector with a grid of equidistant values (currently `seq(0.01, 10, length = 51)`). Here we stick with the default grid because the computational workload is not at all as high as in the current implementation of kernel adjusted density estimation of the previous sections. If a scalar is provided for `Sigma` no adaptation is performed, but only the computation of the regression estimator for that σ. Otherwise the optimization of the adaptive method will be performed, but only on the given σ-grid, i.e., no (possibly refining) numerical minimization is currently implemented. (There is currently also no automatic plotting provided by `kare`.)

```
> data <- na.omit(with(airquality,
+                      data.frame(x = Temp, y = Ozone)))
> fit1 <- kare(x.points = 75, data = data)
```

`kare`'s return value is a list of eight components if `x.points` receives only a scalar as above:

```
> str(fit1)

List of 8
 $ x          : num 75
 $ y          : num 16.8
 $ sigma.adap: num 3.21
 $ msehat.min: num 1.5
 $ Sigma      : num [1:51] 0.01 0.21 0.41 0.609 0.809 ...
```

```
 $ Bn          : num [1:51] 23.2 19.9 18.5 14.1 11.1 ...
 $ Vn2         : num [1:51] 2.72 2.19 3.04 3.31 3.6 ...
 $ MSE         : num [1:51] 541 397 345 201 127 ...
```

Component `x` contains the x-value at which the regression estimator was computed, `y` the value of the estimator, `sigma.adap` the value of $\hat{\sigma}_x$, and `msehat.min` the pertaining value of $\widehat{\mathrm{MSE}}(\hat{\sigma}_x)$. `Sigma` contains the σ-grid on which the minimization process was performed while `Bn`, `Vn2`, and `MSE` contain vectors with the estimators of variance, bias, and MSE, respectively, on that σ-grid, e.g., to be able to draw them for visualisation purposes (as will be demonstrated shortly).

If `x.points` receives a vector of $k > 1$ elements (as below), `kare` returns a list with the same component names, but then `x` contains the k-vector $(x_1, \ldots, x_k)$ in `x.points` (at which the regression estimator was computed), `y` the vector of the k estimator values $(\hat{m}_n(x_1), \ldots, \hat{m}_n(x_k))$, `sigma.adap` the vector of the $\hat{\sigma}_{x_j}$, $j = 1, \ldots, k$, and `msehat.min` the pertaining values of $\widehat{\mathrm{MSE}}(\hat{\sigma}_{x_j})$. `Sigma` contains the σ-grid of length, say, s on which the minimization process was performed while `Bn`, `Vn2`, and `MSE` contain $(s \times k)$-matrices with the estimators of variance, bias, and MSE, respectively, on that σ-grid in their columns (which correspond to the k x-values). Let's execute `kare` for this case and present the structure of its return value:

```
> xgrid <- seq(55, 100, by = 0.5)
> str(fit <- kare(x.points = xgrid, data = data))

List of 8
 $ x          : num [1:91] 55 55.5 56 56.5 57 57.5 58 58.5 ...
 $ y          : num [1:91] 6 6 6 6 ...
 $ sigma.adap: num [1:91] 3.21 3.81 4.81 6.4 ...
 $ msehat.min: num [1:91] 0.332 0.332 0.332 0.32 ...
 $ Sigma      : num [1:51] 0.01 0.21 0.41 0.609 ...
 $ Bn         : num [1:51, 1:91] 37.9 28.6 11.3 11.4 ...
 $ Vn2        : num [1:51, 1:91] 3.09 2.26 1.38 2.9 ...
 $ MSE        : num [1:51, 1:91] 1438 820 129 132 ...
```

Using the returned information, Fig. 15.10 was created with the code below and displays the regression estimator (top) on the x-grid as well as the estimators of squared bias, variance, and MSE on the σ-grid for one selected x-value in the two lower plots).

```
>  # Open a graphics device and preparing its layout for 3 plots to come:
> par(mfrow = c(3, 1), mar = c(3, 3, 2, 0.1), mgp = c(1.6, 0.5, 0),
+   tcl = -0.3, cex.main = 1.4)
>  # The scatter plot of the"raw data":
> plot(y ~ x, data = data, xlim = range(data$x, fit$x),
+   ylim = range(data$y, fit$y, na.rm = TRUE),
+   main = bquote(n == .(nrow(data))), xlab ="Temp", ylab ="Ozone")
>  # Overlay the graph of the obtained estimator on the x-grid:
> lines(x = fit$x, y = fit$y, col ="red")
```

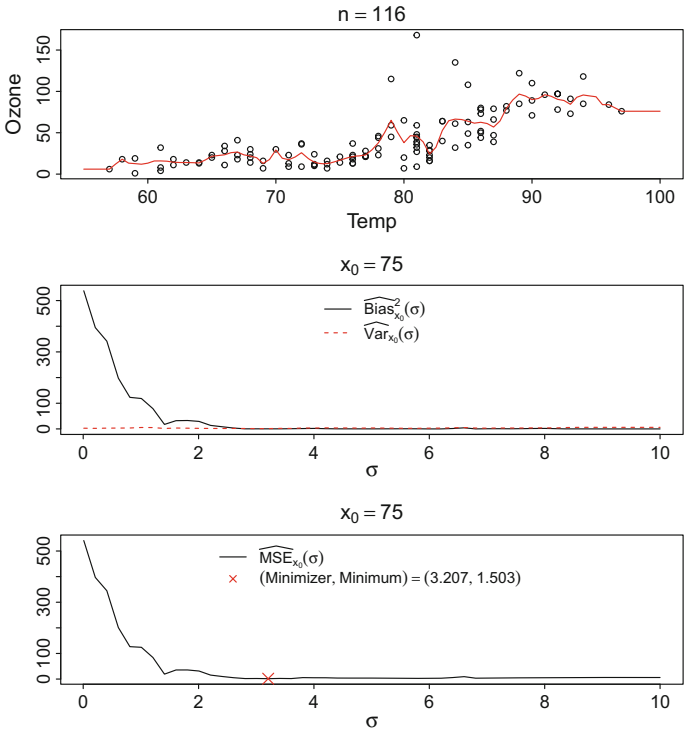


Fig. 15.10 *Top* Kernel adjusted regression estimator for the ozone data as produced by `kare`. *Middle* estimators of squared bias and variance on the σ-grid for one selected x-value (reported in the plot title). *Bottom* estimator of MSE on the σ-grid for one selected x-value (reported in the plot title) together with the detected minimum. Note how the estimated MSE is here dominated by the squared estimated bias

```
>  # Draw the estimators of (Bias_x(sigma))^2 and Var_x(sigma) on
>  # the sigma-grid for one selected x:
> ix <- 41   # Index of selected point of x-grid
> with(fit,
+   matplot(Sigma, cbind(Bn[, ix]^2, Vn2[, ix]), type = "l", lty = 1:2,
+     col = c("black","red"), xlab = expression(sigma), ylab = "",
+     main = bquote(x[0] == .(x[ix]))))
>  # Legend for the estimators:
> legend("top", lty = 1:2, col = c("black","red"), bty = "n", cex = 1.2,
+   legend = c(expression(paste(widehat(plain(Bias))[x[0]]^2, (sigma))),
+              expression(widehat(plain(Var))[x[0]](sigma))))

>  # Draw the estimator of MSE_x(sigma) on the sigma-grid together
>  # with the point indicating the detected minimum, and a legend:
> with(fit, {
+   plot(Sigma, MSE[, ix], type = "l", xlab = expression(sigma),
+     ylab = "", main = bquote(x[0] == .(x[ix])))
+   points(sigma.adap[ix], msehat.min[ix], pch = 4, col = "red", cex = 2)
+   legend("top", lty = c(1, NA), pch = c(NA, 4), col = c("black","red"),
```

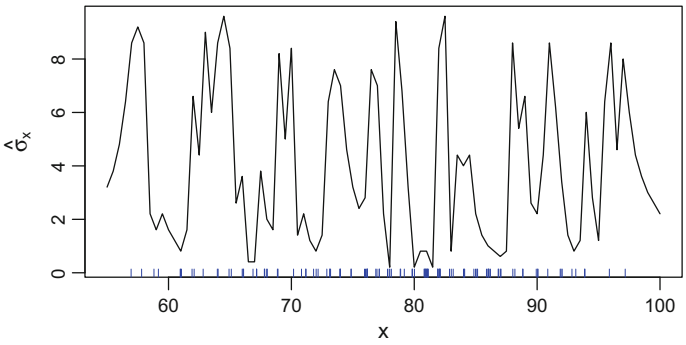


Fig. 15.11 Kernel adjusted regression estimation for the ozone data: $\hat{\sigma}_x$ versus x on an x-grid of 91 equidistant points, overlaid with a rug plot of the (slightly jittered) regressor data. (Remember that also the Y_is enter the variance estimator and hence affect $\hat{\sigma}_x$, so that this simple display is certainly far from telling the "whole story")

```
+     legend = c(expression(widehat(plain(MSE))[x[0]](sigma)),
+       substitute(group("(", list(plain(Minimizer), plain(Minimum)),")")
+                  == group("(", list(x, y),")"),
+         list(x = signif(sigma.adap[ix], 4),
+              y = signif(msehat.min[ix], 4)))), bty = "n", cex = 1.2)
+ })
```

The fine grid of x-values used for `fit` allows to plot the graph of $\hat{\sigma}_x$ versus x, e.g., to gain some empirical insight into how the minimizer $\hat{\sigma}_x$ "depends" on x and the regressor distribution. The following code snippet produced Fig. 15.11. (Note that the added rug plot of the regressor data is here slightly jittered to reduce overplotting of tied values.)

```
> with(fit, plot(x, sigma.adap, type = "l",
+                ylab = expression(hat(sigma)[x])))   # Fig. 11
> set.seed(2017)                      # Reproducibly jittered rug
> rug(jitter(data$x), col = "blue") # plot of regressor data.
```

A few **remarks** concerning the implementation of the estimation method:

- There is a function named `weights_ES2012` that implements the weights of (15.19) for the pre-computed quantities $(x - X_1)/h, \ldots, (x - X_n)/h$ and $(\theta - X_1)/h, \ldots, (\theta - X_n)/h$ where x is the single (!) location for which the weights are to be computed, θ is the location parameter, the X_i's are the regressor values, and h is the bandwidth.
- Based on `weights_ES2012` the bias estimator of (15.21) and the variance estimator of (15.22) are computed by functions `bias_ES2012` and `var_ES2012`, respectively. They expect the same arguments as function `weights_ES2012`, but the first one needs in addition the pre-computed vector $(m_n(X_1) - m_n(x), \ldots, m_n(X_n) - m_n(x))$ and the latter one the vector $((Y_1 - m_n(x))^2, \ldots, (Y_n - m_n(x))^2)$.

Concluding remarks regarding future package versions: The main focus shall be an increase in computing efficiency, considering both mathematical and programming tools to achieve this. Another goal is to supply numerical minimization also in `kare` to refine its grid search. Harmonizing `kade`'s and `kare`'s return values and their argument lists with respect to contents and structure is desirable, but would brake backwards compatibility which needs to be considered carefully. Finally, implementing the L_2-approach of kernel adjusted density estimation is on the to-do list.

Acknowledgements Thanks to the two referees whose constructive criticism and suggestions helped to improve the paper considerably.

References

Bowman, A.W., Azzalini, A.: sm: nonparametric smoothing methods. R package version 2.2-5.4 (2014). Available from http://www.stats.gla.ac.uk/~adrian/sm or http://azzalini.stat.unipd.it/Book_sm or https://CRAN.R-project.org/package=sm. Last accesses: November 2016.

Duong, T., Wand, M.: feature: Local Inferential Feature Significance for Multivariate Kernel Density Estimation. R package version 1.2.13. (2015). Available from https://CRAN.R-project.org/package=feature. Last access: November 2016.

Duong, T.: ks: Kernel Smoothing. R package version 1.10.4 (2016). Available from https://CRAN.R-project.org/package=ks. Last access: November 2016.

Eichner, G., Stute, W.: Kernel adjusted nonparametric regression. J. Stat. Plan. Infer. **142**, 2537–2544 (2012) doi:10.1016/j.jspi.2012.03.011.

Eichner, G., Stute, W.: Rank transformations in kernel density estimation. J. Nonpar. Stat. **25**, 427–445 (2013) doi:10.1080/10485252.2012.760737

Eichner, G., Stute, W.: Rank-Based Kernel Smoothing – L_2-approach. Talk presented at the 12^{th} Workshop on Stochastic Models, Statistics and Their Applications in Wroclaw, Poland, February 2015.

Feluch, W., Koronacki, J.: A note on modified cross-validation in density estimation. Comput. Statist. Data Anal. **13**, 143–151 (1992)

Guidoum, A.C.: kedd: Kernel estimator and bandwidth selection for density and its derivatives.R package version 1.0.3 (2015). Available from http://CRAN.R-project.org/package=kedd. Last access: November 2016.

Härdle, W.: Applied nonparametric regression. Cambridge Univ. Press, Cambridge (1990)

Hayfield, T., Racine, J.S.: Nonparametric Econometrics: The np Package. J. Stat. Software **27**(5) (2008). http://www.jstatsoft.org/v27/i05

Herrmann, E., packaged for R and enhanced by Maechler, M.: lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth, R package version 1.1-8 (2016). Available from http://CRAN.R-project.org/package=lokern. Last access: February 16, 2017.

Nadaraya, E.A.: On estimating regression. Theory Prob. Appl. **10**, 186–190 (1964)

Parzen, E.: On the Estimation of a Probability Density Function and the Mode. Ann. Math. Statist. **33**, 1065–1076 (1962)

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Last access: December 2016.

Rosenblatt, M.: Remarks on some Nonparametric Estimates of a Density Function. Ann. Math. Statist. **27** 832–837 (1956)

Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society series B, **53**, 683–690 (1991)

Silverman, B. W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall, London (1986)

Srihera, R., Stute, W.: Kernel Adjusted Density Estimation. Statistics Probability Letters **81**, 571–579 (2011) doi:10.1016/j.spl.2011.01.013

Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Springer New York (2002)

Wand, M.: KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2.23-15 (2015). Available from https://CRAN.R-project.org/package=KernSmooth. Last access: November 2016.

Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman & Hall, London (1995)

Watson, G.S.: Smooth regression analysis. Sankhyā, Series A **26**, 359–372 (1964)

Limiting Experiments and Asymptotic Bounds on the Performance of Sequence of Estimators

16

Debasis Bhattacharya and George G. Roussas

16.1 Introduction

The notion of "statistical experiment" was first introduced by Blackwell (1951) for the mathematical description of the observed data in a probabilistic framework. Later, it was studied extensively by Le Cam (1964, 1986) in framing the general principles of asymptotic theory of statistics. A statistical experiment (or a statistical model) is defined by $E = (X, A, P_\theta, \theta \in \Theta)$, where X is the sample space of the statistical data, A is a σ-field of subsets of X, P_θ is a probability measures on A, which depends on the value of the unknown parameter θ, and Θ is the parameter space, or the set of all possible theories. We can think of each θ as a theory that associates a stochastic model P_θ with the observation process to be carried out by the experimenter. For non-asymptotic (or exact) models, the best statistical decisions are sought based on a finite amount of statistical data. Unfortunately, an explicit solution of a finite sample problem is available only for simple models. On the other hand, for a large amount of statistical data, probabilistic and statistical laws like the law of large numbers, and the central limit theorem begin to work, and that allow us to think about the possibility of approximating the model under consideration by some simpler "limit models". The original decision problem then is reduced to: "How do we construct asymptotically optimal decisions, if we know the structure of the optimal decisions for the limit model?". In fact, statistical decision theory discusses the problems of construction of optimal decisions for a given statistical experiment, whereas the questions of comparison, approximation, convergence, etc. for different statistical

D. Bhattacharya (✉)
Visva-Bharati University, Santiniketan, India
e-mail: debasis_us@yahoo.com

G.G. Roussas
University of California, Davis, California, USA

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_16

experiments are discussed under the asymptotic theory of statistical experiments. Let $E_n = (X_n, A_n, P_{\theta,n}, \theta \in \Theta), n \geq 1$, integer, be statistical experiments, where n is the size or dimension of the statistical data. For the classical (i.e., independent) framework: $X_n = R^n$, $A_n = B(R^n)$, the Borel σ-field over R^n, $P_{\theta,n} = P_\theta \times ... \times P_\theta$. One possible way to define the idea of convergence of the experiments $E_n, n \geq 1$, to some limit experiment $E = (X, A, P_\theta, \theta \in \Theta)$, is by considering random functions, such as the likelihood ratios $\Lambda_n(\theta, \theta_n)$ and $\Lambda(\theta, \theta_n)$, instead of the families of measures $P_{\theta,n}$ and P_θ. The parameter point θ_n is defined at the beginning of the preliminaries section. Basically, the asymptotic problems in statistics revolve largely around the idea of approximating a family of probability measures $(P_\theta, \theta \in \Theta)$ by other families $(Q_\theta, \theta \in \Theta)$ that may be better known or more tractable. Le Cam in his 1960 seminal paper introduced the concept of "Limit Experiments", which states, in effect, that, if one is interested in studying the asymptotic properties of a converging sequence of experiments, it is enough to prove the result for the limit experiment. Then the corresponding limiting result for the sequence of experiments will follow. In other words, it provides an absolute standard for what can be achieved asymptotically by a sequence of experiments. No sequence of experiments or statistical procedures can be asymptotically better than the best procedure in the limit experiment. In case of a sequence of tests, the best limiting power function is the best power function in the limit experiment. A sequence of estimators converges to a best estimator in the limit experiment, etc. The asymptotic behavior of the log-likelihood ratio function, corresponding to an arbitrary but fixed parameter point "θ" and a neighboring point was first studied by Le Cam (1960, 1972) under the independent and identical paradigm.

In this connection, mention may be made of the works of Hájek (1962) and Hájek and Šidák (1967), where contiguity results have been employed in the context of rank tests. Those two works elaborated on the so-called Le Cam's three lemmas and thus contributed to familiarizing researchers with these concepts (see Hájek and Šidák (1967), pages 202–210). In Hájek (1970), the author derived a deep result having substantial impact in asymptotic statistical theory, developed by means of contiguity considerations. Coincidentally, the same result was published in the same year by Inagaki (1970). Another reference which should not be omitted here is the book by Strasser (1985). This book is endowed with the novel feature of bridging well-known results in classical statistics with more general settings and current (for up to the 1980s) research methodology. In classical statistics, one is faced with one experiment—typically, a parametric model—and the objective is to provide optimal solutions to an assortment of problems. As long as the sample size n is fixed, this objective obtains only for a very few particular cases. This necessitates asymptotic methods, by letting n tend to infinity. A rough description of the procedure involved is this: The specific experiment that one is presented with is embedded in a convergence sequence of experiments, and then one concentrates on the study of the limit experiment instead of the original one. This book confines itself to the case that the limit experiments are Gaussian shifts. In implementing the above described procedure, the author exploits skillfully almost the entirety of Le Cam's powerful results, spread over thirty years of research effort. In a sense, this

book may be looked upon as a precursor to Le Cam's own book (Le Cam (1986)). Of course, in the process, the results of many other researchers are utilized, such as Wald (1943, 1949, 1950), Blackwell (1947, 1951, 1953), etc. In the context of comparison of experiments, Torgersen (1991) has provided a thorough study of the subject of comparison of experiments, and discussed most of the relevant references up to the year 1988. Le Cam's work has a prominent presence in the study, which is set mostly in the framework of normed vector lattices and their special cases, the L-spaces and the M-spaces. The concepts of deficiency, sufficiency, and completeness—as defined by Le Cam—play a significant role in the study. There does not seem to be a separate concrete treatment of locally asymptotically mixed normal (*LAMN*) and locally asymptotically quadratic (*LAQ*) experiments.

Later on, shifts in the paradigm took place. Various studies related to the asymptotic behavior of the log-likelihood ratio function have been made by various authors under different models and set-ups. As we can expect, and which is also very logical, the direction of the shift was towards dependent observations. First, for Markov dependent processes, and later for general stochastic processes, Roussas in a series of papers investigated the asymptotic behavior of the log-likelihood ratio function (for all related references see Roussas (1972)). After that, Jeganathan (1980), Swensen (1980), Basu and Bhattacharya (1988, 1990, 1992), Roussas and Bhattacharya (2002, 2008, 2009), and others studied the asymptotic behavior of the log-likelihood ratio function from various view-points. Also, Fabian and Hannan (1987) is a relatively long paper containing a wealth of important results. Their detailed description would take us too far afield. We only mention that their contribution consists in generalizing, in the framework of locally asymptotically mixed normal families, previously derived results under local asymptotic normality. It has been observed (Le Cam (1960, 1986), Roussas (1972)) that, under suitable conditions, the limiting distribution of the sequence of log-likelihood ratio processes, approximately and in the neighborhood of any parameter point, belongs to an exponential family. Some implications of this observation have been discussed in Chaps. 3–6 in Roussas (1972). It is to be noted that, if the two parameter points are sufficiently apart from each other, any decent statistical procedure will be able to differentiate between them. An issue arises when the parameter points are close to each other, and yet the respective probability measures are distinctly different. In such cases, the approach has got to be asymptotic in nature, and it will utilize the concept of contiguity and its applications (Le Cam (1986), Roussas (1972)). With the above motivation, several basic optimality results for a locally asymptotically normal (*LAN*) family of distributions, under i.i.d. and general dependence set-up have been obtained. The exponential approximation, Hájek-Inagaki representation theorem, and related results for the Markovian set-up and under a general dependence set-up have been discussed in Roussas and Soms (1972) and Roussas and Bhattacharya (2007), respectively. The results have been obtained for non-random sample size, and also when the sampling is based on a random number of random variables.

Subsequently, it was observed that there exist a number of processes, where the *LAN* conditions are violated. For example, critical and super-critical Galton-Watson branching processes, critical and super-critical autoregressive processes, Ornstein-

Uhlenbeck processes, etc. (Davis (1985), Jeganathan (1982), Le Cam and Yang (2000)). In such cases, it has been observed that the sequence of log-likelihood ratio processes satisfies more general conditions known as locally asymptotically mixture of normal or locally asymptotically mixed normal (*LAMN*), or locally asymptotically quadratic (*LAQ*) conditions. As a result, the limit experiment no longer belongs to an *LAN* family, but to an *LAMN*, or an *LAQ* family. Naturally, investigators started to obtain optimal results for *LAMN* and *LAQ* families of distributions. Results in this direction can be found in Roussas and Bhattacharya (2002, 2007, 2008, 2009, 2010, 2011) and Basu and Bhattacharya (1999), Bhattacharya and Basu (2006), Davis (1985), Basawa and Scott (1983), Jeganathan (1982, 1995), Le Cam and Yang (2000); see also Taniguchi and Kakizawa (2000) and van der Vaart (1998).

In a paper of Bhattacharya and Roussas (2001), the exponential approximation result for the randomly stopped *LAMN* experiment was derived. In that paper, it has been observed that the sequence of randomly stopped *LAMN* experiments can be approximated in the L_1—norm sense by another experiment which is a member of a curved exponential family, in the sense of Efron (1975). This, in turn, implies that the sequences of random vectors and matrices involved in the distribution of an *LAMN* experiment form a sequence of locally asymptotically (differentially) sufficient statistics (see pages 80–81 in Roussas (1972) for a definition). The exponential approximation result derived in Bhattacharya and Roussas (2001) can be used to study the asymptotic behavior of sequential estimates of parameters, asymptotic properties of risk functions, performance of sequential tests, etc.

The familiar Hájek-Inagaki (Hájek (1970), Inagaki (1970)) representation of the asymptotic distribution of a sequence of sequential estimates of parameters for general stochastic processes has been discussed in Roussas and Bhattacharya (2008) and a similar result under the *LAMN* framework has been the topic of discussion of Roussas and Bhattacharya (2009). An example where the limit experiment is neither *LAN* nor *LAMN*—although, in general, is translation invariant—is provided by the paper of Koul and Pflug (1990). The purpose of the present paper is to provide an overview of selected available results in this area on asymptotic theory of statistical inference, which covers the concept of limit experiments and asymptotic bounds on the performance of a sequence of estimators.

Throughout the entire paper, we use the following notation: For a vector $\mathbf{y} \in \mathbf{R}^k$, $\mathbf{y}'$ denotes the transpose of $\mathbf{y}$, and for a square matrix D, $|D|$ denotes the determinant of D, $\|D\|$ denotes the norm of D, defined by the square root of the sum of squares of its elements. If P and Q are two probability measures on a measurable space (X, A), then dP/dQ denotes the Radon-Nikodym derivative of the Q-continuous part of P with respect to Q. If p and q are the probability density functions of $P$and Q, respectively, with respect to some σ-finite measure λ, then $||P - Q|| = \int |p - q| d\lambda$ is the L_1-norm. The symbol "$\Rightarrow$" denotes the convergence in distribution, whereas the symbol $\overset{P_{\theta,n}}{\rightarrow}$ denotes convergence in $P_{\theta,n}$-probability. Unless otherwise stated, the expectation of a random variable is to be understood under θ. Also to avoid repetitions, it is stated that all the limits are taken as n or subsequences of $\{n\}$ tend to infinity.

The paper is organized as follows: Sect. 16.2 introduces the preliminaries, which describe the technical notation and state assumptions required to develop the main results. In the same section, two examples are given, where the underlying assumptions hold. In the following section, Sec. 16.3, the results for non-random sample size are presented. In Sect. 16.4, the results for random sample size are stated and some justifications are provided.

16.2 Preliminaries

Let $X_1, X_2, ..., X_n$ be the first n random variables from a discrete time-parameter stochastic process; the random variables are defined on the probability space (X, A, P_θ) and take values in (S, F), where S is a Borel subset of a Euclidean space, and F is the σ-field of Borel subsets of S. Let $\theta \in \Theta$, where Θ is the underlined parameter space. It is assumed that Θ is an open subset of R^k. It is also assumed that the joint probability law of any finite number of such random variables has some known functional form except for the unknown parameter θ involved in the distribution. Let A_n be the σ-field induced by $X_1, X_2, ..., X_n$, and let $P_{\theta,n}$ be the restriction of P_θ to A_n. It is assumed that, for $j \geq 2$, a regular conditional probability measure of the distribution of X_j, given $(X_1, X_2, ..., X_{j-1})$, is absolutely continuous with respect to a σ-finite measure μ_j with corresponding (probability) density $f_j(\theta) = f_j(x_j|x_1, x_2, ..., x_{j-1}; \theta)$, and the distribution of X_1 is absolutely continuous with respect to a σ-finite measure μ_1 with corresponding density $f_1(\theta) = f_1(x_1, \theta)$. Let $\{\theta_n\}$ be a sequence of local alternatives of θ, where $\theta_n = \theta_n(h) = \theta + \delta_n^{-1}h$, $h \in R^k$, $\{\delta_n\}$ is a sequence of norming factors, such that δ_n is a $k \times k$ positive definite matrix with $||\delta_n^{-1}|| \to 0$. Here δ_n may depend on θ but is independent of the observations; $\theta, \theta_n \in \Theta$. Let $P_{\theta,n}$ and $P_{\theta_n,n}$ be mutually absolutely continuous for all θ and θ_n.Then the sequence of likelihood ratios is given by:

$$L_n(X_1, X_2, ..., X_n; \theta, \theta_n) = L_n(\theta, \theta_n) = \frac{dP_{\theta_n,n}}{dP_{\theta,n}} = \frac{\prod_{j=1}^{n} f_j(\theta_n)}{\prod_{j=1}^{n} f_j(\theta)}, \tag{16.1}$$

and on account of (16.1), the corresponding sequence of log-likelihood ratios is given by:

$$\begin{aligned}\Lambda_n(X_1, X_2, ..., X_n; \theta, \theta_n) &= \Lambda_n(\theta, \theta_n) = \log L_n(\theta, \theta_n)\\ &= \sum_{j=1}^{n} \log \frac{f_j(\theta_n)}{f_j(\theta)} = \sum_{j=1}^{n} 2\log \frac{f_j^{\frac{1}{2}}(\theta_n)}{f_j^{\frac{1}{2}}(\theta)} = \sum_{j=1}^{n} 2\log(1+\eta_{nj}(\theta, h))\\ &= \sum_{j=1}^{n} \{2\eta_{nj}(\theta, h) - \alpha_{nj}^* \eta_{nj}^2(\theta, h)\},\ 0 \le \alpha_{nj}^* \le 1\\ &= \sum_{j=1}^{n} \{(\eta_{nj}(\theta, h)+1)^2 - 1\} - \sum_{j=1}^{n} (\alpha_{nj}^* + 1)\eta_{nj}^2(\theta, h),\\ &= \sum_{j=1}^{n} U_{nj}(\theta, h) - \sum_{j=1}^{n} (\alpha_{nj}^* + 1)\eta_{nj}^2(\theta, h), \qquad (16.2)\end{aligned}$$

where

$$\eta_{nj}(\theta, h) = \frac{f_j^{\frac{1}{2}}(\theta_n)}{f_j^{\frac{1}{2}}(\theta)} - 1, \text{ and } U_{nj}(\theta, h) = U_{nj} = (\eta_{nj}(\theta, h)+1)^2 - 1.$$

Clearly,

$$E(U_{nj}|A_{j-1}) = 0, \text{ for } j \ge 1, \text{ where } A_0 = \phi.$$

Thus, $\{U_{nj}\}$ is a martingale difference sequence.

Under a standard set of assumptions (see Roussas and Bhattacharya (2011)), it can be shown that there exists a sequence of k- dimensional random vectors $\{\Delta_n(\theta)\}$ and a sequence of $k \times k$ symmetric almost sure (a.s.) positive definite random matrices $\{T_n(\theta)\}$ such that the log-likelihood ratio function, as defined in (16.2), can be approximately written as a sum of two terms: A term $h'\Delta_n(\theta)$, which is linear in the local parameter h, and a term $-\frac{1}{2}h'T_n(\theta)h$, which is quadratic in h; i.e., for every $h \in R^k$,

$$\Lambda_n(\theta, \theta_n) - (h'\Delta_n(\theta) - \tfrac{1}{2}h'T_n(\theta)h) \to 0 \text{ in } P_{\theta,n}\text{-probability.} \qquad (16.3a)$$

Further,

$$L(\Delta_n(\theta), T_n(\theta)|P_{\theta,n}) \Rightarrow L(\Delta(\theta), T(\theta)|P_\theta), \qquad (16.3b)$$

where $T(\theta)$ is an a.s. positive definite random matrix and $\Delta(\theta)$ is a random vector whose conditional distribution (under P_θ), given $T(\theta)$, is $N_k(0, T(\theta))$. (See also Fabian and Hannan (1987).)

If the matrices in the quadratic term, i.e.,$T_n(\theta)$, converge to a non-random matrix $T(\theta)$, then the sequence of log-likelihood ratios belongs to the *locally asymptotically normal* (*LAN*) *family*. In this case, $\Delta(\theta) \sim N(0, T(\theta))$ and $T(\theta)$ is a non-random positive definite matrix.

Let the random vectors $\Delta_n = \Delta_n(\theta)$ in Eqs. (16.3a) and (16.3b) be represented in the form:

$$\Delta_n(\theta) = T_n^{\frac{1}{2}}(\theta) W_n(\theta), \tag{16.4}$$

such that

$$L(\Delta_n(\theta), T_n(\theta)|P_{\theta,n}) \Rightarrow L(\Delta(\theta), T(\theta)|P_\theta),\ \Delta(\theta) = T^{\frac{1}{2}}(\theta) W,$$

where $T(\theta)$ is an a.s. positive definite random matrix and $W \sim N_k(0, I)$ independent of $T(\theta)$. Then the sequence of experiments is *locally asymptotically mixture of normal* (*LAMN*). Clearly, under *LAMN* conditions, the distribution of $\Delta(\theta)$, given $T(\theta)$, is $N(0, T(\theta))$, where $T(\theta)$ is as above. Given $T(\theta)$, $E(h'\Delta(\theta))^2 = h'T(\theta)h$. It is to be noted that, under the *LAMN* set-up, the distribution of $T(\theta)$ does not depend on the local parameter h; i.e., $L(T_n(\theta)|P_{\theta_n,n})$ with $\theta_n = \theta + \delta_n^{-1}h$, has a limit distribution independent of h. In general, $\Delta_n(\theta)$ and $T_n(\theta)$, being dependent on θ, are not statistics, and they are not so useful elements for inferential purposes. For the definition of *LAQ* and the difference of the *LAQ* conditions from those of *LAMN* and *LAN*, the reader is referred to Le Cam and Yang (2000, pages 120–121) and Jeganathan (1995). However, it is noteworthy that *LAN* implies *LAMN* and that *LAMN* implies *LAQ*. It is to be observed that the representation of $\Delta_n(\theta)$, as the one stated in (16.4), does not hold in the *LAQ* set-up. Those points θ at which *LAN* or *LAMN* conditions do not hold are called "critical points". Sometimes, in an *LAQ* expansion of $\Lambda_n(\theta, \theta_n)$, it can be seen that $\Delta(\theta)$ and $T(\theta)$ appearing in the expansion are functionals of a Brownian motion or a Gaussian process. Then those models are called *locally asymptotically Brownian functional* (*LABF*) or *locally asymptotically Gaussian functional* (*LAGF*). To be more specific, the sequence of models is *LABF* if

$$\Delta(\theta) = \int_0^1 F_t dB_t,\ T(\theta) = \int_0^1 F_t^2 dt,$$

where B_t is a standard Brownian motion and F_t is a predictable process. The sequence of models is *LAMN* if

$$\Delta(\theta) = T^{\frac{1}{2}}(\theta) B_1 \text{ and } T(\theta) \text{ is random},$$

and it is *LAN* if

$$\Delta(\theta) = T^{\frac{1}{2}}(\theta) B_1 \text{ and } T(\theta) \text{ is a non-random quantity.}$$

An example of *LAQ*, which is not *LAMN*, is given in Le Cam and Yang (2000, page 121).

For the *LAQ* model, using contiguity (see Definition 1) of two sequences $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$, with $\theta_n = \theta + \delta_n^{-1}h$, we will have

$$E[\exp(h'\Delta(\theta) - \frac{1}{2}h'T(\theta)h)] = 1, \quad \text{for all } h, \tag{16.5}$$

where $\Delta(\theta)$ and $T(\theta)$ are as they appear in (16.3b).

In order to obtain various asymptotic results, under the *LAQ* model, the following representation of the expression $\exp(h'\Delta(\theta) - \frac{1}{2}h'T(\theta)h)$, appeared in (16.5), is convenient:

$$\exp(h'\Delta - \frac{1}{2}h'Th) = \frac{\exp\{-\frac{1}{2}(T^{-\frac{1}{2}}\Delta - T^{\frac{1}{2}}h)'(T^{-\frac{1}{2}}\Delta - T^{\frac{1}{2}}h)\}}{\exp\{-\frac{1}{2}(T^{-\frac{1}{2}}\Delta)'(T^{-\frac{1}{2}}\Delta)\}}. \tag{16.6}$$

The representation given in (16.6) is actually a ratio of two multivariate normal densities under the following $(\gamma, \Gamma; \phi)$ model:

$$\frac{dP_{\theta_n,n}}{dP_{\theta,n}} = \frac{\phi(\gamma(x) - \Gamma(x)h)}{\phi(\gamma(x))},$$

where $\phi = \phi(x)$ denotes a standard multivariate normal density.

As defined earlier, let $\theta_n = \theta + \delta_n^{-1}h$ and let $f_j(\theta) = f_j(X_j|\mathbf{X}_{j-1};\theta)$, where for convenience, the same notation $f_j(\theta)$ is used when the observed values are replaced by the random variables.

Define $\xi_{nj}(.;\theta,h)$ by:

$$\xi_{nj}(.;\theta,h) = f_j^{\frac{1}{2}}(.|\mathbf{X}_{j-1};\theta_n) - f_j^{\frac{1}{2}}(.|\mathbf{X}_{j-1};\theta).$$

Then assume that there exists a k-dimensional random vector $\xi_j(\theta)$ such that:

$$\sum_{j=1}^{k} E_\theta[\int (\xi_{nj}(x_j;\theta,h) - \frac{1}{2}h'\delta_n^{-1}\xi_j(\theta))^2 d\mu_j] \to 0.$$

Set $\eta_j(\theta) = \xi_j(\theta)/f_j^{\frac{1}{2}}(\theta)$ and assume that $E_\theta[\eta_j(\theta)|A_{j-1}] = 0$ a.s. $[P_\theta]$, $j \geq 1$. Let

$$\xi_j(\theta) = \frac{\frac{\partial}{\partial\theta}f_j(\theta)}{f_j^{1/2}(\theta)} \text{ so that } \eta_j(\theta) = \frac{\partial}{\partial\theta}f_j(\theta)/f_j(\theta),$$

where

$$\left(\frac{\partial}{\partial\theta}f_j(\theta)\right)' = \left(\frac{\partial}{\partial\theta_1}f_j(\theta), ..., \frac{\partial}{\partial\theta_k}f_j(\theta)\right).$$

The important results stated in the following sections require a specification of different quantities involved. This is done below.
The norming matrices δ_n are such that

$$\delta_n' \delta_n = \sum_{j=1}^{n} E_\theta[\eta_j(\theta)\eta_j'(\theta)].$$

Under the i.i.d. set-up, $\delta_n' \delta_n = nI_k$, so that $\delta_n^{-1} = \frac{1}{\sqrt{n}}$ and the "local neighborhoods" of θ become $\theta_n = \theta + \delta_n^{-1}h = \theta + \frac{1}{\sqrt{n}}h$.

$$T_n(\theta) = \delta_n^{-1}\{\sum_{j=1}^{n} E_\theta[\eta_j(\theta)\eta_j'(\theta)|A_{j-1}]\}\delta_n^{-1},$$

and

$$W_n(\theta) = T_n^{-\frac{1}{2}}(\theta)[\delta_n^{-1}\sum_{j=1}^{n}\eta_j(\theta)].$$

For each $n \geq 1$, the quantities $\sum_{j=1}^{n} E_\theta[\eta_j(\theta)\eta_j'(\theta)]$ and $\sum_{j=1}^{n} E_\theta[\eta_j(\theta)\eta_j'(\theta)|A_{j-1}]$ are generally called the Fisher information (*FI*) matrix and conditional Fisher information (*CFI*) matrix, respectively. Under the *LAN* set-up, the ratio of *CFI* to *FI* converges to 1 as $n \to \infty$, but under *LAMN* and *LAQ* set-ups, the said ratios converge to a random variable.

Under $P_{\theta_n,n}$, the limiting distribution of $(\Delta_n(\theta), T_n(\theta))$ is such that the marginal distribution of the second component is independent of h, and the conditional distribution of $\Delta(\theta)$, given $T(\theta)$, is $N(T(\theta)h, T(\theta))$, since $L(\Delta_n(\theta), T_n(\theta)|\, P_{\theta_n,n}) \Rightarrow L(\Delta(\theta), T(\theta)|\, P_\theta)$, with $\Delta(\theta) = T^{1/2}(\theta)W + T(\theta)h$, where $T(\theta)$ and W are independent and $W \sim N(0, I_k)$.

It follows that $E_\theta[\Delta(\theta)|T(\theta)] = T(\theta)h$ and $E_\theta[T^{-1/2}(\theta)W|T(\theta)] = E_\theta[\frac{\Delta(\theta)}{T(\theta)}|T(\theta)] = h$ (see, for example, van der Vaart (1998), Theorem 9.8, or Le Cam and Yang (2000), page 118).

For illustrative purposes, we consider the following two examples, where the *LAMN* properties hold, and therefore all the relevant results apply. See, however, the paper of Koul and Pflug (1990), where the authors, in the absence of *LAN* and *LAMN* properties, exploit the translation invariance of the limit experiment to construct an adaptive estimator for the autoregressive parameter in the framework of an explosive autoregression model.

Example 1 Explosive autoregressive process of first order.

Here the process consists of random variables X_j, $j \geq 0$, generated as follows:

$$X_j = \theta X_{j-1} + \varepsilon_j,\ X_0 = 0,\ |\theta| > 1, \tag{16.7}$$

where the ε_j's are inependent random variables distributed as $N(0,1)$. These random variables, as defined in (16.7), form a Markov process with transition p.d.f. as that of $N(\theta X_{j-1}, 1)$, so that

$$f_j(\theta) = f(x_j|x_{j-1};\theta) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x_j - \theta\, x_{j-1})^2].$$

In Basu and Bhattacharya (1988), it has been checked that the relevant assumptions hold, so that the underlying family of probability measures is *LAMN* (see also, Basawa and Prakasa Rao (1980), Basawa and Scott (1983), Greenwood and Wefelmeyer (1993, page 110), van der Vaart (1998, pages 135–136)).

The key quantities here are:

$$\delta_n^{-1}(\theta) = \delta_n^{-1} = \tfrac{(\theta^2-1)^2}{\theta^n}, \text{ so that } \theta_n = \theta + \tfrac{(\theta^2-1)h}{\theta^n},$$

$$\xi_j(\theta) = f_j'(\theta)/f_j^{\frac{1}{2}}(\theta),$$

where

$$f_j'(\theta) = \frac{\partial}{\partial\theta} f(x_j|x_{j-1};\theta) = \frac{x_{j-1}(x_j - \theta\, x_{j-1})}{\sqrt{2\pi}} \exp[-\tfrac{1}{2}(x_j - \theta\, x_{j-1})^2],$$

so that

$$\eta_j(\theta) = \frac{\xi_j(\theta)}{f_j^{\frac{1}{2}}(\theta)} = \frac{f_j'(\theta)}{f_j(\theta)} = x_{j-1}(x_j - \theta\, x_{j-1}).$$

It follows that

$$T_n(\theta) = \frac{(\theta^2-1)}{\theta^{2n}} \sum_{j=1}^n X_{j-1}^2,\ W_n(\theta) = (\sum_{j=1}^n X_{j-1}\varepsilon_j)/(\sum_{j=1}^n X_{j-1}^2)^{\frac{1}{2}},$$

$$\Delta_n(\theta) = \frac{(\theta^2-1)}{\theta^n}(\sum_{j=1}^n X_{j-1}\varepsilon_j).$$

Furthermore, it is seen that the log-likelihood ratio is given by:

$$\Lambda_n(\theta,\theta_n) = \frac{(\theta^2-1)h}{\theta^n} \frac{(\sum_{j=1}^n X_{j-1}(X_j - \theta X_{j-1})}{(\sum_{j=1}^n X_{j-1}^2)^{\frac{1}{2}}} - \frac{(\theta^2-1)^2h^2}{2\theta^{2n}}(\sum_{j=1}^n X_{j-1}^2) + o_{P_{\theta,n}}(1).$$

From results found in Basawa and Brockwell (1984, pages 164–165) and Greenwood and Wefelmeyer (1993, page 110), it is observed that

$$L(T_n(\theta)|P_{\theta.n}) \Rightarrow L(T(\theta)|P_\theta) = \chi_1^2,$$
$$L(\Delta_n(\theta), T_n(\theta)|P_{\theta,n}) \Rightarrow L(\Delta(\theta), T(\theta)|P_\theta),$$

with $\Delta(\theta) = T^{\frac{1}{2}}(\theta)W$, where $T^{\frac{1}{2}}(\theta) \sim N(0,1)$, $W \sim N(0,1)$, and $T^{\frac{1}{2}}(\theta)$ and W are independent (all under P_θ).

Example 2 Super-critical Galton-Watson branching process with geometric offspring distribution.

Here the process consists of random variables X_j, $j \geq 0$, generated as follows:

$$f_j(\theta) = f(x_j|x_{j-1};\theta) = (1-\tfrac{1}{\theta})^{x_j - x_{j-1}}(\tfrac{1}{\theta})^{x_{j-1}},\ \theta > 1. \tag{16.8}$$

The geometric offspring distribution is given by:

$$P(X_1 = j) = \theta^{-1}(1-\theta^{-1})^{j-1},\ j = 1,\ 2,\ ...,\ 1 < \theta < \infty,$$

such that

$$E(X_1) = \theta \text{ and } V(X_1) = \sigma^2(\theta) = \theta(\theta-1).$$

Using (16.8), the key quantities are:

$\delta_n^{-1} = \frac{\theta^{\frac{1}{2}}(\theta-1)}{\theta^{n/2}}$, so that $\theta_n = \theta + \frac{\theta^{\frac{1}{2}}(\theta-1)}{\theta^{n/2}}$,

$\xi_j(\theta) = f_j'(\theta)/f_j^{\frac{1}{2}}(\theta)$, where $f_j'(\theta) = \frac{\partial}{\partial\theta} f_j(\theta) = f_j(\theta).\frac{\partial}{\partial\theta}\log f_j(\theta)$,

and $\frac{\partial}{\partial\theta}\log f_j(\theta) = \frac{(X_j - \theta X_{j-1})}{\theta(\theta-1)}$, so that, $\eta_j(\theta) = (X_j - \theta X_{j-1})/\theta(\theta-1)$.

It follows that:

$$T_n(\theta) = \frac{(\theta-1)}{\theta^n}\sum_{j=1}^{n} X_{j-1},$$

$$W_n(\theta) = \sum_{j=1}^{n}(X_j - \theta X_{j-1})/[\theta(\theta-1)\sum_{j=1}^{n} X_{j-1}]^{\frac{1}{2}},$$

and

$$\Delta_n(\theta) = \sum_{j=1}^{n}(X_j - \theta X_{j-1})/\theta^{\frac{(n-1)}{2}}.$$

Furthermore, using (16.8), the log-likelihood function can be written as:

$$\Lambda_n(\theta,\theta_n) = \frac{h}{\theta^{\frac{(n-1)}{2}}}\sum_{j=1}^{n}(X_j - \theta X_{j-1}) - \frac{\theta(\theta-1)^2h^2}{2\theta^n}(\sum_{j=1}^{n} X_{j-1}) + o_{P_{\theta,n}}(1).$$

It can be seen that:
$L(T_n(\theta)|P_{\theta.n}) \Rightarrow L(T(\theta)|P_\theta)$,
$L(\Delta_n(\theta), T_n(\theta)|P_{\theta.n}) \Rightarrow L(\Delta(\theta), T(\theta)|P_\theta)$, with $\Delta(\theta) = T^{\frac{1}{2}}(\theta)W$,
$L(\Delta_n(\theta)|P_{\theta.n}) \Rightarrow L(T^{\frac{1}{2}}(\theta)W|P_\theta)$,
where $T(\theta)$ is an exponential random variable under P_θ with unit mean, $W \sim N(0, 1)$, and $T^{\frac{1}{2}}(\theta)$ and W are independently distributed random variables. See also Basawa and Prakasa Rao (1980, pages 22–25), Basawa and Scott (1983, pages 2–3), for further readings.

The example pertaining to the *LAQ* model will be the unit root autoregressive process as described in Example 1 with $\theta = 1$. For the key quantities, log-likelihood function, and the asymptotic distribution of different statistics of interest, see Roussas and Bhattacharya (2011, pages 273–274, Example 17.3).

16.3 Results for Non-random Sample Size

We are now ready to state some basic results for the limit experiments when the sample size is non-random. These are the following:

Result 1 Let $\Lambda_n(\theta, \theta_n)$, $\Delta_n(\theta)$ and $T(\theta)$ be the quantities appearing in 3(a) and 3(b). Then, under a set of standard assumptions (see Roussas (1972), Roussas and Bhattacharya (2011)), we have:

(i) $\Lambda_n(\theta, \theta_n) - h'\Delta_n(\theta) \to -\frac{1}{2}h'T(\theta)h$ in $P_{\theta,n}$-probability.
(ii) $L(\Delta_n(\theta)|P_{\theta,n}) \Rightarrow L(\Delta(\theta)|P_\theta)$, where $\Delta(\theta) \sim N(0, T(\theta))$.
(iii) $L(\Lambda_n(\theta, \theta_n)|P_{\theta,n}) \Rightarrow L(\Lambda(\theta)|P_\theta)$, where $\Lambda(\theta) \sim N(-\frac{1}{2}h'T(\theta)h, h'T(\theta)h)$.

Result 2 In the notation of Result 1, and under the same set of assumptions:

(i) $\Lambda_n(\theta, \theta_n) - h'\Delta_n(\theta) \to -\frac{1}{2}h'T(\theta)h$ in $P_{\theta_n,n}$-probability.
(ii) $L(\Delta_n(\theta)|P_{\theta_n,n}) \Rightarrow L(\Delta(\theta)|P_\theta)$, where $\Delta(\theta) \sim N(T(\theta)h, T(\theta))$.
(iii) $L(\Lambda_n(\theta, \theta_n)|P_{\theta_n,n}) \Rightarrow L(\Lambda(\theta)|P_\theta)$, where $\Lambda(\theta) \sim N(\frac{1}{2}h'T(\theta)h, h'T(\theta)h)$.

Result 2 can be obtained from Result 1 without much effort at all, because of the contiguity of the sequences $\{P_{\theta,n}\}$and $\{P_{\theta_n,n}\}$, and by using Le Cam's third lemma (see also Corollary 7.2, page 35, in Roussas (1972)).

The concept of contiguity, introduced and developed by Le Cam (1960) and was extensively applied by Roussas (1972), refers to two sequences of probability measures. Contiguity is concerned with the "closeness" or "nearness" and is defined as follows:

Definition 1 For $n = 1, 2, ...$, let P_n and P'_n be two probability measures defined on measurable space $(X, \mathbf{A}_n)$. Then the sequences $\{P_n\}$ and $\{P'_n\}$ are said to be *contiguous* if for any $A_n \in \mathbf{A}_n$, $P_n(A_n) \to 0$ implies $P'_n(A_n) \to 0$, and vice versa.

Here we record the following interesting observations; for a detail and lucid description of different features of contiguity see Roussas (1972).

(i) For any $\mathbf{A}_n$-measurable random variable T_n, $T_n \to 0$ in P_n-probability if and only if $T_n \to 0$ in P'_n-probability.

(ii) L_1-norm implies contiguity but not the converse.

(iii) Contiguity is a weaker measure of "closeness" of two sequences of probability measures than that provided by the L_1 (or *sup*) - norm convergence.

Result 3 With $\theta_n = \theta + \delta_n^{-1}h$, let $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ be two sequences of probability measures, and let $\Lambda_n = \log \frac{dP_{\theta_n,n}}{dP_{\theta,n}}$. Assume that $\{L(\Lambda_n|P_{\theta,n})\}$ converges to $N(\mu, \sigma^2)$. Then the two sequences $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ are contiguous if and only if $\mu = -\frac{\sigma^2}{2}$.

Result 3 follows from Le Cam's third lemma (see also Corollary 7.2, page 35, in Roussas (1972)), which states that, if $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ are contiguous and

$$\begin{aligned} &L(\Lambda_n|P_{\theta,n}) \Rightarrow L(\Lambda) \\ &L(\Lambda_n|P_{\theta_n,n}) \Rightarrow L(\Lambda'), \end{aligned} \tag{16.9}$$

then the distribution of Λ', appearing in (16.9), is determined by:

$$\frac{dF_{\Lambda'}}{dF_{\Lambda}} = e^{\lambda}. \tag{16.10}$$

Let $L(\Lambda_n|P_{\theta,n}) \Rightarrow L(\Lambda) = F$, where F is $N(\mu, \sigma^2)$, and $L(\Lambda_n|P_{\theta_n,n}) \Rightarrow L(\Lambda') = G$.

Then, because of contiguity of $P_{\theta,n}$ and $P_{\theta_n,n}$, $\mu = -\frac{\sigma^2}{2}$, and G is $N(\frac{\sigma^2}{2}, \sigma^2)$.

All these results follow from the facts that $G(d\lambda) = e^{\lambda}F(d\lambda)$ and $\int G(d\lambda) = 1$.

Now, $\int G(d\lambda) = 1$ implies $\int e^{\lambda}F(d\lambda) = 1$,

or $\int \frac{1}{\sigma\sqrt{2\pi}} e^{\lambda}.e^{-\frac{1}{2\sigma^2}(\lambda-\mu)^2} d\lambda = 1$ implies $\sigma^4 + 2\mu\sigma^2 = 0$; i.e., $\mu = -\frac{\sigma^2}{2}$,

and $G(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\lambda}.e^{-\frac{1}{2\sigma^2}(\lambda+\frac{\sigma^2}{2})^2} d\lambda = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\lambda-\frac{\sigma^2}{2})^2} d\lambda.$

Thus, if we have $L(\Delta_n(\theta)|P_{\theta,n}) \Rightarrow N(0, T(\theta))$, and $\Lambda_n(\theta, \theta_n) - h'\Delta_n(\theta) \to -\frac{1}{2}h'T(\theta)h$ in $P_{\theta,n}$- probability for every $h \in R^k$, then we will have $L(\Lambda_n(\theta, \theta_n)|P_{\theta,n}) \Rightarrow N(-\frac{1}{2}h'T(\theta)h, h'T(\theta)h)$. Using contiguity of $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ and the relation given in (16.10) we get:

$$L(\Lambda_n(\theta, \theta_n)|P_{\theta_n,n}) \Rightarrow N(\tfrac{1}{2}h'Th, h'T(\theta)h),$$

and

$$L(\Delta_n(\theta)|P_{\theta_n,n}) \Rightarrow N(T(\theta)h, h'T(\theta)h).$$

Under the *LAMN* model, the term $T_n(\theta)$ appearing in (16.3a) stays random in the limit. The unconditional distribution of the log-likelihood ratio function can be shown to be a mean and variance mixture of a normal distribution, the mixing variable being $T(\theta)$.

Result 4 If the sequence of experiments $\{E_n\}$ satisfies the *LAMN* conditions at $\theta \in \Theta$, then for every $h \in R^k$ we have:

$$L(\Lambda_n(\theta, \theta_n), \Delta_n(\theta), T_n(\theta)|P_{\theta,n}) \Rightarrow L(h'T^{1/2}(\theta)W - \frac{1}{2}h'T(\theta)h, T^{1/2}(\theta)W, T(\theta)),$$

and

$$L(\Lambda_n(\theta, \theta_n), \Delta_n(\theta), T_n(\theta)|P_{\theta_n,n}) \Rightarrow L(h'T^{1/2}(\theta)W + \frac{1}{2}h'T(\theta)h, T^{1/2}(\theta)W + T(\theta)h, T(\theta)),$$

where W is a $k \times 1$ standard normal vector independent of $T(\theta)$.

Result 5 If the sequence of experiments $\{E_n\}$ satisfies the *LAMN* conditions at $\theta \in \Theta$, then for every $h \in R^k$ we have :

$$L(T_n^{-1/2}(\theta)\Delta_n(\theta)|P_{\theta,n}) \Rightarrow L(W),$$

and

$$L(T_n^{-1/2}(\theta)\Delta_n(\theta)|P_{\theta_n,n}) \Rightarrow L(W + T^{1/2}(\theta)h),$$

where $W$is a $k \times 1$ standard normal vector independent of $T(\theta)$.

Result 6 If the sequence of experiments $\{E_n\}$ satisfies the *LAMN* conditions at $\theta \in \Theta$, then for every $h \in R^k$ the joint convergence of $(\Delta_n(\theta), T_n(\theta))$is as follows:

$$L(\Delta_n(\theta), T_n(\theta)|P_{\theta,n}) \Rightarrow L(\Delta(\theta), T(\theta)), \text{ where } \Delta(\theta) = T^{1/2}(\theta)W,$$

and

$$L(\Delta_n(\theta), T_n(\theta)|P_{\theta_n,n}) \Rightarrow L(\tilde{\Delta}(\theta), T(\theta)), \text{ where } \tilde{\Delta}(\theta) = T^{1/2}(\theta)W + T(\theta)h.$$

Here $W$is a $k \times 1$ standard normal vector independent of $T(\theta)$.

Justification of the Results 4-6 can be found in Basu and Bhattacharya (1988, 1990, 1992), Le Cam and Yang (2000), Jeganathan (1995), Davis (1985) and Roussas and Bhattacharya (2010).

All the results stated above are instrumental in deriving several optimality results related to the *LAN*, *LAMN* and *LAQ* families of distributions. The exponential approximation result holds for the *LAN* and *LAMN* models with respect to a certain truncated version of $\Delta_n(\theta)$ (Roussas and Bhattacharya 2002; and Bhattacharya and Roussas, 2001). It is to be noted, however, that the approximating family, under

the *LAMN* framework, no longer belongs to a standard exponential family, but to a curved exponential family, as defined in Efron (1975). Roughly speaking, an exponential family is curved when the dimensionality of the sufficient statistics for θ is larger than the dimensionality of θ. For example, the normal family $N(\theta, \theta^2)$, $\theta \in R$, is a curved exponential family. In Result 4, the limiting distribution of the log-likelihood ratio function is a member of curved exponential family. However, the independence of $T(\theta)$ and W, appearing in the limiting distribution, and the fact that $W \sim N(0, I_k)$ turn the limiting distribution into a normal distribution, when conditioned upon $T(\theta) = t$. This fact is exploited in making statistical inference in the limit experiment, and then transposing it to the original experiment. Optimal properties of different statistical procedures can be studied, which are based on this idea of conditioning on the mixing variable. Efficient tests for *LAMN* experiments have been derived in Roussas and Bhattacharya (2010) and Bhattacharya and Roussas (1999).

Asymptotic efficiency of an estimate can be determined by following various routes. The Weiss-Wolfowitz approach (see, e.g., Wolfowitz (1965), Weiss and Wolfowitz (1967)) is based on the asymptotic concentration of probabilities over certain classes of sets. Wald's approach (Wald 1939, 1947) was directed towards measuring the risk of estimation under an appropriate loss function. The idea of measuring asymptotic efficiency of Wald was extended by Hájek (1972). Ibragimov and Has'minskii (1981) employed the concept of a locally asymptotically normal experiment (see pages 7 and 120–123) in the context of statistical estimation. Fabian and Hannan (1982) considered locally asymptotically normal families, and under some additional very mild assumptions, they constructed locally asymptotically minimax estimates, and provided a condition under which an estimate is locally asymptotically minimax adaptive. Also, they showed that a lower bound, due to Hájek (1972), is not attained, and a new sharp lower bound was obtained. Schick (1988) considered the problem of estimating a finite dimensional parameter in the presence of an arbitrary nuisance parameter in the framework of locally asymptotically mixed normal families. A lower bound for the local asymptotic risk of an estimate was obtained, and sufficient conditions were given for this bound to be achieved for bounded loss functions. Furthermore, a necessary condition was given for the existence of adaptive estimates, and also a necessary condition for an estimate to be adaptive. The concept of studying asymptotic efficiency based on large deviations has been recommended by Basu (1956) and Bahadur (1960, 1967). The search for an asymptotically efficient estimator is done in two steps: First, obtain a bound for the limit (lim sup or lim inf) of a certain desirable quantity, the risk of estimation under an appropriate loss function, say, for a wide class of competing estimators; and second, find an estimator in the target class for which this bound is attained. In the following subsection a search for an efficient estimator is presented.

16.3.1 Asymptotic Bounds on the Performance of Sequence of Estimators

The convolution theorem and the local asymptotic minimax theorem may be used decisively with regards to the asymptotic performance of a sequence of estimators. They replace the Cramér-Rao lower bound for the variance of unbiased estimators in an asymptotic sense. The convolution theorem is applicable to a class of regular estimator sequence, a normalized version of which is assumed to converge weakly to a probability measure (see, e.g., relation (3.1), page 136, in Roussas (1972)), and states that those estimators are asymptotically distributed as the convolution of a normal random variable and an independent noise variable. As a result, the limiting measure is most concentrated when it is restricted to its normal component, and diffused otherwise. The local asymptotic minimax theorem is not restricted to any particular class of estimator sequences, and gives a lower bound for the limit of the minimax risk over a shrinking neighborhood of the true distribution.

In asymptotic theory of estimation, local asymptotic minimax risk is often viewed as an effective measure of asymptotic optimality of an estimator. Minimax risk bounds for the estimator of a parameter, under *LAQ* family of distributions, are discussed below. For details of the derivations, see Basu and Bhattacharya (1999).

Let L be the class of loss functions $\ell : R^k \to [0, \infty)$ satisfying the conditions:

(i) $\ell(x) = \ell(|x|)$.

(ii) $\ell(0) = 0$.

(iii) $\ell(x) \leq \ell(y)$ if $|x| \leq |y|$.

Actually, L contains all bowl-shaped symmetric loss functions with zero minimum loss.

Here, an asymptotic minimax risk bound for an estimator of the parameter θ for some $l \in L$ is determined. Note that L contains the majority of the loss functions usually considered in the literature (Le Cam, 1986; Le Cam and Yang, 2000). Let $\hat{\theta}_n$ be an estimator of the parameter $\theta_n = \theta + \delta_n^{-1}h$. Then the loss in estimating θ_n by $\hat{\theta}_n$ is:

$$\ell(\hat{\theta}_n, \theta_n) = \ell(\hat{\theta}_n - \theta_n) = \ell(\hat{\theta}_n - (\theta + \delta_n^{-1}h)).$$

Let $E_h = E_{\theta+\delta_n^{-1}h}$, $R_n(\theta) = \delta_n(\hat{\theta}_n - \theta_n)$, and let $Q(\cdot)$ be the multivariate normal density with mean vector zero and dispersion matrix I; also, let λ be the Lebesgue measure on the Borel subsets of K_α, where K_α is a cube in R^k whose vertices have coordinates $\pm\alpha$, so that the k-dimensional volume of K_α is $\int_{K_\alpha} \lambda_\alpha(dh) = (2\alpha)^k$.

Then the following result, Result 7, holds under the assumptions stated below:

1. $L(\Delta_n(\theta), T_n(\theta), R_n(\theta)|P_{\theta,n}) \to L(\Delta(\theta), T(\theta), R(\theta))$.
2. $R_n(\theta)$ is tight in the sense that $\{P_{\theta,n}\}$ is tight for fixed θ and for all n.
3. $R_n(\theta) - T_n^{-1}(\theta)\Delta_n(\theta)$ is $o_P(1)$ in $P_{\theta,n}$-probability.
4. For some positive ε_α and C_α with $\varepsilon_\alpha \to 0$, $C_\alpha\varepsilon_\alpha \to \infty$:
   (a) $\int P_{\theta+\delta_n^{-1}h}\{|T^{-1}(\theta)\Delta(\theta)| > \alpha - C_\alpha\}\lambda_\alpha(dh) \to 0$ as $\alpha \to \infty$,
   (b) $\int P_{\theta+\delta_n^{-1}h}\{|T^{\frac{1}{2}}(\theta)| \leq \varepsilon_\alpha\}\lambda_\alpha(dh) \to 0$ as $\alpha \to \infty$.

Result 7 Suppose the sequence of experiments $\{E_n\}$ is *LAQ* at θ, $\ell \in L$ is bounded and $\{\theta_n\}$ is any sequence of estimators. Then:

$$\lim_{\alpha\to\infty} \liminf_{n\to\infty} \sup_{h\in R} E_{\theta+\delta_n^{-1}h}\ell(\delta_n T_n^{\frac{1}{2}}(\theta)(\hat{\theta}_n - (\theta + \delta_n^{-1}h))) \geq E\ell(B_1),$$

where B_1 is $N(0, I_k)$.

The above result implies that the minimax risk of any estimator is bounded by the risk of an asymptotically normally distributed estimator. Hence, any asymptotically normally distributed estimator will be minimax in a *LAQ* model also. A result similar to Result 7, pertaining to finding the asymptotic minimax bounds for sequential estimators of parameters in a locally asymptotically quadratic family, can be found in Basu and Bhattacharya (1999).

It has been noted that the symmetric loss structure as considered above may not be applicable to all situations. Indeed, there are situations, where the loss can be different for equal amounts of over- estimation and under-estimation. Estimation errors of the same magnitude, but of opposite signs, may not have equal consequences. For example, in the *AR*(16.1) process, consider the situations where the autoregressive parameter θ, whose actual value is 0.95, say, is estimated first by $\hat{\theta} = 0.85$ and second by $\hat{\theta} = 1.05$. Though the absolute magnitude of error in estimating θ, $|e| = 0.10$, is the same in both cases, the consequences of the estimation are significantly different. In the first case, that is when $\hat{\theta} = 0.85$, the process is a stationary process, but in the second case, that is when $\hat{\theta} = 1.05$, the process is an explosive process. There are many other situations, where researcher's favorite symmetric loss function may not be appropriate. For a detail discussion on this issue, see Bhattacharya et al. (2002) and Roychowdhury and Bhattacharya (2008, 2010). Local asymptotic minimax risk bounds for estimators of θ in a locally asymptotically mixture of normal experiments under asymmetric loss has been discussed in Bhattacharya and Basu (2006). For lack of space, we omit recording the results here, and refer the interested readers to the above cited paper.

16.4 Results Under Random Sample Size

Let $\{\nu_n, n \geq 1\}$ be a sequence of stopping times; that is, a sequence of non-negative integer-valued random variables defined on the process$\{X_n, n \geq 1\}$,tending non-decreasingly to ∞ a.s. $[P_\theta]$, and such that, for each n, $(\nu_n = m) \in A_m, m \geq 1$. It is also assumed that $\{\nu_n/n \to 1\}$ in P_θ-probability. Let $A_{\nu_n} = \sigma(X_1, X_2, ..., X_{\nu_n})$ be the σ-field induced by the random variables $X_1, X_2, ..., X_{\nu_n}$, and let $P_{\theta,\nu_n} = \tilde{P}_{\theta,n}$ be the restriction of P_θ to A_{ν_n}.

If $P_{\theta,n}$ and $P_{\theta^*,n}$ are mutually absolutely continuous (that is, $P_{\theta,n} \approx P_{\theta^*,n}$), for $\theta, \theta^* \in \Theta$, $n \geq 1$, then $\tilde{P}_{\theta,n} \approx \tilde{P}_{\theta^*,n}$, and the Radon-Nikodym derivative of $\tilde{P}_{\theta^*,n}$

with respect to $P_{\theta,n}$ is given by:

$$\frac{d\tilde{P}_{\theta^*,n}}{d\tilde{P}_{\theta,n}} = L_{\nu_n}(\theta, \theta^*) = \prod_{j=1}^{\nu_n} [f_j(\theta^*)/f_j(\theta)]. \tag{16.11}$$

Using (16.11), the randomly stopped likelihood ratios are given by:

$$L_{\nu_n}(\theta, \theta_n) = \prod_{j=1}^{\nu_n} [f_j(\theta_n)/f_j(\theta)], \text{ for } \theta, \theta_n \in \Theta, \tag{16.12}$$

and therefore, from (16.12),

$$\Lambda_{\nu_n}(\theta, \theta_n) = \sum_{j=1}^{\nu_n} \log[f_j(\theta_n)/f_j(\theta)], \tag{16.13}$$

where $\theta_n = \theta + \delta_n^{-1}h$, $h \in R^k$ and $||\delta_n^{-1}|| \to 0$.

All the results recorded below are obtained under the basic assumption that the underlying family of probability measures is *LAMN*. That is, for each $\theta \in \Theta$, the sequence of probability measures $\{P_{\theta,n}; \theta \in \Theta\}$, $n \geq 1$, is *LAMN* at θ.

Result 8 Let $\Lambda_{\nu_n}(\theta, \theta_n)$ be given by (16.13), and let $T_{\nu_n}(\theta)$ and $W_{\nu_n}(\theta)$ be as below:

$$T_{\nu_n}(\theta) = \delta_n^{-1} \sum_{j=1}^{\nu_n} E_\theta[\eta_j(\theta)\eta_j'(\theta)|A_{j-1}]\}\delta_n^{-1},$$

and

$$W_{\nu_n}(\theta) = T_{\nu_n}^{-\frac{1}{2}}(\theta)[\delta_n^{-1} \sum_{j=1}^{\nu_n} \eta_j(\theta)].$$

Then, for each $\theta \in \Theta$, the sequence $\{\tilde{P}_{\theta,n}; \theta \in \Theta, n \geq 1\}$, of families of probability measures is *LAMN* at θ; that is,

$$\Lambda_{\nu_n}(\theta, \theta_n) - [h' T_{\nu_n}^{\frac{1}{2}}(\theta) W_{\nu_n}(\theta) - \frac{1}{2}h' T_{\nu_n}(\theta)h] \to 0 \text{ in } \tilde{P}_{\theta,n}\text{- probability for every } h \in R^k, \tag{16.14}$$

and

$$L(W_{\nu_n}(\theta), T_{\nu_n}(\theta)|\tilde{P}_{\theta,n}) \Rightarrow L(W, T(\theta)|P_\theta), \tag{16.15}$$

where $W \sim N(0, I_k)$ and W is independent of $T(\theta)$.

Recall that $\theta_n = \theta + \delta_n^{-1}h$, $h \in R^k$ and θ, although arbitrary in Θ, is kept fixed. In order to emphasize this fact and the dependence of $\tilde{P}_{\theta_n,n}$, as well as of the log-likelihood on h, we write

$$\tilde{P}_{h,n} = \tilde{P}_{\theta_n,n},\ \Lambda_{\nu_n}(h) = \Lambda_{\nu_n}(\theta, \theta_n). \tag{16.16}$$

Then the following results hold:

Result 9 Let $\Lambda_{\nu_n}(h)$ be as in (16.16). Then the sequence $\{L(\Lambda_{\nu_n}(h)|\tilde{P}_{\theta,n})\}$ is relatively compact (equivalently, tight), and

$$L[\Lambda_{\nu_n}(h)|\tilde{P}_{\theta,n}] \Rightarrow L(\Lambda(h)|P_\theta),\ \text{with}\ \Lambda(h) = h'T^{\frac{1}{2}}(\theta)W - \frac{1}{2}h'T(\theta)h,$$

where W and $T(\theta)$ are as in Result 8.

Result 10 For every $h \in R^k$, the sequences of probability measures $\{\tilde{P}_{\theta,n}\}$ and $\{\tilde{P}_{h,n}\}$ are contiguous.

Proofs of the convergence results (16.14)-(16.15) and Results 9-10 can be found in Basu and Bhattacharya (1988, 1990, 1992).

Result 11 With $\Lambda_{\nu_n}(h)$ and $\tilde{P}_{h,n}$ as given in (16.16), and for every $h \in R^k$, the following result holds:

$$\Lambda_{\nu_n}(h) - [h'T_{\nu_n}^{\frac{1}{2}}(\theta)W_{\nu_n}(\theta) - \frac{1}{2}h'T_{\nu_n}(\theta)h] \to 0 \text{ in } \tilde{P}_{h,n}\text{-probability.} \tag{16.17}$$

Result 11 follows from Results 8 and 10.

Now, let $T_{\nu_n}(\theta)$ and $W_{\nu_n}(\theta)$ be as in Result 8, and for a sequence $0 < k_n \to \infty$, let $W_{\nu_n}^{k_n}(\theta)$ be a truncated version of $W_{\nu_n}(\theta)$ defined by:

$$W_{\nu_n}^{k_n}(\theta) = W_{\nu_n}(\theta)I(|\Delta_{\nu_n}(\theta)| \le k_n),$$

where

$$\Delta_{\nu_n}(\theta) = T_{\nu_n}^{\frac{1}{2}}(\theta)W_{\nu_n}(\theta).$$

By means of $W_{\nu_n}^{k_n}(\theta)$, and for every $h \in R^k$, define a curved exponential probability measure $Q_{\nu_n}(h) = \tilde{Q}_{h,n}$ by:

$$\tilde{Q}_{h,n}(A) = \tilde{C}_{h,n}\int_A \exp[h'T_{\nu_n}^{\frac{1}{2}}(\theta)W_{\nu_n}^{k_n}(\theta) - \frac{1}{2}h'T_{\nu_n}h]\,d\tilde{P}_{\theta,n},\ \text{for}\ A \in A_{\nu_n}, \tag{16.18}$$

where $\tilde{C}_{h,n}$ is the norming constant, so that

$$\frac{d\tilde{Q}_{h,n}}{d\tilde{P}_{\theta,n}} = \tilde{C}_{h,n} \exp[h' T_{\nu_n}^{\frac{1}{2}}(\theta) W_{\nu_n}^{k_n}(\theta) - \frac{1}{2} h' T_{\nu_n}(\theta) h]. \tag{16.19}$$

Then the following result holds:

Result 12 With $\tilde{Q}_{h,n}(A)$ as defined in (16.18) and (16.19), and on the basis of the convergence result stated in (16.17), the following assertions hold:

(i) $\sup[||\tilde{P}_{h,n} - \tilde{Q}_{h,n}||; h \in B$, a bounded subset of $R^k] \to 0$.

(ii) $\sup[||\tilde{C}_{h,n} - 1||; h \leq b] \to 0$, for every $b > 0$.

(iii) $||\tilde{P}_{h_n,n} - \tilde{Q}_{h_n,n}|| \to 0$ for every bounded sequence $\{h_n\}$ in R^k satisfying $\theta + \delta_n^{-1} h_n \in \Theta$.

The proof of the assertions can be found in Bhattacharya and Roussas (2001).

This section is concluded with the following remarks.

Remark 1 As it has been remarked before, from relations (16.18) and (16.19), it follows that $\tilde{Q}_{h,n}$ belongs to a curved exponential family. However, the conditional probability measure of $\tilde{Q}_{h,n}$, given $T(\theta)$, belongs to an exponential family. This fact suggests that the conditional approach, applied to a curved exponential family, might be used in the same way that an exponential family is used for certain purposes of statistical inference (see, for example, pages 113–127, in Roussas (1972)). This idea of conditional inference for *LAMN* model has also been indicated in Sweeting (1992) and Basawa and Brockwell (1984).

16.4.1 A Convolution Representation Theorem

In this section, we will discuss the issue related to the representation of the asymptotic distribution of a regular sequence of sequential estimates of θ (see relation (16.20) below), when properly normalized, in the *LAMN* framework. This representation has far reaching consequences on statistical inferential procedures about θ, and in particular, the behavior of estimates from asymptotic efficiency view-point. Under this set-up, the representation theorem states that the limiting distribution of any regular sequence of sequential estimates of θ must be a normal distribution $N(0, T^{-1}(\theta))$, for a given $T(\theta)$, convoluted with some other probability measure, and therefore it has less concentration than the normal distribution involved in the limit distribution. Hence the Hájek-Inagaki representation theorem in the *LAMN* framework, holds conditionally on $T(\theta)$. Specifically, we have the following theorem.

Result 13 Let θ be an arbitrary but fixed point in Θ, and suppose that the sequence of families $\{\tilde{P}_{\theta,n}; \theta \in \Theta\}, n \geq 1$, satisfies the *LAMN* conditions at $\theta \in \Theta$, with $k \times k$ random matrices $T_{\nu_n}(\theta)$ and the $k \times 1$ random vectors $W_{\nu_n}(\theta)$ (as defined

in Result 8). Let $\{V_{\nu_n}\}, n \geq 1$, be a sequence of k-dimensional random vectors of regular sequential estimates of θ; that is, for every $h \in R^k$,

$$L[\delta_n(V_{\nu_n} - \theta_n), T_{\nu_n}(\theta)|\tilde{P}_{\theta_n,n}] \Rightarrow L(V(\theta), T(\theta)|P_\theta) = L(\theta), \text{ say}. \tag{16.20}$$

Let $L_{T(\theta)}$ be a regular conditional probability measure of $V(\theta)$, given $T(\theta)$. Then:

$$L_{V(\theta)|T(\theta)} = L_{T(\theta)} = L_1 * L_2, \text{ a.s. } P_\theta, \tag{16.21}$$

where $L_1 = N(0, T^{-1}(\theta))$ and L_2 is the conditional distribution (under P_θ) of $(V(\theta) - T^{-\frac{1}{2}}(\theta)W)$, given $T(\theta)$, and $W \sim N(0, I_k)$ is independent of $T(\theta)$.

Outline of Proof Here, what we are aiming at showing is that the characteristic function (ch.f.) of the conditional limiting distribution of $V_{\nu_n}(\theta)$, given $T(\theta)$, is the product of two suitable ch.f's. Then the composition (or convolution) theorem on page 193 in Loève (1963), or Theorem 6, page 212 in Roussas (2014) would apply and give the result. The result is proved for $k = 1$, since the derivations are easier to describe in R. The modifications required for $k > 1$ are basically notational, and can be implemented without much difficulty.

Let u, v and h be real numbers, and observe that the joint ch.f. of $\delta_n(V_{\nu_n} - \theta_n)$ and $T_{\nu_n}(\theta)$ is given by:

$$\begin{aligned} &E[\exp(iu\delta_n(V_{\nu_n} - \theta_n) + ivT_{\nu_n}|\tilde{P}_{\theta_n,n})] \\ &= E[\exp(iu\delta_n(V_{\nu_n} - \theta) - iuh + ivT_{\nu_n}|\tilde{P}_{\theta_n,n}], \end{aligned} \tag{16.22}$$

since

$$iu\delta_n(V_{\nu_n} - \theta_n) = iu\delta_n(V_{\nu_n} - \theta - \delta_n^{-1}h) = iu\delta_n(V_{\nu_n} - \theta) - iuh.$$

However,

$$\begin{aligned} &E[\exp(iu\delta_n(V_{\nu_n} - \theta) + ivT_{\nu_n}|\tilde{P}_{\theta_n,n}] \\ &= E[\exp(iu\delta_n(V_{\nu_n} - \theta) + ivT_{\nu_n} + \Lambda_{\nu_n}|\tilde{P}_{\theta,n}], \text{ since } \frac{d\tilde{P}_{\theta_n,n}}{d\tilde{P}_{\theta,n}} = \exp(\Lambda_{\nu_n}). \end{aligned} \tag{16.23}$$

Now, using (16.22) and (16.23), we have,

$$\begin{aligned} &E[\exp(iu\delta_n(V_{\nu_n} - \theta_n) + ivT_{\nu_n}|\tilde{P}_{\theta_n,n}] \\ &= E[\exp(iu\delta_n(V_{\nu_n} - \theta) + ivT_{\nu_n} + \Lambda_{\nu_n}|\tilde{P}_{\theta,n}]\exp(-iuh). \end{aligned} \tag{16.24}$$

At this point, let

$$\psi_n(u, v, h) = E[\exp(iu\delta_n(V_{\nu_n} - \theta) + ivT_{\nu_n} + \Lambda_{\nu_n}|\tilde{P}_{\theta_n,n}]. \tag{16.25}$$

Since the exponential approximation result holds for the LAMN experiments under sequential sampling (Bhattacharya and Roussas (2001)), there exists a sequence $\{\Lambda^*_{\nu_n}\}$, where $\Lambda^*_{\nu_n}$ is a suitably truncated version of Λ_{ν_n}. It is to be noted that $\Delta_n(\theta)$ has appeared in the definition of *LAMN* experiments, and $\Delta_{\nu_n}(\theta)$ is obtained from $\Delta_n(\theta)$ just by replacing n by ν_n. Actually,

$$\Delta_{\nu_n}(\theta) = T^{\frac{1}{2}}_{\nu_n}(\theta) W_{\nu_n}(\theta).$$

Now, let us define

$$\phi_n(u, v, h) = E[\exp(iu\delta_n(V_{\nu_n} - \theta) + ivT_{\nu_n} + h\Delta^*_{\nu_n} - \frac{h^2}{2} T_{\nu_n} | \tilde{P}_{\theta_n,n}]. \quad (16.26)$$

Then, using (16.25) and (16.26), it can be shown that:

$$\psi_n(u, v, h) - \phi_n(u, v, h) \to 0 \text{ in } P_\theta\text{-probability.} \quad (16.27)$$

Again, from (16.24) and (16.27), we have:

$$E[\exp(iu\delta_n(V_{\nu_n} - \theta_n) + ivT_{\nu_n} | \tilde{P}_{\theta_n,n})]$$

$$= \exp(-iuh)\psi_n(u, v, h)$$

$$= \exp(-iuh)\phi_n(u, v, h) + o(1) \to \exp(-iuh)\phi(u, v, h), \quad (16.28)$$

where

$$\phi(u, v, h) = E[\exp(iuV + ivT + hT^{\frac{1}{2}} W - \frac{h^2}{2} T | P_\theta]. \quad (16.29)$$

Now, from the conditions of the theorem, and for every $h \in R$, we have:

$$L(\delta_n(V_{\nu_n} - \theta_n), T_{\nu_n} | \tilde{P}_{\theta_n,n}) \Rightarrow L(V(\theta), T(\theta) | P_\theta),$$

$$E[\exp(iu\delta_n(V_{\nu_n} - \theta_n) + ivT_{\nu_n} | \tilde{P}_{\theta_n,n})]$$

$$\to E[\exp(iuV + ivT | P_\theta)] = \phi_n(u, v, 0). \quad (16.30)$$

Thus, from (16.28) and (16.29), we obtain the following equation valid for all real u, v and h:

$$\phi(u, v, 0) = \exp(-iuh)\phi(u, v, h). \quad (16.31)$$

From (16.29), (16.30) and (16.31),

$$E_\theta[\exp(ivT)]E_\theta[\exp(iuV)|T] = E_\theta[\exp(ivT)]E_\theta[\exp(iuV + hT^{\frac{1}{2}} W - \frac{h^2}{2} T - iuh)|T]. \quad (16.32)$$

Now, using the uniqueness of Fourier transforms, we get,

$$E_\theta[\exp(iuV)|T=t] = E_\theta[\exp(iuV + ht^{\frac{1}{2}}W - \frac{h^2}{2}t - iuh)|T=t] \text{ a.s.}[Q_\theta], \tag{16.33}$$

where Q_θ is the distribution of T under P_θ.

The right hand-side function in (16.33), being looked upon as a function of h, may be shown to be analytic (due to Lemma 3.2, page 140, in Roussas (1972)), so that the equality in (16.33) holds when we replace h by $-it^{-1}\ u$ and find the following, for every u:

$$\begin{aligned} E_\theta[\exp(iuV)|T=t] &= E_\theta[\exp(iuV - iut^{-\frac{1}{2}}W + \frac{u^2}{2t} - t^{-1}u^2)|T=t] \\ &= E_\theta[\exp(iu(V - t^{-\frac{1}{2}}W)|T=t]\exp(-t^{-1}\frac{u^2}{2}) \text{ a.s. } [Q_\theta]. \end{aligned} \tag{16.34}$$

Now, relation (16.34) and the Convolution Theorem 6 in Roussas (2014), page 212, yield:

$$L_{V|T} = L_1{}^* L_2, \text{ a.s.}[P_\theta],$$

where $L_1 = N(0, T^{-1}(\theta))$, L_2 is the conditional distribution of the random variable $(V - T^{-\frac{1}{2}}W)$, given T, and W follows the $N(0,1)$ distribution and is independent of T. For the details of the proof, the reader is referred to Roussas and Bhattacharya (2009).

In a concluding remark, it is mentioned that such convolution result in an *LAQ* framework remains unexplored until now, and it can be an interesting problem of future research.

Acknowledgements An early draft of the paper was carefully reviewed by two referees. They made some useful comments and suggestions on the inclusion of additional references, which were adopted and implemented in this version of the paper, and they are gratefully acknowledged herewith.

References

Bahadur, R.R. (1960): On the asymptotic efficiency of tests and estimates, *Sankhyā*, 22, 229-252.

Bahadur, R.R. (1967): Rates of convergence of estimates and test statistics. *Ann. Math. Statist.*, 38, 303-324.

Basawa, I.V. and Brockwell, P.J. (1984): Asymptotic conditional inference for regular nonergodic models with an application to autoregressive processes. *Ann. Statist*, 12, 161-171.

Basawa, I.V. and Prakasa Rao, B.L.S. (1980): *Statistical Inference for Statistical Processes*, Academic Press.

Basawa, I.V. and Scott, D.J. (1983): *Asymptotic optimal inference for non-ergodic models*. Lecture Notes in Statistics, 17, Springer-Verlag.

Basu, A.K. and Bhattacharya, D. (1988): Local asymptotic mixed normality of log-likelihood based on stopping times. *Calcutta Statist. Assoc. Bull*., 37, 143-159.

Basu, A.K. and Bhattacharya, D. (1990): Weak convergence of randomly stopped log-likelihood ratio statistics to mixed Gaussian process. *Calcutta Statist. Assoc. Bull*., 39, 137-149.

Basu, A.K. and Bhattacharya, D. (1992): On the asymptotic non-null distribution of randomly stopped log-likelihood ratio statistic. *Calcutta Statist. Assoc. Bull*., 42, 255-260.

Basu, A. K. and Bhattacharya, D. (1999): Asymptotic minimax bounds for sequential estimators of parameters in a locally asymptotically quadratic family. *Brazilian Journal of Probability and Statistics*, 13, 137-148.

Basu, D. (1956): The concept of asymptotic efficiency. *Sankhyā*, 17, 193-196.

Bhattacharya, D. and Basu, A. K. (2006): Local asymptotic minimax risk bounds in a locally asymptotically mixture of normal experiments under asymmetric loss. *IMS Lecture Notes-Monograph Series*, 49, 312-321.

Bhattacharya, D. and Roussas, G. G. (1999): On asymptotically efficient tests for autoregressive process. *Stoch. Model Appl*., 2(1), 17-30.

Bhattacharya, D. and Roussas, G. G. (2001): Exponential approximation for randomly stopped locally asymptotically mixture of normal experiments. *Stoch. Model Appl*., 4(2), 56-71.

Bhattacharya, D., Samaneigo, F.J. and Vestrup, E. M. (2002): On the comparative performance of Bayesian and classical point estimators under asymmetric loss, *Sankhyā*, Ser. B, 64, 230-266.

Blackwell, D. (1947): Conditional expectation and unbiased sequential estimation. *Ann. Math. Stat*., 18, 105-110.

Blackwell, D. (1951): Comparison of experiments. *Proc. 2nd Berkeley Symp. Math. Stat. Probab.*, 1, 93-102.

Blackwell, D. (1953): Equivalent comparisons of experiments. *Ann. Math. Stat*., 24, 265-272.

Davis, R.B. (1985): Asymptotic inference when the amount of information is random. *Proc. Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, 2, L. Le Cam, and R. A. Olson eds., 841-864, Wadsworth, Monterey, California.

Efron, B. (1975): Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist*., 3(6), 1189-1242.

Fabian, V. and Hannan, J. (1982): On estimation and adaptive estimation for locally asymptotically normal families. *Z.Wahrschelnlichkeitstheoric Und Verw. Gebiete*, 59 (4), 459- 478.

Fabian, V. and Hannan, J. (1987): Local asymptotic behavior of densities. *Statist. Decisions*, 5 (1-2), 105-138.

Greenwood, P. and Wefelmeyer, W. (1993): Asymptotic minimax of a sequential estimator for a first order autoregressive model. *Stochastics and Stochastic Reports*, 38, 49-65.

Hájek, J. (1962): Asymptotically most powerful rank-order tests. *Ann. Math. Statist*., 33, 1124 -1147.

Hájek, J. (1970): A characterization of limiting distributions of regular estimates. *Z.Wahrschelnlichkeitstheoric Und Verw. Gebiete*, 14, 323-330.

Hájek, J. (1972): Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.,* University of California Press, Berkeley, 1, 175-194.

Hájek, J. and Šidák, Z. (1967): *Theory of rank tests*. Academic Press, New York-London.

Ibragimov, I. A. and Has'minskii, R. Z. (1981): *Statistical Estimation. Asymptotic Theory*, Translated from the Russian by Samuel Kotz. Applications of Mathematics, Springer-Verlag, New York-Berlin.

Inagaki, N. (1970): On the limiting distribution of sequence of estimators with uniformity property. *Ann. Inst. Statist. Math*., 22, 1-13.

Jeganathan, P. (1980): An extension of a result of L. Le Cam concerning asymptotical normality. *Sankhyā, Ser. A*, 42, 146-160.

Jeganathan, P. (1982): On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. *Sankhyā, Ser. A*, 44, 173-212.

Jeganathan, P. (1995): Some aspects of asymptotic theory with applications to time series models. *Econometric Theory*, 2, 818-887.

Koul, H.L. and Pflug, G. Ch.(1990): Weakly adaptive estimators in explosive autoregression. *Ann. Statist*., 18(2), 939-960.

Le Cam, L. (1960): Locally asymptotically normal families of distributions. *Univ. of California Publications in Statistics*, 3, 37-98.

Le Cam, L. (1964): Sufficiency and approximate sufficiency. *Ann. Math. Stat*., 35, 1419-1455.

Le Cam, L. (1972): Limit of experiments. *Proc. 6th Berkeley Symp. Math. Stat. Probab.*, 1, 245-261.

Le Cam, L. (1986): *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York.

Le Cam, L. and Yang, G.L. (2000): Asymptotics in Statistics, Some Basic Concepts. *Lecture notes in Statistics*, Springer-Verlag.

Loève, M. (1963): *Probability Theory* (3rd Edition).Van Nostrand, Princeton.

Roussas, G.G. (1972): *Contiguity of Probability Measures*: *Some Applications in Statistics*. Cambridge Univ. Press.

Roussas, G.G. (2014): *An Introduction to Measure Theoretic Probability* (2nd Edition). Academic Press/ Elsevier, Boson, Massachusetts.

Roussas, G.G. and Bhattacharya, D. (2002): Exponential approximation of distributions. *Teoriya Imovirnostey ta Matematichna Statystika*, 66, 108-120. Also, in *Theory of Probability and Mathematical Statistics*, 66, 119–132 (English version, 2003).

Roussas, G.G. and Bhattacharya, D. (2007): Asymptotic expansions, exponential approximation and the Hájek-Inagaki representation theorem under a general dependence set-up. *Proceedings of 20th Panhelenic Statistical Conference (Invited Keynote address)*, *Nikosia*, Cyprus, 45-65.

Roussas, G.G. and Bhattacharya, D. (2008): Hájek-Inagaki representation theorem under a general stochastic process framework, based on stopping times. *Stat. Probab. Lett.,* 78, 2503- 2510.

Roussas, G.G. and Bhattacharya, D. (2009): Hájek-Inagaki convolution representation theorem for randomly stopped locally asymptotically mixed normal experiments. *Stat. Infer.* Stoch. Process, 12, 185-201.

Roussas, G.G. and Bhattacharya, D. (2010): Asymptotically optimal tests under a general dependence set-up. *Jour. of Statistical Research*, 44(1), 57-83.

Roussas, G.G. and Bhattacharya, D. (2011): Revisiting local asymptotic normality (LAN) and passing on to local asymptotic mixed normality (LAMN) and local asymptotic quadratic (LAQ) experiments. Chapter 17 in *Advances in Directional and Linear Statistics*, A festschrift for Sreenivasa Rao Jammalamadaka, Eds. Martin T. Wells and Ashis Sengupta, Springer-Verlag, Berlin, 253-280.

Roussas, G.G. and Soms, A. (1972): On the exponential approximation of a family of probability measures and representation theorem of Hájek-Inagaki. *Ann. Inst. Statist. Math*., 25, 27-39.

Roychowdhury, S. and Bhattacharya, D. (2008): On the Performance of Estimators of Parameter in Autoregressive Model of Order One and Optimal Prediction under Asymmetric Loss. *Model Assisted Statistics and Applications*. 3(3), 225-232.

Roychowdhury, S. and Bhattacharya, D. (2010): On Estimation of Regression Parameter under Asymmetric Loss, *Journal of Applied Probability and Statistics*. 5(2), 161-168.

Schick, A. (1988): On estimation in LAMN families when there are nuisance parameters present. *Sankhyā*, *Ser. A*, 50(2),249-268.

Strasser, H. (1985): *Mathematical Theory of Statistics. Statistical experiments and asymptotic decision theory*. De Gruyter Studies in Mathematics, 7, Walter de Gruyter and Co., Berlin.

Sweeting, T. (1992): *Asymptotic ancillarity and conditional inference for stochastic processes*. *Ann. Statist.*, 20(1), 580-589.

Swensen, A. (1980): *Asymptotic inference for a class of stochastic processes*. Ph.D. thesis, Univ. of California, Berkeley.

Taniguchi, M. and Kakizawa, Y. (2000): *Asymptotic theory of statistical inference for time series*. Springer Series in Statistics, Springer-Verlag.

Torgersen, E. (1991): *Comparison of Statistical Experiments*. Encyclopedia of Mathematics and its Applications, 36, Cambridge University Press, Cambridge.
van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press.
Wald, A. (1939): Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist*. 10, 299-326.
Wald, A. (1943): Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc*., 54, 426-482.
Wald, A. (1947): An essentially complete class of admissible decision functions. *Ann. Math. Statist*. 18, 549-555.
Wald, A. (1949): Note on consistency of the maximum likelihood estimate. *Ann. Math. Statist*., 20, 595-601.
Wald, A. (1950): *Statistical Decision Functions*. John Wiley and Sons, New York.
Weiss, L. and Wolfowitz, J. (1967): Maximum probability estimators. *Ann. Inst. Statist. Math.*, 19, 193-206.
Wolfowitz, J. (1965): Asymptotic efficiency of the maximum likelihood estimator. *Theor. Probab. Appl*., 10, 247-260.

Part IV
Mathematical Finance

Risk Bounds and Partial Dependence Information

17

Ludger Rüschendorf

17.1 Introduction

For the evaluation of risks there are several structural and dependence models in use. The risk vector $X = (X_1, \ldots, X_n)$ is described typically by specified marginal distributions and by some copula model describing the dependence structure. Alternatively there are several structural models like factor models in common use to describe the connection between the risks. Several basic statistical methods and techniques have been developed to construct estimators of the dependence structure like the empirical copula function (see Rüschendorf 1976; Deheuvels 1979; Stute 1984) or the tail empirical copula as estimator for the (tail-)copula function. Similarly various estimators for dependence parameters as for the tail dependence index, for Spearmann's ϱ or for Kendall's τ have been introduced and used to test hypotheses on the dependence structure. (see f.e. Rüschendorf 1974; Genest et al. 1995, 2009). In many applications however there are not enough data available to use these methods in a reliable way. As a consequence there is a considerable amount of model risk when using these methods in an uncritical way. Many instances of these problems have been documented in the recent literature.

In recent years a lot of effort has been undertaken to base risk bounds only on reliable information available from the data, arising from history or from external sources. In particular the case where only information on the marginals is available while the dependence structure is completely unknown has been considered in detail

L. Rüschendorf (✉)
Albert-Ludwigs-Universität Freiburg, Abteilung für Mathematische Stochastik,
Eckerstrasse 1, 79104 Freiburg, Germany
e-mail: ruschen@stochastik.uni-freiburg.de
URL: http://www.stochastik.uni-freiburg.de/rueschendorf

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_17

starting with the paper of Embrechts and Puccetti (2006). In this paper we concentrate on the risk of the aggregated portfolio, where the aggregation is given by the sum $\sum_{i=1}^n X_i$.

In the first section we give a brief review of this development. In the following sections we describe several recent approaches to introduce additional dependence information and structural information in order to tighten the risk bounds. In particular we consider higher order marginals, positive resp. negative dependence restrictions, independence information, variance and higher order moment bounds and partially specified risk factor models. The general insight obtained is that positive dependence information allows to increase lower risk bounds but typically not to decrease the upper risk bounds. Negative dependence information on the other hand allows to decrease upper risk bounds but typically does not increase the lower risk bounds.

17.2 VaR and TVaR Bounds with Marginal Information

Let $X = (X_1, \dots, X_n)$ be a risk vector with marginals $X_i \sim F_i$, $1 \le i \le n$. Then the sharp tail risk bounds without dependence information are given by

$$M(s) = \sup_{X_i \sim F_i} P\left(\sum_{i=1}^n X_i \ge s\right) \quad \text{and} \quad m(s) = \inf_{X_i \sim F_i} P\left(\sum_{i=1}^n X_i \ge s\right). \tag{17.1}$$

Similarly, for the Value at Risk of the sum $S = \sum_{i=1}^n X_i = S_n$ we define the sharp VaR bounds

$$\overline{\mathrm{VaR}}_\alpha = \sup_{X_i \sim F_i} \mathrm{VaR}_\alpha(S) \quad \text{and} \quad \underline{\mathrm{VaR}}_\alpha = \inf_{X_i \sim F_i} \mathrm{VaR}_\alpha(S). \tag{17.2}$$

The dependence uncertainty (DU-)interval is defined as the interval $[\,\underline{\mathrm{VaR}}_\alpha, \overline{\mathrm{VaR}}_\alpha\,]$. Dual representations of (sharp) upper and lower bounds were given in Embrechts and Puccetti (2006) and in Puccetti and Rüschendorf (2012a). In some homogeneous cases i.e. for risk vectors with identical marginal distributions, exact sharp bounds were derived in Wang and Wang (2011) and extended in Puccetti and Rüschendorf (2013) resp. Puccetti et al. (2013) and in Wang (2014). Since the dual bounds are difficult to evaluate in higher dimensions in the inhomogeneous case the development of the rearrangement algorithm (RA) in Puccetti and Rüschendorf (2012a) and in extended form in Embrechts et al. (2013) was an important step to approximate the sharp VaR bounds in a reliable way also in high dimensional examples.

As a result it has been found that the DU-interval typically is very wide. The comonotonic sum $S^c = \sum_{i=1}^n X_i^c$ is typically not the worst dependence structure and often the worst case VaR exceeds the comonotonic VaR denoted as VaR^+ by a factor of 2 or more as shown f.e. in the following two examples (see Table 17.1 and Fig. 17.1). A detailed discussion of these effects is given in Embrechts et al. (2013).

Table 17.1 VaR bounds, $n = 648$, $F_i = \text{Pareto}(2)$, $1 \le i \le n$

α	$\underline{\text{VaR}}_\alpha$ (RA range)	VaR_α^+ (exact)	$\overline{\text{VaR}}_\alpha$ (exact)	$\overline{\text{VaR}}_\alpha$ (RA range)
0.99	530.12 – 530.24	5832.00	12302.00	12269.74 – 12354.00
0.995	562.33 – 562.50	8516.10	17666.06	17620.45 – 17739.60
0.999	608.08 – 608.47	19843.56	40303.48	40201.48 – 40467.92

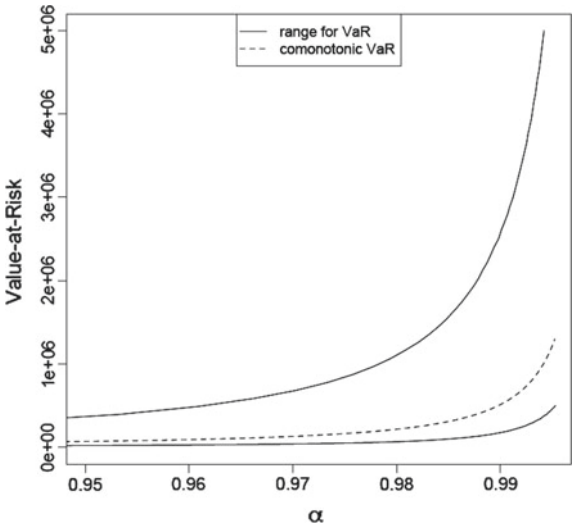


Fig. 17.1 VaR bounds, $d = 8$, risk bounds for operational risk data with marginal Generalized Pareto distributions (GPD) from Moscadelli (2004)

The following theorem gives simple to calculate *unconstrained* bounds for the VaR in terms of the TVaR resp. the LTVaR defined as

$$\text{TVaR}_\alpha(X) = \frac{1}{\alpha}\int_{1-\alpha}^{1} \text{VaR}_u(X)du \quad \text{resp.} \quad \text{LTVaR}_\alpha(X) = \frac{1}{\alpha}\int_0^{\alpha} \text{VaR}_u(X)du. \tag{17.3}$$

Theorem 1 (unconstrained bounds)

$$\begin{aligned} A := \sum_{i=1}^{n} \text{LTVaR}_\alpha(X_i) &= \text{LTVaR}_\alpha(S_n^c) \\ &\le \text{VaR}_\alpha(S_n) \le \text{TVaR}_\alpha(S_n) \\ &\le \text{TVaR}_\alpha(S_n^c) = \sum_{i=1}^{n} \text{TVaR}_\alpha(X_i) =: B \end{aligned}$$

For these bounds (see Wang and Wang 2011; Puccetti and Rüschendorf 2012a; Bernard et al. 2015a).

Puccetti and Rüschendorf (2014) found the astonishing result, that the sharp VaR bounds are asymptotically equivalent to the unconstrained TVaR bounds in Theorem 1 (in the homogeneous case under some regularity conditions)

$$\overline{\mathrm{VaR}}_\alpha \sim \mathrm{TVaR}_\alpha(S_n^c) \quad \text{and} \quad \underline{\mathrm{VaR}}_\alpha \sim \mathrm{LTVaR}_\alpha(S_n^c) \ \text{ as } n \to \infty, \tag{17.4}$$

meaning that the quotients converge to 1 as $n \to \infty$. This result was then extended to the inhomogeneous case in Puccetti et al. (2013) and Wang (2014). The worst case dependence structure has negative dependence in the upper part of the distribution. Construction of this mixing (negatively dependent) part is an interesting task in itself. As a result one obtains tools to determine VaR bounds also for the high dimensional and for the general inhomogeneous case based on marginal informations only. The bounds however are typically too wide to be applicable in practise. As consequence it is necessary to include further information on the dependence structure in order to obtain tighter risk bounds.

17.3 Higher Dimensional Marginals

The class of all possible dependence structures can be restricted if some higher dimensional marginals are known. Let $\mathcal{E}$ be a system of subsets J of $\{1, \dots, n\}$ and assume that for $J \in \mathcal{E}$, $F_{X_J} = F_J$ is known. The class

$$\mathcal{F}_\mathcal{E} = \mathcal{F}(F_J;\, J \in \mathcal{E}) \subset \mathcal{F}(F_1, \dots, F_n) \tag{17.5}$$

resp. the corresponding class of distributions $\mathcal{M}_\mathcal{E}$ is called generalized Fréchet class. In some applications e.g. some two-dimensional marginals additionally to the one-dimensional marginals might be known. The relevant tail risk bounds then are given by

$$M_\mathcal{E}(s) = \sup\{P(S \geq s);\, F_X \in \mathcal{F}_\mathcal{E}\} \text{ and } m_\mathcal{E}(s) = \inf\{P(S \geq s);\, F_X \in \mathcal{F}_\mathcal{E}\}. \tag{17.6}$$

Under some conditions a duality result corresponding to the simple marginal case has been established under the assumption $\mathcal{M}_\mathcal{E} \neq \emptyset$ for various classes of functions ϕ as e.g. upper semicontinuous functions (see Rüschendorf 1984, 1991a; Kellerer 1988). The duality theorem then takes the form:

$$M_\mathcal{E}(\phi) = \sup\left\{\int \phi dP;\, P \in M_\mathcal{E}\right\} = \inf\left\{\sum_{J\in\mathcal{E}} \int f_J dP_J;\, \sum_{J\in\mathcal{E}} f_J \circ \pi_J \geq \phi\right\}. \tag{17.7}$$

The dual problem is however not easy to determine. Note that by definition $M_{\mathcal{E}} = M_{\mathcal{E}}(\phi_s)$, where $\phi_s(x) = \mathbb{1}_{[s,\infty]}(\sum_{i=1}^n X_i)$. For specific classes of indicator functions one can use the duality result to connect up with Bonferroni type bounds.

Let $(E_i, \mathcal{A}_i)$, $1 \le i \le n$ be measurable spaces and let for $J \in \mathcal{E}$, $P_J \in M^1(E_J, \mathcal{A}_J)$ be a marginal system, i.e. P_J are probability measures on $(E_J, \mathcal{A}_J)$ with $(E_J, \mathcal{A}_J) = \bigotimes_{j \in J} (E_j, \mathcal{A}_j)$. The following class of improved Fréchet bounds, i.e. bounds for a marginal class with additional dependence restrictions, was given in Rüschendorf (1991a).

Proposition 1 (Bonferroni type bounds) *Let $(E_i, \mathcal{A}_i)$, $1 \le i \le n$, $(P_J, J \in \mathcal{E})$ be a marginal system. For $A_i \in \mathcal{A}_i$ and $A_J = \prod_{j \in J} A_j$ the following estimates hold:*

1. $M_{\mathcal{E}}(A_1 \times \cdots \times A_n) \le \min_{J \in \mathcal{E}} P_J(A_J)$
2. *In the case that $\mathcal{E} = J_2^n = \{(i, j); i, j \le n\}$, and with $q_i = P_i(A_i^c)$, $q_{ij} = P_{ij}(A_i^c \times A_j^c)$ it holds:*

$$M_{\mathcal{E}}(A_1 \times \cdots \times A_n) \le 1 - \sum q_i + \sum_{i<j} q_{ij} \tag{17.8}$$

$$m_{\mathcal{E}}(A_1 \times \cdots \times A_n) \ge 1 - \sum q_i + \sup_{\tau \in T} \sum_{(i,j) \in \tau} q_{ij}, \tag{17.9}$$

where T is the class of all spanning trees of G_n, the complete graph of $\{1, \ldots, n\}$.

Part 1. yields improved Fréchet bounds compared to the usual Fréchet bounds with marginal information only. Part 2. relates Fréchet bounds to Bonferroni bounds of higher order, and implies in particular improved bounds for the distribution function.

For particular cases of decomposable systems also conditional bounds were given in Rüschendorf (1991a) and applied to risk bounds in Embrechts et al. (2013). For non-overlapping systems $\mathcal{E} = \{J_1, \ldots, J_m\}$ with $J_k \cap J_i = \emptyset$ for $i \ne k$ define $Y_r := \sum_{i \in J_r} X_i$, $H_r := F_{Y_r}$, $r = 1, \ldots, m$ and $\mathcal{H} = \mathcal{F}(H_1, \ldots, H_m)$. Then consider

$$M_{\mathcal{H}}(s) = \sup\{P(Y_1 + \cdots + Y_m \ge s);\ F_Y \in \mathcal{H}\} \quad \text{and}$$
$$m_{\mathcal{H}}(s) = \inf\{P(Y_1 + \cdots + Y_m \ge s);\ F_Y \in \mathcal{H}\}$$

where F_Y are the distribution functions of $(Y_1, \ldots, Y_m)$.

$M_{\mathcal{H}}$ and $m_{\mathcal{H}}$ are tail bounds corresponding to a simple marginal system with marginals H_i.

Proposition 2 (non-overlapping systems) *For a non-overlapping marginal system $\mathcal{E}$, holds:*

$$M_{\mathcal{E}}(s) = M_{\mathcal{H}}(s) \quad \text{and} \quad m_{\mathcal{E}}(s) = m_{\mathcal{H}}(s). \tag{17.10}$$

The following extension to general marginal systems was given in Embrechts and Puccetti (2010); Puccetti and Rüschendorf (2012a). Let $\eta_i := \#\{J_r \in \mathcal{E}; i \in J_r\}$, $1 \le i \le n$. For a risk vector X with $F_X \in \mathcal{F}_{\mathcal{E}}$ define:

$$Y_r := \sum_{i \in J_r} \frac{X_i}{\eta_i}, \quad H_r := F_{Y_r}, \quad r = 1, \dots, m.$$

$\mathcal{H} = \mathcal{F}(H_1, \dots, H_m)$ denotes the corresponding Fréchet class.

Proposition 3 reduced Fréchet bounds *Let $\mathcal{F}_{\mathcal{E}} \neq \emptyset$ be a consistent marginal system such that $\mathcal{M}_{\mathcal{E}} \neq \emptyset$. Then for $s \in \mathbb{R}$ holds*

$$M_{\mathcal{E}}(s) \le M_{\mathcal{H}}(s) \quad \text{and} \quad m_{\mathcal{E}}(s) \ge m_{\mathcal{H}}(s). \tag{17.11}$$

In comparison to the non-overlapping case the bounds in (17.11) are not sharp in general but they can be determined numerically. The RA algorithm can be used to calculate the reduced Fréchet bounds $M_{\mathcal{H}}$ and $m_{\mathcal{H}}$. In order to apply the reduced bounds in Propositions 2 and 3 it is enough to know the partial sum distributions H_r instead of the whole multivariate marginal distributions F_{J_r}.

Also generalized weighting schemes of the form

$$Y_r^{\alpha} = \sum_{i=1}^{m} \alpha_i^r X_i, \text{ with } \alpha_i^r > 0 \text{ iff } i \in J_r \text{ and } \sum_{r=1}^{m} \alpha_i^r = 1$$

have been introduced, leading to a parametrized family of bounds.

The magnitude of reduction of the reduced VaR bound with higher order marginal information given by $\mathcal{M}_{\mathcal{E}}$ which we denote $\overline{\mathrm{VaR}}_{\alpha}^{r}$ compared to the unconstrained upper bound $\overline{\mathrm{VaR}}_{\alpha}$ and the comonotonic VaR^{+} depends on the structure of the marginals. In the following example we assume that there are $n = 600$ Pareto(2) risks and that the two-dimensional marginals are comonotonic in case A) and independent in case B). The results confirm the intuition, that in case A) the improvement is moderate while in case B) it is considerable (see Fig. 17.2 and Table 17.2).

As a result it is found that higher order marginals may lead to a considerable reduction of VaR bounds, when the known higher dimensional marginals do not specify strong positive dependence. For various applications like in insurance applications however this kind of higher oder marginals information F_{J_r} or H_r may not be available.

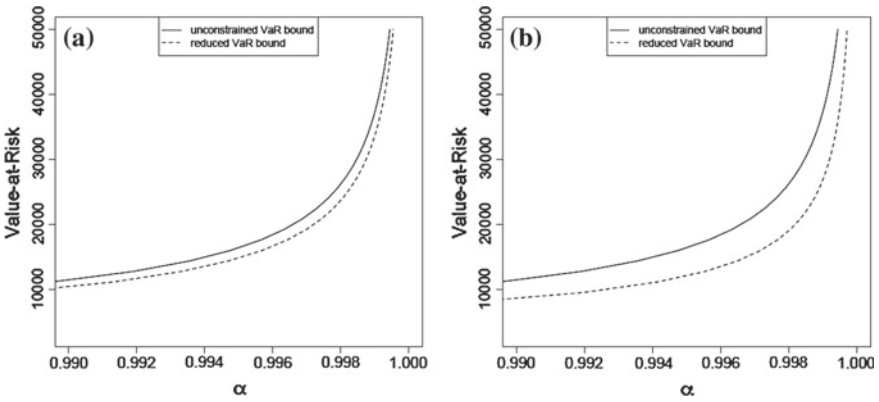


Fig. 17.2 Reduced bounds $n = 600$ Pareto(2) variables, A $\sim$ comonotone $F_{2j-1,2j}$ marginals, B $\sim$ independent $F_{2j-1,2j}$ marginals

Table 17.2 Reduced bounds as in Fig. 17.2

α	VaR_α^+	$\overline{\mathrm{VaR}}_\alpha^{r,\mathrm{A}}$	$\overline{\mathrm{VaR}}_\alpha^{r,\mathrm{B}}$	$\overline{\mathrm{VaR}}_\alpha(L)$
0.99	5400.00	10309.14	8496.13	11390.00
0.995	7885.28	14788.71	12015.04	16356.42
0.999	18373.67	33710.3	26832.2	37315.70

17.4 Risk Bounds with Variance and Higher Order Moment Constraints

In several applications like in typical insurance applications it may be possible to have information available on bounds for the variance or for higher order moments of the portfolio. Consider therefore information of the form:

$$X_i \sim F_i,\ i \le i \le n \quad \text{and} \quad \mathrm{Var}(S_n) \le s^2. \tag{17.12}$$

Alternatively also partial information on some of the covariances $\mathrm{Cov}(X_i, X_j)$ may be available. The corresponding optimization problems

$$\begin{aligned} M &= M(s^2) = \sup\{\mathrm{VaR}_\alpha(S_n);\ S_n \text{ satisfies (12)}\} \text{ and} \\ m &= m(s^2) = \inf\{\mathrm{VaR}_\alpha(S_n);\ S_n \text{ satisfies (12)}\} \end{aligned} \tag{17.13}$$

have been considered in Bernard et al. (2015a). A variant of the Cantelli bounds then is given as follows:

Theorem 2 (VaR bounds with variance information) *Let* $\alpha \in (0,1)$ *and* $\mathrm{Var}(S_n) \le s^2$, *then*

$$a := \max\left(\mu - s\sqrt{\frac{\alpha}{1-\alpha}}, A\right) \le m \le \mathrm{VaR}_\alpha(S_n)$$
$$\le M \le b := \min\left(\mu + s\sqrt{\frac{\alpha}{1-\alpha}}, B\right) \quad \textit{where } \mu = \mathrm{ES}_n\,. \tag{17.14}$$

The bounds in (17.14) are simple to evaluate and depend only on the variance bound s, on the mean μ as well as on the unconstrained bounds A, B.

The VaR bounds and the convex order worst case dependence structure depend on convex order minima in the upper and in the lower part $\{S_n \ge \mathrm{VaR}_\alpha(S_n)\}$ resp. $\{S_n < \mathrm{VaR}_\alpha(S_n)\}$ of the distribution of S_n. This is described in the following proposition (cf. Bernard et al. 2015a). Let for $X_i \sim F_i$, $q_i(\alpha)$ denote the upper α-quantile of X.

Proposition 4 *Let* $X_i \sim F_i$, $F_i^\alpha \sim F_i/[q_i(\alpha), \infty)$ *and let* X_i^α, $Y_i^\alpha \sim F_i^\alpha$, *then:*

(a) $$M = \sup_{X_i \sim F_i} \mathrm{VaR}_\alpha\left(\sum_{i=1}^n X_i\right) = \sup_{Y_i^\alpha \sim F_i^\alpha} \mathrm{VaR}_0\left(\sum_{i=1}^n Y_i^\alpha\right)$$

(b) If $S^\alpha = \sum_{i=1}^n Y_i^\alpha \le_{\mathrm{cx}} \sum_{i=1}^n X_i^\alpha$, *then*

$$\mathrm{VaR}_0\left(\sum_{i=1}^n X_i^\alpha\right) \le \mathrm{VaR}_0(S^\alpha) = \operatorname{ess\,inf}\left(\sum_{i=1}^n Y_i^\alpha\right) \le B$$

Thus maximizing of VaR corresponds to maximizing the minimal support over all $Y_i \sim F_i^\alpha$ and it is implied by convex order. This connection is intuitively explainable. An extreme dependence structure for the maximization is obtained when the random variables are mixable in the upper resp. the lower part of the distribution. Here mixable means that a coupling of the random variables can be found on these parts such that the sum is constant in these parts. In the following Fig. 17.3 this is applied to the quantile function in the comonotonic case and leads to an increase of the upper resp. decrease of the lower value of VaR if the distribution of S_n is mixable on the upper resp. lower part of the distribution.

The connection to the convex order gives the motivation for the extended rearrangement algorithm (ERA) a variant of the RA (see Fig. 17.4). This algorithm consists of two alternating steps:

1. choice of domain, starting from largest α-domain
2. rearrangement in the upper α-part and in the lower $1-\alpha$-part
3. check if the variance constraint is fulfilled
4. shift the domain and iterate

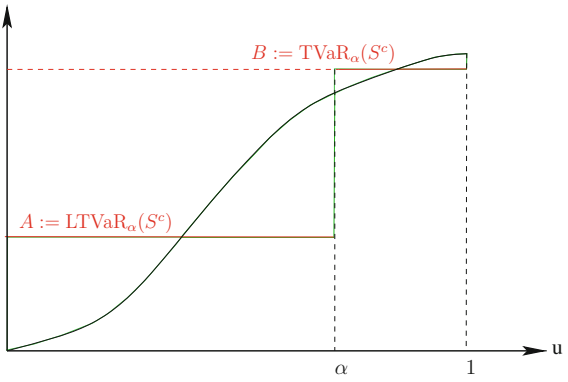


Fig. 17.3 VaR bounds and convex order

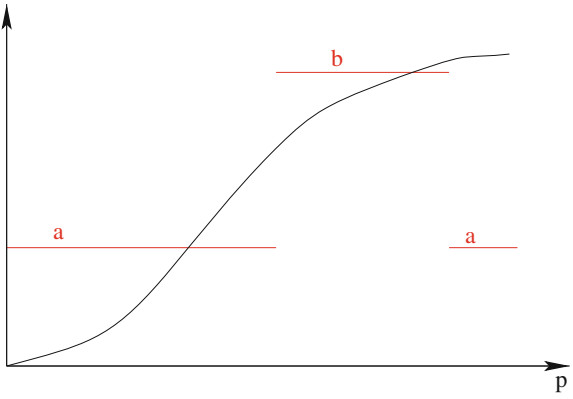


Fig. 17.4 ERA algorithm

Also a variant of the algorithm has been introduced which uses self determined splits of the domain. The following Table 17.3 compares for a portfolio of $n = 100$ Pareto(3) distributed risks the approximate sharp bounds (m, M) calculated by the ERA for various variance restrictions, determined by constant pairwise correlations ϱ with the VaR bounds (a, b) and the unconstrained bounds (A, B).

We find considerable improvements over the unconstrained bounds (A, B) for small variance levels. Since the ERA bounds correspond to valid dependence structures and are close to the theoretical bounds (a, b) this shows that the bounds (a, b) are good and also that the ERA works well.

In an application to a credit risk portfolio of $n = 10000$ binomial loans $X_j \sim \mathcal{B}(1, p)$ with default probability $p = 0.049$ and variance $s^2 = np(1-p) + n(n-1)$ $p(1-p)\varrho^D$ where the default correlation is $\varrho^D = 0.0157$, Bernard et al. (2015a) compared the unconstrained and constrained bounds with some standard industry models like KMV, Beta and Credit Metrics. Table 17.4 shows the improvement of the variance constrained bounds and also the still considerable dependence uncertainty. It raises some doubts on the reliability of the standard models used in practice.

Table 17.3 VaR bounds and ERA with unconstrained bounds for Pareto(3) variables, $n = 100$

(m, M)	$\varrho = 0$	$\varrho = 0.15$	$\varrho = 0.3$		
$\text{VaR}_{0.95}$	(47.96; 84.72)	(42.48; 188.9)	(39.61; 243.3)		
$\text{VaR}_{0.99}$	(48.99; 129.5)	(46.61; 366.0)	(45.36; 489.5)		
$\text{VaR}_{0.995}$	(49.23; 162.8)	(47.54; 499.1)	(46.68; 671.5)		
(a, b)	$\varrho = 0$	$\varrho = 0.15$	$\varrho = 0.3$	(A, B)	
$\text{VaR}_{0.95}$	(47.96; 84.74)	(42.48; 188.9)	(39.61; 243.4)	$\text{VaR}_{0.95}$	(36.46; 303.3)
$\text{VaR}_{0.99}$	(48.99; 129.6)	(46.59; 367.3)	(45.33; 491.7)	$\text{VaR}_{0.99}$	(44.47; 577.6)
$\text{VaR}_{0.995}$	(49.23; 162.9)	(47.54; 500.0)	(46.65; 676.3)	$\text{VaR}_{0.995}$	(46.33; 741.1)

Table 17.4 VaR bounds compared to some standard models (KMV, Beta, Credit Metrics)

	(A, B) (%)	(a, b) (%)	(m, M) (%)	KMV (%)	Beta (%)	Credit metrics (%)
$\text{VaR}_{0.8}$	(0; 24.50)	(3.54; 10.33)	(3.63; 10)	6.84	6.95	6.71
$\text{VaR}_{0.9}$	(0; 49.00)	(4.00; 13.04)	(4.00; 13)	8.51	8.54	8.41
$\text{VaR}_{0.95}$	(0; 98.00)	(4.28; 16.73)	(4.32; 16)	10.10	10.01	10.11

Table 17.5 VaR bounds with higher order moment constraints $\varrho = 0.10$, $n = 100$, models as in Table 17.4

$q =$	KMV	Comon.	Unconstrained	$K = 2$	$K = 3$	$K = 4$
0.95	340.6	393.3	(34.0; 2083.3)	(97.3; 614.8)	(100.9; 562.8)	(100.9; 560.6)
0.99	539.4	2374.1	(56.5; 6973.1)	(111.8; 1245.0)	(115.0; 941.2)	(115.9; 834.7)
0.995	631.5	5088.5	(89.4; 10119.9)	(114.9; 1709.4)	(117.6; 1177.8)	(118.5; 989.5)

It is found that the amount of reduction of the VaR bounds can be considerable when the variance bound s^2 is small enough. Additional higher order moment restrictions of the form $ES_n^k \le c_k$, $2 \le k \le K$ are considered in Bernard et al. (2014, 2017). Table 17.5 shows the potential of higher order moments in a specific case for a corporate portfolio.

The variance resp. moment restriction is a global negative dependence assumption. Therefore one can expect from this assumption a reduction of the upper VaR bounds as shown in the examples. The effect on an improvement of lower bounds is of minor magnitude.

17.5 Dependence/Independence Information

How does positive, negative or independence information influence risk bounds? A weak notion of positive dependence is the positive orthant dependence (POD). X is

called *positive upper orthant dependent* (PUOD) if

$$\overline{F}_X(x) = P(X \geq x) \geq \prod_{i=1}^{n} P(X_i \geq x_i) = \prod_{i=1}^{n} \overline{F}_i(x_i).$$

X is called *positive lower orthant dependent* (PLOD) if

$$F_X(x) \geq \prod_{i=1}^{n} F_i(x_i), \ \forall x.$$

X is POD if X is PLOD and PUOD.

More generally for $F = F_X$, $\overline{F} = \overline{F}_X$ let G be an increasing function with $F^- \leq G \leq F^+$; F^-, F^+ the Fréchet bounds and let H be a decreasing function with $\overline{F}^- \leq H \leq \overline{F}^+$. Further let $\leq_{\text{uo}}$, $\leq_{\text{lo}}$ denote the upper resp. lower orthant ordering. Then

$G \leq F$ is a *positive dependence restriction* on the lower tail probabilities and
$H \leq \overline{F}$ is a *positive dependence restriction* on the upper tail probabilities.

In the case that G is a distribution function and H is a survival function these conditions correspond to ordering conditions w.r.t. $\leq_{\text{lo}}$ resp. $\leq_{\text{uo}}$. In the case that $G(x) = \prod F_i(x_i)$, these conditions together are equivalent to X being POD.
Similarly: $F \leq H$, $\overline{F} \leq H$ are *negative dependence restrictions*.

These kind of restrictions have been discussed in a series of papers, as in Williamson and Downs (1990), Denuit et al. (1999), Denuit et al. (2001), Embrechts et al. (2003) Rüschendorf (2005), Embrechts and Puccetti (2006), Puccetti and Rüschendorf (2012a). As a result the following improved standard bounds are obtained (see Puccetti and Rüschendorf 2012a).

Theorem 3 (positive dependence restriction, improved standard bounds) *Let X be a risk vector with marginals $X_i \sim F_i$. Let G be an increasing function with $F^- \leq G \leq F^+$ and let H be a decreasing function with $\overline{F}^- \leq H \leq \overline{F}^+$. Then*

(a) If $G \leq F_X$, then

$$P\left(\sum_{i=1}^{d} X_i \leq s\right) \geq \bigvee G(s); \tag{17.15}$$

(b) If $H \leq \overline{F}_X$, then

$$P\left(\sum_{i=1}^{d} X_i < s\right) \leq 1 - \bigvee H(s); \tag{17.16}$$

(c) If F is POD, then

$$\bigvee\left(\prod_{i=1}^{d} F_i\right)(s) \le P\left(\sum_{i=1}^{d} X_i \le s\right),$$
$$P\left(\sum_{i=1}^{d} X_i < s\right) \le 1 - \bigvee\left(\prod_{i=1}^{d} \overline{F}_i\right)(s), \tag{17.17}$$

where with $U(s) := \left\{x \in \mathbb{R}^n; \sum_{i=1}^{n} x_i = s\right\}$, $\bigwedge G(s) := \inf_{x \in U(s)} G(x)$ *is the G-infimal convolution,* $\bigvee H(s) := \sup_{x \in U(s)} H(x)$ *is the G-supremal convolution.*

Bignozzi et al. (2015) considered the following specific type of model assumption to explore the consequences of this kind of dependence assumptions. Let the risk vector $X = (X_1, \dots, X_n)$ have marginals $F_i = F_{X_i}$ and assume that $\{1, \dots, n\} = \bigcup_{j=1}^{k} I_j$ is a split into k subgroups. Let $Y = (Y_1, \dots, Y_n)$ be a random vector, that satisfies

$$F_Y(x) = \prod_{j=1}^{k} \min_{i \in I_j} G_j(x_i), \tag{17.18}$$

i.e. Y has k independent homogeneous subgroups and the components within the subgroup I_j are comonotonic. The basic assumption made is

$$Y \le X \tag{17.19}$$

where $\le$ is the upper or lower positive orthant ordering $\le_{\text{uo}}$ resp. $\le_{\text{lo}}$.

In case $F_i = G_j$ for $i \in I_j$ and $k = n$, (17.19) is equivalent to X being PUOD resp. PLOD. As k decreases the assumption is getting stronger and for $k = 1$ it amounts to the strictest assumption that X is comonotonic. In Bignozzi et al. (2015) an analytic expression for the upper and lower bounds $\text{VaR}_\alpha^{\text{ub}}$, $\text{VaR}_\alpha^{\text{lb}}$ under this assumption is given. It turns out that as expected the upper VaR bounds are only slightly improved. The lower bounds are improved strongly if k is relatively small. For $k = n$ there is no improvement of the unconstrained lower VaR bounds $\underline{\text{VaR}}_\alpha$. The POD assumption alone is too weak to lead to improved lower bounds (see Table 17.6).

Similar conclusions are also obtained for inhomogeneous cases.

A stronger notion of positive dependence is the (sequential) positive cumulative dependence (PCD) defined by

$$P\left(\sum_{i=1}^{k-1} X_i > t_1 \mid X_k > t_2\right) \ge P\left(\sum_{i=1}^{k-1} X_i > t_1\right), \quad 2 \le k \le n \tag{17.20}$$

Table 17.6 n homogeneous Pareto(2) risks, split into $\frac{n}{k}$ subgroups of equal size

$n = 8$	$k = 1$		$k = 2$	$k = 4$	$k = 8$
α	$\underline{\text{VaR}}_\alpha$	$\text{VaR}_\alpha^{\text{lb}}$	$\text{VaR}_\alpha^{\text{lb}}$	$\text{VaR}_\alpha^{\text{lb}}$	$\text{VaR}_\alpha^{\text{lb}}$
0.990	9.00	72.00	36.00	18.00	9.00
0.995	13.14	105.14	52.57	26.28	13.14

This is a sequential version of the PCD notion in Denuit et al. (2001). Similarly, (sequential) negative cumulative dependence (NCD) is defined if "$\leq$" holds in (17.20).

From the PCD assumption one obtains the following convex ordering result, where $X \leq_{\text{cx}} Y$ means that $Ef(X) \leq Ef(Y)$ for all convex functions f such that $f(X)$, $f(Y)$ are integrable.

Proposition 5 *Let* $S_n^\perp = \sum_{i=1}^n X_i^\perp$ *denote the independent sum with* $X_i^\perp \sim F_i$.

(a) If X is PCD, then $S_n^\perp \leq_{\text{cx}} S_n$
(b) If X is NCD, then $S_n \leq_{\text{cx}} S_n^\perp$

This result implies as consequence the following VaR resp. TVaR bounds.

Corollary 1 (positive dependence restriction) *If X is PCD, then*

(a) $\text{TVaR}_\alpha(S_n^\perp) \leq \text{TVaR}_\alpha(S_n)$
(b) $\text{LTVaR}_\alpha(S_n^\perp) \leq \text{LTVaR}_\alpha(S_n) \leq \text{VaR}_\alpha(S_n) \leq \text{TVaR}_\alpha(S_n^c)$

The stronger PCD notion implies improvements of the lower bounds for VaR and for TVaR. Under the corresponding negative dependence assumption one obtains improvements of the upper bounds.

Proposition 6 (negative dependence restriction) *If X is NCD, then*

(a) $S_n \leq_{\text{cx}} S_n^\perp$ *and*
(b) $\text{VaR}_\alpha(S_n) \leq \text{TVaR}_\alpha(S_n) \leq \text{TVaR}_\alpha(S_n^\perp)$

Remark 1 A stronger positive dependence ordering between any two random vectors X and Y, the WCS = the weakly conditionally in sequence ordering was introduced in Rüschendorf (2004).

$$X \leq_{\text{wcs}} Y \text{ implies that } \sum_{i=1}^n X_i \leq_{\text{cx}} \sum_{i=1}^n Y_i. \tag{17.21}$$

Table 17.7 $n = 8$, Gamma distributed risks, 4 Gamma (2, 1/2), 4 Gamma (4, 1/2)

$n = 8$	Unconstrained		$k = 1$	$k = 2$	$k = 4$	$k = 8$
α	$\underline{\mathrm{ES}}_\alpha$	$\overline{\mathrm{ES}}_\alpha$	$\mathrm{ES}_\alpha^{\mathrm{lb}}$	$\mathrm{ES}_\alpha^{\mathrm{lb}}$	$\mathrm{ES}_\alpha^{\mathrm{lb}}$	$\mathrm{ES}_\alpha^{\mathrm{lb}}$
0.990	12.00	38.27	38.27	29.15	23.29	19.56
0.995	12.00	41.64	41.64	31.15	24.52	20.33

This ordering notion allows to pose more general kinds of positive (negative) dependence restrictions and to compare not only to the independent case. Several examples for applications of this ordering are given in that paper.

In the subgroup example the WCS condition is strong enough to imply strongly improved lower bounds for $k \leq n$ subgroups also in the case that $k = n$ (see Table 17.7).

The reduction of the DU-spread in this example ranges from about 28% for $k = 8$ to 65% for $k = 2$.

A particular relevant case of reduction of the VaR bounds arises under the independence assumption I) which was discussed in Puccetti et al. (2015).

I) The subgroups $I_1, \ldots, I_k$ are independent.

In this case we can represent the sum S as an independent sum

$$S = \sum_{i=1}^{k} Y_i \quad \text{where} \quad Y_i = \sum_{j \in I_i} X_j. \tag{17.22}$$

We denote by $S^{c,k} = \sum_{i=1}^{k} Y_i^c$ the comonotonic version of the sum and by $\overline{\mathrm{VaR}}_\alpha^{\mathrm{I}}$ the sharp upper bound for VaR_α with this independence information.

Theorem 4 *Under the independence assumption* I) *holds:*

$$\begin{aligned} a^{\mathrm{I}} := \mathrm{LTVaR}_\alpha(S^{c,k}) &\leq \underline{\mathrm{VaR}}_\alpha^{\mathrm{I}} \leq \overline{\mathrm{VaR}}_\alpha^{\mathrm{I}} \\ &\leq b^{\mathrm{I}} := \mathrm{TVaR}_\alpha(S^{c,k}). \end{aligned}$$

Note that the upper and lower bounds $a^{\mathrm{I}}, b^{\mathrm{I}}$ can be calculated numerically by Monte Carlo simulation. As consequence one obtains strongly improved VaR bounds $a^{\mathrm{I}}, b^{\mathrm{I}}$ compared to the sharp VaR bounds as is demonstrated for a Pareto example in Table 17.8.

The bounds in Theorem 4 have also been extended to the case of partial independent substructures which appear to be realistic models in several important

Table 17.8 $n = 50$, Pareto(3) variables

$(a^{\mathrm{I}}, b^{\mathrm{I}})$	$k = 1$	$k = 2$	$k = 5$	$k = 25$	$k = 50$	$(\underline{\mathrm{VaR}}_\alpha; \overline{\mathrm{VaR}}_\alpha)$
$\alpha = 0.990$	(18.23; 153.72)	(20.21; 116.32)	(22.03; 81.54)	(23.76; 48.57)	(24.15; 41.09)	(18.24; 153.3)
$\alpha = 0.995$	(22.24; 297.84)	(23.14; 208.2)	(23.92; 132.28)	(24.59; 65.87)	(24.73; 51.98)	(22.26; 297.64)

Table 17.9 comparison of b^{I}, VaR_α^+, and $\overline{\mathrm{VaR}}_\alpha$ for a insurance portfolio, $n = 11$

α	b^{I}	VaR_α^+	$\overline{\mathrm{VaR}}_\alpha$
0.990	147.34 – 149.66	168.37	209.59
0.995	173.37 – 176.96	202.89	249.55
0.999	250.41 – 262.47	304.63	367.70

applications like in hierarchical insurance models (containing several independencies). It has been applied to a real insurance example in dimension $n = 11$ and with $k = 4$ independent subgroups.

Let $I_1, \ldots, I_4$ be risks which are modeled in the insurance company $I_1 =$ {market-, credit-, insurance-, business-, asset-, non life-, reput.-, and life risk} by Gaussian marginals. Further denote by $I_2 =$ {reinsurance risk}, $I_3 =$ {operational risk} risks which are modeled by log-Normal distributions and finally let $I_4 =$ {catastrophic risk} be a risk modeled by a Pareto distribution. The independence assumption leads to a considerable reduction of approximatively 30% of the upper risk bound (see Table 17.9) which is even a strong improvement over the comonotonic case.

An analysis shows that in this example the independence information is dominating the variance information, i.e. the independence bounds improve on the variance based bounds. The results in this example yield upper risk bounds which are based on reliable information and are acceptable for the application considered.

17.6 Partially Specified Risk Factor Models

In Bernard et al. (2016) risk bounds are discussed under additional structural information. It is assumed that the risk vector is described by a

$$\textbf{factor model}: \quad X_j = f_j(Z, \epsilon_j), \quad 1 \le j \le n \tag{17.23}$$

where Z is a systemic risk factor and ϵ_j are individual risk factors. It is assumed that the joint distributions H_j of (X_j, Z) are known $1 \le j \le n$, but the joint distribution of (ϵ_j) and Z is not specified as is done in the usual factor models. Therefore, this

describes partially specified risk factor models without the usual assumptions of conditional independence of (ϵ_j) given the risk factor Z.

In particular the marginal distributions $F_{j|z}$ of X_j given $Z = z$ are known. The set of admissible models consistent with this partial specification is denoted by $A(H)$ where $H = (H_j)$. The idea underlying this approach is that the common risk factor Z should reduce the DU-interval. This model assumption reduces the upper VaR bounds $\overline{\mathrm{VaR}}_\alpha^f$ over the class of admissible models if Z generates negative dependence and it increases the lower VaR bounds $\underline{\mathrm{VaR}}_\alpha^f$ when Z induces positive dependence.

The partially specified factor model can be described by a mixture representation $X = X_Z$ with $X_z = (X_{j,z}) \in A(F_z)$, $F_z = (F_{j|z})$, where Z and $(X_{j,z})$ are independent. Then

$$F_S = \int F_{S_z}\, dG(z) \quad \text{with } G \sim Z. \tag{17.24}$$

Let $q_z(\alpha) = \mathrm{VaR}_\alpha(S_z)$ denote the VaR of S_z at level α and define for $\gamma \in \mathbb{R}^1$, $\gamma_z = q_z^{-1}(\gamma)$ the inverse γ-quantile of S_z i.e. the amount of probability chosen from $\{Z = z\}$. Further define

$$\gamma^*(\beta) = \inf\left\{\gamma \in \mathbb{R};\ \int \gamma_z\, dG(z) \geq \beta\right\}. \tag{17.25}$$

From the mixture representation in (17.23) the following mixture representation of $\mathrm{VaR}_\alpha(S_Z)$ and of the worst case $\overline{\mathrm{VaR}}_\alpha^f$ w.r.t. the admissible class is derived.

Theorem 5 (worst case VaR in partially specified factor model) *For $\alpha \in (0,1)$ holds:*

(a) $\mathrm{VaR}_\alpha(S_Z) = \gamma^*(\alpha)$
(b)

$$\overline{\mathrm{VaR}}_\alpha^f = \overline{\gamma}^*(\alpha) = \inf\left\{\gamma;\ \int \overline{\gamma}_z dG(z) \geq \alpha\right\}, \tag{17.26}$$

where $\overline{q}_z(\alpha) = \overline{\mathrm{VaR}}_\alpha(S_z)$, $\overline{\gamma}_z = (\overline{q}_z)^{-1}(\gamma)$ is the worst case inverse γ-quantile.

The mixture representation in (17.26) has an obvious intuitive meaning. It is however in general not simple to calculate. For that purpose it is useful to replace the conditional VaR's in formula (17.26) by conditional TVaR's which are easy to calculate, i.e. define

$$t_z(\beta) = \mathrm{TVaR}_\beta(S_z^c) = \sum_{j=1}^n \mathrm{TVaR}_\beta(X_{j,z}). \tag{17.27}$$

Then $q_z(\beta) \leq t_z(\beta)$ and we obtain

$$\overline{\gamma}^*(\beta) \leq \gamma_t^*(\beta) = \inf\left\{\gamma;\ \int t_z^{-1}(\gamma) dG(z) \geq \beta\right\}. \tag{17.28}$$

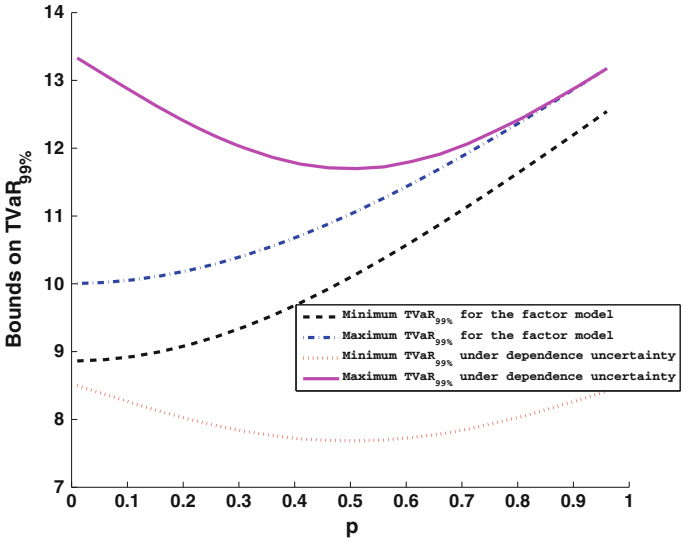


Fig. 17.5 TVaR reduction in partially specified risk factor model, reduction of DU-spread in dependence on p

As a result this estimate from above leads to the following corollary.

Corollary 2 (TVaR bounds for the partially specified risk factor model)

(a)

$$\overline{\mathrm{VaR}}_\alpha^f = \overline{\gamma}^*(\alpha) \le \gamma_t^*(\alpha). \tag{17.29}$$

(b) With $T_z^+ := \mathrm{TVaR}_U(S_z^c)$, $U \sim U(0,1)$, *the following representation holds*

$$\mathrm{VaR}_\alpha(T_Z^+) = \gamma_t^*(\alpha). \tag{17.30}$$

The expression in (17.30) is well suited for Monte Carlo simulations and thus for the numerical calculation of upper bounds for $\overline{\mathrm{VaR}}_\alpha^f$. The following example confirms the idea of the influence of the systemic risk factor Z on the reduction of the DU-spread.

Example 1 Consider the case $n = 2$ where

$$X_1 = (1-Z)^{-1/3} - 1 + \epsilon_1$$
$$X_2 = p\left((1-Z)^{-1/3} - 1\right) + (1-p)\left(Z^{-1/3} - 1\right) + \epsilon_2$$

where $Z \sim U(0,1)$, $\epsilon_i \sim$ Pareto(4) and $p \in [0,1]$ is a dependence parameter. For small p the common risk factor produces strong negative dependence, for large p

it produces strong positive dependence. Therefore, for $p \approx 0$ we expect a strong reduction of the upper risk bounds; for $p \approx 1$ we expect a strong improvement of the lower risk bound. This is confirmed in Fig. 17.5 for the case $\alpha = 0.90$.

Similar reduction results are also obtained at other confidence levels α for VaR and hold true also in higher dimensional examples (see Bernard et al. 2016). For strong negative dependence we see a strong reduction of the upper bounds, for strong positive dependence induced by the common risk factor Z we obtain a strong improvement of the lower bound. But for all possible values of the dependence parameter p the reduction of the DU-spread is of similar order. In our example above it is of order of 60–70% which is due to the dominant influence of the common risk factor Z.

The consideration of partially specified risk factor models is a flexible and effective tool to reduce DU-spreads. The magnitude of the reduction amounts to the influence of the common risk factor Z. Examples of particular interest for applications are the *Bernoulli mixture* models for credit risk portfolios where the conditional distributions $F_{i|z}$ of X_i given $Z = z$ are given by $B(1, p_i(z))$. Common models for financial portfolios are the multivariate normal mean-variance mixture models of the form

$$X_i = \mu_i + \gamma_i Z + \sqrt{Z}\varrho_i \epsilon_i, \quad 1 \le i \le n \tag{17.31}$$

where Z is a stochastic factor and ϵ_i are standard normal distributed. These models include many of the standard and well established marginal distributions in finance like Variance Gamma, hyperbolic or Normal Inverse Gaussian distributions. In our partially specified factor model we dismiss with the usual Gaussian dependence among the ϵ_i.

The results on partially specified risk factor models described above can be extended to more general *mixture models*. Let $D = D_1 + D_2 + D_3$ be a decomposition of the state space D of Z. Assume that for states $z \in D_1$ of the risk factor Z we have available a precise model P_z^1 for the risk vector X given $Z = z$ while for states $z \in D_2$ we have available the conditional distributions $F_z = (F_{j|z})$ i.e. the partially specified distributions. For $z \in D_3$ we only have available marginal information (G_j). As result we obtain a mixture model of the form

$$P^X = \int_{D_1} P_z^1 \, dP^Z(z) + \int_{D_2} P_z^2 \, dP^Z(z) + p_3 P^3 \tag{17.32}$$

with P_z^1 completely specified for $z \in D_1$, $P_z^2 \in A(F_z)$ for $z \in D_2$ and $P^3 \in A(G_j)$. With $p_i = P(Z \in D_i)$ the model in (17.32) has three components

$$P^X = p_1 P^1 + p_2 P^2 + p_3 P^3, \tag{17.33}$$

where the (normalized) first component P^1 is explicitly modeled, the second one P^2 contains partially specified risk factor information, and the third one P^3 contains only marginal information.

Since

$$P\left(\sum_{j=1}^{n} X_j \geq t\right) = \sum_{i=1}^{3} p_i P^i \left(\sum_{j=1}^{n} X_j \geq t\right) \tag{17.34}$$

we obtain the sharp tail risk bound for this extended mixture model

$$\overline{M}(t) = p_1 P^1 \left(\sum_{j=1}^{n} X_j \geq t\right) + p_2 \int_{D_2} \overline{M}_{2,z}(t) dP^Z(z) + p_3 \overline{M}_3(t), \tag{17.35}$$

where $\overline{M}_{2,z}(t)$ is the constrained tail risk bound in D_2 and $\overline{M}_3(t)$ is the marginal tail risk bound in D_3. The convex sharp upper bound in this model is given by

$$S = \sum_{i=1}^{n} X_i \leq_{\text{cx}} I(Z \in D_1) F_1^{-1}(U) + I(Z \in D_2) S_{2,Z}^c + I(Z \in D_3) S_3^c, \tag{17.36}$$

where F_1 is the distribution function of $\sum_{j=1}^{n} X_i$ under P^1, $S_{2,z}^c = \sum_{j=1}^{n} F_{j|z}^{-1}(U)$ and $S_3^c = \sum_{j=1}^{n} G_j^{-1}(U)$ are the conditional resp. unconditional comonotonic vectors, $U \sim U(0, 1)$ independent of Z. The formula in (17.36) implies directly sharp upper bounds for the Tail Value at Risk of S.

Also the TVaR upper bounds in Corollary 2 generalize to this extended mixture model since they are based only on the convex ordering properties as in (17.36).

An interesting case of this general model is the case where $D = \{0, 1\}$ and where for $z = 0$ we have an exact model in the central part of the distribution in $\mathbb{R}^n$ and for $z = 1$ we have only marginal information. The model has been suggested and analyzed in Bernard and Vanduffel (2015). In particular, the reduction of tail risk of the distribution of S for moderate levels α by the exactly modeled central part of the distribution is of practical relevance.

17.7 Conclusion

Sharp risk bounds for portfolios where only marginal information is available can be calculated by the RA-algorithm. They are however typically to wide to be usable in applications. Therefore, various further reductions of the VaR bounds have been proposed in the literature and are discussed in this paper. These are based on additional dependence or structural information.

Higher order marginals may give a good reduction of th DU-bounds when available. Variance constraints and also higher order moment constraints are often available and yield a good reduction when the constraints are small enough.

Partial dependence information together with structural information on subgroups can lead to interesting improvements, when the dependence notion used is strong enough. The weak positive orthant dependence (POD) alone is not sufficient. Of particular interest for applications is to include some (structural) independence information on the underlying model.

A particular flexible method to introduce relevant structural information is based on partially specified risk factor models. These models can be used based on realistic model information and often give a considerable improvement of the DU-spread depending on the magnitude of the influence of the common risk factor. We also briefly describe in this paper an extension of this approach to a more general class of mixture models.

References

C. Bernard and S. Vanduffel. A new approach to assessing model risk in high dimensions. *Journal of Banking and Finance*, 58:166–178, 2015.

C. Bernard, X. Jiang, and S. Vanduffel. Note on 'Improved Fréchet bounds and model-free pricing of multi-asset options' by Tankov (2011). *Journal of Applied Probability*, 49(3):866–875, 2012.

C. Bernard, Y. Liu, N. MacGillivray, and J. Zhang. Bounds on capital requirements for bivariate risk with given marginals and partial information on the dependence. *Dependence Modeling*, 1:37–53, 2013.

C. Bernard, M. Denuit, and S. Vanduffel. Measuring portfolio risk under partial dependence information. *Social Science Research Network*, 2014. doi:10.2139/ssrn.2406377.

C. Bernard, L. Rüschendorf, and S. Vanduffel. Value-at-Risk bounds with variance constraints. *Journal of Risk and Insurance*, 2015a. Preprint (2013), available at http://ssrn.com/abstract=2342068.

C. Bernard, L. Rüschendorf, S. Vanduffel, and J. Yao. How robust is the Value-at-Risk of credit risk portfolios? *Finance & Stochastics*, 21:60–82, 2017. doi:10.1080/1351847X.2015.1104370.

C. Bernard, L. Rüschendorf, S. Vanduffel, and R. Wang. Risk bounds for factor models. *Social Science Research Network*, 2016. doi:10.2139/ssrn.2572508.

V. Bignozzi, G. Puccetti, and L. Rüschendorf. Reducing model risk via positive and negative dependence assumptions. *Insurance: Mathematics and Economics*, 61(1):17–26, 2015.

P. Deheuvels. La fonction de dépendance empirique et ses propriétés. *Académie royale de Belgique, Bulletin de la classe des sciences*, 65(5):274–292, 1979.

M. Denuit, J. Genest, and É. Marceau. Stochastic bounds on sums of dependent risks. *Insurance: Mathematics and Economics*, 25(1):85–104, 1999.

M. Denuit, J. Dhaene, and C. Ribas. Does positive dependence between individual risks increase stop loss premiums. *Insurance: Mathematics and Economics*, 28:305–308, 2001.

P. Embrechts and G. Puccetti. Bounds for functions of dependent risks. *Finance and Stochastics*, 10(3):341–352, 2006.

P. Embrechts and G. Puccetti. Bounds for the sum of dependent risks having overlapping marginals. *Journal of Multivariate Analysis*, 101(1):177–190, 2010.

P. Embrechts, A. Höing, and A. Juri. Using copulae to bound the Value-at-Risk for functions of dependent risks. *Finance and Stochastics*, 7(2):145–167, 2003.

P. Embrechts, G. Puccetti, and L. Rüschendorf. Model uncertainty and VaR aggregation. *Journal of Banking and Finance*, 37(8):2750–2764, 2013.

P. Embrechts, G. Puccetti, L. Rüschendorf, R. Wang, and A. Beleraj. An academic response to Basel 3.5. *Risk*, 2(1):25–48, 2014.

P. Embrechts, B. Wang, and R. Wang. Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics*, 19(4):763 –790, 2015.

C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

C. Genest, B. Rémillard, and D. Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213, 2009.

D. Hunter. An upper bound for the probability of a union. *Journal of Applied Probability*, 13:597–603, 1976.

H. G. Kellerer. Measure theoretic versions of linear programming. *Mathematische Zeitschrift*, 198(3):367–400, 1988.

M. Moscadelli. The modelling of operational risk: experience with the analysis of the data collected by the basel committee. Working Paper 517, Bank of Italy – Banking and Finance Supervision Department, 2004.

G. Puccetti and L. Rüschendorf. Bounds for joint portfolios of dependent risks. *Statistics & Risk Modeling*, 29(2):107–132, 2012a.

G. Puccetti and L. Rüschendorf. Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics*, 236(7):1833–1840, 2012b.

G. Puccetti and L. Rüschendorf. Sharp bounds for sums of dependent risks. *Journal of Applied Probability*, 50(1):42–53, 2013.

G. Puccetti and L. Rüschendorf. Asymptotic equivalence of conservative VaR- and ES-based capital charges. *Journal of Risk*, 16(3):3–22, 2014.

G. Puccetti, B. Wang, and R. Wang. Complete mixability and asymptotic equivalence of worst-possible VaR and ES estimates. *Insurance: Mathematics and Economics*, 53(3):821–828, 2013.

G. Puccetti, L. Rüschendorf, D. Small, and S. Vanduffel. Reduction of Value-at-Risk bounds via independence and variance information. *Forthcoming in Scandinavian Actuarial Journal*, 2015.

L. Rüschendorf. On one sample rank order statistics for dependent random variables. In J. Kozesnik, editor, *Transactions of the Seventh Prague Conference, Volume B*, pages 449–456. Springer, 1974.

L. Rüschendorf. Asymptotic distributions of multivariate rank order statistics. *The Annals of Statistics*, 4(5):912–923, 1976.

L. Rüschendorf. On the minimum discrimination information theorem. *Statistics & Decisions, Supplement Issue*, 1:263–283, 1984.

L. Rüschendorf. Bounds for distributions with multivariate marginals. In K. Mosler and M. Scarsini, editors, *Stochastic orders and decision under risk*, volume 19 of *IMS Lecture Notes*, pages 285–310. 1991a.

L. Rüschendorf. Fréchet bounds and their applications. In G. Dall'Aglio, S. Kotz, and G. Salinetti, editors, *Advances in Probability Distributions with Given Marginals*, volume 67 of *Mathematics and Its Applications*, pages 151–188. Springer, 1991b.

L. Rüschendorf. Comparison of multivariate risks and positive dependence. *Journal of Applied Probability*, 41:391–406, 2004.

L. Rüschendorf. Stochastic ordering of risks, influence of dependence, and a.s. constructions. In N. Balakrishnan, I. G. Bairamov, and O. L. Gebizlioglu, editors, *Advances on Models, Characterization and Applications*, pages 19–56. Chapman and Hall/CRC, 2005.

L. Rüschendorf. *Mathematical Risk Analysis*. Springer Series in Operations Research and Financial Engineering. Springer, 2013.

W. Stute. The oscillation behavior of empirical processes: The multivariate case. *The Annals of Probability*, 12(2):361–379, 1984.

B. Wang and R. Wang. The complete mixability and convex minimization problems with monotone marginal densities. *Journal of Multivariate Analysis*, 102(10):1344–1360, 2011.

R. Wang. Asymptotic bounds for the distribution of the sum of dependent random variables. *Journal of Applied Probability*, 51(3):780–798, 2014.

R. Wang, L. Peng, and J. Yang. Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics*, 17:395–417, 2013.

R. C. Williamson and T. Downs. Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2):89–158, 1990.

Shot-Noise Processes in Finance

18

Thorsten Schmidt

18.1 Introduction

Shot-noise processes constitute a well-known tool for modelling sudden changes (*shots*), followed by a typical following pattern (*noise*). In this regard, they are more flexible than other approaches simply utilizing jumps and this led to many applications in physics, biology and, with an increasing interest, in finance. Quite remarkably, shot-noise effects were already introduced in the early 20th century, see Schottky (1918); Campbell (1909a, b), sometimes also referred to as Schottky-noise. First fundamental treatments were only developed many years later with Rice (1944, 1945, 1977). Applications of shot-noise processes also arise in insurance mathematics, marketing, and even astronomy—see the survey article Bondesson (2004). The first appearances in a finance context seem to be Samorodnitsky (1996); Chobanov (1999) while in insurance mathematics these class of processes were studied even earlier, see Klüppelberg et al. (2003) for literature in this regard.

In a general form, denote by $0 < T_1 < T_2 < \dots$ the arrival times of the shots, and by $(H(., T) : T \in \mathbb{R}_{\geq 0})$ a family of stochastic processes representing the noises, then a *shot-noise process* S is given by the superposition

$$S_t = \sum_{i \geq 1} \mathbb{1}_{\{t \leq T_i\}} H(t, T_i), \quad t \geq 0; \tag{18.1}$$

an example at this level of generality can be found in Schmidt and Stute (2007). Of course, absolute convergence of the sum needs to be justified, typically by making

[1]The works stem from different authors, Stephen Oswald Rice and John Rice.

T. Schmidt (✉)
Department of Mathematical Stochastics, University Freiburg, Eckerstr. 1,
79104 Freiburg, Germany
e-mail: Thorsten.Schmidt@stochastik.uni-freiburg.de

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_18

assumptions on the arrival times together with suitable restrictions on the noise processes. For the consideration of stationarity, the process is often extend to the full real line.

At this level of generality, shot-noise processes extend compound Poisson processes significantly and neither need to be Markovian nor semimartingales. While the definition in (18.1) is very general, more restrictions will be needed to guarantee a higher level of tractability. In this paper we will focus on shot-noise processes which are semimartingales. To the best of our knowledge all articles, including Rice (1977) and many others, assume that the noises are i.i.d. and independent from the arrival times of the shots. The most common assumption even leads to a piecewise deterministic Markov process: this is the case, if the noise process are given by $H(t, T_i) = U_i e^{-a(t-T_i)}$ with i.i.d. $(U_i)_{i\geq 1}$, independent from $(T_i)_{i\geq 1}$, and $a \in \mathbb{R}$. We will show later that this is essentially the only example where Markovianity is achieved. More general cases allow for different decay, as for example a power-law decay, see e.g. Lowen and Teich (1990), or do not assume a multiplicative structure for the jump heights (U_i). These cases can be summarized under the assumption that

$$H(t, T_i) = G(t - T_i, U_i), \qquad t \geq 0,\ i \geq 1, \tag{18.2}$$

with some general random variables (U_i) and a suitable (deterministic) function G.

The obtained class of processes is surprisingly tractable, and the reason for this is that the Fourier and Laplace transforms of S are available in explicit form, depending on the considered level of generality. Even integration does not leave the class, a property shared by affine processes and of high importance for applications in interest rate markets and credit risk, see Gaspar and Schmidt (2010).

A branch of literature considers limits of shot-noise processes when the intensity of the shot arrivals increases and show, interestingly, that limits of this class of processes have fractional character, see Lane (1984); Lowen and Teich (1990); Klüppelberg and Kühn (2004), and the early studies in insurance mathematics

The application of shot-noise processes to the modelling of consumer behaviour has been suggested in Kopperschmidt and Stute (2009, 2013), wherein also the necessary statistical tools have been developed. The key in this approach is that i.i.d. shot-noise processes are at hand which allows a good access to statistical methodologies.

In the financial and insurance community they have been typically used to efficiently model shock effects, see for example Dassios and Jang (2003); Albrecher and Asmussen (2006); Schmidt and Stute (2007); Altmann et al. (2008); Jang et al. (2011); Scherer et al. (2012), and references therein. Besides this, in Moreno et al. (2011) an estimation procedure in a special class of shot-noise processes utilizing the generalized method of moments (GMM) is developed.

The paper is organized as follows: in Sect. 18.2 we introduce a suitably general formulation of shot-noise processes and derive their conditional characteristic function. Moreover, we study the connection to semimartingales and Markov processes. Proposition 2 proves that exponential decay is equivalent to Markovianity of the shot-noise process. In Sect. 18.3 we propose a model for stocks having a shot-noise

component. After the study of equivalent and absolutely continuous measure changes we obtain a drift condition implying absence of arbitrage and give an example where independence and stationarity of increments holds under the objective and the equivalent martingale measure.

18.2 Shot-Noise Processes

Our focus will lie on shot-noise processes satisfying (18.2) and the detailed study of this flexible class. Consider a filtered probability space $(\Omega, \mathscr{F}, \mathbb{F}, \mathbb{P})$ where the filtration $\mathbb{F} = (\mathscr{F}_t)_{t\geq 0}$ satisfies the usual conditions, i.e. $\mathbb{F}$ is right-continuous and $A \subset B \in \mathscr{F}$ with $\mathbb{P}(B) = 0$ implies $A \in \mathscr{F}_0$. By $\mathscr{O}$ and $\mathscr{P}$ we denote the optional, respectively predictable, σ-fields, generated by the càdlàg, respectively càg, processes.

We will allow for a marked point process as driver,[2] generalizing previous literature. In this regard, consider a sequence of increasing stopping times $0 < T_1 < T_2 < \ldots$ and a sequence of d-dimensional random variables $U_1, U_2, \ldots$. The double sequence $Z = (T_i, U_i, i \geq 1)$ is called *marked point process*. Such processes are well-studied in the literature and we refer to Brémaud (1981) for further details and references. We consider one-dimensional shot-noise processes only, a generalization to more (but finitely many) dimensions is straightforward; for the more general case see e.g. Bassan and Bona (1988) for shot-noise random fields.

Definition 1 If $Z = (T_i, U_i, i \geq 1)$ is a marked point process and $G : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ a measurable function, we call a stochastic process $S = (S_t)_{t\geq 0}$ having the representation

$$S_t = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \leq t\}} G(t - T_i, U_i), \qquad t \geq 0, \tag{18.3}$$

a *shot-noise process*. If Z has independent increments we call S an *inhomogeneous shot-noise process* and if the increments are moreover identically distributed S is called *standard shot-noise process*.

The *classical shot-noise process* is obtained when g does not depend on (U_i), see Bondesson (2004) for links to the rich literature on this class. Time-inhomogeneous Lévy processes have independent increments and hence may serve as a useful class of driving processes; see Jacod and Shiryaev (2003) for an in-depth study of processes with independent increments and, for example, Sato (1999); Cont and Tankov (2004) for a guide to the rich literature on Lévy processes. The interest in driving processes

[2] We consider here for simplicity $\mathbb{R}^d$ as mark space, while $\mathbb{R}^d$ can be replaced by a general Lusin space, see Björk et al. (1997) in this regard.

beyond processes with independent increments can be traced back to Ramakrishnan (1953); Smith (1973); Schmidt (1987)—only under additional assumptions explicit formulae can be obtained.

Note that absolute convergence of the infinite sum in (18.3) is implicit in our assumption and needs not be true in general. However, when the stopping times $(T_i)_{i\geq 1}$ have no accumulation point, this will always hold. A precise definition of this technical fact will utilize the relation to random measures and the associated compensators, which we introduce now.

To the marked point process Z we associate an integer-valued random measure μ on $\mathbb{R}_+ \times \mathbb{R}^d$ by letting

$$\mu_t(A) = \mu([0,t] \times A) := \sum_{i\geq 1} \mathbb{1}_{\{U_i \in A\}} \mathbb{1}_{\{T_i \leq t\}}, \quad t \geq 0 \tag{18.4}$$

for any $A \in \mathscr{B}(\mathbb{R}^d)$. Sometimes we also write $Z = (Z_t)_{t\geq 0}$ for the stochastic process given by $Z_t = \sum_{i\geq 1} U_i \mathbb{1}_{\{T_i \leq t\}} = \mu([0,t], \mathbb{R}^d)$. As usual, we define $\tilde{\Omega} = \Omega \times \mathbb{R}_{\geq 0} \times \mathbb{R}^d$, $\tilde{\mathscr{P}} = \mathscr{P} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^d$, and $\tilde{\mathscr{O}} = \mathscr{O} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^d$. A $\tilde{\mathscr{O}}$-measurable function W on $\tilde{\Omega}$ is called *optional*. For an optional function W and a random measure μ we define

$$W * \mu_t = \int_{[0,t]\times\mathbb{R}^d} W(s,x)\mu(ds,dx), \quad t \geq 0,$$

if $\int_{[0,t]\times\mathbb{R}^d} |W(s,x)|\mu(ds,dx)$ is finite, and $W * \mu_t = +\infty$ otherwise.

From Definition 1, we obtain that a shot-noise process S has the representation

$$S_t = \int_0^t \int_{\mathbb{R}^d} G(t-s,x)\mu(ds,dx), \quad t \geq 0$$

and in Lemma 1 we prove the for ous relevant fact that, if g is absolutely continuous, then S is a semimartingale.

The *compensator* of μ is the unique, $\mathbb{F}$-predictable random measure ν such that

$$\mathbb{E}[W * \mu_\infty] = \mathbb{E}[W * \nu_\infty]$$

for any non-negative $\tilde{\mathscr{P}}$-measurable function W on $\tilde{\Omega}$, see Theorem II.1.8 in Jacod and Shiryaev (2003).

Some properties of the marked point process Z can be determined from the compensator: if ν is deterministic (i.e. does not depend on ω), then Z has independent increments. Moreover, if the compensator additionally does not depend on time, i.e. $\nu(dt,dx) = \nu(dx)dt$, then Z also has stationary increments, hence is a Lévy process.

Example 1 (*Exponential decay*) An important special case is the well-known case when the decay is exponential. We will later show that this is essentially the only case when S is Markovian. Consider $d = 1$, assume that $\nu([0,t], \mathbb{R}) < \infty$ for all $t \geq 0$

and denote $Z_t = \sum_{T_i \leq t} U_i$, $t \geq 0$. When $G(t, x) = xe^{-bt}$, we obtain $\partial_x G(t, x) = -bG(t, x)$ and $G(0, x) = x$, such that, by Itô's formula,

$$S_t = \int_0^t -bS_u du + Z_t.$$

Hence, if Z has independent increments, then S is a Markov process, in particular, an Ornstein-Uhlenbeck process.

We give some further useful specifications of shot-noise processes to be used in the following.

Example 2 Specific choices of the noise function G lead to processes with independent increments, Markovian, and non-Markovian processes.

(i) A *jump to a new level* (with $d = 1$ and $G(t, x) = x$). Then $Z = S$ and S has the same properties, as for example independent and stationary increments, such that S is a Lévy process.
(ii) We say that S has *power-law decay* when

$$G(t, x) = \frac{x}{1 + ct}$$

with some $c > 0$. This case allows for long-memory effects and heavy clustering, compare Moreno et al. (2011). In this case, the noise decay is slower than for the exponential case and the effect of the shot persists for longer time in the data.
(iii) If the decay parameter is random, we obtain the important class of *Random decay*. For example, let $d = 2$ and

$$G(t, (u, v)) = u \exp(-vt).$$

Clearly, jump height and decay size can be dependent, see also Schmidt and Stute (2007).

Shot-noise processes offer a parsimonious and flexible framework as we illustrate in the following example. In general, shot-noise processes are not necessarily semimartingales: indeed, this is the case if $t \mapsto G(t, x)$ is of infinite variation for all x (or for at least some x).

The following result, which is well-known for standard shot-noise processes, gives the conditional characteristic function of S. This is a key result to the following applications to credit risk. We give a proof using martingale techniques which is suitable for our setup.

Proposition 1 *Assume that S is a shot-noise process, $\nu(dt, dx)$ does not depend on ω, and $\nu([0,T], \mathbb{R}^d) < \infty$. Then, for any $0 \le t \le T$ and $\theta \in \mathbb{R}$,*

$$\mathbb{E}\Big[e^{i\theta S_T}|\mathscr{F}_t\Big] = e^{i\theta \int_0^t \int_{\mathbb{R}^d} G(T-s,x)\mu(ds,dx)} \cdot \exp\left(\int_t^T \int_{\mathbb{R}^d} \Big(e^{i\theta G(T-s,x)} - 1\Big)\nu(ds,dx)\right). \tag{18.5}$$

Proof For fixed T and θ, we define

$$Z_t := \exp\left(i\theta \int_0^t \int_{\mathbb{R}^d} G(T-u,x)\mu(du,dx)\right), \quad 0 \le t \le T.$$

By the Itô formula we obtain that

$$Z_t = 1 + \int_0^t Z_{s-}\Big(e^{i\theta G(T-s,x)} - 1\Big)\mu(ds,dx), \quad 0 \le t \le T.$$

We set $\varphi(t) := \mathbb{E}[Z_t]$ for $t \in [0,T]$. Then

$$\begin{aligned}\varphi(T) &= \mathbb{E}\Big[e^{i\theta S_T}\Big] \\ &= 1 + \mathbb{E}\left[\int_0^T Z_{t-} \int_{\mathbb{R}^d} \Big(e^{i\theta G(T-t,x)} - 1\Big)\nu(dt,dx) + M_T\right] \\ &= 1 + \int_0^T \varphi(t-)F(dt), \quad 0 \le t \le T,\end{aligned}$$

where M is a martingale and $F(t) = \int_0^t \int_{\mathbb{R}^d} \Big(e^{i\theta G(T-t,x)} - 1\Big)\nu(dt,dx)$, $0 \le t \le T$ is an increasing function with associated measure $F(dx)$. The unique solution of this equation is given by

$$\begin{aligned}\varphi(T) &= \exp(F(T)) \\ &= \exp\left(\int_0^T \int_{\mathbb{R}^d} \Big(e^{i\theta G(T-t,x)} - 1\Big)\nu(dt,dx)\right).\end{aligned}$$

Finally, we observe that for $0 \le t \le T$,

$$\mathbb{E}\Big[e^{i\theta S_T}|\mathscr{F}_t\Big] = e^{i\theta \int_0^t \int_{\mathbb{R}^d} G(T-s,x)\mu(ds,dx)} \cdot \mathbb{E}\left[\exp\left(i\theta \int_t^T \int_{\mathbb{R}^d} G(T-s,x)\mu(ds,dx)\right)|\mathscr{F}_t\right].$$

As Z has independent increments by assumption, the conditional expectation is in fact an ordinary expectation which can be computed as above and we obtain the desired result.

The first part of (18.5) corresponds to the noise of already occured shots (at time t). The second part denotes the expectation of future jumps in S. By the application of iterated conditional expectations, Proposition 1 also allows to compute the finite-dimensional distributions of S.

Example 3 (*The standard shot-noise process*) If Z is a compound Poisson process, then $\nu(ds, dx) = \lambda F(dx)ds$ where λ is the arrival rate of the jumps, which itself are i.i.d. with distribution F. The classical proof of the above results uses that the jumping times of a Poisson process have the same distribution as order statistics of uniformly distributed random variables, see p. 502 in Rolski et al. (1999). In this case the proof simplifies to

$$\begin{aligned}
\mathbb{E}\Big[e^{i\theta S_T}\Big] &= \mathbb{E}\Big[\sum_{n\geq 1} \mathbb{1}_{\{T_n \leq T, T_{n+1} > T\}} e^{i\theta \sum_{j=1}^{n} G(t-T_i, U_i)}\Big] \\
&= e^{-\lambda T}\frac{(\lambda T)^n}{n!}\prod_{j=1}^{n}\frac{1}{T}\int_0^T\int_{R^d} e^{i\theta G(t-s,u)}F(du)ds \\
&= \exp\Big(-\lambda T + \lambda\int_0^T\int_{R^d} e^{i\theta G(t-s,u)}F(du)ds\Big) \\
&= \exp\Big(\int_0^T\int_{R^d}(e^{i\theta G(t-s,u)} - 1)\lambda F(du)ds\Big).
\end{aligned}$$

A conditional version is obtained in an analogous manner.

Remark 1 (*On the general case*) What can be said when ν is not deterministic? In fact, for the proof we need to compute

$$\mathbb{E}\Big[\exp\Big(i\theta\int_t^T\int_{\mathbb{R}^d} G(T-u,x)\mu(du,dx)\Big)|\mathscr{F}_t\Big]. \tag{18.6}$$

For this we need to obtain the *exponential compensator* of μ given $\mathscr{F}_t$, i.e. the $\mathscr{F}_t$-measurable random measure γ^t, such that

$$(6) = \exp\Big(i\theta\int_t^T\int_{\mathbb{R}^d} G(T-u,x)\gamma^t(du,dx)\Big).$$

Exponential compensators for semimartingales were introduced in Kallsen and Shiryaev (2002) and play an important rôle in interest rate theory, compare? We will show later that for affine shot-noise processes we will be able to compute the exponential compensator efficiently, see Example 4 where we study a self-exciting shot-noise process.

The following result, taken from Schmidt (2014), gives sufficient conditions which yield that S is a semimartingale.

Lemma 1 *Fix $T > 0$ and assume that $G(t,x) = G(0,x) + \int_0^t g(s,x)ds$ for all $0 \le t \le T$ and all $x \in \mathbb{R}^d$. If*

$$\int_0^T \int_{\mathbb{R}^d} (g(s,x))^2 \nu(ds,dx) < \infty, \tag{18.7}$$

$\mathbb{P}$-a.s., then $(S_t)_{0\le t\le T}$ is a semimartingale.

For the convenience of the reader we repeat the proof of this result.

Proof Under condition (18.7), we can apply the stochastic Fubini theorem in the general version given in Theorem IV.65 in Protter (2004). Observe that

$$\begin{aligned} S_t &= \int_0^t \int_{\mathbb{R}^d} \int_s^t g(u-s,x)du\,\mu(ds,dx) + \int_0^t \int_{\mathbb{R}^d} G(0,x)\mu(ds,dx) \\ &= \int_0^t \int_0^s \int_{\mathbb{R}^d} g(u-s,x)\mu(ds,dx)\,du + \int_0^t \int_{\mathbb{R}^d} G(0,x)\nu(ds,dx) + M_t, \end{aligned} \tag{18.8}$$

with a local martingale M. This is the semimartingale represenation of S and hence S is a semimartingale.

It is possible to generalize this result to the case where $G(t,x) = G(0,x) + \int_0^t g(s,x)dA(s)$ with a process A of finite variation. Here, however, we do not make use of such a level of generality—see Jacod and Shiryaev (2003), Proposition II.2.9. for details on the choice of A.

Moreover, a characterization of semimartingales when starting from the more general formulation in (18.1) is possible using similar methodologies, see Schmidt and Stute (2007) for an example.

Remark 2 Having a driver Z which has independent and stationary increments may be a limitation in some applications. It is straightforward to allow for more general driving processes. For example, consider a filtration $\mathbb{G} = (\mathcal{G}_t)_{t\ge 0}$ satisfying the usual conditions. Let ν be a $\mathcal{G}_0$-measurable random measure on $[0,T] \times \mathbb{R}^d$ such that for any open set A in $\mathbb{R}^k$,

$$\mathbb{P}\Big(\sum_{T_i \in (s,t]} \mathbb{1}_{\{X_i \in A\}} = k \,\Big|\, \mathcal{G}_s \Big) = e^{-\nu((s,t]\times A)} \frac{(\nu((s,t]\times A))^k}{k!}.$$

If $X_1, X_2, \ldots$ are i.i.d. and independent of $\mathbb{G}$, then Z is a $\mathbb{G}$-*doubly stochastic marked Poisson process*, Intuitively, given $\mathcal{G}$, Z is a (time-inhomogeneous) Poisson process with $\mathcal{G}_0$-measurable jumps. This is a so-called initial enlargement of filtration, compare Bielecki et al. (2000) or Jeanblanc and Rutkowski (2000) for an introduction into this field. Doubly-stochastic marked Poisson processes in credit risk modeling

have also been considered in Gaspar and Slinko (2008), however not in a shot-noise setting.

Example 4 (An affine self-exciting shot-noise process) Inspired by Errais et al. (2010) we consider the two-dimensional affine process $X = (N, \lambda)^\top$ where

$$d\lambda_t = \kappa(\theta - \lambda_t)dt + dN_t \tag{18.9}$$

and N is a counting process with intensity λ. In this case the compensator of N is given by $\nu_N(dt, dx) = \lambda_t \delta_1(dx)dt$; δ_a denoting the Dirac measure at the point a. Following Keller-Ressel et al. (2013), the process X is a two-dimensional affine process with state space $\mathbb{N}_0 \times \mathbb{R}_{\geq 0}$ when $\theta \geq 0$. Hence its conditional distribution is given in exponential affine form, i.e.

$$\mathbb{E}[e^{iuX_T}|\mathscr{F}_t] = \exp\Big(\phi(T-t,u) + \langle \psi(T-t,u), X_t\rangle\Big),$$

for all $u \in \mathbb{R}^2$ and the coefficients ϕ and ψ solve the generalized Riccati equations

$$\begin{aligned}
\partial_t \phi(t,u) &= \kappa\theta\psi_2(t,u)\\
\partial_t \psi_1(t,u) &= 0\\
\partial_t \psi_2(t,u) &= -\kappa\psi_2(t,u) + \exp(\psi_2(t,u) + \psi_1(t,u)) - 1
\end{aligned}$$

with the boundary conditions $\phi(0,u) = 0$ and $\psi(0,u) = u$ (see Proposition 3.4 in Keller-Ressel et al. (2013)). Hence $\psi_1(t,u) = u_1$. Observe that λ is a shot-noise process (when $\lambda_0 = \theta = 0$): the solution of (18.9) is

$$\lambda_t = e^{-\kappa t}\lambda_0 + \theta(1 - e^{-\kappa t}) + \sum_{i=1}^{N_t} e^{-\kappa(t-T_i)},$$

where we denoted by $T_1, T_2, \ldots$ the jumping times of N. Hence, for $\lambda_0 = \theta = 0$, λ is an (affine and Markovian) shot-noise process.

18.2.1 Markovianity

Proposition 1 allows us to draw a connection to *affine* processes. This processes have been studied intensively in the literature because of their high tractability. Letting $G(t,x) = xe^{-bt}$ implies that $G(t+s,x) = G(t,x)e^{-bs}$ which is the key to Markovianity. Then,

$$\begin{aligned}
e^{i\theta\int_0^t\int_{\mathbb{R}^d} G(T-s,x)\mu(ds,dx)} &= e^{i\theta\int_0^t\int_{\mathbb{R}^d} e^{-b(T-t)}G(t-s,x)\mu(ds,dx)}\\
&= e^{i\theta\int_0^t\int_{\mathbb{R}^d} e^{-b(T-t)}G(t-s,x)\mu(ds,dx)}\\
&= e^{i\theta e^{-b(T-t)}\int_0^t\int_{\mathbb{R}^d} G(t-s,x)\mu(ds,dx)} = e^{i\theta e^{-b(T-t)}S_t},
\end{aligned}$$

such that

$$\mathbb{E}\Big[e^{i\theta S_T}\,\Big|\,\mathscr{F}_t\Big] = \exp\left(\int_t^T \int_{\mathbb{R}^d} \Big(e^{i\theta G(T-s,x)} - 1\Big)\nu(ds,dx)\right) \cdot e^{i\theta e^{-b(T-t)}S_t}$$
$$=: \exp(\phi(t,T,\theta) + \psi(t,T,\theta)S_t),$$

which is the exponential-affine structure classifying affine processes. While for affine processes ϕ and ψ are determined via solutions of generalized Riccati equations, in the shot-noise case we obtain a simpler integral-representation. Similar in spirit, we obtain that under Markovianity, many expectations simplify considerably, as the following result illustrates.

Corollary 1 *Consider an inhomogeneous shot-noise process S with $G(t,x) = xe^{-bt}$ and $E|X_i| < \infty, i \geq 1$. Then, for $T > t$,*

$$\mathbb{E}[S_T|\mathscr{F}_t] = e^{-b(T-t)}S_t + \mathbb{E}\Big[\sum_{T_i\in(t,T]} X_i e^{-b(T-T_i)}\Big].$$

We now focus our attention on the important question of Markovianity of shot-noise processes. Typically, shot-noise processes are not Markovian. Still, from a computational point of view Markovianity could be preferable. Proposition 2 provides a clear classification when the decay function satisfies $G(t,x) = xH(t)$: then Markovianity is equivalent to an *exponential decay*. In more general cases one typically looses Markovianity.

Proposition 2 *Consider a standard shot-noise process S where $G(t,x) = xH(t)$ with a càdlàg function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}$. Assume that there exists an $\epsilon > 0$ such that $(0,\epsilon] \subset H(\mathbb{R}^+)$. Then S is Markovian, if and only if there exist $a, b \in \mathbb{R}$ such that*

$$H(t) = ae^{-bt}.$$

Proof First, consider the case where the shot-noise process S is Markovian (with respect to the filtration $\mathbb{F}$). For $s > t$, we have that

$$\mathbb{E}[S_s|\mathscr{F}_t] = \sum_{T_i\leq t} U_n H(s-T_i) + \mathbb{E}\Big[\sum_{T_i\in(t,s]} U_n H(s-T_i)\,\Big|\,\mathscr{F}_t\Big]. \tag{18.10}$$

As Z has independent and stationary increments, we obtain that

$$\mathbb{E}\Big[\sum_{T_i\in(t,s]} U_n H(s-T_i)\,\Big|\,\mathscr{F}_t\Big] = \mathbb{E}\Big[\sum_{T_i\in(0,s-t]} U_n H(s-t-T_i)\Big]$$

is a deterministic function (and hence does not depend on ω). From Markovianity it follows that $\mathbb{E}[S_s|\mathscr{F}_t] = \mathbb{E}[S_s|S_t] =: \tilde{F}(t,s,S_t)$ for all $0 \leq s \leq t$, where $\tilde{F}$ is

a measurable function. Hence, we obtain the existence of a measurable function $F : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}$, such that

$$\sum\nolimits_{T_i \le t} U_n H(s - T_i) = F(t, s, S_t) \tag{18.11}$$

a.s. for all $0 \le t \le s$. If $\mathbb{P}(U_1 = 0) = 1$ the claim holds with $a = 0$. Otherwise choose non-zero u such that (18.11) holds with $U_1, U_2, \ldots$ replaced by u. W.l.o.g. consider $u = 1$. In particular, $F(t, s, H(t - T_1)) = H(s - T_1)$ holds a.s. As in Remark 3 we condition on $N_t = n$ and obtain that

$$F(t, s, \sum_{i=1}^{n} H(t - \eta_i)) = \sum_{i=1}^{n} H(s - \eta_i) = \sum_{i=1}^{n} F(t, s, H(t - \eta_i)) \tag{18.12}$$

with probability one, where η_i are i.i.d. $U[0, s]$. As $(0, \epsilon] \subset H(\mathbb{R}^+)$,

$$F(t, s, x_1 + \cdots + x_n) = \sum_{i=1}^{n} F(t, s, x_i) \tag{18.13}$$

for all $x_1, x_2, \ldots \mathbb{R}^+$ and $n \ge 1$ except for a null-set with respect to the Lebesgue measure. Note with h being càd so is F in the third coordinate and we obtain that (18.13) holds for all $x_1, x_2, \cdots \in \mathbb{R}^+$. Hence, F is additive such that $F(t, s, x) = F(t, s, 1)x$ (see Theorems 5.2.1 and 9.4.3 in Kuczma and Gilányi (2009)) for all $x \in \mathbb{R}^+$.

Next, we exploit

$$F(t, s, 1)H(t - u) = H(s - u)$$

for all $0 \le u \le t \le s$ to infer properties of h. First, $u = 0$ gives $F(t, s, 1)H(t) = H(s)$ and so $H(0) \ne 0$ because otherwise $H(s)$ would vanish for all $s \ge 0$ which contradicts $(0, \epsilon] \subset H(\mathbb{R}^+)$. Next, $u = t$ gives $H(s - t) = F(t, s, 1)H(0)$ such that

$$H(s - t)H(t) = F(t, s, 1)H(0)H(t) = H(s)H(0).$$

This in turn yields that $f := H(t)/H(0)$ satisfies

$$f(x + y) = f(x)f(y).$$

Then f is additive and measurable and hence continuous. The equation is a multiplicative version of Cauchy's equation and hence $f(x) = e^{-bx}$, see Theorem 13.1.4 in Kuczma and Gilányi (2009) such that we obtain $H(x) = H(0)e^{-bx}$.

For the converse, note that if $G(t) = ae^{-bt}$, then

$$\sum_{T_i \le t} U_n G(s - T_i) = G(s - t) \sum_{T_i \le t} U_n G(t - T_i),$$

and hence (18.10) yields that S is Markovian.

Remark 3 For Markovianity it is necessary that $U_1, U_2, \ldots$ are independent *and* identically distributed. Merely for the sake of the argument, assume that $U_1, U_2 \in \{0, 1, 2\}$ and $0 = U_3 = U_4, \ldots$. If $t > T_1$ and $S_t = 2$ the distribution of S_{t+1} depends not only on S_t but also on the number of jumps before t and so it is not Markovian.

18.3 The Application to Financial Markets

Shot-noise processes have been applied to the modelling of stock markets and to the modelling of intensities, which is useful in credit risk and insurance mathematics. Following the works Altmann et al. (2008); Schmidt and Stute (2007); Moreno et al. (2011) we consider the application to the modelling of stocks. The main idea is to extend the Black-Scholes-Merton framework by a shot-noise component.

In this regard, we denote by X the price process of the stock. As in the previous section, a marked point process $Z = (T_i, X_i)_{i \geq 1}$ with mark space $\mathbb{R}^d$ and a noise function $g : \mathbb{R}_{\geq 0} \times \mathbb{R}^d$ determine the shot-noise component. Additionally, there is a one-dimensional Brownian motion W, which is independent of Z, and $\sigma > 0$. The integer-valued random measure μ counts the jumps of the marked point process Z, see Eq. 18.4. Alltogether, we assume that

$$X(t) = X(0) \exp\left(\mu t + \sigma W(t) \frac{\sigma^2 t}{2} + \int_0^t \sum_{T_i \leq s} g(s - T_i, U_i) ds + \sum_{T_i \leq t} G(0, U_i)\right), \; t \geq 0. \tag{18.14}$$

To guarantee absence of arbitrage, one has to find an equivalent martingale measure. However, for statistical estimation of the model it is important to have a nice structure of the process under the risk-neutral measure. It turns out that this is not the case for the minimal martingale measure, studied in Schmidt and Stute (2007), and our goal is it to classify all martingale measures by a drift condition and give some hints of possible choices of nice martingale measures. The first important step will therefore be to classify all equivalent measures.

18.3.1 Equivalent Measure Changes

In this section we study all equivalent measure changes which apply to different settings of shot-noise processes. We consider an initial filtration $\mathscr{H} \subset \mathscr{F}_0$.

(A1) $\mathscr{F} = \mathscr{F}_{\infty-}$ and $\mathbb{F}$ is the smallest filtration for which μ is optional and $\mathscr{H} \subset \mathscr{F}_0$.

(A2) The measure ν is absolutely continuous with respect to the Lebesgue-measure, i.e. there is a kernel, which we denote again by $\nu(t, dx)$ such that

$$\nu(dt, dx) = \nu(t, dx) dt.$$

We set

$$\xi_n := \inf(t : \int_0^t \int_{\mathbb{R}^d} (1 - \sqrt{Y(s,u)})^2 \nu(s,du)ds \geq n) \quad \text{for all } n.$$

Proposition 3 *Assume that* **(A1)** *and* **(A2)** *hold and* $\mathbb{P}' \ll \mathbb{P}$. *Then there exists a* $\mathscr{P} \otimes \mathbb{R}^d$*-measurable non-negative function* Y *such that the density process* Z *of* $\mathbb{P}'$ *relative to* $\mathbb{P}$ *coincides with*

$$Z_t^n = e^{-\int_0^{t\wedge\xi_n} \int_{\mathbb{R}^d} (Y(s,u)-1)\nu(s,du)ds} \prod_{T_n \leq t} Y(T_n, U_n) \tag{18.15}$$

Moreover, Z *is a (possibly explosive) marked point process under* $\mathbb{P}'$ *and its compensator w.r.t.* $\mathbb{P}'$ *is* $Y(t,u)\nu(t,du)dt$.

Proof We apply Theorem III.5.43 in Jacod and Shiryaev (2003) and refer to their notation for this proof. All references in this proof rever to Jacod and Shiryaev (2003). Note that because the compensator of Z is absolutely continuous,

$$\hat{Y}_t = \int_{\mathbb{R}^d} Y(t,u)\nu(\{t\},du) = 0$$

(compare Eq. III.5.2) and therefore σ given in Eq. III.5.6 satisfies $\sigma = \infty$. Furthermore, the process H given in Eq. III.5.7 computes to

$$H_t = \int_0^t \int_{\mathbb{R}^d} (1 - \sqrt{Y(s,u)})^2 \nu(s,du)ds.$$

A priory we do not have that H is finite, so that following III.5.9 we define $\xi_n := \inf(t : H_t \geq n)$ and define N^{ξ_n} by

$$N_t^{\xi_n} := \int_0^{t\wedge\xi_n} (Y - 1)\,(\mu(ds,du) - \nu(s,du)ds).$$

Proposition III.5.10 yields that there exists a unique N which coincides with N^{G^n} at least on all random intervals $[0, \xi_n]$, $n \geq 1$. Theorem III.5.43 yields that under our assumptions the density Z coincides with Z^n as inspection of formula III.5.21 shows. This gives our claim.

The main tool is the following result which considers the stronger case of equivalent measures.

Theorem 1 *Assume that* **(A1)** *and* **(A2)** *hold and* $\mathbb{P} \sim \mathbb{P}'$. *Then*

$$\int_0^t \int_{\mathbb{R}^d} Y(s,u)\nu(s,du)ds < \infty \tag{18.16}$$

$\mathbb{P}'$*-almost surely for all* $t > 0$ *and the density* Z *is given by*

$$Z_t = e^{-\int_0^t \int_{\mathbb{R}^d} (Y(s,u)-1)G(s,du)ds} \prod_{T_n \le t} Y(T_n, U_n), \quad t \ge 0. \tag{18.17}$$

Proof As $\mathbb{P}$ and $\mathbb{P}'$ are equivalent and we consider only non-explosive marked point processes, $\mathbb{P}'(\lim_{n\to\infty} T_n = \infty) = 1$. Hence $\int_0^t \int_{\mathbb{R}^d} \nu(s,du)ds < \infty$ for all $t > 0$, almost surely with respect to $\mathbb{P}$ and $\mathbb{P}'$.

Also, $\int_0^t \int_{\mathbb{R}^d} Y(s,u)\nu(s,du)ds < \infty$: let $A_t := \{\omega \in \Omega : \int_0^t \int_{\mathbb{R}^d} Y(s,u)\nu(s,du)ds = \infty\}$ be a set with positive probability. Then, Z vanishes on A_t and so $\mathbb{P}$ is not equivalent to $\mathbb{P}$ which gives a contradiction. Because $Y\nu$ is non-negative $A_t \subset A_{t+\epsilon}$ for all $\epsilon > 0$ and (18.16) follows $\mathbb{P}'$-almost surely for all $t > 0$. Finally, note that

$$\int_0^t \int_{\mathbb{R}^d} (1 - \sqrt{Y(s,u)})^2 \nu(s,du)ds \le \int_0^t \int_{\mathbb{R}^d} (1+Y)\nu(s,du)ds < \infty.$$

Hence the ξ_n in Proposition 3 tend to infinity with probability 1. Then (18.15) together with Proposition III.5.10 in Jacod and Shiryaev (2003) gives (18.17).

We have the following important result: the shot-noise property is preserved under an absolutely continuous (and hence also under an equivalent) change of measure.

Corollary 2 *Assume that* **(A1)** *and* **(A2)** *hold and* $\mathbb{P}' \ll \mathbb{P}$. *If* S *is a shot-noise process under* $\mathbb{P}$, *then* S *is a shot-noise process under* $\mathbb{P}'$.

Proof The result follows immediately from the Definition 1 together with Proposition 3: under $\mathbb{P}'$, the representation (18.3) of course still holds and by Proposition 3 states that $Z = (T_i, U_i)_{i\ge 1}$ is a marked point process under $\mathbb{P}'$.

We will see that additional useful properties, like independent increments are not preserved under the change of measure, such that the specific structures of the shot-noise process under both measures can be substantially different.

18.3.2 Preserving Independent Increments

For tractability reasons one often considers shot-noise processes driven by a marked point process which has independent increments. If the increments are moreover stationary, the associated process is a Lévy process. We cover both cases in this section.

Theorem 2 *Assume that* $\mathbb{P} \sim \mathbb{P}'$. *Let the density process of* $\mathbb{P}'$ *relative to* $\mathbb{P}$ *be of the form* (18.17).

1. *If* Z *has independent increments under* $\mathbb{P}$ *and* $\mathbb{P}'$, *then* Y *is deterministic.*
2. *If* Z *has independent and stationary increments under* $\mathbb{P}$ *and* $\mathbb{P}'$, *then* Y *is deterministic and does not depend on time.*

Proof Z is a process with independent increments (PII), if and only if its compensator is deterministic, see Jacod and Shiryaev (2003). Hence, if Z is a PII under $\mathbb{P}$, then $\nu(\omega, t, dx) = \nu(t, dx)$ is deterministic. By Theorem 1, Z has a deterministic compensator under $\mathbb{P}'$ if and only if $Y(\omega, t, u)\nu(t, du)$ is deterministic and hence $Y(\omega, t, u) = Y(t, u)$ is deterministic. Stationarity is equivalent to ν being independent of time and so (ii) follows analogously.

Example 5 (*The Esscher measure*) Consider a generic n-dimensional stochastic process X. Then the Esscher measure Esscher (1932) is given by the density

$$Z_t = \frac{e^{hX_t}}{\mathbb{E}(e^{hX_t})}$$

where $h \in \mathbb{R}^d$ is chosen in such a way that Z is a martingale. Esche and Schweizer (2005) showed that the Esscher measure preserves the Lévy property, in a specific context. It is quite immediate that if applied to a model for stock-prices driven by shot-noise processes this property will not hold in general. Dassios and Jang (2003) applied the Esscher measure to Markovian shot-noise processes.

Example 6 (*The minimal martingale measure*) The minimal martingale measure as proposed in Föllmer and Schweizer (1990) for a certain class of shot-noise processes has been analysed in Schmidt and Stute (2007). It can be described as follows: consider the special semimartingale X in its semimartingale decomposition $X = A + M$ where A is an increasing process of bounded variation and M is a local martingale. Assume that there exists a process ℓ which satisfies

$$A_t = \int_0^t \ell_s d\langle M \rangle_s.$$

Then the density of the minimal martingale measure with respect to $\mathbb{P}$ is given by

$$Z = \mathscr{E}\left(\int_0^{\cdot} \ell_{s-} dM_s\right).$$

Here $\mathscr{E}$ denotes the Doleans-Dade stochastic exponential, i.e. Z is the solution of $dZ_t = Z_{t-}\ell_{t-}dM_t$. The minimal martingale measure need not exist in general. From

(18.14), proceeding as in the proof of Proposition 4.1 in Schmidt and Stute (2007), we obtain that

$$\ell_{t-} = \frac{1}{X_{t-}} \frac{\mu + \sum_{T_i<t} g(t - T_i, U_i) + \int_{\mathbb{R}^d} (e^{G(0,x)} - 1)\nu(t, ds)}{\int_{\mathbb{R}^d} (e^{G(0,x)} - 1)^2 \nu(t, ds)}.$$

Conditions which ensure that the minimal martingale measure is indeed a probability measure can be found in Schmidt and Stute (2007).

From Theorem 2 it is clear, that the minimal martingale measure will not preserve independent increments of Z—a property which makes this measure less tractable for financial applications. In the following section, we propose an alternative to this approach.

18.3.3 The Drift Condition

We consider the equivalent measure $\mathbb{P}' \sim \mathbb{P}$ and assume that **(A1)** and **(A2)** hold. Then Theorem 1 gives the relationship between both measures and Z is again a marked point process und $\mathbb{P}'$. The compensator of μ under $\mathbb{P}'$ is given by

$$\nu'(dt, tx) = \nu'(t, dx)dt = \nu(t, dx)Y(t, x)dt.$$

By the equivalent change of measure, there exists a market price of risk ξ, such that $W' = W + \xi$ is a $\mathbb{P}'$-Brownian motion, see Jacod and Shiryaev (2003), Theorem III.3.24.

We assume that discounting takes place via a bank account with constant short rate r.

Theorem 3 *The equivalent measure $\mathbb{P}'$ is a (local) martingale measure, if*

$$r = \mu - \sigma\xi_t + \int_0^t \int_{\mathbb{R}^d} g(t - s, x)\mu(ds, dx) + \int_{\mathbb{R}^d} \left(e^{G(0,x)} - 1\right)\nu'(t, dx) \tag{18.18}$$

$d\mathbb{P} \otimes dt$-almost surely for all $t \geq 0$.

Proof We first derive the semimartingale representation of X. By Itô's formula and (18.8),

$$\begin{aligned} dX_t = X_{t-}\left(\mu dt + \sigma dW_t + \int_0^t \int_{\mathbb{R}^d} g(t - s, x)\mu(ds, dx)\right) \\ + \int_{\mathbb{R}^d} X_{t-}\left(e^{G(0,x)} - 1\right)\mu(dt, dx). \end{aligned}$$

The equivalent change of measure allows to introduce a drift ξ to the Brownian motion, such that $W' = W + \xi$ is a $\mathbb{P}'$-Brownian motion. Compensating μ with the $\mathbb{P}'$-compensator gives the result.

It is apparent that typically there will be many solutions of the drift condition. With a view on tractability it is reasonable to impose that the marked point process Z has independent (and possibly stationary) increments under $\mathbb{P}$ and $\mathbb{P}'$. From Theorem 2 it follows that this is the case if the function Y is deterministic (and does not depend on time). Then, from Eq. (18.18) we obtain that

$$\xi_t = \sigma^{-1}\Big(\mu + \int_0^t \int_{\mathbb{R}^d} g(t-s,x)\mu(ds,dx) + \int_{\mathbb{R}^d} \Big(e^{G(0,x)} - 1\Big) Y(t,x)\nu(t,dx)\Big).$$

Example 7 (*Independent and stationary increments under both measures*) Fix a finite time horizon T^* and assume that Z has independent and stationary increments, i.e. $\nu(t,dx) = \lambda F(dx)$ where F is the distribution of U_1 and $\lambda > 0$ is the arrival rate of the jumps. Assume that F' is equivalent to F, i.e. $F'(dx) = \eta(x)F(dx)$ and $\lambda' > 0$. Then an equivalent change of measure is obtained via $Y(t,x) = \frac{\lambda'}{\lambda}\eta(x)$. In this case, the arrival rate of jumps under $\mathbb{P}'$ is λ' and the jumps sizes are again i.i.d. with distribution F'. Assume that $\int e^{G(0,x)} F'(dx) < \infty$ and let ξ be such that

$$\xi_t = \sigma^{-1}\Big(\mu + m_1 + \int_0^t \int_{\mathbb{R}^d} g(t-s,x)\mu(ds,dx)\Big) \tag{18.19}$$

with $m_1 := \int_{\mathbb{R}^d} \Big(e^{G(0,x)} - 1\Big)\lambda' F'(dx)$. If furthermore the process

$$\Big(\mathscr{E}\Big(\int_0^t \xi_s dW_s\Big)\Big)_{0\le t\le T^*}$$

is a true martingale, then $\mathbb{P}'$ is an equivalent (local) martingale measure.

References

Albrecher, H. and Asmussen, S. (2006), 'Ruin probabilities and aggregate claims distributions for shot noise Cox processes', *Scandinavian Actuarial Journal* (2), 86–110.

Altmann, T., Schmidt, T. and Stute, W. (2008), 'A shot noise model for financial assets', *International Journal of Theoretical and Applied Finance* **11**(1), 87–106.

Bassan, B. and Bona, E. (1988), Shot noise random fields, *in* 'Biomathematics and related computational problems (Naples, 1987)', Kluwer Acad. Publ., Dordrecht, pp. 423–427.

Bielecki, T. R., Jeanblanc, M. and Rutkowski, M. (2000), Modeling default risk: Mathematical tools, *in* 'Fixed Income and Credit risk modeling and Management', New York University, Stern School of Business, Statistics and Operations Research Department, workshop, pp. 171–269.

Björk, T., Kabanov, Y. and Rungaldier, W. (1997), 'Bond market structure in the presence of marked point processes', *Mathematical Finance* **7**, 211–239.

Bondesson, L. (2004), *Shot-Noise Processes and Distributions*, John Wiley & Sons, Inc.

Brémaud, P. (1981), *Point Processes and Queues*, Springer Verlag. Berlin Heidelberg New York.

Campbell, N. (1909a), 'Discontinuities in light emission', *Proc. Cumbr. Phil. Sot.* **15**, 310–328.

Campbell, N. (1909b), 'The study of discontinuous phenomena', *Proc. Cumbr. Phil. Sot.* **15**, 117–136.

Chobanov, G. (1999), 'Modeling financial asset returns with shot noise processes', *Mathematical and Computer Modelling* **29**(10), 17–21.

Cont, R. and Tankov, P. (2004), *Financial Modelling with Jump Processes*, Chapman & Hall.

Dassios, A. and Jang, J. (2003), 'Pricing of catastrophe reinsurance & derivatives using the Cox process with shot noise intensity', *Finance and Stochastics* **7**(1), 73–95.

Errais, E., Giesecke, K. and Goldberg, L. R. (2010), 'Affine point processes and portfolio credit risk', *SIAM Journal on Financial Mathematics* **1**, 642–665.

Esche, F. and Schweizer, M. (2005), 'Minimal entropy preserves the Lévy property: how and why', *Stochastic Process. Appl.* **115**(2), 299–327.

Esscher, F. (1932), 'On the probability function in the collective theory of risk', *Skandinavisk Aktuarietidskrift* **15**, 175–95.

Föllmer, H. and Schweizer, M. (1990), Hedging of contingent claims under incomplete information, *in* M. H. A. Davis and R. J. Elliott, eds, 'Applied Stochastic Analysis', Vol. 5, Gordon and Breach, London/New York, pp. 389–414.

Gaspar, R. M. and Schmidt, T. (2010), 'Credit risk modelling with shot-noise processes', *working paper*.

Gaspar, R. M. and Slinko, I. (2008), 'On recovery and intensity's correlation - a new class of credit risk models', *Journal of Credit Risk* **4**, 1–33.

Jacod, J. and Shiryaev, A. (2003), *Limit Theorems for Stochastic Processes*, 2nd edn, Springer Verlag, Berlin.

Jang, J., Herbertsson, A. and Schmidt, T. (2011), 'Pricing basket default swaps in a tractable shot noise model', *Statistics and Probability Letters* **8**, 1196–1207.

Jeanblanc, M. and Rutkowski, M. (2000), Modeling default risk: an overview, *in* 'Mathematical Finance: theory and practice Fudan University', Modern Mathematics Series, High Education press, Beijing, pp. 171–269.

Kallsen, J. and Shiryaev, N. A. (2002), 'The cumulant process and esscher's change of measure', *Finance and Stochastics* **6**(4), 397–428.

Keller-Ressel, M., Schachermayer, W. and Teichmann, J. (2013), 'Regularity of affine processes on general state spaces', *Electron. J. Probab.* **18**, 17 pp.

Klüppelberg, C. and Kühn, C. (2004), 'Fractional Brownian motion as a weak limit of Poisson shot noise processes—with applications to finance', *Stochastic Processes and their Applications* **113**(2), 333–351.

Klüppelberg, C., Mikosch, T. and Schärf, A. (2003), 'Regular variation in the mean and stable limits for poisson shot noise', *Bernoulli* **9**(3), 467–496.

Kopperschmidt, K. and Stute, W. (2009), 'Purchase timing models in marketing: a review', *AStA Advances in Statistical Analysis* **93**(2), 123–149.

Kopperschmidt, K. and Stute, W. (2013), 'The statistical analysis of self-exciting point processes', *Statistica Sinica* **23**(3), 1273–1298.

Kuczma, M. and Gilányi, A. (2009), *An Introduction to the Theory of Functional Equations and Inequalities*, Springer-Verlag, New York.

Lane, J. A. (1984), 'The central limit theorem for the poisson shot-noise process', *Journal of Applied Probability* **21**, 287–301.

Lowen, S. B. and Teich, M. C. (1990), 'Power-law shot noise', *IEEE Transactions on Information Theory* **36**(6), 1302–1318.

Moreno, M., Serrano, P. and Stute, W. (2011), 'Statistical properties and economic implications of jump-diffusion processes with shot-noise effects', *European Journal of Operational Research* **214**(3), 656–664.

Protter, P. (2004), *Stochastic Integration and Differential Equations*, 2nd edn, Springer Verlag. Berlin Heidelberg New York.

Ramakrishnan, A. (1953), 'Stochastic processes associated with random divisions of a line', *Proc. Cambridge Philos. Soc.* **49**, 473–485.

Rice, J. (1977), 'On generalized shot noise', *Advances in Applied Probability* **9**, 553–565.

Rice, S. O. (1944), 'Mathematical analysis of random noise', *The Bell System Technical Journal* **23**, 282–332.

Rice, S. O. (1945), 'Mathematical analysis of random noise', *The Bell System Technical Journal* **24**, 46–156.

Rolski, T., Schmidli, H., Schmidt, V. and Teugels, J. (1999), *Stochastic Processes for Insurance and Finance*, John Wiley & Sons. New York.

Samorodnitsky, G. (1996), *A class of shot noise models for financial applications*, Springer New York, pp. 332–353.

Sato, K.-I. (1999), *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press.

Scherer, M., Schmid, L. and Schmidt, T. (2012), 'Shot-noise multivariate default models', *European Actuarial Journal* **2**, 161–186.

Schmidt, T. (2014), 'Catastrophe insurance modeled by shot-noise processes', *Risks* **2**, 3–24.

Schmidt, T. and Stute, W. (2007), 'General shot-noise processes and the minimal martingale measure', *Statistics & Probability Letters* **77**, 1332–1338.

Schmidt, V. (1987), 'On joint queue-length characteristics in infinite-server tandem queues with heavy traffic', *Advances in Applied Probability* **19**(2), 474–486.

Schottky, W. (1918), 'Über spontane Stromschwankungen in verschiedenen Elektrizitätsleitern', *Annalen der Physik* **362**(23), 541–567.

Smith, W. (1973), 'Shot noise generated by a semi-Markov process', *J. Appl. Probability* **10**, 685–690.

A Lévy-Driven Asset Price Model with Bankruptcy and Liquidity Risk

19

Patrick Bäurer and Ernst Eberlein

19.1 Introduction

Standard models for asset prices do not take the possibility of bankruptcy of the underlying company into account. In real markets, however, there are plenty of cases where a listed company went bankrupt with the consequence of a total loss of the invested capital. Figure 19.1 shows an example. It is the purpose of this paper to expand an approach such that bankruptcy can occur. As underlying asset price model $S = (S_t)_{t\geq 0}$ we choose an exponential model which is driven by a Lévy process $L = (L_t)_{t\geq 0}$. A second Lévy process $Z = (Z_t)_{t\geq 0}$ is used as driver for the hazard rate which determines the default time. The asset price jumps to zero when this event happens.

It is a well-known fact that there is a strong negative dependence between the value of the asset and the probability of default of the corresponding company. Figure 19.3 shows a striking example where we plotted CDS quotes of the German energy company E.ON against its stock price. In order to take this dependence into account in the modeling approach which will be developed, the process Z is not only used for the definition of the time point of default, but enters as an additional driver into the equation for the asset price. Negative dependence is generated via a minus sign in front of Z. The remaining terms in the definition of S are determined by the fact that the discounted asset price should be a martingale.

Earlier approaches where bankruptcy of the underlying company is taken into account are Davis and Lischka (2002), Andersen and Buffum (2004), Linetsky (2006)

P. Bäurer (✉) · E. Eberlein
Mathematical Institute, University of Freiburg, Freiburg Im Breisgau, Germany
e-mail: p.baeurer@gmx.de

E. Eberlein
e-mail: eberlein@stochastik.uni-freiburg.de

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_19

and Carr and Madan (2010). In these papers the driving process is a standard Brownian motion and the hazard rate of bankruptcy is chosen as a decreasing function of the stock price. A particular parsimonious specification for such a function is given by a negative power of the stock price. In order to improve the performance Carr and Madan (2010) use a stochastic volatility model and jointly employ price data on credit default swaps (CDSs) and equity options to simultaneously infer the risk neutral stock dynamics in the presence of the possibility of default.

Since we will use European option prices to calibrate the model, a Fourier-based valuation formula is derived. Several types of options are discussed explicitly. In order to get prices expressed as expectations in a form which is convenient from the point of view of numerics, the survival measure is introduced. The effect of the measure change is that expectations are those of a standard payoff function. Calibration is done with L being a normal inverse Gaussian (NIG) and the independent process Z being a Gamma process. As an alternative to the Fourier-based valuation method we derive also the corresponding partial integro-differential equations (PIDEs). In the last section we show that the defaultable asset price approach which is exposed here, provides also an appropriate basis for the recently developed two price theory. The latter allows to get bid and ask prices and thus to model in addition the liquidity component of the market.

19.2 The Defaultable Asset Price Model

A standard model for the price process $(S_t)_{t\geq 0}$ of a traded asset which goes back to Samuelson (1965) is given by

$$S_t = S_0 e^{X_t} \tag{19.1}$$

where $X = (X_t)_{t\geq 0}$ is a Brownian motion. This approach represented an essential improvement on the initial Bachelier (1900) model where S had been a Brownian motion itself. The main differences are that asset prices according to (1) are positive and behave in a multiplicative or geometric way. The geometric Brownian motion became well-known as the basis for the celebrated option pricing formula due to Black and Scholes (1973) and Merton (1973). A from the point of view of distributional assumptions more realistic modeling was achieved by replacing Brownian motion by jump-type Lévy processes like hyperbolic Lévy motions, see Eberlein and Keller (1995), Eberlein and Prause (2002) and Eberlein (2001). Similar results were obtained by using the class of Variance Gamma Lévy processes as seen in Madan and Seneta (1990), Madan and Milne (1991) and Carr et al. (2002). A virtually perfect adjustment of theoretical to real option prices across all strikes and maturities was achieved by using Sato processes (Carr et al. 2007).

In this paper, the asset price model (19.1) is enhanced by including the possibility of default. A meaningful dependence structure between the asset price and the probability of default is introduced. Since we shall use this model for valuation, the

specification is done a priori in a risk-neutral setting, i.e. we assume the underlying measure P to be risk-neutral. The economic objects to be modeled are

- the hazard rate λ as a nonnegative stochastic process with càdlàg paths, which describes the behaviour of the default time τ,
- the asset price S as a nonnegative stochastic process with càdlàg paths.

We want the asset price S to be negatively dependent on the hazard rate λ. Therefore, we use two sources of randomness

(1) a Lévy process $Z = (Z_t)_{t\geq 0}$ as driver of the hazard rate λ,
(2) an independent Lévy process $L = (L_t)_{t\geq 0}$, which represents the market noise of the asset price.

In general a Lévy process is an $\mathbb{R}^d$-valued, adapted stochastic process $X = (X_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathscr{F}, \mathbb{F} = (\mathscr{F}_t)_{t\geq 0}, P)$ which starts at zero and has independent and stationary increments. Any Lévy process is characterised by its Lévy triplet (b, c, ν_X), where $b \in \mathbb{R}^d$, c is a symmetric nonnegative $d \times d$ matrix and ν_X is a measure on $\mathbb{R}^d$, called the Lévy measure of X. The characteristic function of X_1 is given in its Lévy-Khintchine representation as follows

$$E[e^{i\langle u, X_1\rangle}] = \exp\left[i\langle u, b\rangle - \frac{1}{2}\langle u, cu\rangle + \int [e^{i\langle u,x\rangle} - 1 - i\langle u, h(x)\rangle]\nu_X(dx)\right].$$

If a random vector X has an exponential moment of order $v \in \mathbb{R}^d$, i.e. if $E[e^{\langle v, X\rangle}]$ is finite, we write $v \in \mathbb{EM}_X$ and in this case $E[e^{\langle z, X\rangle}]$ can be defined for all $z \in \mathbb{C}^d$ with $Re(z) \in \mathbb{EM}_X$. For Lévy processes X we have under the proper moment assumption that $E[e^{\langle z, X_t\rangle}] = e^{t\theta_X(z)}$, where

$$\theta_X(z) := \log E[e^{\langle z, X_1\rangle}] = \langle z, b\rangle + \frac{1}{2}\langle z, cz\rangle + \int [e^{\langle z,x\rangle} - 1 - \langle z, h(x)\rangle]\nu_X(dx)$$

is called the cumulant function of X. Since $\mathbb{EM}_{X_t}$ is independent of t for Lévy processes we use $\mathbb{EM}_X$ in this case to express that the moment condition holds for every t. The existence of exponential moments implies the finiteness of moments of arbitrary order, in particular the finiteness of the expectation. The latter entails that the truncation function h can be chosen to be the identity, i.e. $h(x) = x$. With the following lemma we are able to calculate explicitly the expectations of exponentials of stochastic integrals with respect to a Lévy process.

Lemma 1 *Let X be a Lévy process such that $[-M_X(1+\varepsilon), M_X(1+\varepsilon)]^d \subset \mathbb{EM}_X$ for constants $M_X, \varepsilon > 0$. If $f : \mathbb{R}_+ \to \mathbb{C}^d$ is a complex-valued, continuous function such that $|Re(f^i)| \leq M_X$ $(i = 1, ..., d)$, then*

$$E\left[\exp\left(\int_0^t f(s)dX_s\right)\right] = \exp\left(\int_0^t \theta_X(f(s))ds\right).$$

Proof This is a straightforward extension of Lemma 3.1. in Eberlein and Raible (1999). A proof can be found in Kluge (2005). □

In the following we shall only use one-dimensional Lévy processes.

Example 1 A very flexible and useful subclass of Lévy processes is given by the normal inverse Gaussian (NIG) processes, which are generated by the NIG distribution with the simple characteristic function

$$\varphi_{NIG}(u) = e^{iu\mu} \frac{\exp(\delta\sqrt{\alpha^2 - \beta^2})}{\exp(\delta\sqrt{\alpha^2 - (\beta + iu)^2})}$$

and the four parameters $\mu, \beta \in \mathbb{R}$, $\delta > 0$ and $\alpha > |\beta| \geq 0$.

Example 2 The Gamma process, generated by the Gamma distribution, is an increasing Lévy process. The Gamma distribution has the parameters $p, b > 0$ and the characteristic function

$$\varphi_\Gamma(u) = \left(\frac{b}{b - iu}\right)^p.$$

The default time $\tau : \Omega \to [0, \infty]$ is constructed via

$$\tau = \inf\{t \geq 0 \mid e^{-\Gamma_t} \leq \xi\}.$$

where $\Gamma_t := \int_0^t \lambda_s ds$ is the integral over the hazard rate $\lambda = (\lambda_t)_{t\geq 0}$, a nonnegative $\mathbb{F}$-adapted process with càdlàg paths and ξ is a uniformly distributed random variable on $[0, 1]$, independent of $\mathbb{F}$. This is the so-called intensity-based approach of default modelling. Details can be found in Bielecki and Rutkowski (2004). We need three properties of this construction:

1. One can easily show that

$$P(t < \tau \mid \mathcal{F}_t) = e^{-\Gamma_t}. \tag{19.2}$$

   Thus, the survival probability can be calculated to be $P(t < \tau) = E\left[e^{-\Gamma_t}\right]$.
2. If $(M_t)_{t\geq 0}$ is a nonnegative $\mathbb{F}$-martingale, then

$$\left(M_t \mathbb{1}_{\{\tau > t\}} e^{\Gamma_t}\right)_{t\geq 0}$$

   follows a $\mathbb{G}$-martingale. $\mathbb{G} = (\mathcal{G}_t)_{t\geq 0}$ is defined by $\mathcal{G}_t := \mathcal{F}_t \vee \mathcal{H}_t$, where $\mathcal{H}_t := \sigma(\{\tau \leq u \mid u \leq t\})$ is the filtration which carries the information about the default time.

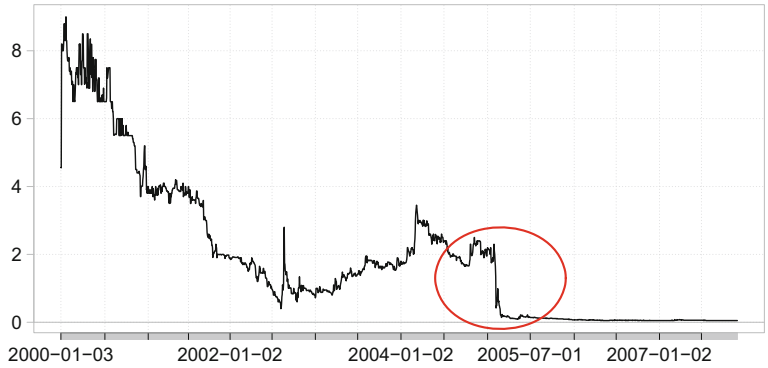


Fig. 19.1 The bankruptcy of Walter Bau

3. For the t-survival measure

$$P^t(A) := P(A \mid t < \tau),$$

which is the measure P conditioned on no default until t, one gets $P^t \ll P$ and

$$\frac{dP^t|_{\mathscr{F}_t}}{dP|_{\mathscr{F}_t}} = \frac{e^{-\Gamma_t}}{E[e^{-\Gamma_t}]}. \tag{19.3}$$

Now we are ready to specify the asset price model in the form

$$S_t = S_0 \exp\Big[rt + L_t - \zeta Z_t + \omega t + \Gamma_t\Big]\mathbb{1}_{\{t<\tau\}} \tag{19.4}$$

with a constant r, representing the continuously compounded interest rate. Default is modeled by a single jump to zero at time point τ. This reflects the idea of no recovery for shareholders. This assumption seems to be reasonable if we look at the history of bankruptcies. As an example, the time series of stock prices showing the bankruptcy of the former German company Walter Bau is represented in Fig. 19.1. Effectively, the default event, marked by the ellipse, is a jump to zero. In the sequel, this model will be denoted the **Defaultable Asset Price Model** (**DAM**).

The term $-\zeta Z_t$ models the dependency between credit risk and asset price with an additional parameter $\zeta \geq 0$. A surge of the default probability leads to a decline of the asset price. A generalisation to a more complex functional dependence structure $-f(Z_t)$ is possible and in line with the pricing methods below. The simple form $-\zeta Z_t$ was chosen for convenience.

Since we want $(S_t)_{t\geq 0}$ to be a martingale after discounting, the reason for the term $\omega t + \Gamma_t$ is a mathematical one. Using the well-known fact that $e^{X_t}/E[e^{X_t}]$ is a martingale for a process X with independent increments, we can choose the constant ω such that $\exp[L_t - \zeta Z_t + \omega t]$ is an $\mathbb{F}$-martingale:

$$\omega = -\log E[e^{L_1}] - \log E[e^{-\zeta Z_1}] = -\theta_L(1) - \theta_Z(-\zeta).$$

Thus, as indicated before, the discounted price process

$$e^{-rt}S_t = S_0 \exp\left[L_t - \zeta Z_t + \omega t\right] \cdot e^{\Gamma_t} \mathbb{1}_{\{t<\tau\}}$$

is a $\mathbb{G}$-martingale. This ensures that the considered financial market model is arbitrage-free, cf. Delbaen and Schachermayer (2006).

For the existence of $\omega \in \mathbb{R}$, we need the conditions

(i) $1 \in \mathbb{EM}_L$.
(ii) $-\zeta \in \mathbb{EM}_Z$.

A similar type of model for pricing convertible bonds was introduced by Davis and Lischka (2002). Their model, driven by a Brownian Motion $(W_t)_{t\geq 0}$ with volatility σ, is

$$S_t = S_0 \exp\left[rt + \sigma W_t - \frac{1}{2}\sigma^2 t + \int_0^t \lambda_s ds\right] \mathbb{1}_{\{t<\tau\}},$$

where $(\lambda_s)_{s\geq 0}$ is the hazard rate corresponding to the default time τ. This model approach was enhanced by Andersen and Buffum (2004), Linetsky (2006) and Carr and Madan (2010). Their idea of getting a reasonable dependence structure between credit risk and asset price was a different one. They choose the hazard rate as a function of the asset price, for example

$$\lambda_s = \lambda(S_s) = \alpha S_s^{-p},$$

which leads to a stochastic integral equation. Our approach, which is also an enhancement of this model, avoids this. Thus, we get a more direct analytical access.

As a model for the hazard rate $(\lambda_t)_{t\geq 0}$, we choose a positive Ornstein-Uhlenbeck (OU) process driven by an increasing Lévy process $(Z_t)_{t\geq 0}$ which is assumed to be independent of L

$$d\lambda_t = \kappa(\mu - \lambda_t)dt + dZ_t. \qquad (\kappa, \mu \geq 0). \tag{19.5}$$

This kind of processes moves up by the jumps of Z and then declines exponentially as if there is a restoring force measured by the parameter κ, see Fig. 19.2. One main advantage is the analytical tractability, see for example Barndorff-Nielsen and Shephard (2001) or Cont and Tankov (2004), where OU processes are used as stochastic volatility models for financial assets. Schoutens and Cariboni (2009) investigated OU processes already as hazard rate models.

The upward jumps can be interpreted as bad news about the firm, like a profit alert, an essential loss of capital or a failed project. Other reasons could be major events or even catastrophes with consequences for a whole industrial sector or the global economy. Examples are the burst of the Dot-com bubble in 2000, the terror attacks of 9/11, the collapse of Lehman Brothers in 2008 or the Fukushima disaster in 2011. Hazard rates are not directly observable, but CDS quotes also reflect the default

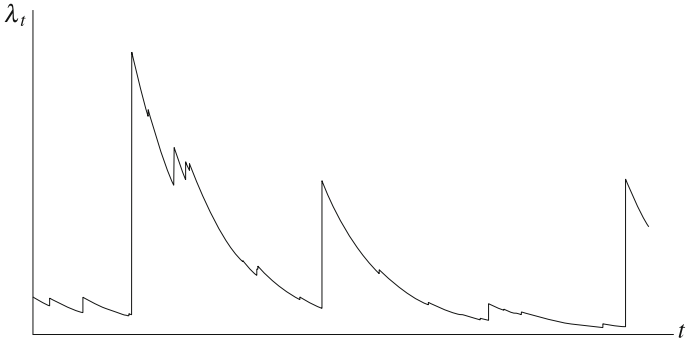


Fig. 19.2 OU process driven by a Γ process

probability. Hence, the time evolution of hazard rates and short time CDS quotes should look quite similar. We take the one-year CDS quotes of the German energy company E.ON SE as an example, see Fig. 19.3. There are two big jumps, one after the collapse of Lehman Brothers (left line) and one when the German government resolved the nuclear phase-out a few months after the Fukushima disaster (middle line). We can conclude that the model approach (19.5) looks quite reasonable in view of this example. The relation between the upward jumps of the CDS quotes and the downward movement of the stock price is clearly visible.

The explicit expression for (19.5) is

$$\lambda_t = \lambda_0 e^{-\kappa t} + \mu(1 - e^{-\kappa t}) + \int_0^t e^{\kappa(s-t)} dZ_s. \tag{19.6}$$

Using Fubini's Theorem for stochastic integrals, cf. Theorem 64 in Chapter IV of Protter (2005), we get for the hazard process

$$\Gamma_t = \Gamma_t^d + \int_0^t \gamma_s^t dZ_s \tag{19.7}$$

where we used the abbreviations

$$\Gamma_t^d := \frac{\lambda_0}{\kappa}(1 - e^{-\kappa t}) + \mu\left(t + \frac{e^{-\kappa t}}{\kappa} - \frac{1}{\kappa}\right)$$
$$\gamma_s^t := \frac{1 - e^{-\kappa(t-s)}}{\kappa}.$$

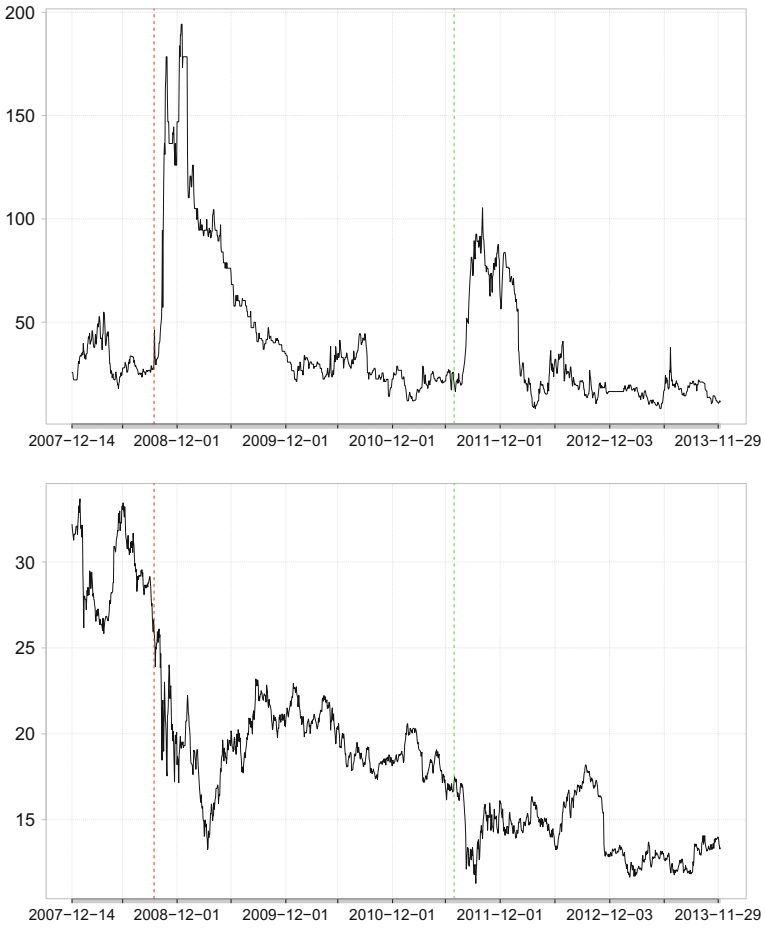


Fig. 19.3 One-year CDS quotes (*top*) and stock price (*bottom*) of the German energy company E.ON SE. The *left* line marks the collapse of Lehman Brothers, the *middle* line the German nuclear phase-out after the Fukushima disaster

For the numerical calculation of the survival probability $P(t < \tau) = E[e^{-\Gamma_t}]$, we can now use Lemma 1

$$\begin{aligned} E[e^{-\Gamma_t}] &= e^{-\Gamma_t^d} E\left[\exp\left(-\int_0^t \gamma_u^t dZ_u\right)\right] \\ &= e^{-\Gamma_t^d} \exp\left(\int_0^t \theta_Z(-\gamma_u^t) du\right), \end{aligned} \tag{19.8}$$

where θ_Z is the cumulant function of Z. To obtain (19.8), we need the assumptions

(iii) There are constants $M_Z, \varepsilon > 0$ such that $\pm M_Z(1+\varepsilon) \in \mathbb{EM}_Z$.
(iv) κ satisfies $\frac{1}{\kappa} \leq M_Z$.

This kind of model cannot be adjusted to an exogenously given survival function $t \mapsto P(t < \tau) = E[e^{-\Gamma_t}]$. The survival function can be recovered from CDS quotes using the methods described in Madan et al. (2004).

The same problem is known from short rate models for the term structure of interest rates (for an overview see the book of Brigo and Mercurio (2001)). The famous Vasicek (1977) model is not able to incorporate the current yield curve. Hull and White (1990) overcame this drawback by making one parameter in the Vasicek model time-dependent. The same idea could be used to extend (19.5) in the following way

$$d\lambda_t = \kappa(\mu(t) - \lambda_t)dt + dZ_t.$$

19.3 Option Pricing

In this section, we price some European options under the Defaultable Asset Price Model. We define the $\mathbb{F}$-adapted semimartingale

$$X_t := \log S_0 + rt + L_t - \zeta Z_t + \omega t + \Gamma_t$$

such that $S_t = e^{X_t}\mathbb{1}_{\{t<\tau\}}$ and use the Fourier-based valuation method as given in Eberlein et al. (2010). This leads to the equation

$$E_Q[f(X_T)] = \frac{1}{2\pi}\int \varphi^Q_{X_T}(u - iR)\widehat{f}(iR - u)du, \tag{19.9}$$

where $\widehat{f}$ denotes the Fourier transform of f, which is defined by $\widehat{f}(u) = \int e^{iux} f(x)dx$ and where $\varphi^Q_{X_T}$ denotes the extended characteristic function of X_T under the probability measure Q. $R \in \mathbb{R}$ is a constant that must satisfy

(C1) $g \in L^1_{bc}(\mathbb{R}) = \{h \in L^1(\mathbb{R}) \mid h \text{ bounded and continuous}\}$,
(C2) $R \in \mathbb{EM}_{X_T}$,
(C3) $\widehat{g} \in L^1(\mathbb{R})$,

where $g(x) := e^{-Rx} f(x)$. The key point of (19.9) is the separation of the function f from the distribution Q_{X_T} of X_T.

In order to use the Fourier-based method within the Defaultable Asset Price Model one has to separate the indicator $\mathbb{1}_{\{t<\tau\}}$ from the payoff function. This means that we only consider payoff functions f which can be written as

$$f(S_T) = f(\mathbb{1}_{\{T<\tau\}}e^{X_T}) = \mathbb{1}_{\{T\geq\tau\}} f_1(X_T) + \mathbb{1}_{\{T<\tau\}} f_2(X_T) \tag{19.10}$$

for functions f_1 and f_2, that satisfy the assumptions for the valuation formula (19.9).

Lemma 2 *Let f be a payoff function of an option with maturity $T > 0$ which satisfies (19.10). Then the following formula holds*

$$E[f(S_T)] = E[f_1(X_T)] - E\left[e^{-\Gamma_T}\right] E_T[f_1(X_T)] + E\left[e^{-\Gamma_T}\right] E_T[f_2(X_T)] \tag{19.11}$$

where $E_T := E_{P^T}$ is the expectation under the survival measure P^T.

Proof For this calculation, we use the change-of-numeraire technique with the survival measure P^T

$$\begin{aligned}
E[f(S_T)] &\overset{(19.10)}{=} E\left[\mathbb{1}_{\{T\geq\tau\}} f_1(X_T)\right] + E\left[\mathbb{1}_{\{T<\tau\}} f_2(X_T)\right] \\
&= E\left[(1-\mathbb{1}_{\{T<\tau\}}) f_1(X_T)\right] + E\left[\mathbb{1}_{\{T<\tau\}} f_2(X_T)\right] \\
&= E\left[f_1(X_T) E\left[(1-\mathbb{1}_{\{T<\tau\}}) \mid \mathscr{F}_T\right]\right] + E\left[f_2(X_T) E\left[\mathbb{1}_{\{T<\tau\}} \mid \mathscr{F}_T\right]\right] \\
&\overset{(19.2)}{=} E[f_1(X_T)] - E\left[e^{-\Gamma_T} f_1(X_T)\right] + E\left[e^{-\Gamma_T} f_2(X_T)\right] \\
&\overset{(19.3)}{=} E[f_1(X_T)] - E\left[e^{-\Gamma_T}\right] E_T[f_1(X_T)] + E\left[e^{-\Gamma_T}\right] E_T[f_2(X_T)].
\end{aligned}$$

□

The elements on the right side of (19.11) can be calculated numerically. $E[e^{-\Gamma_T}]$ can be calculated by using Lemma 1. For the calculation of the expectations $E_T[f(X_T)]$ under the survival measure P^T for different functions f, we use (19.9). We shall calculate the extended characteristic function $\varphi_{X_T}^{P^T}$ of X_T under the survival measure P^T. We begin with a generic lemma of stochastic analysis.

Lemma 3 *Let X and Y be two independent semimartingales and H be a deterministic process with left-continuous paths. Then the processes X and $(\int_0^t H_s dY_s)_{t\geq 0}$ are independent as well.*

Proof Fix $t \geq 0$ and define

$$H_t^n := \mathbb{1}_{\{0\}} H_0 + \sum_{k=1}^{2^n} \mathbb{1}_{](k-1)\frac{t}{2^n}, k\frac{t}{2^n}]} H_{k\frac{t}{2^n}}.$$

For each $n \geq 1$ and each $t' \geq 0$, $X_{t'}$ is independent from

$$\int_0^t H_s^n dY_s = \sum_{k=1}^{2^n} H_{k\frac{t}{2^n}} (Y_{k\frac{t}{2^n}} - Y_{(k-1)\frac{t}{2^n}}).$$

$\int_0^t H_s^n dY_s$ is a Riemann approximation for the stochastic integral $\int_0^t H_s dY_s$, i.e.

$$\int_0^t H_s^n dY_s \to \int_0^t H_s dY_s$$

in probability, see Proposition I.4.44 in Jacod and Shiryaev (2003). Independence is transferred to the stochastic limit, cf. Proposition 1.13 in Sato (1999), and thus the assertion follows. □

Lemma 4 *Let $R > 1$ ($R < 0$ resp.) such that*

(v) $R \in \mathbb{EM}_L$, *i.e.* $E[e^{RL_T}]$ *exists for all* $T \geq 0$,
(vi) $\max\{\zeta R,\ \frac{R-1}{\kappa} - \zeta R\} \leq M_Z$ ($\max\{-\zeta R,\ \zeta R - \frac{R-1}{\kappa}\} \leq M_Z$ *resp.).*

Then $M^T_{X_T}(R) = E_T[e^{RX_T}]$ *exists, i.e. assumption (C2) of* (19.9) *is satisfied.*

Proof Using Lemma 3, we obtain

$$\begin{aligned} M^T_{X_T}(R) &= E_T[\exp(RX_T)] \\ &= \text{const.} \cdot E_T[\exp(RL_T)\exp(-\zeta R Z_T + R\Gamma_T)] \\ &\overset{(19.3)}{=} \text{const.} \cdot E[\exp(RL_T)\exp(-\zeta R Z_T + (R-1)\Gamma_T)] \\ &= \text{const.} \cdot M_{L_T}(R) \cdot E\left[\exp\left(\int_0^T (R-1)\gamma^T_s - \zeta R\, dZ_s\right)\right]. \end{aligned}$$

(vi) implies $|(R-1)\gamma^T_s - \zeta R| \leq M_Z$, and thus the existence of the last factor. □

To use (19.9), we need to calculate the extended characteristic function $\varphi^{P^T}_{X_T}$ of X_T under P^T. We abbreviate

$$\begin{aligned} d_t &:= \ln S_0 + rt + \omega t \\ D_t(x) &:= \frac{\exp[x(d_t + \Gamma^d_t) - \Gamma^d_t]}{E[e^{-\Gamma_t}]}, \end{aligned}$$

and obtain for all $x \in \mathbb{C}$ with $\mathrm{Re}(x) = R$

$$\begin{aligned} E_T[e^{xX_T}] &= e^{xd_T} E_T[e^{x(L_T - \zeta Z_T + \Gamma_T)}] \\ &= e^{xd_T} E\left[\frac{e^{-\Gamma_T}}{E[e^{-\Gamma_T}]} e^{x(L_T - \zeta Z_T + \Gamma_T)}\right] \\ &= D_T(x) E\left[e^{xL_T} e^{\int_0^T x\gamma^T_s - x\zeta - \gamma^T_s dZ_s}\right] \\ &\overset{(\star)}{=} D_T(x) E\left[e^{xL_T}\right] E\left[e^{\int_0^T x\gamma^T_s - x\zeta - \gamma^T_s dZ_s}\right] \\ &= D_T(x) \exp\left[T \cdot \theta_L(x)\right] \exp\left[\int_0^T \theta_Z(x\gamma^T_s - x\zeta - \gamma^T_s) ds\right], \end{aligned}$$

where we have used Lemma 3 in equation ($\star$). In the last step of this calculation, we used Lemma 1. The requirement

$$\left|\mathrm{Re}(x\gamma_s^t - x\zeta - \gamma_s^t)\right| \leq M_Z$$

is satisfied by the assumptions of Lemma 4. Hence, we have for all $u \in \mathbb{R}$ and suitable $R \in \mathbb{R}$

$$\begin{aligned}\varphi_{X_T}^{P^T}(u - iR) &= E_T[e^{(R+iu)X_T}] \\ &= D_T(R+iu)\exp[T\cdot\theta_L(R+iu)]\exp\left[\int_0^T \theta_Z((R+iu)\gamma_s^T - (R+iu)\zeta - \gamma_s^T)ds\right].\end{aligned} \tag{19.12}$$

Example 3 In the case of a call option, we have $f(x) = (e^x - K)^+$, i.e.

$$\widehat{f}(z) = \frac{K^{1+iz}}{iz(1+iz)}, \qquad \mathrm{Im}(z) \in (1, \infty).$$

Conditions (C1) and (C3) are fulfilled for $R > 1$. The payoff function is of type (19.10) with $f_1 \equiv 0$ and $f_2(x) = (e^x - K)^+$. For the put option, where $f(x) = (K - e^x)^+$, we have

$$\widehat{f}(z) = \frac{K^{1+iz}}{iz(1+iz)}, \qquad \mathrm{Im}(z) \in (-\infty, 0).$$

Conditions (C1) and (C3) are fulfilled for $R < 0$. We have $f_1 \equiv K$ and $f_2(x) = (K - e^x)^+$. By using (19.11), we obtain the call prices

$$C_0(T, K) = e^{-rT}E[e^{-\Gamma_T}]E_T[(e^{X_T} - K)^+] \tag{19.13}$$

and the put prices

$$P_0(T, K) = e^{-rT}\left[E[e^{-\Gamma_T}]E_T[(K - e^{X_T})^+] + K(1 - E[e^{-\Gamma_T}])\right]. \tag{19.14}$$

Example 4 The payoff function of a digital call option with barrier $B > 0$ and maturity $T > 0$ is $f(x) = \mathbb{1}_{\{x>B\}}$, i.e. it is of type (19.10) with $f_1 \equiv 0$ and $f_2 = \mathbb{1}_{\{e^x>B\}}$. We use (19.11) and obtain

$$E\left[e^{-rT}\mathbb{1}_{\{S_T>B\}}\right] = e^{-rT}E[e^{-\Gamma_T}]E_T\left[\mathbb{1}_{\{e^{X_T}>B\}}\right].$$

The Fourier transform of f_2 is

$$\widehat{f_2}(z) = -\frac{B^{iz}}{iz} \quad \text{for} \quad \mathrm{Im}(z) > 0.$$

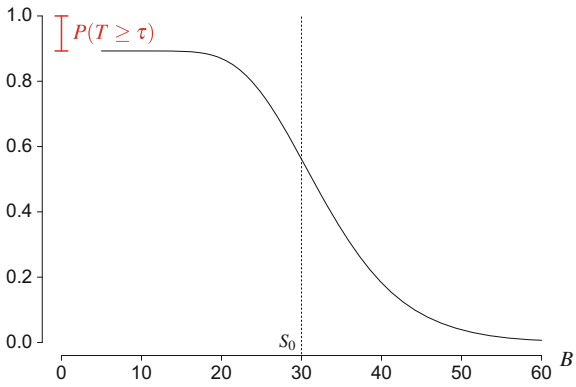


Fig. 19.4 Prices of digital call options with barrier B

The assumptions for applying (19.9) are satisfied for $R > 0$, cf. Eberlein et al. (2010). For the digital put option, we have

$$E\left[e^{-rT}\mathbb{1}_{\{S_T<B\}}\right] \overset{(19.11)}{=} e^{-rT}\left(1 - E[e^{-\Gamma_T}] + E[e^{-\Gamma_T}]E_T\left[\mathbb{1}_{\{e^{X_T}<B\}}\right]\right).$$

The Fourier transform of $f_2(x) = \mathbb{1}_{\{e^x<B\}}$ is

$$\widehat{f_2}(z) = \frac{B^{iz}}{iz} \quad \text{for} \quad \mathrm{Im}(z) < 0.$$

In this case, we need $R < 0$. To give a numerical example, we take $S_0 = 30$, $T = 260$ and the parameters

$$\begin{array}{lll} \alpha = 50.0 & \beta = -0.1 & \delta = 0.012 \\ p = 0.0035 & b = 66 & \kappa = 0.11 \\ & \zeta = 9.0 & \end{array} \qquad (*)$$

which correspond to a one-year default probability of about 10.7 %. The results can be seen in Fig. 19.4. The main difference to a non-defaultable model is that the prices tend to $1 - P(T \geq \tau)$ for $B \searrow 0$ and not to 1.

Example 5 The payoff of a self-quanto call option with strike $K > 0$ is $e^x(e^x - K)^+$, i.e. we have

$$e^{-rT}E\left[\mathbb{1}_{\{T<\tau\}}e^{X_T}(e^{X_T} - K)^+\right] = e^{-rT}E[e^{-\Gamma_T}]E_T\left[e^{X_T}(e^{X_T} - K)^+\right].$$

The Fourier transform of $f_2(x) = e^x(e^x - K)^+$ is

$$\widehat{f_2}(z) = \frac{K^{2+iz}}{(1+iz)(2+iz)} \quad \text{for} \quad \mathrm{Im}(z) > 2.$$

For a self-quanto put option with payoff $e^x(K - e^x)^+$ we have

$$e^{-rT} E\left[\mathbb{1}_{\{T<\tau\}} e^{X_T}(K - e^{X_T})^+\right] = e^{-rT} E[e^{-\Gamma_T}] E_T\left[e^{X_T}(K - e^{X_T})^+\right].$$

The Fourier transform of $f_2(x) = e^x(K - e^x)^+$ is the same as above, but for $\mathrm{Im}(z) < 1$.

For calculating expectations $E[f(S_T)]$, we can also use Monte Carlo simulations, i.e. we can simulate the random variable S_T for example N times and approximate $E[f(S_T)]$ by $\frac{1}{N}\sum_{i=1}^{N} f(s_T^i)$, where $(s_T^i)_{i=1,\ldots,N}$ denotes a simulated sample of S_T. For the pathwise simulation of the Defaultable Asset Price Model

$$S_t = S_0 \exp\left[rt - qt + L_t - \zeta Z_t + \omega t + \Gamma_t\right]\mathbb{1}_{\{t<\tau\}},$$

we have to be able to simulate the Lévy processes L_t and Z_t pathwise. This means, that it is necessary to simulate whole paths $(S_t)_{0\le t\le T}$ if we want to create a simulation for S_T. If we have to do that already, with only little additional effort one can price path-dependent options or options with different maturities $T_k \le T$ $(k = 1, \ldots, n)$ simultaneously.

Example 6 An Asian option is a derivative, whose payoff depends on the average price

$$\overline{S}_T := \frac{1}{T}\int_0^T S_t dt$$

of the underlying price process $(S_t)_{0\le t\le T}$. We simulate the price path on an equidistant time grid $0 = t_0 < t_1 < \ldots < t_n = T$. The simulated value $\overline{s}_T^i$ of the average price is then given as the mean

$$\overline{s}_T^i = \frac{1}{n}\sum_{k=0}^{n} s_{t_k}^i$$

of the simulated prices $(s_{t_k}^i)_{k=0,\ldots,n}$ for each simulation $i \in \{1, \ldots, N\}$. Figure 19.5 shows an example.

19.4 Calibration

Calibration is conducted by minimising the sum of the squared differences between observed market prices and model prices

$$\mathrm{SD}(\alpha) := \sum_j \left(\pi_j^{\mathrm{Model}}(\alpha) - \pi_j^{\mathrm{Market}}\right)^2$$

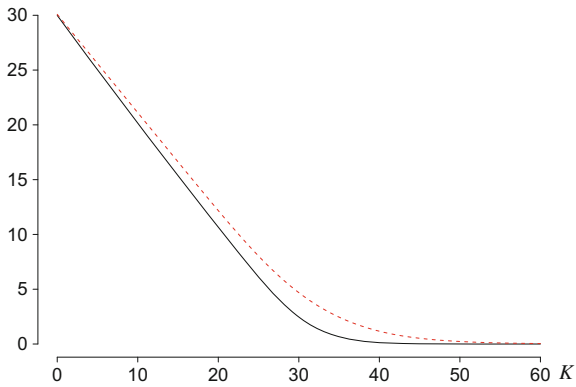


Fig. 19.5 Prices of average price calls with payoff $(\overline{S}_T - K)^+$ (*solid line*). For comparison, prices of ordinary calls (*dashed line*)

over the model parameters $\alpha = (\alpha_1, ..., \alpha_n)$ in a parameter space $A_1 \times ... \times A_n \subset \mathbb{R}^n$. This space is given by constraints on the mathematical model. In our case, we have to consider the parameter spaces of the processes L and Z and have to check the conditions (i)–(vi).

We choose a NIG$(\alpha, \beta, \delta, \mu)$ process for L and a $\Gamma(p, b)$ process for Z as an example. This leads to a model with the seven parameters

$\alpha > 0, \quad \beta \in (-\alpha, \alpha), \quad \delta > 0$	for the NIG process
$p, \quad b > 0$	for the Γ process
$\kappa \geq 0$	for the OU restoring force
$\zeta \geq 0$	as dependence parameter.

We note here that the drift parameter μ of the NIG process is redundant. The reason is the martingale setting. If L_1 is NIG-distributed, then $L_1 - \log E[e^{L_1}]$ is also NIG-distributed, but independent of μ.

The model assumptions (i)–(vi) can be reduced to restrictions on the process parameters. For the NIG process L, we have $\mathbb{EM}_L = (-\alpha - \beta, \alpha - \beta)$ and for the Γ process Z, we get $\mathbb{EM}_Z = (-\infty, b)$. Consequently we can convert the conditions to

(i) $1 < \alpha - \beta$
(ii) $-\zeta < b$
(iii) is always satisfied
(iv) $\frac{1}{\kappa} < b$
(v) $1 < R < \alpha - \beta$
(vi) $\max\{\zeta R, \ \frac{R-1}{\kappa} - \zeta R\} < b \quad (\max\{-\zeta R, \ \zeta R - \frac{R-1}{\kappa}\} < b$ resp.),

which can all be checked easily.

We calibrate all parameters, i.e. the parameters for L, the credit parameters and the dependence parameter ζ, to the option price surface. Hence, we obtain the required risk-neutral parameters of the model which are needed to price other financial products based on this asset. Accordingly, we can extract credit risk information about the

firm from option quotes. This enables us to calculate default probabilities. Alternatively, one could calibrate the credit parameters to the CDS term structure, fix them and calibrate the remaining ones using option prices.

We consider the stocks of the European banks BNP Paribas, Commerzbank, Credit Agricole, Credit Suisse, Deutsche Bank, UBS and UniCredit and look at the corresponding call prices on March 20, 2014. We restrict ourselves to calls with expiration date T_1 in December 2014 and T_2 in December 2015. As a riskless interest rate, we take the EONIA rate. The current stock prices are dividend-adjusted via

$$S_0 \quad \rightsquigarrow \quad S_0 - e^{-rT_D} \cdot D,$$

where we take the estimated or promised dividend payment of each bank for D and the day following the annual general assembly for T_D. The results of the calibrations can be found in Tables 19.1 and 19.2.

In Fig. 19.6, we observe a virtually perfect fit of the DAM to the real market data of BNP Paribas.

Table 19.1 Calibration results 1

	BNP Paribas		Commerzbank		Credit Agricole		Credit Suisse	
	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
α	53.0	52.6	50.3	49.9	45.2	46.1	45.8	44.0
β	−0.09	−0.05	−0.23	−0.17	−0.10	0.03	−0.08	−0.1
δ	0.0087	0.0091	0.0229	0.0213	0.0088	0.0095	0.0056	0.0060
p	0.00218	0.00182	0.00134	0.00122	0.004	0.00366	0.00312	0.00244
b	51	81	91	101	90	119	78	112
κ	0.162	0.402	0.47	0.402	0.16	0.234	0.18	0.25
ζ	5.0	5.0	5.5	5.5	4.6	5.1	4.0	3.0

Table 19.2 Calibration results 2

	Deutsche Bank		UBS		UniCredit	
	T_1	T_2	T_1	T_2	T_1	T_2
α	61.3	60.4	69.0	69.1	45.0	45.0
β	−0.95	−1.1	−0.5	−0.8	−3.2	−3.2
δ	0.0109	0.0106	0.0120	0.0110	0.013	0.013
p	0.00314	0.00276	0.0028	0.0025	0.0022	0.0020
b	87	126	142	144	154	146
κ	0.182	0.26	0.28	0.27	0.16	0.18
ζ	3.5	3.8	3.0	3.5	6.0	5.8

19.5 A Differential Equation for the Option Pricing Function

In the former sections, the calculation of the desired expectation $E[f(S_T)]$ is accomplished by combining the change of measure with the Fourier-based valuation method. Now we shall investigate another common method, namely pricing by solving a partial integro-differential equation (PIDE). The procedure is the following. Write the martingale $E[f(S_T) \mid \mathscr{F}_t]$ as a C^2-function g of t and some underlying process $V_t = (V_t^1, ..., V_t^d)$

$$E[f(S_T) \mid \mathscr{F}_t] = g(V_t, t). \tag{19.15}$$

We assume that the processes V^i are special semimartingales, i.e. they possess a (unique) decomposition $V^i = V_0 + M^i + A^i$ with a local martingale M^i and a predictable process A^i with paths of bounded variation. By applying Itô's formula we obtain

$$\begin{aligned} g(V_t, t) = g(V_0, 0) &+ \sum_{i \le d} \int_0^t \partial_i g(V_{s-}, s) dV_s^i + \int_0^t \partial_{d+1} g(V_{s-}, s) ds \\ &+ \frac{1}{2} \sum_{i,j \le d} \int_0^t \partial_{ij} g(V_{s-}, s) d\langle (V^i)^c, (V^j)^c \rangle_s \\ &+ \sum_{s \le t} \left[g(V_s, s) - g(V_{s-}, s) - \sum_{i \le d} \partial_i g(V_{s-}, s) \Delta V_s^i \right]. \end{aligned} \tag{19.16}$$

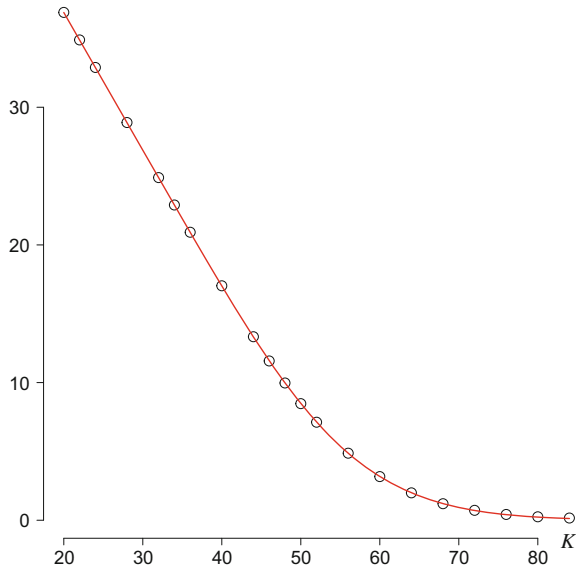


Fig. 19.6 Quoted call prices of BNP Paribas (*circles*) and the model prices (*line*) after the calibration

$g(V_t, t)$ is a special semimartingale, but also a martingale by (19.15). Consequently, any decomposition

$$g(V_t, t) = g(V_0, 0) + M_t + A_t$$

with a local martingale M and a predictable process A with paths of bounded variation has to satisfy $A \equiv 0$. Expanding and sorting the the right-hand side of (19.16) in this sense leads to the desired PIDE

$$\begin{aligned} 0 = &\sum_{i \le d} \int_0^t \partial_i g(V_{s-}, s) dA_s^i + \int_0^t \partial_{d+1} g(V_{s-}, s) ds \\ &+ \frac{1}{2} \sum_{i,j \le d} \int_0^t \partial_{ij} g(V_{s-}, s) d\langle (V^i)^c, (V^j)^c \rangle_s \\ &+ \int_{[0,t] \times \mathbb{R}^d} \left[g(V_{s-} + x, s) - g(V_{s-}, s) - \sum_{i \le d} \partial_i g(V_{s-}, s) x \right] (\mu^V)^p (ds, dx), \end{aligned} \tag{19.17}$$

where $(\mu^V)^p$ is the predictable compensator of the jump measure μ^V of V, cf. Theorem II.1.8 in Jacod and Shiryaev (2003). The boundary condition is set at the maturity date T of the contingent claim

$$g(x^1, ..., x^d, T) = f(l(x^1, ..., x^d)),$$

where l is the function, such that $S_T = l(V_T^1, ..., V_T^d)$. Solving the PIDE (numerically) on $\mathbb{R}^d \times [0, T]$ gives us the desired value

$$E[f(S_T)] = g(V_0, 0).$$

The boundary condition determines the solution $g(x, t)$ at the end of the considered time interval $[0, T]$, but the value we are looking for is the one at the beginning.

In order to apply this approach to the Defaultable Asset Price Model

$$S_t = \exp\Big[\log S_0 + rt + L_t - \zeta Z_t + \omega t + \Gamma_t \Big] \mathbb{1}_{\{t<\tau\}} = e^{X_t} \mathbb{1}_{\{t<\tau\}},$$

we firstly have to take care of the indicator function $\mathbb{1}_{\{t<\tau\}}$. Therefore, we shall only consider payoff functions f of type (19.10), i.e. we assume that

$$f(S_T) = f(\mathbb{1}_{\{T<\tau\}} e^{X_T}) = \mathbb{1}_{\{T \ge \tau\}} f_1(X_T) + \mathbb{1}_{\{T<\tau\}} f_2(X_T)$$

for functions f_1 and f_2. As seen before, most of the common payoff functions have this form. In this case, we can eliminate the indicator function $\mathbb{1}_{\{t<\tau\}}$ in the time-0 pricing formula

$$\begin{aligned}
\pi_0 &= e^{-rT} E[f(S_T)] \overset{(19.10)}{=} e^{-rT} E[\mathbb{1}_{\{T\geq\tau\}} f_1(X_T) + \mathbb{1}_{\{T<\tau\}} f_2(X_T)] \\
&= e^{-rT}\{E[E[\mathbb{1}_{\{T\geq\tau\}} f_1(X_T) \mid \mathscr{F}_T]] + E[E[\mathbb{1}_{\{T<\tau\}} f_2(X_T) \mid \mathscr{F}_T]]\} \\
&= e^{-rT}\{E[f_1(X_T) E[\mathbb{1}_{\{T\geq\tau\}} \mid \mathscr{F}_T]] + E[f_2(X_T) E[\mathbb{1}_{\{T<\tau\}} \mid \mathscr{F}_T]]\} \\
&= e^{-rT}\{E[f_1(X_T)(1 - e^{-\Gamma_T})] + E[f_2(X_T) e^{-\Gamma_T}]\} \\
&= e^{-rT} E[f_1(X_T)(1 - e^{-\Gamma_T}) + f_2(X_T) e^{-\Gamma_T}] =: e^{-rT} E[\tilde{f}(X_T, \Gamma_T)].
\end{aligned}$$

In the next step, we write the martingale $E[\tilde{f}(X_T, \Gamma_T) \mid \mathscr{F}_t]$ as a function of the processes

$$V_t^1 := L_t, \quad V_t^2 := Z_t, \quad V_t^3 := Y_t := \int_0^t e^{\kappa s} dZ_s \quad \text{and} \quad t.$$

We remark here that $e^{-r(T-t)} E[\tilde{f}(X_T, \Gamma_T) \mid \mathscr{F}_t]$ does not represent the option price at time t. It is only an auxiliary function that is needed for the calculation of π_0. The correct option price at time t would be given by $e^{-r(T-t)} E[\tilde{f}(X_T, \Gamma_T) \mid \mathscr{G}_t]$.

Lemma 5 *Let $(X_t)_{t\geq 0}$ be a semimartingale with independent increments and let $f : [0, \infty) \to \mathbb{R}$ be a locally bounded, deterministic and left-continous function. Then the semimartingale $(Y_t)_{t\geq 0}$ defined by*

$$Y_t := \int_0^t f(s) dX_s$$

has independent increments as well.

Proof Due to Theorem II.4.15 in Jacod and Shiryaev (2003), there is a version of the characteristics of X, which is deterministic. The characteristics of Y can be calculated by only using the characteristics of X and the function f, see Proposition IX.5.3 in Jacod and Shiryaev (2003). Consequently, there is a version of the characteristics of Y, which is deterministic. So Theorem II.4.15 gives us the intended result. □

Lemma 6 *The conditional expectation $E[\tilde{f}(X_T, \Gamma_T) \mid \mathscr{F}_t]$ is a function of L_t, Z_t, Y_t and t*

$$E[\tilde{f}(X_T, \Gamma_T) \mid \mathscr{F}_t] = g(L_t, Z_t, Y_t, t). \tag{19.18}$$

Proof First of all, we note that Γ_t is a function of Z_t, Y_t and t

$$\Gamma_t = \Gamma_t^d + \int_0^t \frac{1 - e^{-\kappa(t-s)}}{\kappa} dZ_s = \Gamma_t^d + \frac{1}{\kappa}\Big[Z_t - e^{-\kappa t} \int_0^t e^{\kappa s} dZ_s\Big],$$

and that $\Gamma_T - \Gamma_t$ is a function of $Z_T - Z_t$, $Y_T - Y_t$, Y_t and t

$$\begin{aligned}\Gamma_T - \Gamma_t &= \Gamma_T^d - \Gamma_t^d + \frac{1}{\kappa}\big[Z_T - Z_t - e^{-\kappa T} Y_T + e^{-\kappa t} Y_t\big] \\ &= \Gamma_T^d - \Gamma_t^d + \frac{1}{\kappa}\big[Z_T - Z_t - (e^{-\kappa T} - e^{-\kappa t}) Y_t - e^{-\kappa T}(Y_T - Y_t)\big].\end{aligned}$$

Consequently,

$$\begin{aligned}X_T =& \log S_0 + rT + \omega T + L_T - \zeta Z_T + \Gamma_T \\ =& \log S_0 + rT + \omega T + L_T - L_t + L_t - \zeta(Z_T - Z_t + Z_t) \\ &+ \Gamma_T - \Gamma_t + \Gamma_t\end{aligned}$$

is a function of

(a) the increments $L_T - L_t$, $Z_T - Z_t$, $Y_T - Y_t$,
(b) the random variables L_t, Z_t, Y_t and t.

L and Z are Lévy processes, and so Lemma 5 shows that all increment terms under (a) are independent of $\mathscr{F}_t$. The terms under (b) are $\mathscr{F}_t$-measurable. Hence, we get the intended result

$$\begin{aligned}E[\tilde{f}(X_T, \Gamma_T) \mid \mathscr{F}_t] &= E[\widehat{f}(L_T - L_t, Z_T - Z_t, Y_T - Y_t, L_t, Z_t, Y_t, t) \mid F_t] \\ &= E[\widehat{f}(L_T - L_t, Z_T - Z_t, Y_T - Y_t, x, y, z, t)]\big|_{x=L_t, y=Z_t, z=Y_t}.\end{aligned}$$

□

Theorem 1 *Assume that the function $g(x, y, z, t)$, defined in (19.18), is of class $C^2(\mathbb{R}^4)$ and that L_1 and Z_1 have a finite first moment. Then g satisfies the following integro-differential equation*

$$\begin{aligned}0 =& E[L_1]\partial_1 g + E[Z_1]\partial_2 g + E[Z_1]e^{\kappa t}\partial_3 g + \partial_4 g + \frac{1}{2}c_L \partial_{11} g \\ &+ \int_{\mathbb{R}} [g(x + \xi, y, z, t) - g - \xi \partial_1 g]\nu_L(d\xi) \\ &+ \int_{\mathbb{R}} [g(x, y + \xi, z + e^{\kappa t}\xi, t) - g - \xi \partial_2 g - e^{\kappa t}\xi \partial_3 g]\nu_Z(d\xi)\end{aligned} \tag{19.19}$$

with boundary condition

$$g(x, y, z, T) = f_1(b_2(x, y, z, T))(1 - e^{-b_1(x,y,z,T)}) + f_2(b_2(x, y, z, T))e^{-b_1(x,y,z,T)},$$

where we have abbreviated $g = g(x, y, z, t)$ and

$$\begin{aligned}b_1(x, y, z, t) &:= \Gamma_t^d + \frac{1}{\kappa}(y - e^{-\kappa t} z), \\ b_2(x, y, z, t) &:= \log S_0 + rt + \omega t + x - \zeta y + b_1(x, y, z, t).\end{aligned}$$

ν_L and ν_Z are the Lévy measures of the processes L and Z. c_L denotes the variance of the Brownian part of L.

Proof We denote $V_t = (V_t^1 = L_t, V_t^2 = Z_t, V_t^3 = Y_t)$ and apply Itô's formula (19.16), cf. Theorem I.4.57 in Jacod and Shiryaev (2003). The existence of the first moment gives us a simple semimartingale representation for the Lévy process L

$$L_t = L_t - tE[L_1] + tE[L_1] =: M_t^L + tE[L_1].$$

As a consequence, we obtain the semimartingale representation of the stochastic integral $\int H_s dL_s$

$$\int_0^t H_s dL_s = \int_0^t H_s dM_s^L + E[L_1] \int_0^t H_s ds,$$

where H is a locally bounded predictable process. The first summand is a local martingale, cf. I.4.34 (b) in Jacod and Shiryaev (2003). We are interested in the second one, which is a predictable process with paths of bounded variation. The same procedure can be applied to the increasing Lévy process Z. Therefore, we get the representations

$$\begin{aligned}
\int_0^t H_s dZ_s &= \int_0^t H_s dM_s^Z + E[Z_1] \int_0^t H_s ds \ , \\
\int_0^t H_s dY_s &= \int_0^t H_s e^{\kappa s} dZ_s = \int_0^t H_s e^{\kappa s} dM_s^Z + E[Z_1] \int_0^t H_s e^{\kappa s} ds \ .
\end{aligned}$$

Since Z is an increasing Lévy process, we have $Z^c \equiv 0$ and also $Y^c \equiv 0$. Thus, the second term of Itô's formula is simplified considerably

$$\frac{1}{2} \sum_{i,j \le d} \int_0^t \partial_{ij} g(V_{s-}, s) d\langle (V^i)^c, (V^j)^c \rangle_s = \frac{1}{2} c_L \int_0^t \partial_{11} g(V_{s-}, s) ds.$$

The jump term in Itô's formula can be written in terms of the jump measure $\mu^{(L,Z)}$ of the two-dimensional Lévy process (L, Z)

$$\begin{aligned}
&\sum_{s \le t} \left[g(V_{s-} + \Delta V_s, s) - g(V_{s-}, s) - \sum_{i \le d} \partial_i g(V_{s-}, s) \Delta V_s^i \right] \\
&= \sum_{s \le t} \Big[g(L_{s-} + \Delta L_s, Z_{s-} + \Delta Z_s, Y_{s-} + e^{\kappa s} \Delta Z_s, s) - g(V_{s-}, s) \\
&\qquad - \partial_1 g(V_{s-}, s) \Delta L_s - \partial_2 g(V_{s-}, s) \Delta Z_s - \partial_3 g(V_{s-}, s) e^{\kappa s} \Delta Z_s \Big] \\
&= \int_{[0,t] \times \mathbb{R}^2} \Big[g(L_{s-} + x, Z_{s-} + y, Y_{s-} + e^{\kappa s} y, s) - g(V_{s-}, s) \\
&\qquad - \partial_1 g(V_{s-}, s) x - \partial_2 g(V_{s-}, s) y - \partial_3 g(V_{s-}, s) e^{\kappa s} y \Big] \mu^{(L,Z)}(ds, (dx, dy)).
\end{aligned}$$

The semimartingale representation of this type of integral is

$$W * \mu^V = \underbrace{W * \mu^V - W * (\mu^V)^p}_{\text{martingale}} + \underbrace{W * (\mu^V)^p}_{\text{pred. + bounded variation}},$$

cf. Theorem II.1.8. in Jacod and Shiryaev (2003). So, we have to investigate the predictable compensator of the jump measure $\mu^{(L,Z)}$, which is

$$\left(\mu^{(L,Z)}\right)^p(\omega; dt, (dx, dy)) = dt \otimes \nu_{(L,Z)}(dx, dy),$$

where $\nu_{(L,Z)}$ is the Lévy measure of (L, Z). Since the processes L and Z are independent, $\nu_{(L,Z)}$ is supported on the union of the coordinate axes and we can write

$$\nu_{(L,Z)}(A) = \nu_L(A_x) + \nu_Z(A_y),$$

where $A_x := \{(x, 0) \,|\, x \in A\}$ is the projection on the x-axis and $A_y := \{(0, y) \,|\, y \in A\}$ the projection on the y-axis. This result can be found in Sato (1999), E 12.10.(i) or Cont and Tankov (2004), Proposition 5.3. Consequently, each two-dimensional integral w.r.t. $\nu_{(L,Z)}$ is the sum of two one-dimensional integrals

$$\int g(x, y)\nu_{(L,Z)}(dx, dy) = \int g(x, 0)\nu_L(dx) + \int g(0, y)\nu_Z(dy). \tag{19.20}$$

As a result, the predictable and bounded variation part of the jump term is

$$\begin{aligned}
&\int_{[0,t]\times\mathbb{R}^2} \Big[g(L_{s-} + x, Z_{s-} + y, Y_{s-} + e^{\kappa s} y, s) - g(V_{s-}, s) \\
&\quad - \partial_1 g(V_{s-}, s)x - \partial_2 g(V_{s-}, s)y - \partial_3 g(V_{s-}, s)e^{\kappa s} y\Big] ds \otimes \nu_{(L,Z)}(dx, dy) \\
&= \int_0^t \int_{\mathbb{R}} \Big[g(L_{s-} + x, Z_{s-}, Y_{s-}, s) - g(V_{s-}, s) - \partial_1 g(V_{s-}, s)x\Big] \nu_L(dx) \\
&\quad + \int_{\mathbb{R}} \Big[g(L_{s-}, Z_{s-} + y, Y_{s-} + e^{\kappa s} y, s) - g(V_{s-}, s) \\
&\qquad - \partial_2 g(V_{s-}, s)y - \partial_3 g(V_{s-}, s)e^{\kappa s} y\Big] \nu_Z(dy)\ ds.
\end{aligned}$$

If we now zero all the predictable parts of Itô's formula with bounded variation, we obtain

$$0 = \int_0^t H(L_{s-}, Z_{s-}, Y_{s-}, s)ds \qquad (\forall t \geq 0)$$

for

$$\begin{aligned} H(x,y,z,t) :=& E[L_1]\partial_1 g + E[Z_1]\partial_2 g + E[Z_1]e^{\kappa t}\partial_3 g + \partial_4 g + \frac{1}{2}c_L \partial_{11} g \\ &+ \int_{\mathbb{R}} [g(x+\xi,y,z,t) - g - \xi\partial_1 g]\nu_L(d\xi) \\ &+ \int_{\mathbb{R}} [g(x,y+\xi,z+e^{\kappa t}\xi,t) - g - \xi\partial_2 g - e^{\kappa t}\xi\partial_3 g]\nu_Z(d\xi), \end{aligned}$$

where we wrote for short $g = g(x, y, z, t)$. By continuity, $H(x, y, z, t)$ has to be zero for every $t \geq 0$, every $x \in S(L_t)$, every $y \in S(Z_t)$ and every $z \in S(Z_t)$, whereby $S(X)$ denotes the support of the random variable X. This is the desired Eq. (19.19). □

In many cases, we have $S(L_t) = \mathbb{R}$ and $S(Z_t) = S(Y_t) = \mathbb{R}_+$, such that we have to solve equation (19.19) for $x \in \mathbb{R}$, $y, z \in \mathbb{R}_+$ and $t \in \mathbb{R}_+$.

To apply the stated theorem, we have to verify that the function g, defined in (19.18), is of class $C^2(\mathbb{R}^4)$. The validity of this condition depends on the specific processes L and Z and on the payoff function f of the claim which we consider. Cont and Voltchkova (2005) investigated a similar issue in the simpler case of exponential Lévy models. The problem is more complicated in our model setting and is not pursued in this paper.

19.6 Two Price Theory

In the classical risk-neutral valuation theory for financial derivatives it is implicitly assumed that the product is traded in a perfectly liquid market, which means that it can be bought and sold at once within the trading session and that this does not cause any substantial price movement. Typical examples for assets which are traded in rather liquid markets are shares of big listed companies, the corresponding plain vanilla options on these shares and government bonds of countries with a high rating. Neglecting processing, inventory and transaction costs of the market makers, in these markets the law of one price prevails, which means that the price for buying an asset is the same as the one for selling it.

In reality however there are two prices, one for buying from the market—the ask price—and one for selling to the market—the bid price. "The difference between these two prices can be quite large and may have little connection to processing, inventory, transactions costs or information considerations. The differences instead reflect the very real and substantial costs of holding unhedgeable risks in incomplete markets."[1]. In particular a large part of the products financial institutions are dealing with are very specialised. The markets for these over-the-counter (OTC) traded

[1] Cherny and Madan (2010), Introduction, p. 1150

structured products are very narrow with the consequence of large spreads between bid and ask prices.

Cherny and Madan (2010) started to develop a two price theory, which models bid and ask prices in a way which takes the cost of unhedgeable risks into account. In classical financial mathematics, cf. Delbaen and Schachermayer (2006), the price $\pi_0(X)$ of a derivative with discounted payoff X is calculated via

$$\pi_0(X) = E_P[X],$$

where P is a risk-neutral pricing measure. This formula is now substituted by the non-linear pricing formulas

$$b(X) = \inf_{Q \in \mathscr{D}} E_Q[X]$$
$$a(X) = \sup_{Q \in \mathscr{D}} E_Q[X]$$

for the bid and the ask price of an asset with discounted payoff X. $\mathscr{D}$ is a convex set of probability measures which contains a risk-neutral measure P. The size of $\mathscr{D}$ is related to the degree of uncertainty (liquidity) in the market under consideration. With increasing uncertainty more measures (scenarios) should be added to the set. Conversely, $\mathscr{D}$ could be shrunk when the uncertainty in the market decreases. Details and a vivid explanation of this can be found in Cherny and Madan (2010).

Under slight additional assumptions, namely comonotonicity and law-invariance, these two values can be calculated using concave distortions Ψ, more exactly

$$b(X) = \int_{\mathbb{R}} y \Psi(F_X(dy)) \tag{19.21}$$
$$a(X) = -\int_{\mathbb{R}} y \Psi(F_{-X}(dy)), \tag{19.22}$$

where F_X is the distribution function of X under P. Very useful parametrized families of distortions $(\Psi_\gamma)_{\gamma \geq 0}$ are presented in the following example.

Example 7 The MINVAR-family of distortions is defined by

$$\Psi_\gamma^{\mathsf{MI}}(y) := 1 - (1-y)^{\gamma+1}, \quad \gamma \geq 0, \ y \in [0,1].$$

Another family is given by

$$\Psi_\gamma^{\mathsf{MA}}(y) := y^{\frac{1}{\gamma+1}}, \quad \gamma \geq 0, \ y \in [0,1]$$

and is called MAXVAR. One possible combination of MINVAR and MAXVAR is

$$\Psi_\gamma^{\mathsf{MAMI}}(y) := (1-(1-y)^{\gamma+1})^{\frac{1}{1+\gamma}}, \quad \gamma \geq 0, \ y \in [0,1]$$

and is called MAXMINVAR. The other possible combination is

$$\Psi_\gamma^{\mathsf{MIMA}}(y) := 1 - (1 - y^{\frac{1}{\gamma+1}})^{\gamma+1}, \quad \gamma \geq 0, \ y \in [0,1]$$

and is called MINMAXVAR.

The existence of the integrals in (19.21) and (19.22) is not discussed in Cherny and Madan (2010). It depends on the payoff X and the used distortion Ψ. The existence under the four introduced distortions is ensured, if X possesses exponential moments, as seen in the following proposition.

Proposition 1 *Let X be a random variable with $E[e^{tX}] < \infty$ for $|t| \leq t_0$. Then the integrals (19.21) and (19.22) exist for the distortion families Ψ^{MA}, Ψ^{MI}, Ψ^{MAMI}, Ψ^{MIMA} and any $\gamma \geq 0$.*

Proof The assumption implies that the distribution function F_X of X decays exponentially. We consider the left tail of Ψ^{MA}

$$\int_{-\infty}^{0} \Psi_\gamma^{\mathsf{MA}}(F_X(y))dy \leq \int_{-\infty}^{0} \Psi_\gamma^{\mathsf{MA}}(Ce^{t_0 y})dy = C^{\frac{1}{1+\gamma}} \int_{-\infty}^{0} e^{\frac{t_0}{1+\gamma} y} dy < \infty$$

and the left tail of Ψ^{MI}

$$\begin{aligned}
\int_{-\infty}^{0} \Psi_\gamma^{\mathsf{MI}}(F_X(y))dy &\leq \int_{-\infty}^{0} \Psi_\gamma^{\mathsf{MI}}(Ce^{t_0 y})dy \\
&= \int_{-\infty}^{0} 1 - (1 - Ce^{t_0 y})^{1+\gamma} dy \\
&\leq C_1 + \int_{-\infty}^{-d^2} 1 - (1 + (1+\gamma)(-Ce^{t_0 y}))dy \\
&= C_1 + \int_{-\infty}^{-d^2} (1+\gamma)Ce^{t_0 y} dy < \infty,
\end{aligned}$$

where we have used Bernoulli's inequality

$$(1+x)^r \geq 1 + rx \qquad (x > -1, \ r \geq 1).$$

The same arguments show the statement for the right tails of Ψ^{MI} and Ψ^{MA} and for both tails of the distortion families Ψ^{MAMI} and Ψ^{MIMA}. □

Example 8 Since the payoff $P = (K - S_T)^+$ of a put option always possesses exponential moments if $S_T \geq 0$, the bid and ask prices always exist and are given by

$$a_\gamma(P) = \int_0^K \Psi_\gamma(F_{S_T}(x))dx \tag{19.23}$$

$$b_\gamma(P) = \int_0^K (1 - \Psi_\gamma(1 - F_{S_T}(x)))dx. \tag{19.24}$$

The payoff $C = (S_T - K)^+$ of a call option does not possess exponential moments in general for nonnegative random variables S_T. Consider $S_T = S_0 \exp(Y)$ for a random variable Y with exponential moment at $u_0 > 1$. Let Ψ be the MINVAR-family of distortions. Then the integrals (19.21) and (19.22) exist for every $\gamma \geq 0$ and we get

$$a_\gamma(C) = \int_K^\infty \Psi_\gamma(1 - F_{S_T}(x))dx \tag{19.25}$$

$$b_\gamma(C) = \int_K^\infty (1 - \Psi_\gamma(F_{S_T}(x)))dx. \tag{19.26}$$

Let Ψ be the MAXVAR-, MAXMINVAR- or MINMAXVAR-family of distortions. Then the integrals exist for every $\gamma \in [0, u_0 - 1)$ and the formulas (19.25) and (19.26) are in force for $\gamma \in [0, u_0 - 1)$. The proofs are similar to that of Proposition 1. Details can be found in Bäurer (2015).

We now apply the two price theory to the Defaultable Asset Price Model and derive bid and ask prices for options. As a consequence, we get prices for which market, credit and liquidity risk is taken into account. The bid and ask price formulas (19.21) and (19.22) depend on the distribution function F_X of the option payoff X. In many cases, it can be reduced to a dependence on F_{S_T}, the distribution function of the underlying S_T, cf. Example 8. In the DAM, the distribution function

$$F_{S_T}(x) = P(T \geq \tau) + P(e^{X_T} \leq x \text{ and } T < \tau)$$

of the asset price S_T is not known explicitly, because of the dependence between X_T and τ. Nevertheless one can calculate the desired values numerically. Using Lemma 1, the quantities $P(T < \tau)$ and $P(T \geq \tau) = 1 - P(T < \tau)$ are given by a simple integral

$$P(T < \tau) = E[e^{-\Gamma_T}] = e^{-\Gamma_T^d} \exp\left(\int_0^T \theta_Z(-\gamma_u^T)du\right).$$

We use the T-survival measure $P^T(A) := P(A \mid T < \tau)$ to determine

$$\begin{aligned} P(e^{X_T} \le x \text{ and } T < \tau) &= P(e^{X_T} \le x \mid T < \tau) \cdot P(T < \tau) \\ &= P^T(e^{X_T} \le x) \cdot P(T < \tau). \end{aligned}$$

The probability $P^T(e^{X_T} \le x)$ can be calculated numerically by Fourier inversion

$$\begin{aligned} P^T(e^{X_T} \le x) &= P^T(X_T \le \log(x)) \approx P^T(C \le X_T \le \log(x)) \\ &= \frac{1}{2\pi} \int \frac{e^{-itC} - e^{-it\log(x)}}{it} \varphi_{X_T}^{P^T}(t) dt, \end{aligned} \tag{19.27}$$

where the constant $C \in \mathbb{R}$ has to be chosen properly. $\varphi_{X_T}^{P^T}$ is the characteristic function of X_T under P^T, which can be calculated by integration via (19.12). Thus, the computational cost for calculating the distribution function at one point is that of two simple integrations and one double integration.

Alternatively, we can compute the distribution function F_{S_T} by Monte Carlo simulations. We can then also assess the bid and ask prices for path-dependent options.

For the existence of the integrals in (19.21) and (19.22), we often need the existence of exponential moments of

$$X_T := \log S_0 + rT + L_T - \zeta Z_T + \omega T + \Gamma_T.$$

Lemma 7 *Suppose that*

(I) L_T *has an exponential moment of order* $u_0 > 0$.
(II) Z_T *has an exponential moment of order* $u_0[(\frac{1}{\kappa} - \zeta) \vee \zeta]$.

Then X_T *has an exponential moment of order* u_0.

Proof First we observe that $|\gamma_s^T - \zeta| \le (\frac{1}{\kappa} - \zeta) \vee \zeta$ and therefore we can conclude

$$\begin{aligned} E[\exp(u_0 X_T)] &= \text{const.} \cdot E\left[\exp\left(u_0 L_T - u_0 \zeta Z_T + u_0 \Gamma_T\right)\right] \\ &= \text{const.} \cdot E[\exp(u_0 L_T)] E\left[\exp\left(\int_0^T \gamma_s^T u_0 - \zeta u_0 dZ_s\right)\right] \\ &\le \text{const.} \cdot M_{L_T}(u_0) \cdot E\left[\exp\left(\int_0^T u_0 |\gamma_s^T - \zeta| dZ_s\right)\right] \\ &\le \text{const.} \cdot M_{L_T}(u_0) \cdot E\left[\exp\left(u_0 \left[\left(\frac{1}{\kappa} - \zeta\right) \vee \zeta\right] Z_T\right)\right] < \infty. \end{aligned}$$

□

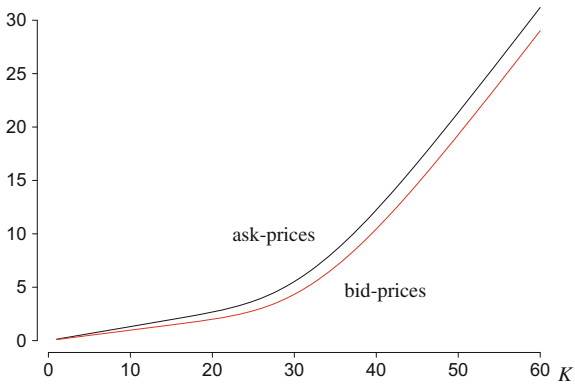


Fig. 19.7 Bid and ask prices of a put with $S_0 = 30$, DAM with parameters $(**)$, $T = 260$, $\gamma = 0.1$, MAXVAR

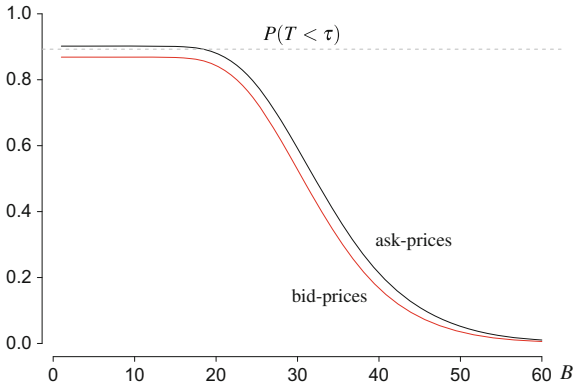


Fig. 19.8 Bid and ask prices of a digital call with $S_0 = 30$, DAM with parameters $(**)$, $T = 260$, $\gamma = 0.1$, MAXVAR

Example 9 For pricing calls and puts, we can use (19.23), (19.24), (19.25) and (19.26). Suppose X_T has an exponential moment at $u_0 > 1$. If Ψ is the MINVAR-family of distortions, then the integrals in (19.25) and (19.26) exist for every $\gamma \geq 0$. If Ψ is the MAXVAR-, MAXMINVAR- or MINMAXVAR-family of distortions, then the integrals exist for every $\gamma \in [0, u_0 - 1)$. A numerical example with the parameter set

$$\begin{array}{llll} \alpha = 50.0 & \beta = -0.1 & \delta = 0.012 & \\ p = 0.0035 & b = 66 & \kappa = 0.11 & (**) \\ & \zeta = 9.0. & & \end{array}$$

is shown in Fig. 19.7.

Example 10 For a digital call option with barrier $B > 0$ and payoff $X = \mathbb{1}_{\{S_T > B\}}$, we can use the simple formulas

$$a_\gamma(X) = \Psi_\gamma(1 - F_{S_T}(B)) \quad \text{and}$$
$$b_\gamma(X) = 1 - \Psi_\gamma(F_{S_T}(B)).$$

Figure 19.8 shows a numerical example. For this option, there are no constraints concerning the integrability.

References

Andersen, L. and D. Buffum (2004). Calibration and implementation of convertible bond models. *The Journal of Computational Finance 7*, 1–34.

Bachelier, L. (1900). *Théorie de la Spéculation*. Ph. D. thesis, École Normale Superieure Paris.

Barndorff-Nielsen, O. and N. Shephard (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B 63(2)*, 167–241.

Bäurer, P. (2015). *Credit and Liquidity Risk in Lévy Asset Price Models*. Ph. D. thesis, Universität Freiburg.

Bielecki, T. and M. Rutkowski (2004). *Credit Risk: Modeling, Valuation and Hedging*. (2. ed.). Springer.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy 81(3)*, 637–654.

Brigo, D. and F. Mercurio (2001). *Interest Rate Models - Theory and Practice*. Springer.

Carr, P., H. Geman, D. Madan, and M. Yor (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business 75(2)*, 305–332.

Carr, P., H. Geman, D. Madan, and M. Yor (2007). Self-decomposability and option pricing. *Mathematical Finance 17(1)*, 31–57.

Carr, P. and D. Madan (2010). Local volatility enhanced by a jump to default. *SIAM Journal of Financial Mathematics 1(1)*, 2–15.

Cherny, A. and D. Madan (2010). Markets as a counterparty: An introduction to conic finance. *International Journal of Theoretical and Applied Finance 13(08)*, 1149–1177.

Cont, R. and P. Tankov (2004). *Financial Modelling with Jump Processes*. Chapman and Hall/CRC.

Cont, R. and E. Voltchkova (2005). Integro-differential equations for option prices in exponential Lévy models. *Finance and Stochastics 9(3)*, 299–325.

Davis, M. and F. Lischka (2002). Convertible bonds with market risk and credit risk. In D. Y. R. Chan, Y-K. Kwok and Q. Zhang (Eds.), *Applied Probability*, Studies in Advanced Mathematics, pp. 45–58. American Mathematical Society/International Press.

Delbaen, F. and W. Schachermayer (2006). *The Mathematics of Arbitrage*. Springer.

Eberlein, E. (2001). Application of generalized hyperbolic Lévy motions to finance. In O. Barndorff-Nielsen, T. Mikosch, and S. Resnick (Eds.), *Lévy Processes: Theory and Applications*, pp. 319–336. Birkhäuser.

Eberlein, E., K. Glau, and A. Papapantoleon (2010). Analysis of Fourier transform valuation formulas and applications. *Applied Mathematical Finance 17(3)*, 211–240.

Eberlein, E. and U. Keller (1995). Hyperbolic distributions in finance. *Bernoulli 1(3)*, 281–299.

Eberlein, E. and K. Prause (2002). The generalized hyperbolic model: Financial derivatives and risk measures. In H. Geman, D. Madan, S. Pliska, and T. Vorst (Eds.), *Mathematical Finance: Bachelier Congress 2000*, Springer Finance, pp. 245–267. Springer.

Eberlein, E. and S. Raible (1999). Term structure models driven by general Lévy processes. *Mathematical Finance 9(1)*, 31–53.

Hull, J. and A. White (1990). Pricing interest rate derivative securities. *The Review of Financial Studies 3(4)*, 573–592.

Jacod, J. and A. Shiryaev (2003). *Limit Theorems for Stochastic Processes* (2. ed.). Springer.

Kluge, W. (2005). *Time-inhomogeneous Lévy Processes in Interest Rate and Credit Risk Models.* Ph. D. thesis, Universität Freiburg.

Linetsky, V. (2006). Pricing equity derivatives subject to bankruptcy. *Mathematical Finance 16(2)*, 255–282.

Madan, D., M. Konikov, and M. Marinescu (2004). Credit and basket default swaps. *The Journal of Credit Risk 2(1)*, 67–87.

Madan, D. and F. Milne (1991). Option pricing with V.G. martingale component. *Mathematical Finance 1(4)*, 39–55.

Madan, D. and E. Seneta (1990). The variance gamma (V.G.) model for share market returns. *Journal of Business 63(4)*, 511–524.

Merton, R. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science 4(1)*, 141–183.

Protter, P. E. (2005). *Stochastic Integration and Differential Equations* (2. ed.). Springer.

Samuelson, P. (1965). Rational theory of warrant pricing. *Industrial Management Review 6*(2), 13–32.

Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Schoutens, W. and J. Cariboni (2009). *Lévy Processes in Credit Risk*. Wiley Finance.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics 5(2)*, 177–188.

Effects of Regime Switching on Pricing Credit Options in a Shifted CIR Model

20

L. Overbeck and J. Weckend

20.1 Introduction

This paper studies the effects of regime switching in interest rate and single-name credit risk modeling in the context of Cox and Ross (1985) (CIR) processes. The focus is on the price implication for CDS options.

Figure 20.1 shows the absolute daily spread changes of the most liquid default product in Europe—the iTraxx® Main 5 year.

One way to interpret the inhomgeneous picture in Fig. 20.1 can be based on changing distributions of the spread changes in different time intervals over the whole period. There is at least one regime switch at the beginning of the credit crisis in the mid of 2007. Over a longer time period there are more regime switches, and there are economic cycles with irregular economy state changes (e.g., Hamilton (1989)). This observation gave the motivation to consider credit models with regime switching for pricing credit options.

Usually there are two classes of credit models. For a general overview on pricing models for credit derivatives Bielecki and Rutkowski (2004) may serve as a reference. The first one is the reduced form model, cf. e.g. Lando (1998); Duffie and Singleton (1999), to which the CIR family of models belongs. The second one is the class of structural models. They assume an underlying structure which causes the default event. The default time τ_D is for example modelled as the first hitting time of the firm value process A, i.e. $\tau_D = \inf\{t \leq 0 | A_t \leq D_0\}$ where the default point D_0 might be

L. Overbeck (✉)
Institute of Mathematics, University of Gießen, 35392 Gießen, Germany
e-mail: ludger.overbeck@math.uni-giessen.de

J. Weckend
University of Gießen, Gießen, Germany
e-mail: johannes.weckend@gmail.com

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_20

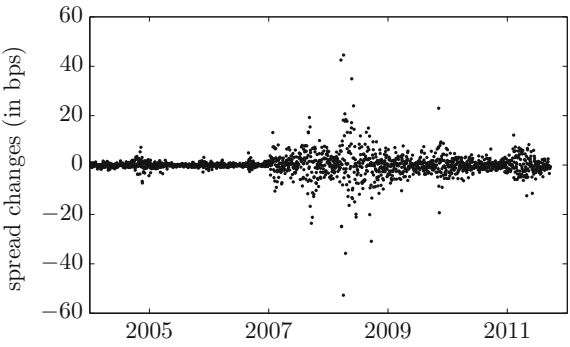


Fig. 20.1 iTraxx®Main 5 year historic spread changes (6th July 2012, Copyright©2012 Markit Group Limited)

derived from balance sheet data. The reduced form models reduce the analysis to the direct modeling of the default intensity of τ_D. Mathematically τ_D is a Cox process with stochastic intensity $x(t), t \geq 0$. In the CIR model, which is again a subclass of affine models (e.g., Duffie and Singleton (1999)), this intensity follows the so-called square-root process.

Since structural models are very difficult to calibrate to the term structure of defaults of a counterparty, in the area of single name credit derivative pricing, reduced form models are usually preferred and implemented, cf. e.g. Brigo et al. (2013). According to this reference the CIR-EJ++ is a widely used and accepted model and serves as a starting point in the present paper.

Because of the above mentioned observation we want to analyse how the pricing will change if we incorporate regime switching into the picture.

In addition to the default intensity the modeling of the interest component is done by the short rate model CIR++. For interest rate it prevents negative rates by definition, is analytically tractable and matches the term structure exactly. This model is an extension of the CIR model with a deterministic shift introduced by Brigo and Mercurio (2001). In Brigo and Mercurio (2006) it is called shifted square-root diffusion model in case of default. Due to the large jumps in the credit market, the model is extended by an exponential jump component. Exponential jumps keep the property of positivity, but do not enable the model to reflect large down movements of the default intensity which are also visible in specific market environments.

The changes of regimes are handled in the model by introducing a hidden Markov process to simulate the possibility of different distributions depending on the states. If there are two states only these may be interpreted as good and bad economy, but a finer grid of states is possible as well. This regime switching component affects both, the interest and the default component. Usually, in case of a bad economy, the interest rate decreases and the default intensity increases, and vice versa in a good economy.

Upon introduction of the regime switching into the CIR++ process the analytical tractability is lost. The calibration to the term structure is done by the deterministic shift, the volatility calibration is not possible in an analytical way.

20.2 Regime Switching CIR

Figure 20.1 indicates that the distribution of iTraxx® Main spread changes has significantly changed between the time period before and after mid of 2007. Constant component models would omit this economic behavior by using one distribution only for the risk factor. The regime switching models offer one option to overcome this problem. In these models a state variable is introduced. The distribution of the process thus, becomes state dependent. This state variable can be modeled as a Markov process.

The regime switching models can be traced back to the early work of Lindgren (1978) and became popular after the seminal work of Hamilton (1988). There are many papers on this topic as, among others, Gray (1996) or Ang and Bekaert (2002). Most papers on regime switching have no pricing background. In this paper the effect and the importance of the regime switching will be shown.

The model for the interest rate r and the default intensity λ are in its most general form CIR-EJ++ models, i.e. CIR models with exponential distributed jumps. "++" stands for the shift in order to calibrate the current term structure, cf. Brigo and Mercurio (2001) without the EJ component and Brigo and El-Bachir (2006) including the jump component:

$$y(t) = x^{\alpha}(t) + \varphi^{\mathrm{CIR}}(t;\alpha) \tag{20.1}$$

Here $x^{\alpha}(t)$ is CIR process with exponential jumps defined by

$$dx^{\alpha}(t) = \kappa\left(\theta - x^{\alpha}(t)\right)dt + \sigma\sqrt{x^{\alpha}(t)}dW(t) + J(\gamma)\,dN(\varsigma) \tag{20.2}$$

where $dN(\varsigma)$ represents a homogeneous Poisson process with constant intensity $\varsigma > 0$ (jump arrival rate) and N is independent of the Brownian motion W. J is exponentially distributed with a positive mean γ. The deterministic function $\varphi^{\mathrm{CIR}}(t;\alpha)$ is chosen to match the initial term structure as usual. The parameter vector is $\alpha = (x(0), \kappa, \theta, \sigma, \varsigma, \gamma)$. The process x^{α} is always positive if the Feller condition $\left(2\kappa\theta > \sigma^2\right)$ is satisfied.

An example of a probability density function of CIR and CIR-EJ is shown in Fig. 20.2 where the parameter vector (0.0165, 0.4, 0.026, 0.14, 0.25, 0.15) is applied with a time horizon of one year. The last two components of the parameter vector are only used in the CIR-EJ model. As expected, the exponential jump component shifts the density to higher values.

The CIR++ process and the CIR-EJ++ process are also called shifted square-root diffusion (SSRD) process and shifted square-root jump diffusion (SSRJD) process, respectively, as described in Brigo and Mercurio (2006).

Now we come to the regime switching.We use the generator matrix approach described in Elliott et al. (1995) to determine the state process and its transitions.The economic state variable is modeled by an $\mathbb{F}$-adapted continuous-time Markov chain process $S(t)$ with a finite state space $\mathcal{S} := (s_1, \ldots, s_Z)$. As in Elliott et al. (2005) the state space is described by a finite set of unit vectors $\{e_1, \ldots, e_Z\}$ where $e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^Z$ without loss of its generality.

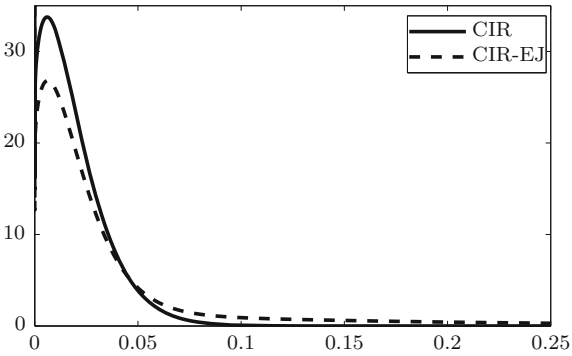


Fig. 20.2 Probability density function for CIR and CIR-EJ

$p_{ij}(t,T) := \mathbb{P}\left(S(T) = e_j \mid S(t) = e_i\right)$ denotes the transition probabilities of S from state e_i to state e_j for all times $t \leq T, i, j = 1, \ldots, Z$, $Q(t,T) := \left[p_{ij}(t,T)\right]$. Let $A(t) = \left[a_{ij}(t)\right]_{i,j=1,\ldots,Z}$ denote the generator matrix of the Markov chain process. The real matrix $A(t)$ satisfies the usual requirements as $a_{ii}(t) = -\sum_{j \neq i} a_{ij}(t)$ and $a_{ij}(t) \geq 0$ for all $i \neq j$.

Assumption 1 Each parameter in the parameter vector is itself a vector of length Z (number of states). The short rate r is assumed to be a RS CIR++ model given by[1]

$$r(t) = x_r^{\alpha}(t) + \varphi^{\text{RS CIR}}(t;\alpha) \text{ where}$$

$$dx_r^{\alpha}(t) = \langle \kappa_r, S(t) \rangle \left(\langle \theta_r, S(t) \rangle - x_r^{\alpha}(t) \right) dt + \langle \sigma_r, S(t) \rangle \sqrt{x_r^{\alpha}(t)} dW(t)$$

with parameter vector $\alpha = \left(x_r^{\alpha}(0), \kappa_r, \theta_r, \sigma_r\right) \in \mathbb{R} \times \mathbb{R}^Z \times \mathbb{R}^Z \times \mathbb{R}^Z$.

Assumption 2 The default intensity λ is assumed to be a RS CIR++ or a RS CIR-EJ++ by

$$\lambda(t) = x_{\lambda}^{\beta}(t) + \varphi^{\text{RS CIR-EJ}}(t;\beta) \text{ where}$$

$$\begin{aligned} dx_{\lambda}^{\beta}(t) = {} & \langle \kappa_{\lambda}, S(t) \rangle \left(\langle \theta_{\lambda}, S(t) \rangle - x_{\lambda}^{\beta}(t) \right) dt + \langle \sigma_{\lambda}, S(t) \rangle \sqrt{x_{\lambda}^{\beta}(t)} d\bar{W}(t) \\ & + J\left(\langle \gamma_{\lambda}, S(t) \rangle \right) d\bar{N}\left(\langle \varsigma_{\lambda}, S(t) \rangle \right) \end{aligned}$$

with parameter vector $\beta = \left(x_{\lambda}^{\beta}(0), \kappa_{\lambda}, \theta_{\lambda}, \sigma_{\lambda}, \varsigma_{\lambda}, \gamma_{\lambda}\right) \in \mathbb{R} \times \mathbb{R}^Z \times \mathbb{R}^Z \times \mathbb{R}^Z \times \mathbb{R}^Z \times \mathbb{R}^Z$. The last two parameters exist only in the RS CIR-EJ++ model.

[1] $\langle ., . \rangle$ is a scalar product in $\mathbb{R}^Z$ that for any $a, b \in \mathbb{R}^Z$ is: $\langle a, b \rangle = \sum_{i=1}^{Z} a_i b_i$.

Assumption 3 The Brownian motions W and $\bar{W}$ are correlated according to

$$d\left\langle W(t), \bar{W}(t)\right\rangle = \rho\, dt$$

The jump component is assumed to be independent of the Brownian motions.

The state space process S is independent of the Brownian motions W and $\bar{W}$ and of the jump component $\bar{N}$.

The deterministic shift is again chosen to match the exact term structure as in the CIR++ or CIR-EJ++ model. Besides the starting point of the process, each parameter is a vector of length Z. The other way round each state $i = 1, \ldots, Z$ has one parameter vector with constants $(x(0), \kappa_i, \theta_i, \sigma_i, \varsigma_i, \gamma_i)$. The Feller condition is satisfied if it is satisfied for each of the Z one-dimensional parameter vectors.

The interest rate has no jump component because it plays a minor role only in the valuation of credit products. Moreover, in the world of interest rates a jump component with positive shocks only is not reasonable. In the credit world shocks are possible which cause a positive jump of the default intensity. For example the default intensity of a firm jumps in case of bad news. The downward movement after such bad news is mostly smoother which is covered by the diffusion part.

20.3 Results

We now present results on the effect of the different models on Credit Default Swap option valuation which are based on a tree implementation of the regime switching model, cf. Overbeck and Weckend (2017).

A credit default swap (CDS) is a swap of premium payments K at times $T_{a+1}, \ldots, T_b$ in exchange for a single protection payment LGD $(= 1 - \text{RR})$ at default time τ, provided that $T_a < \tau \leq T_b$. To simplify the forthcoming formulas the notional is set to one. The CDS can be split into two parts, such as, the default leg for the default payments and the premium leg for the insurance payments. The discounted payoffs at time t are equal to

$$\mathbf{1}_{\{T_a<\tau\leq T_b\}} D(t,\tau)\,\text{LGD} \tag{20.3}$$

resp.

$$K\left(\sum_{i=a+1}^{b} D(t,T_i)\,\alpha_i \mathbf{1}_{\{\tau\geq T_i\}} + D(t,\tau)\left(\tau - T_{\beta(\tau)-1}\right)\mathbf{1}_{\{T_a<\tau<T_b\}}\right) \tag{20.4}$$

where $\beta(t) = \min\{k \mid T_k > t\}$ is the next date in the tenor structure after t, thus, $t \in \left[T_{\beta(t)-1}, T_{\beta(t)}\right)$. $D(t,T)$ is the stochastic discount factor at t for time T.

The price of a CDS is the expectation of the difference between its two legs (protection seller view) and given by

$$\mathrm{CDS}_{a,b}(t;K) = \mathbb{E}\Bigg[K\left(\sum_{i=a+1}^{b} D(t,T_i)\,\alpha_i \mathbf{1}_{\{\tau \geq T_i\}} + D(t,\tau)\left(\tau - T_{\beta(\tau)-1}\right)\mathbf{1}_{\{T_a<\tau<T_b\}}\right) - \mathbf{1}_{\{T_a<\tau\leq T_b\}} D(t,\tau)\,\mathrm{LGD}\,\Big|\,\mathcal{F}_t\Bigg]$$

where $\mathcal{F}_t$ describes the information available at t.

A credit default swap option or credit default swaption is an option on a CDS. It gives the holder the right to enter a CDS at its beginning T_a at a predefined level (strike) K. There are two different types of options, to sell or buy protection. The price of a call option (CallCDS) is given by

$$\mathrm{CallCDS}_{a,b}(t;K) := \mathbb{E}\left\{\mathbf{1}_{\{\tau>T_a\}} D(t,T_a)\cdot\left[-\mathrm{CDS}_{a,b}(T_a;K)\right]^{+} \mid \mathcal{F}_t\right\}.$$

The calibrations used in the examples were carried out in the spirit of market models (as in Schönbucher (2000), Jamshidian (2004), Brigo and Morini (2005)). This is possible since in case without regime switching semi-analytic formula for standard products are available. For more details of the calibration and implementation we refer to the Ph.D. thesis Weckend (2014). The programming language MATLAB® is used to obtain these results.

The market data of the 6th July 2012 is the taken for the short rate, the default intensity is based on Allianz market data on that date.[2] All parameters are summarized in Table 20.1. Here the first three rows belong to the interest rate component.

The regime switching model is presented for two and for three regimes. For the two and the three regimes one parameter vector is given in Table 20.1 only. The reason is that the one and two regime parameter vectors are taken additionally for the two and three regime models respectively. The parameters in the one-regime case calibrated to the option prices are adjusted to match the Feller condition. These single regime parameter is used as start regime in any case.

The parameters in the second and third regime are chosen to reflect different market conditions accordingly to a basic statistical analysis as in Weckend (2014). That means they are not calibrated to market data. In the non-jump models the mean-reversion level and the volatility are adapted and for the interest rate model the speed of mean-reversion as well. On the jump model only the jump components are changed to separate the effects.

The second regime reflects a better economy state than the first one. The third regime represents a very bad economy state with very high volatilities and high

[2] We do not present more recent data, since the calibration to negative interest rates seems to require an extension of the model and some additonal calibration procedures. This we want to avoid in order to focus on the regime switching.

Table 20.1 Parameters in RS CIR model

model	variable	$x(0)$	κ	θ	σ	ς	γ
CIR++	α	0.0003	0.0707	0.032	0.065	-	-
2-RS CIR++	α	0.0003	0.0707	0.05	0.01	-	-
3-RS CIR++	α	0.0003	0.707	0.02	0.12	-	-
CIR++	β	0.0055	0.2	0.035	0.11	-	-
2-RS CIR++	β	0.0055	0.2	0.02	0.05	-	-
3-RS CIR++	β	0.0055	0.2	0.07	0.3	-	-
CIR-EJ++	β	0.0038	0.0307	0.03	0.04	0.02	0.1445
2-RS CIR-EJ++	β	0.0038	0.0307	0.03	0.04	0.01	0.05
3-RS CIR-EJ++	β	0.0038	0.0307	0.03	0.04	0.06	0.25

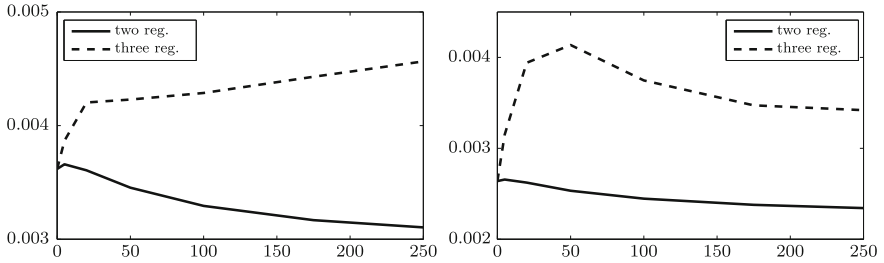
default intensities. The applied generator matrices of the transition probabilities are given by

$$A = \begin{pmatrix} -0.014 & 0.014 \\ 0.021 & -0.021 \end{pmatrix}, \quad A = \begin{pmatrix} -0.024 & 0.014 & 0.01 \\ 0.015 & -0.0151 & 0.0001 \\ 0.011 & 0.0001 & -0.0111 \end{pmatrix}$$

in the two and three regime models, respectively. The parameter vectors in the different states does satisfy the Feller condition. The condition that the model forward curve is below the market forward curve is satisfied for the applied transition probabilities. This is necessary to have a positive deterministic shift function keeping the overall short rate process positive.

As already announced all different models are applied to the CDS call option. The maturity date of the option is the 20th December 2012, the final CDS date is the 20th June 2017. The strike is chosen to be at the money. For the CIR default intensity model and in case of deterministic interest rates the analytical price is given by 36 bps, in the CIR-EJ model by 27bps. The correlation between interest rate and default intensity is assumed to be zero.

In Fig. 20.3 the prices are shown in relation to a multiplier of the generator matrix. The specified generator matrix is used per year in case of the multiplier being one. The statistics in Weckend (2014) are calculated on a daily basis. Therefore the factor goes up to 250 business days which means that the generator matrix is rescaled to a daily basis. A factor of 250 results in very high transition probabilities which are not in line with economic cycles and with durations of several years at least. But



(a) RS CIR model (b) RS CIR-EJ model

Fig. 20.3 CDS option prices

its application is useful to show some effects of regime switching and its impact on CDS option prices.

In Fig. 20.3a the regime switching model is shown without jumps. The lower volatility parameter in the second state leads to a lower price for the two regime case as expected. The small hump in the beginning can be explained by the volatility through the regime switching. The three regime case is exactly vice versa. The additional volatile regime leads to increased prices by an increased multiplicative factor. The slope is much higher in the beginning than for factors higher than 20. For higher factors the process is more like a mixture process and the slope becomes smaller.

The effects in case of the RS CIR-EJ++ model are presented in Fig. 20.3b. The shapes are very similar to the non-jump model parameters in Fig.20.3a. There is again a hump on the two regime case and afterwards slightly decreasing prices. The prices in the three regime case strongly increase in the beginning due to higher probabilities for the third regime in which the CDS option has a higher value. At certain point in time the prices decrease due to the effect of very fast regime switches.

In this example it is shown that using a regime switching models for option pricing has a significant impact.

20.4 Conclusion

Looking at the historic CDS data suggests that the distribution of the time series cannot be handled using only one parameter vector for all time periods. Based on a rough analysis of historical data (Weckend (2014)), we implement a model with three different states to explain the spread returns occurring on the iTraxx® Main and CDX® IG historic time series.

More specially in this paper default intensities and interest rates are modeled using a regime switching CIR model (RS CIR). The basic model definition and the impact on CDS option prices are given. In these examples the price differences between a single regime CIR model and RS CIR models are very significant.

For further model development, the calibration of the RS CIR model to market data is a necessary step. This will require a (semi-) analytic solution of the bond option prices in case of regime switching.

References

Ang, A. & G. Bekaert (2002). Regime switches in interest rates. *Journal of Business and Economic Statistics*, 20(2), 163–182.

Bielecki, T. R. & M. Rutkowski (2004). *Credit risk: Modeling, valuation and hedging*. Springer Science+Business Media, 2nd edition.

Brigo, D. & N. El-Bachir (2006). Credit derivatives pricing with a smile-extended jump stochastic intensity model. http://www.defaultrisk.com/pp_crdrv_96.htm.

Brigo, D. & F. Mercurio (2001). A deterministic-shift extension of analytically-tractable and time-homogeneous short-rate models. *Finance and Stochastics*, 5(3), 369–387.

Brigo, D. & F. Mercurio (2006). *Interest rate models - theory and practice: With smile, inflation and credit*. Springer Science+Business Media, 2nd edition.

Brigo, D. & M. Morini (2005). CDS market formulas and models. http://www.defaultrisk.com/pp_crdrv171.htm.

Brigo, D., M. Morini & A. Pallavicini (2013). *Counterparty Credit Risk, Collateral and Funding*. Wiley Finance.

Cox, J. C., J. E. Ingersoll Jr & S. A. Ross (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2), 385–407.

Duffie, D. & K. J. Singleton (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies*, 12(4), 687–720.

Elliott, R. J., L. Aggoun & J. B. Moore (1995). *Hidden markov models: Estimation and control*. Springer Science+Business Media, 1st edition.

Elliott, R. J., L. Chan & T. K. Siu (2005). Option pricing and Esscher transform under regime switching. *Annals of Finance*, 1(4), 423–432.

Gray, S. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1), 27–62.

Hamilton, J. D. (1988). Rational-expectations econometric analysis of changes in regime. *Journal of Economic Dynamics and Control*, 12, 385–423.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384.

Jamshidian, F. (2004). Valuation of credit default swaps and swaptions. *Finance and Stochastics*, 8(3), 343–371.

Lando, D. (1998). On Cox processes and credit risky securities. *Review of Derivatives Research*, 2(2), 99–120.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, 5(2), 81–91.

Overbeck, L. & J. Weckend (2017). Regime switching CIR tree. Preprint 2017.

Schönbucher, P. J. (2000). A Libor market model with default risk. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=261051.

Weckend, J. (2014). *Shifted regime switching CIR diffusion tree for credit options*. Ph.D. thesis, Justus-Liebig-University Giessen.

Part V
Gender Gap Analysis

Hierarchical Organizations and Glass Ceiling Effects

21

María Paz Espinosa and Eva Ferreira

JEL codes: C73 · J71 · M51

21.1 Introduction

From an empirical viewpoint it is not clear whether there are glass ceiling effects in corporate, political or academic contexts. The results are mixed: sometimes discrimination seems to become worse at the upper levels, while at others it is constant or even improves (Smith, 2012; Jackson and O'Callagham, 2009). These discrepancies may be partly explained by the fact that the notion of glass ceiling effects is not always the same in all papers. But it may also be the case that glass ceilings are present in some markets or for some types of corporate hierarchy but not for others, so that different data sets yield different answers as to whether or not there is a glass ceiling (Baxter and Wright, 2000).

On the other hand, there is clear evidence of gender and race biases in different contexts (Goldin and Rouse, 2000; Moss-Racusin et. al., 2012; and Reuben, Sapienza and Zingales, 2014, among others) and of selection processes that make use of stereotypes (Fershtman and Gneezy, 2001) and could result in a bias in the perception of female qualifications.

M.P. Espinosa (✉)
Departamento de Fundamentos Del Análisis Económico II,
University of the Basque Country, Avenida Lehendakari Aguirre 83, 48015 Bilbao, Spain
e-mail: mariapaz.espinosa@ehu.eus

E. Ferreira
Departamento de Economía Aplicada III & BETS, University of the Basque Country, Avenida Lehendakari Aguirre 83, 48015 Bilbao, Spain
e-mail: eva.ferreira@ehu.eus

D. Ferger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_21

In this paper we explore the conditions for obtaining glass ceiling effects when there is a gender bias at different levels of a hierarchical organization. We show that the explanation for the glass ceiling effect does not need to rely on exogenous greater discrimination at the top levels but on the dynamics of the process, the shape of the hierarchical organization and the distribution of abilities. To isolate the effect of the bias we assume that male and female candidates have the same preferences and outside options, so that there are no labor supply effects (Bertrand, Goldin and Katz, 2010).

The situation for homogeneous individuals has been studied in Espinosa and Ferreira (2015); Espinosa et. al., (2016) and in Ferreira and Stute (2016). These works provide theoretical tools to study the consequences of bias on the long run and on the glass ceiling effects. Here the context is different, individuals are heterogeneous in their abilities and high ability individuals are more productive; the distribution of abilities is the same for men and women, although their perceived abilities may differ when there is a perception bias. In this setup, the selection process involves choosing the best candidates, in particular setting a minimum level, or a quantile, in the space of abilities and selecting candidates above the threshold.

The paper is organized as follows. Section 21.2 presents the main elements of a hierarchical structure and formalizes the concept of a promotion structure in this framework. Section 21.3 characterizes three different notions of glass ceiling effects and their links to the characteristics of a promotion structure. Section 21.4 presents the results for two families of ability distributions, Pareto and Weibull. Finally, Sect. 21.5 discusses the main results and concludes. The proofs are relegated to the Appendix.

21.2 Modeling Discrimination Under Heterogeneus Abilities

We consider selection processes that choose the best candidates in a population with different groups. Thoroughout the paper we refer to these groups as men and women, but they could be defined according to other characteristic such as ethnic origin, etc. The important point is that there is a perception bias against one of the groups; we assume that the distribution, F, of real abilities, X, is the same in the two groups and all individuals show the same willingness to work, so that we can explore the effects of the bias in isolation from any possible labor supply effects. Once in the job, the candidate's productivity is an increasing function of his/her real ability. This setup corresponds to positions where efficiency increases with the qualifications of the elected candidates. The model and the selection procedure are formalized through the following set of assumptions:

The candidates

Assumption *(H.1) (Equal distribution of abilities)* Eligible candidates can be ordered according to their abilities. The abilities of the candidates are independent observations of the same random variable X with distribution F.

We denote the perceived ability by S, that takes a value S^m, when the candidate is a male, and S^f when she is a female. The candidates' abilities are observable but there may be a perception bias:

Assumption *(H.2) (Perception bias)* The perception of male abilities, S^m, has no bias so the perceived and real ability coincide, $S^m = X$, and therefore $P\left(S^m \leq x\right) = F\left(x\right)$, whereas for female candidates the distribution of the perceived abilities, S^f, accounts for some gender bias. More precisely, for female candidates perceived abilities are a function of real abilities: $S^f = \Psi(X)$, where $0 < \Psi(x) \leq x$ and $\Psi'(x) > 0$, so that $P\left(S^m \leq x\right) = F\left(\Psi^{-1}(x)\right)$. The discrepancy between x and $\Psi(x)$ is interpreted as the bias in the selection process and measured as $0 < \Psi(x)/x \leq 1$.

Remark 1 If the bias depends on the level in the hierarchy, we should use $\Psi_l(x)$ However, as mentioned in the introduction, our interest is precisely to analyze the effect of the bias on the glass ceiling effect, even when this bias does not depend on the level.

The hierarchy

We describe the selection processes in a hierarchy with several rungs. The selection process for promotion at each level, l, sets a minimum level of capabilities z_l. All candidates who are perceived to be above that level are promoted.

Assumption *(H.3) (Promotion decision)* For promotion from $l-1$ to l all candidates that are perceived to have abilities greater than z_l are promoted. That is, a candidate is selected if $S \geq z_l$. The candidates competing for level l are all those who were promoted at level $l-1$.

We assume that the hierarchical organization follows the layout of a pyramid:

Assumption *(H.4) (Pyramid organizational structure)* A hierarchy is a sequence $\{z_1, z_2, z_3, ..., z_{l-1}, z_l, ...\} = \{z_l\}_l$, where $z_j < z_{j+1}$ for all j.

Assumption *(H.5) (Continuous abilities)*. We assume that the distribution of abilities is absolutely continuous, which avoids trivial selection processes, that is $F\left(z_j\right) < F\left(z_{j+1}\right)$ for any hierarchy.

Our final interest is to check whether discrimination leads to wider gaps at higher levels in a hierarchical organization or, in other words, whether there is a glass ceiling effect. To that end we look at three different indices. The first one, denoted $p(z_l)$, is just the proportion of women at each level of the hierarchy l which requires ability z_l. Under assumptions *(H.1)* to *(H.5)*,

$$p\left(z_l\right) = P(S = S^f | S \geq z_l) = \frac{P\left(S = S^f\right) P(S \geq z_l | S = S^f)}{P\left(S \geq z_l\right)}$$
$$= \frac{q_l P\left(X \geq \Psi^{-1}(z_l)\right)}{q_l P\left(X \geq \Psi^{-1}(z_l)\right) + (1 - q_l)\, P\left(X \geq z_l\right)}$$

where q_l is the proportion of women in the set of candidates to be promoted at level l, which, by *(H.3)*, is equal to $p\,(z_{l-1})$, the proportion of women at rung $l-1$.

To determine whether there is a glass ceiling effect we must look at the sequence $\{p(z_1), p(z_2), ..., p(z_{l-1}), p(z_l), ...\}$, denoted by $\{p(z_l)\}$. We show that for a given level of ability z_l the proportion $p(z_l)$ is independent of the rest of the hierarchical structure $\{z_1, z_2, ..., z_{l-1}, z_{l+1}, ...\}$. In fact, the proportion $p(z_l)$ depends only on the parameters of the model: the distribution of abilities, F, the bias, $\Psi(x)/x$, and the proportion of women in the population (the pool of candidates), q. This result is stated as follows:

Lemma 1 *Under* $(H.1)$ *to* $(H.5)$ *the proportion of women at each level* l *of the hierarchy,* $p(z_l)$*, depends on the ability threshold* z_l*. It does not depend on the rest of the hierarchy structure* $\{z_1, z_2, ..., z_{l-1}, z_{l+1}, ...\}$.

A promotion structure (PS) is characterized by the bias $\Psi(x)/x$, the initial proportion of women, q, the distribution of abilities, F, and the hierarchy, $\{z_l\}_l$, $(PS) = \{\Psi(x)/x, q, F, \{z_l\}_l\}$. In a promotion structure (PS) a glass ceiling effect is defined as an increasing discrimination gap as l increases. Using the first index, the proportion of women at each level, we define the type 1 glass ceiling as follows.[1].

Definition 1 In a promotion structure (PS), there is a type 1 glass ceiling effect *(GC1)* if $\{p(z_l)\}$ is strictly decreasing in l.

The second index for the discrimination gap looks at the odds of men and women being promoted to level l from the beginning of a career. Consider all female candidates at the beginning of the selection process and denote by $\rho^f\,(z_l)$ the proportion of them that will pass threshold z_l; similarly, $\rho^m\,(z_l)$ is the proportion of men that will pass that same threshold z_l. The partition theorem enables us to write $P\,(S \geq z_l) = P\left(S \geq z_l | S \geq z_{l-j}\right) P\left(S \geq z_{l-j}\right)$, for any previous level $z_{l-j} < z_l$. Thus, from assumption *(H.3),* we have that $\rho^f\,(z_l)$ and $\rho^m\,(z_l)$ are independent of the rest of the hierarchy structure $\{z_1, z_2, z_3, ..., z_{l-1}, z_{l+1}, ...\}$ and in particular of the previous ability levels before level z_l. Therefore, we can write

$$\rho^f\,(z_l) = P\left(S^f \geq z_l\right) = P\left(X \geq \Psi^{-1}(z_l)\right)$$
$$\rho^m\,(z_l) = P\left(S^m \geq z_l\right) = P\,(X \geq z_l)$$

The ratio $\rho^f\,(z_l)\,/\rho^m\,(z_l)$ measures the relative chances of each of the two groups reaching a certain level l in the corporate hierarchy. Since abilities are equally distributed, in the absence of any discrimination $(\Psi^{-1}(z_l) = z_l)$ the index would be 1.

[1]See also Espinosa and Ferreira (2015) for similar definitions of glass ceilings in a homogeneous abilities context

When there is a gender bias ($\Psi^{-1}(z_l) > z_l$), the index will be lower than 1. We are interested in how this discrimination gap changes over the course of a pyramidal hierarchy. Using this ratio, we define the type 2 glass ceiling.

Definition 2 In a promotion structure (PS), there is a type 2 glass ceiling effect *(GC2)* when the sequence $\{\rho^f(z_l)/\rho^m(z_l)\}$ is strictly decreasing in l.

We propose a third measure, frequently used in empirical analysis: women who have already reached high levels find it increasingly difficult to be promoted in comparison with male candidates at the same level. Consider all female candidates who have reached level $l-1$ and denote by $\rho^f_{l-1}(z_l)$ the proportion of them that will pass the threshold z_l; similarly, $\rho^m_{l-1}(z_l)$ is the proportion of men at level $l-1$ that will be selected for level l. Then,

$$\rho^f_{l-1}(z_l) = P\left(S^f \geq z_l | S^f \geq z_{l-1}\right) = P(X \geq \Psi^{-1}(z_l) | X \geq \Psi^{-1}(z_{l-1}))$$
$$\rho^m_{l-1}(z_l) = P\left(S^m \geq z_l | S^m \geq z_{l-1}\right) = P(X \geq z_l | X \geq z_{l-1})$$

The index $\rho^f_{l-1}(z_l)/\rho^m_{l-1}(z_l)$ considers the relative chances of promotion for women and men who have reached a certain level in their careers (as opposed to the chances at the beginning of a career, as in *GC2*). Again, in the absence of any discrimination ($\gamma = 1$) the ratio would be 1, and we look at how this discrimination index moves with l with a constant gender bias $\gamma < 1$.

Definition 3 In a promotion structure (PS), there is a type 3 glass ceiling effect *(GC3)* when the sequence $\{\rho^f_{l-1}(z_l)/\rho^m_{l-1}(z_l)\}$ is strictly decreasing in l.

Since we assume there is a perception bias against one of the groups (women), we are bound to find discrimination. The question is whether a non-increasing bias leads to constant, wider or narrower gender discrimination gaps at higher levels of the hierarchy.

One could conjecture that if the gender bias is constant, the gender gap should also be constant no matter which definition of *GC* is considered. However, we show that this is not the case. Even in this simple framework, for a given promotion structure, the presence of the different types of glass ceiling effects depends crucially on the hierarchy structure and/or the distribution of abilities. The next section presents some general results, and applications to two families of ability distributions, Weibull and Pareto.

21.3 Characterization of Glass Ceiling Effects

In this section we show that type 1 and type 2 glass ceilings do not depend on the particular hierarchy structure $\{z_l\}_l$, but only on the distribution of abilities F and the

biased perception $\Psi(x)$, whereas $GC3$ does depend on the specific hierarchy. The function $g(z) = P\left(X > \Psi^{-1}(z)\right)/P\left(X > z\right)$, which involves the distribution of abilities F and the bias $\Psi(x)/x$, is useful in deriving the results. Using this notation, the general result can be stated as follows:

Proposition 1 *Consider a given promotion structure, characterized by* $(PS) = \{\Psi(x)/x,\ q,\ F,\ \{z_l\}_l\}$. *Then, under Asumptions* (H.1) *to* (H.5),

(i) If $\Psi(x) = x$, *there are no GC effects.*
(ii) If $\Psi(x) < x$, *there is always GC1.*
(iii) If $\Psi(x) < x$ *and* $g(z)$ *is a strictly decreasing function, there is GC2.*
(iv) If $\Psi(x) < x$ *and* $g\left(z_l\right)^2 > g\left(z_{l-1}\right) g\left(z_{l+1}\right)$ *for any* $l > 1$, *there is GC3.*

From this proposition it can be seen that a gender bias, $\Psi(x)/x < 1$, always leads to $GC1$, while $GC2$ also requires $g(z)$ to decrease with z . Moreover, $GC3$ is related to the hierarchy structure and a sufficient condition for $GC3$ is that $\frac{g(z_l)}{g(z_{l-1})}$ be decreasing in l. The following proposition highlights how $GC3$ is related to the hierarchy structure:

Proposition 2 *Fix* $(\Psi(x)/x,\ q,\ F)$ *and consider the class of all* PS *with* $(\Psi(x)/x,\ q,\ F)$. *These promotion structures only differ in the hierarchy* $\{z_l\}_l$, *that is, on the ability requirements for each level* l. *Consider any number of levels* L; *if* g *is strictly monotone, there is a* PS *with* L *levels in the class with* $GC3$ *effects and there is also a* PS *with* L *levels where the GC3 effects are reversed.*

21.4 Glass Ceiling Effects Under Pareto and Weibull Distributions of Ability

In this section we analyze *GC* effects in two applications with different distributions of abilities. In particular, we check whether the conditions on g for the glass ceiling effects hold (Proposition 1). The Weibull and Pareto families of distributions are very common for modeling the distributions of the tails, $P(X \geq z)$, and since the glass ceiling is a concern in the upper tail of the distribution of abilities, both seem to be good choices.

The main difference between the two families is the rate of decay. In the Weibull case it is exponential while in the Pareto case there is a power-law decay. Thus, the Pareto distribution models variables with heavier tails than the Weibull distribution. As shown below, the shape of the rate of decay plays an important role in *GC* effects. We also examine whether there is any role for the hierarchy in the *GC3* effect.

21.4.1 Weibull Case

In this subsection we assume that the ability X follows a distribution in the Weibull family so that $P(X > z) = e^{-(z/\lambda)^a}$, where λ and a are positive parameters. We assume a constant bias so that $\Psi(x) = \gamma X$. We examine which promotion structures $(PS)_W = \{\gamma, q, F_W, \{z_l\}_l\}$, give rise to glass ceiling effects, where W denotes the Weibull distribution function.

Proposition 3 *Under* (H.1) *to* (H.5)*, if* $\gamma < 1$ *and* X *follows a distribution in the Weibull family:*

(i) There are GC1 *and* GC2 *for any promotion structure* $(PS)_W$.
(ii) Given $(PS)_W = \{\gamma, q, F_W, \{z_l\}_l\}$*, there is GC3 whenever* $z_{l+1}^a - z_l^a > z_l^a - z_{l-1}^a$ *for any* l.

It is interesting to discuss the condition for *GC3* in (ii). The result indicates that, even with the same distribution of abilities, promotion structures may or may not lead to *GC3* effects, depending on the ability requirements at the different levels (the hierarchical structure). For instance, in the exponential case ($a = 1$), *GC3* appears whenever the steps $z_l - z_{l-1}$ widen when climbing up the corporate ladder: $z_{l+1} - z_l > z_l - z_{l-1}$. This condition seems to be realistic since in many companies' organizational structures, the jumps between low levels are smaller than at the top rungs.

However, if the steps $z_l - z_{l-1}$ are shorter at the top, not only does *GC3* not appear, but the odds ratio between females and males may reverse at higher levels. This is a very counter-intuitive result since one would expect *GC2* and *GC3* to go in the same direction.

21.4.2 Pareto Distribution

In this subsection, we apply our results to two Pareto distributions, the *first kind* Pareto (FP) and the *generalized* Pareto (GP), and analyze which promotion structures, $(PS)_{FP}$ and $(PS)_{GP}$ respectively, give rise to glass ceiling effects.

First kind Pareto distribution
Consider that ability X follows a *first kind* Pareto distribution (FP), that is, $P(X > z) = (z/k)^a$, where parameters a and k are strictly positive. For this case, the function $g(z) = P(X > z/\gamma)/P(X > z)$ is constant:

$$g(z) = \frac{(z/\gamma k)^a}{(z/k)^a} = \frac{1}{\gamma^a}$$

Therefore, there is no *GC2*. Moreover, since the condition for *GC3* is $g(z_l)^2 < g(z_{l-1})\, g(z_{l+1})$, there is no *GC3* either. This result can be stated as follows:

Proposition 4 *Under* (H.1) *to* (H.5)*, if* $\gamma < 1$ *and X follows a first kind Pareto distribution* (F_{FP})*:*

(i) There is GC1 for any promotion structure $(PS)_{FP} = \{\gamma, q, F_{FP}, \{z_l\}_l\}$.
(ii) Given $(PS)_{FP}$*, there is no GC2 or GC3.*

When abilities follow the first kind Pareto distribution the odds ratio between women and men remains constant throughout the hierarchy, which prevents any type 2 or 3 GC effects.

Generalized Pareto distribution
Now consider that ability X follows a Generalized Pareto distribution, that is,

$$P(X > z) = \begin{cases} \left(1 - \frac{kz}{\sigma}\right)^{1/k} & k \neq 0 \\ \exp\left(-\frac{z}{\sigma}\right) & k = 0 \end{cases} \tag{21.1}$$

with $\sigma > 0$; if $k > 0$, $0 \leq z \leq \sigma/k$; and if $k \leq 0$, $0 \leq z < \infty$. Note that if $k = 0$ then this is the exponential case $\exp(\sigma)$; if $k = 1$, the distribution is uniform $[0, \sigma]$; and if $k < 0$, it is a *second kind* Pareto distribution.

Proposition 5 *Under* (H.1) *to* (H.5)*, if* $\gamma < 1$ *and X follows a generalized Pareto distribution* (F_{GP})*:*

(i) For any $(PS)_{GP} = \{\gamma, q, F_{GP}, \{z_l\}_l\}$*, there is always GC1 and GC2.*
(ii) For any number of rungs L, there is a $(PS)_{GP}$ *with L levels with GC*3*. There is also a* $(PS)_{GP}$ *with L levels where the GC3 effect is reversed.*

21.5 Conclusions

We explore glass ceiling effects in hierarchical organizations. Our results relate the presence of glass ceilings to the structure of the hierarchy and the distribution of abilities. We show that the explanation for glass ceiling effects does not need to rely

on exogenous greater discrimination at top levels but on the requirements for each level in the hierarchy and the characteristics of the distribution of abilities. These results are roughly consistent with the mixed empirical evidence that has found glass ceilings in some contexts but not in others.

As a general result, the type 1 glass ceiling effect appears whenever there is a bias in the selection process, independently of the ability distribution. The type 2 glass ceiling effect, which accounts for unconditional probabilities of climbing the corporate ladder, depends on the ability distribution but not on the hierarchy structure. Finally, the type 3 glass ceiling effect depends on the distribution of abilities and more crucially on the structure of the corporate ladder.

The results for the Weibull distribution point to the existence of all types of glass ceiling effects ($GC1$, $GC2$ and $GC3$) when hierarchies have higher steps at the top than at the bottom of the corporate ladder. However, this result cannot be extended to other distributions.

Acknowledgements We are grateful for comments from Prof. Dr. Winfried Stute, the participants on the International Conference in Probability Theory and Statistics on Tbilisi, 2015 and an anonymous referee. Financial support from MINECO (ECO2015-64467-R and ECO2014-51914-P), the Basque Government (DEUI, IT-783-13) and UPV/EHU (UFI 11/46 BETS) is gratefully acknowledged.

21.6 Appendix

Proof Lemma 1

For the sake of simplicity and without loss of generality consider hierarchy 1 with one level $\{z\}$ and hierarchy 2 with two levels, $\{z_1, z_2\}$, with $z_1 < z_2 = z$.

For the first hierarchy $\{z\}$,

$$p_1(z) = \frac{q\,P\left(X \geq \Psi^{-1}(z)\right)}{q\,P\left(X \geq \Psi^{-1}(z)\right) + (1-q)\,P\left(X \geq z\right)}$$

where q is the proportion of women in the population. For the second hierarchy $\{z_1, z\}$,

$$\begin{aligned} p_2(z) &= \frac{p_1(z_1)P\left(X \geq \Psi^{-1}(z)|X \geq \Psi^{-1}(z_1)\right)}{p_1(z_1)P\left(X \geq \Psi^{-1}(z)|X \geq \Psi^{-1}(z_1)\right) + (1-p_1(z_1))\,P\left(X \geq z|X \geq z_1\right)} \\ &= \frac{q\,P\left(X \geq \Psi^{-1}(z_1)\right)P\left(X \geq \Psi^{-1}(z_1)|X \geq \Psi^{-1}(z_1)\right)}{q\,P\left(X \geq \Psi^{-1}(z_1)\right)P\left(X \geq \Psi^{-1}(z)|X \geq \Psi^{-1}(z_1)\right) + ..} \\ &\quad \frac{}{.. + (1-q)\,P\left(X \geq z_1\right)P\left(X \geq z|X \geq z_1\right)} \\ &= \frac{q\,P\left(X \geq \Psi^{-1}(z)\right)}{q\,P\left(X \geq \Psi^{-1}(z)\right) + (1-q)\,P\left(X \geq z\right)} \\ &= p_1(z) \end{aligned}$$

An induction argument shows the result for any two general hierarchies $\{z_1, z_2, ...z, ...\}$ and $\{y_1, y_2, ...z, ...\}$ with a level z in common. ■

Proof Proposition 1

(i) This is straightforward since, for this case, $P(X > z/\gamma) = P(X > z)$ for all z.

(ii)

$$\begin{aligned} p(z_l) &= P\left(S = S^f | S \geq z_l\right) = \frac{P(S \geq z_l | S = S^f) P\left(S = S^f\right)}{P\left(S \geq z_l\right)} \\ &= \frac{q_l P\left(X \geq \Psi^{-1}(z_l)\right)}{q_l P\left(X \geq \Psi^{-1}(z_l)\right) + (1 - q_l) P\left(X \geq z_l\right)} \\ &= \frac{p(z_{l-1}) P\left(X \geq \Psi^{-1}(z_l)\right)}{p(z_{l-1}) P\left(X \geq \Psi^{-1}(z_l)\right) + (1 - p(z_{l-1})) P\left(X \geq z_l\right)} \end{aligned}$$

From *(H5)*, $P\left(X \geq z_l\right) > P\left(X \geq \Psi^{-1}(z_l)\right)$, so $p(z_l) < p(z_{l-1})$ for any $z_{l-1} < z_l$.

(iii) $GC2$ is characterized by the condition

$$g(z_l) = \frac{P\left(X > \Psi^{-1}(z_l)\right)}{P\left(X > z_l\right)} < \frac{P\left(X > \Psi^{-1}(z_{l-1})\right)}{P\left(X > z_{l-1}\right)} = g(z_{l-1})$$

That is, for a given hierarchy $\{z_l\}_l$ there is $GC2$ if $g\left(z_1\right) > g\left(z_2\right) > ... > g\left(z_l\right) > ....$ If $g(z)$ is a monotone decreasing function, then there is $GC2$ for any hierarchy.

(iv) $GC3$ is characterized by the condition

$$\begin{gathered} \frac{P\left(X > \Psi^{-1}(z_{l+1}) | X > \Psi^{-1}(z_l)\right)}{P\left(X > z_{l+1} | X > z_l\right)} < \frac{P\left(X > \Psi^{-1}(z_l) | X > z_{l-1}/\gamma\right)}{P\left(X > z_l | X > z_{l-1}\right)} \\ \frac{P\left(X > \Psi^{-1}(z_{l+1})\right) / P\left(X > \Psi^{-1}(z_l)\right)}{P\left(X > z_{l+1}\right) / P\left(X > z_l\right)} < \frac{P\left(X > \Psi^{-1}(z_l)\right) / P\left(X > \Psi^{-1}(z_{l-1})\right)}{P\left(X > z_l\right) / P\left(X > z_{l-1}\right)} \\ g(z_{l+1}) / g(z_l) < g(z_l) / g(z_{l-1}) \end{gathered}$$

which is equivalent to the condition $g(z_{l+1}) g(z_{l-1}) < g(z_l)^2$. ■

Proof Proposition 2

Consider that g is strictly decreasing (the increasing case is analogous). Note that $g(z)$ is a continous bounded function which lies in the interval $[0, 1]$. Fix the first point z_1, a low value from the distribution of abilities, with $g(z_1) = \alpha$. Take an increasing sequence of positive values $\{\beta_2 < \beta_3 < \cdots < \beta_L\}$ such that $\sum_{l=2}^{L} \beta_l \leq \alpha$. Then, the corresponding values $\{z_1 < z_2 < \cdots < z_L\}$ such that $g(z_1) = \alpha$ and $g(z_l) = g(z_{l-1}) - \beta_l$, for $l = 2, ..., L$ conform a hierarchy with $GC3$ effects. This is straightforward since $\beta_l < \beta_{l+1}$ implies

$$\begin{aligned} g(z_{l+1}) g(z_{l-1}) &= \left(g(z_l) - \beta_{l+1}\right) \left(g(z_l) + \beta_l\right) \\ &= g(z_l)^2 - \beta_{l+1} g(z_l) + \beta_l g(z_l) - \beta_l \beta_{l+1} \\ &= g(z_l)^2 + g(z_l) \left(\beta_l - \beta_{l+1}\right) - \beta_l \beta_{l+1} < g(z_l)^2 \end{aligned}$$

To construct a hierarchy where the $GC3$ effect is reversed, consider a decreasing sequence of positive values $\{\beta_2 > \beta_3 > \cdots > \beta_L\}$ such that $\sum_{l=2}^{L} \beta_l \leq \alpha/2$, $\beta_{l+1} < \beta_l/2$, and the corresponding values $\{z_1 < z_2 < \cdots < z_L\}$ such that $g(z_1) = \alpha$ and $g(z_l) = g(z_{l-1}) - \beta_l$, for $l = 2, ..., L$. For this hierarchy $g(z_{l+1})g(z_{l-1}) > g(z_l)^2$, since $g(z_l)\,(\beta_l - \beta_{l+1}) > \alpha\beta_l/4 > \beta_l^2/2 > \beta_l\beta_{l+1}$. ■

Proof Proposition 3

(i) From Proposition 1, there is $GC2$ if g is decreasing:

$$\begin{aligned} g\,(z_l) = \frac{e^{-(z_l/\gamma\lambda)^a}}{e^{-(z_l/\lambda)^a}} &< \frac{e^{-(z_{l-1}/\gamma\lambda)^a}}{e^{-(z_{l-1}/\lambda)^a}} = g\,(z_{l-1}) \\ \Leftrightarrow \quad -(z_l/\gamma\lambda)^a + (z_l/\lambda)^a &< -(z_{l-1}/\gamma\lambda)^a + (z_{l-1}/\lambda)^a \\ \Leftrightarrow \quad z_l^a(1/\gamma^a - 1) &> z_{l-1}^a(1/\gamma^a - 1) \\ \Leftrightarrow \quad z_l &> z_{l-1} \end{aligned}$$

which holds for any $(PS)_W$.

(ii) From Proposition 1, there is $GC3$ if $g\,(z_l)^2 > g\,(z_{l-1})\,g\,(z_{l+1})$ for any $l > 2$.

$$\begin{aligned} \frac{e^{-2(z_l/\gamma\lambda)^a}}{e^{-2(z_l/\lambda)^a}} &> \frac{e^{-(z_{l-1}/\gamma\lambda)^a}}{e^{-(z_{l-1}/\lambda)^a}} \frac{e^{-(z_{l+1}/\gamma\lambda)^a}}{e^{-(z_{l+1}/\lambda)^a}} \\ \Leftrightarrow \quad -2\,(z_l/\gamma)^a + 2z_l^a &> -(z_{l-1}/\gamma)^a - (z_{l+1}/\gamma)^a + z_{l-1}^a + z_{l+1}^a \\ \Leftrightarrow \quad -2z_l^a\left(\frac{1}{\gamma^a} - 1\right) &> -z_{l-1}^a\left(\frac{1}{\gamma^a} - 1\right) - z_{l+1}^a\left(\frac{1}{\gamma^a} - 1\right) \\ \Leftrightarrow \quad z_{l+1}^a + z_{l-1}^a &> z_l^a + z_l^a \\ \Leftrightarrow \quad z_{l+1}^a - z_l^a &> z_l^a - z_{l-1}^a \end{aligned}$$

■

Proof Proposition 5

(i) For $GC2$, consider $g(z) = P\,(X > z/\gamma)\,/P\,(X > z)$. For $k = 0$, the exponential case has already been analyzed with the Weibull case. For the rest, it holds that

$$g(z) = \left(1 - \frac{kz}{\gamma\sigma}\right)^{1/k} / \left(1 - \frac{kz}{\sigma}\right)^{1/k}$$

Hence, $g'\,(z) = C(z)\,(\gamma - 1)$ for

$$C(z) = \frac{\left(1 - \frac{kz}{\gamma\sigma}\right)^{\frac{1}{k}-1}\left(1 - \frac{kz}{\sigma}\right)^{\frac{1}{k}-1}}{\sigma\gamma\left(1 - \frac{kz}{\sigma}\right)^{\frac{2}{k}}} > 0$$

so $g'(z) < 0$. Therefore, the function is strictly decreasing, and the result follows from Proposition 1.

(ii) Since $g(z)$ is strictly decreasing, the result follows from Proposition 2. ■

References

Baxter, Janeen, and Erik Olin Wright, 2000. "The glass ceiling hypothesis: A comparative study of the United States, Sweden, and Australia". *Gender and Society* 14(2): 275–294.

Bertrand, Marianne, Claudia Goldin, and Lawrence Katz, 2010. "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors". *American Economic Journal: Applied Economics* 2(3): 228–255.

Espinosa, Maria Paz and Eva Ferreira, 2015. "Gender gap dynamics and glass ceilings". Working paper, Universidad del País Vasco, UPV/EHU.

Espinosa, Maria Paz, Eva Ferreira and Winfried Stute, 2016. "Discrimination, binomials and glass ceiling effects". Springer, Proceedings in Mathematics &Statistics, Vol. 175.

Ferreira, Eva and Winfried Stute, 2016. "Dynamic binomials with an application to gender gap analysis". *Journal of Applied Probability* 53, 82–90.

Fershtman, Chaim and Uri Gneezy. 2001. "Discrimination in a segmented society: An experimental approach". *The Quarterly Journal of Economics* 116: 351–77.

Goldin, Claudia and Cecilia Rouse, 2000. "Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians". *American Economic Review* 90(4): 715–741.

Jackson, Jerlando F. L., and Elizabeth M. O'Callaghan, 2009. "What Do We Know About Glass Ceiling Effects? A Taxonomy and Critical Review to Inform Higher Education Research". *Research in Higher Education* 50:460–482.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman, 2012. "Science faculty's subtle gender biases favor male students". *Proceedings of the National Academy of Sciences (PNAS)* 109(41): 16474–16479.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales, 2014. "How stereotypes impair women's careers in science". *Proceedings of the National Academy of Sciences (PNAS)* 111(12): 4403–4408.

Smith, Ryan A., 2012. "Money, Benefits, and Power : A Test of the Glass Ceiling and Glass Escalator Hypotheses". *The Annals of the American Academy of Political and Social Science* 639: 149–172.