# NONPARAMETRIC ESTIMATION OF THE ANTIMODE AND THE MINIMUM OF A DENSITY FUNCTION

*Hester Loots, M.Sc.*

Thesis submitted to the Faculty of Science in accordance with the requirements for the degree Philosophiae Doctor in the Department of Statistics and Operations Research at Potchefstroom University for Christian Higher Education.

Supervisor: Prof. J.W.H. Swanepoel

POTCHEFSTROOM, SOUTH AFRICA

1995

## ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to the following persons and organisations:

- Prof. J.W.H. Swanepoel, for his valued guidance throughout the study.
- Dr. C.F. de Beer, for valuable discussions and assistance.
- Anton Opperman, Theo Scott and the Department of Statistics, PU for CHE, for the availability of computer facilities.
- Prof. O.C. de Jager, for informative discussions regarding the astrophysical application.
- J.R. Mattox and the Science Support Center of NASA/GSFC, for supplying the astrophysical data.
- The Foundation for Research Development of South Africa, for financial support.
- Prof. A.L. Combrink, for the language editing.
- My husband Jaco and my family, for their continuous interest and support.

*Soli Deo gloria.*

Hester Loots

August 1995

## ABSTRACT

### NONPARAMETRIC ESTIMATION OF THE ANTIMODE AND THE MINIMUM OF A DENSITY FUNCTION

The study of the estimation of the antimode and the minimum of a density function has been neglected in the literature, in spite of their useful applications. The main objective of this thesis is to propose and study nonparametric estimators for these parameters. Strong consistency and limiting distributions are derived. The estimators depend on unknown smoothing parameters. Data-based choices of these smoothing parameters are proposed, using the bootstrap and kernel density estimation techniques. A critical review of data-driven bandwidth selection procedures for kernel density estimation is presented. An extensive Monte Carlo study shows that the small sample behaviour of the newly proposed estimators is very satisfactory. Finally, some applications to real data are discussed.

## OPSOMMING


### NIE-PARAMETRIESE BERAMING VAN DIE ANTIMODUS EN DIE MINIMUM VAN 'N DIGTHEIDSFUNKSIE

Die bestudering van die antimodus en die minimum van 'n digtheidsfunksie het nog weinig aandag in die literatuur geniet. Die hoofdoel van hierdie proefskrif is die voorstel en bestudering van nie-parametriese beramers vir hierdie parameters. Sterk konsekwentheid word aangetoon en limietverdelings word afgelei. Die beramers is afhanklik van onbekende gladstrykingsparameters. Keuses van hierdie gladstrykingsparameters, gegrond op die data, word voorgestel, deur van skoenlus- en kerndigtheidsberamingstegnieke gebruik te maak. 'n Kritiese oorsig van data-gebaseerde gladstrykingsprosedures vir kernberaming word gegee. 'n Uitgebreide Monte Carlo-studie toon dat die kleinsteekproefgedrag van die nuwe voorgestelde beramers baie bevredigend is. Laastens word toepassings op werklike data bespreek.

# Contents

# Chapter 1

# Some direct estimators of the antimode and the minimum of a density function

## 1.1 Introduction

The estimation of the mode and the maximum of a density function has received a considerable amount of attention in the literature during the last few decades. Significant contributions in this area are the results obtained by Parzen (1962), Chernoff (1964), Grenander (1965), Venter (1967), Sager (1975, 1978), Eddy (1980), Romano (1988) and Narayanan and Sager (1989). However, the study of the estimation of the antimode $\theta$ and the minimum $f(\theta)$ of an unknown density function $f$ has been neglected, despite their useful applications. My interest in these two parameters originated from Astrophysics and this study is therefore concluded with relevant real data examples from this field. The behaviour of so-called "maximal spacings" is related to the minimum and to the local behaviour of the density near its minimum. Consequently, the results obtained in this chapter can be applied to obtain new theoretical results for maximal spacings.

Estimators of $f(\theta)$ may be classified as direct or indirect according to their paternity. When the estimator is generated as a by-product from estimating some other quantity,

usually the density $f$ itself, it is called indirect. All the standard density estimators (Wegman (1972), Silverman (1986) and Izenman (1991) have published excellent reviews on nonparametric density estimation methods) provide indirect estimators of $f(\theta)$ by simply minimising the density estimate. In other words, $f$ has to be estimated before $f(\theta)$ can be estimated. The indirect minimum estimator cannot be expressed in closed form. On the other hand, when the estimator is specifically designed for the sole purpose of estimating $f(\theta)$ as a statistical parameter in its own right and can be expressed explicitly, it is called direct. My proposed estimators of $f(\theta)$ can in part be heuristically motivated from the class of density estimators of the histogram type, studied, for example, by Van Ryzin (1973) and Kim and Van Ryzin (1975). However, a special argument enables one to express the estimator in closed form. In this sense the estimator can be viewed as direct. Furthermore, no initial estimation of the density function itself is necessary.

The estimation of $\theta$ may also be classified as indirect or direct. An indirect antimodal estimate is obtained by selecting a value at which a density estimate is minimised. In Chapter 4, the small and moderate sample behaviour of the proposed direct estimators of $\theta$ and $f(\theta)$ are compared with, among others, the indirect estimators based on kernel estimation. The kernel method, introduced by Rosenblatt (1956), is probably the most commonly used density estimation technique and is certainly the best understood mathematically. A background of kernel density estimation in general is given in Chapter 2. It includes a detailed discussion of current choices of the smoothing parameter or bandwidth. The aim of this discussion is to recommend some methods which are probably the best to use currently.

Efron (1979) introduced the well-known resampling procedure called the bootstrap. The bootstrap is a nonparametric computer-orientated technique that is growing more and more popular along with the advancement of computer technology. In Chapter 4 the bootstrap is used to estimate smoothing parameters that appear in the proposed estimators. It is also applied in the construction of confidence intervals for the parameters. Chapter 3 provides a short background of the bootstrap procedure.

In this chapter the direct estimation of the antimode $\theta$ and the minimum $f(\theta)$ of an unknown density function $f$ (with compact support) is studied. The proposed direct

estimators of $\theta$ and $f(\theta)$ are defined in Section 1.3. Strong consistency of the estimators is proved under general conditions. In Section 1.4 almost sure rates of convergence are studied. Interesting and surprising results concerning the limiting distributions of the proposed estimators are derived in Section 1.5. In Section 1.6 a discussion of the relationship between the proposed estimators and maximal spacings is given.

## 1.2 Notation and general assumptions

Let $X_1, X_2, \ldots$, be a sequence of independent and identically distributed random variables on some probability space $(\Omega,\mathcal{F},P)$ with unknown univariate distribution function $F$. Suppose throughout that $F$ is absolutely continuous (with respect to Lebesgue measure) with density $f$. For some finite constants $a$ and $b$, $a < b$, suppose that $f(x) > 0$ for all $x \in [a,b]$ and $f(x) = 0$ otherwise. In this section and the following two sections arguments are used which are reminiscent of those used by Sager (1975), who studied the *mode* of a density function.

**Definition 1.2.1** *The subset $M$ of $[a,b]$ is called the antimodal set of $F$ on $[a,b]$ if*

1. *$f$ is constant on $M$,*
2. *$f(\theta) < f(x)$ for each $x \in [a,b] - M$ and $\theta \in M$,*
3. *for each open set $U$ containing $M$, there exists an $\varepsilon = \varepsilon(U) > 0$ such that $f(x) - \varepsilon \geq f(\theta)$ for each $x \in [a,b] - U$ and $\theta \in M$.*

**Definition 1.2.2** *We say that an absolutely continuous distribution function $F$ satisfies the standard conditions on $[a,b]$ if there is a nonempty antimodal set $M$ in $[a,b]$ such that, for some $\theta \in M$, either*

$$F'_+(\theta) \text{ exists and } f(\theta) = F'_+(\theta), \text{ if } \theta < b, \tag{1.1}$$

*or*

$$F'_-(\theta) \text{ exists and } f(\theta) = F'_-(\theta), \text{ if } \theta > a, \tag{1.2}$$

*where $F'_+(\theta)$ and $F'_-(\theta)$ are the right and left derivatives of $F$ at $\theta$ respectively.*

Denote the order statistics of a random sample $X_1, X_2, \ldots, X_n$ from $F$ by

$$Y_1 \leq Y_2 \leq \ldots \leq Y_n.$$

Let $\{s_n\}$ be a nonrandom sequence of positive integers such that $s_n \to \infty$ as $n \to \infty$. For each $n$, let $K_n$ be a positive integer-valued random variable defined by,

$$Y_{K_n+s_n} - Y_{K_n-s_n} = \max_{s_n+1 \leq j \leq n-s_n} (Y_{j+s_n} - Y_{j-s_n}). \tag{1.3}$$

Note that, since $F$ is absolutely continuous, $K_n$ is unique and $Y_{K_n+s_n} - Y_{K_n-s_n} > 0$ almost surely.

In what follows, $\xrightarrow{a.s.}$ denotes convergence almost surely (with probability one) as $n \to \infty$.

## 1.3 Strong consistency

We propose estimating the antimode $\theta$ by $\hat{\theta}_n$, where $\hat{\theta}_n$ is any statistic satisfying

$$Y_{K_n-s_n} \leq \hat{\theta}_n \leq Y_{K_n+s_n}. \tag{1.4}$$

For example, one can choose

$$\hat{\theta}_n = \tfrac{1}{2}(Y_{K_n-s_n} + Y_{K_n+s_n}) \quad \text{or} \quad \hat{\theta}_n = Y_{K_n}.$$

**Theorem 1.3.1** *Let $F(x)$ satisfy the standard conditions on $[a, b]$, with associated non-empty antimodal set $M$. Suppose*

$$n^{-1}s_n \to 0 \text{ as } n \to \infty, \tag{1.5}$$

$$\sum_{n=1}^{\infty} n\lambda^{s_n} < \infty \quad \text{for all } \lambda, \quad 0 < \lambda < 1, \tag{1.6}$$

*then* $\inf M \leq \liminf_{n\to\infty} Y_{K_n-s_n} \leq \limsup_{n\to\infty} Y_{K_n+s_n} \leq \sup M$ *almost surely.*

First we state a lemma, proved by Sager (1975), that is needed in the proof of Theorem 1.3.1.

**Lemma 1.3.1** *Suppose* $\{l_n\}$ *is a nonrandom sequence of positive integers such that* $l_n \to \infty$, $n^{-1}l_n \to 0$ *as* $n \to \infty$, *and*

$$\sum_{n=1}^{\infty} n\lambda^{l_n} < \infty \quad \textit{for all } \lambda, \quad 0 < \lambda < 1.$$

*Further, let* $S_1, S_2, \ldots,$ *and* $T_1, T_2, \ldots,$ *be sequences of random variables such that* $S_n \leq T_n$ *for each* $n$ *and* $[S_n, T_n]$ *contains exactly* $2l_n + 1$ *of the observations* $X_1, X_2, \ldots, X_n$. *Then,*

$$\frac{F(T_n) - F(S_n)}{F_n(T_n) - F_n(S_n)} \xrightarrow{a.s.} 1 \quad \text{as } n \to \infty,$$

*where* $F_n$ *denotes the empirical distribution function of* $X_1, X_2, \ldots, X_n$.

**Proof of Theorem 1.3.1**

We give the proof when (1.1) holds. The proof for (1.2) is similar.

Choose and fix $\theta \in M$ which satisfies (1.1). For each $n$, let $J_n$ be a discrete random variable defined by the following: if $[\theta, b]$ contains at least $2s_n + 1$ observations, let

$$Y_{J_n+s_n} = \min\{Y_{j+s_n}; j = s_n + 1, \ldots, n - s_n; \theta \leq Y_{j-s_n} \leq Y_{j+s_n} \leq b\}. \tag{1.7}$$

If $[\theta, b]$ contains fewer than $2s_n + 1$ observations, let $J_n = s_n + 1$.

First we note that, since $f(\theta) > 0$, $F$ assigns positive probability to every interval $[\theta, \theta + \varepsilon], \varepsilon > 0$. So, by (1.5) and the strong law of large numbers (SLLN), $Y_{J_n+s_n}$ and $Y_{J_n-s_n}$ converge almost surely to $\theta$.

Consider the following events,

$$\begin{aligned}
\Omega_1 &= [\lim_{n\to\infty}\{F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})\}n(2s_n)^{-1} = 1], \\
\Omega_2 &= [\lim_{n\to\infty}\{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})\}n(2s_n)^{-1} = 1], \\
\Omega_3 &= [\lim_{n\to\infty}\{F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})\}/(Y_{J_n+s_n} - Y_{J_n-s_n}) = f(\theta)], \\
\Omega_4 &= [\lim_{n\to\infty}\{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})\}/(Y_{K_n+s_n} - Y_{K_n-s_n}) = f(\theta)], \\
\Omega_5 &= [\inf M \leq \liminf_{n\to\infty} Y_{K_n-s_n} \leq \limsup_{n\to\infty} Y_{K_n+s_n} \leq \sup M].
\end{aligned} \tag{1.8}$$

The method of proof will be to show that $\Omega_1 \cap \Omega_2 \cap \Omega_3 \subset \Omega_4 \subset \Omega_5$ and that $P(\Omega_1) = P(\Omega_2) = P(\Omega_3) = 1$.

Let $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_3$. Using (1.8) and the fact that

$$0 < Y_{J_n+s_n} - Y_{J_n-s_n} \leq Y_{K_n+s_n} - Y_{K_n-s_n},$$

we have

$$\begin{aligned}
&\frac{\limsup[F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})]/(Y_{J_n+s_n} - Y_{J_n-s_n})}{\limsup[F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})]/(Y_{K_n+s_n} - Y_{K_n-s_n})} \\
&= \limsup\left\{\frac{F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})}{n^{-1}(2s_n)}\right\}\left\{\frac{n^{-1}(2s_n)}{Y_{J_n+s_n} - Y_{J_n-s_n}}\right\}/ \\
&\quad \limsup\left\{\frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{n^{-1}(2s_n)}\right\}\left\{\frac{n^{-1}(2s_n)}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\} \\
&\geq \liminf\left\{\frac{F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})}{n^{-1}(2s_n)}\right\}\limsup\left\{\frac{n^{-1}(2s_n)}{Y_{J_n+s_n} - Y_{J_n-s_n}}\right\}/ \\
&\quad \limsup\left\{\frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{n^{-1}(2s_n)}\right\}\limsup\left\{\frac{n^{-1}(2s_n)}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\} \\
&= \limsup\left\{\frac{n^{-1}(2s_n)}{Y_{J_n+s_n} - Y_{J_n-s_n}}\right\}/\limsup\left\{\frac{n^{-1}(2s_n)}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\} \\
&\geq \liminf\left\{\frac{Y_{K_n+s_n} - Y_{K_n-s_n}}{Y_{J_n+s_n} - Y_{J_n-s_n}}\right\} \\
&\geq 1.
\end{aligned}$$

This implies

$$f(\theta) \geq \limsup_{n\to\infty}\left\{\frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\}.$$

Furthermore,

$$\begin{aligned}
F(Y_{K_n+s_n}) - F(Y_{K_n-s_n}) &= \int_{[Y_{K_n-s_n};Y_{K_n+s_n}]} f(x)dx \\
&\geq \int_{[Y_{K_n-s_n};Y_{K_n+s_n}]} f(\theta)dx \qquad (1.9) \\
&= f(\theta)[Y_{K_n+s_n} - Y_{K_n-s_n}].
\end{aligned}$$

This implies

$$f(\theta) \leq \frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{Y_{K_n+s_n} - Y_{K_n-s_n}},$$

and hence

$$\liminf_{n\to\infty}\left\{\frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\} \geq f(\theta).$$

Hence, we have that

$$\lim_{n\to\infty}\left\{\frac{F(Y_{K_n+s_n})-F(Y_{K_n-s_n})}{Y_{K_n+s_n}-Y_{K_n-s_n}}\right\}=f(\theta).$$

Thus $\omega\in\Omega_4$ which implies that $\Omega_1\cap\Omega_2\cap\Omega_3\subset\Omega_4$.

Using Lemma 1.3.1 we have

$$\begin{aligned}0 &\le \limsup_{n\to\infty}[F(Y_{K_n+s_n})-F(Y_{K_n-s_n})]\\ &= \Big\{\lim_{n\to\infty}[F(Y_{K_n+s_n})-F(Y_{K_n-s_n})]n(2s_n)^{-1}\Big\}\Big\{\limsup_{n\to\infty}(n^{-1}2s_n)\Big\}\\ &= 0\end{aligned}$$

almost surely.

This implies that $F(Y_{K_n+s_n})-F(Y_{K_n-s_n})$ converges almost surely to zero. By (1.9),

$$0\le Y_{K_n+s_n}-Y_{K_n-s_n}\le[F(Y_{K_n+s_n})-F(Y_{K_n-s_n})]/f(\theta),$$

and hence $Y_{K_n+s_n}-Y_{K_n-s_n}$ converges almost surely to zero.

To show that $\Omega_4\subset\Omega_5$, it suffices to show that $\Omega_5^c\subset\Omega_4^c$. Let $\omega\in\Omega_5^c$. Thus there is a subsequence $\{n(j)\}$ such that $[Y_{K_{n(j)}-s_{n(j)}},Y_{K_{n(j)}+s_{n(j)}}]$ lies outside of $(\inf M-\delta,\sup M+\delta)$ for all $j$ large enough and for some $\delta>0$, since $Y_{K_n+s_n}-Y_{K_n-s_n}$ converges to zero.

By Definition 1.2.1(3) this implies that

$$\frac{F(Y_{K_{n(j)}+s_{n(j)}})-F(Y_{K_{n(j)}-s_{n(j)}})}{Y_{K_{n(j)}+s_{n(j)}}-Y_{K_{n(j)}-s_{n(j)}}}\ge f(\theta)+\varepsilon$$

for all large $j$ and for some $\varepsilon>0$. But this implies that $\omega\in\Omega_4^c$.

We now prove the probability statements.

From Lemma 1.3.1 we immediately have $P(\Omega_1)=P(\Omega_2)=1$. To see that $P(\Omega_3)=1$, let

$$C_n=\frac{Y_{J_n+s_n}-\theta}{F(Y_{J_n+s_n})-F(\theta)},$$

and

$$D_n=\frac{Y_{J_n-s_n}-\theta}{F(Y_{J_n-s_n})-F(\theta)}.$$

Write

$$\frac{Y_{J_n+s_n} - Y_{J_n-s_n}}{F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})} = C_n \frac{[F(Y_{J_n+s_n}) - F(\theta)]n(2s_n)^{-1}}{[F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})]n(2s_n)^{-1}} + D_n - D_n \frac{[F(Y_{J_n+s_n}) - F(\theta)]n(2s_n)^{-1}}{[F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})]n(2s_n)^{-1}}. \tag{1.10}$$

Since $Y_{J_n-s_n}$ and $Y_{J_n+s_n}$ converge almost surely to $\theta$, and in view of (1.1), we have that $C_n$ and $D_n$ converge almost surely to $f(\theta)^{-1}$. Thus, setting $S_n = \theta$ and $T_n = Y_{J_n+s_n}$ in Lemma 1.3.1 and using $P(\Omega_1) = 1$, we conclude that the left-hand side of (1.10) converges to $f(\theta)^{-1}$ almost surely. Thus $P(\Omega_3) = 1$. This implies that $P(\Omega_4) = 1$ and $P(\Omega_5) = 1$.

□

The strong consistency of the estimator of $\hat{\theta}_n$ follows directly from the above theorem and we state this as Corollary 1.3.1.

**Corollary 1.3.1** *Under the assumptions of Theorem 1.3.1, if the antimode is unique, i.e., $M = \{\theta\}$, then $Y_{K_n-s_n} \overset{a.s.}{\to} \theta$, $Y_{K_n+s_n} \overset{a.s.}{\to} \theta$ and $\hat{\theta}_n \overset{a.s.}{\to} \theta$ as $n \to \infty$.*

**Remark**

If $\{s_n\}$ is chosen so that $s_n \sim An^{\alpha}$ with $0 < \alpha < 1$ and $A > 0$ then (1.5) and (1.6) hold.

The results obtained in Corollary 1.3.1 can also be derived under different conditions, which do not include the third assumption of Definition 1.2.1 and the assumption that $F$ satisfies the standard conditions on $[a, b]$. In order to do this we first introduce the following definition. Suppose the antimode $\theta \in [a, b]$ is unique.

**Definition 1.3.1** *Let $R_1 > 1$, $R_2 > 1$ and $\delta > 0$ be finite constants. For $a \leq \theta - R_1\delta < \theta - \delta < \theta$ and/or $\theta < \theta + \delta < \theta + R_2\delta \leq b$, define*

$$\alpha(\delta, R_1, R_2) = \frac{\max\{r^+(\delta), 1^+(\delta)\}}{\min\{r^-(R_2\delta), 1^-(R_1\delta)\}},$$

*where*

$$\begin{aligned} r^+(\delta) &= \sup\{f(x); \theta \le x \le \theta + \delta\}, \\ r^-(R_2\delta) &= \inf\{f(x); \theta + R_2\delta \le x \le b\}, \\ 1^+(\delta) &= \sup\{f(x); \theta - \delta \le x \le \theta\}, \\ 1^-(R_1\delta) &= \inf\{f(x); a \le x \le \theta - R_1\delta\}. \end{aligned}$$

*Also, let* $r(\delta, R_2\delta) = r^+(\delta)/r^-(R_2\delta)$ *and* $1(\delta, R_1\delta) = 1^+(\delta)/1^-(R_1\delta)$.

**Theorem 1.3.2** *Suppose (1.5) and (1.6) hold and the antimode* $\theta \in [a, b]$ *is unique.*

*(1) If* $\theta = a$ *and there is a positive constant* $R_2 > 1$ *such that* $r(\delta, R_2\delta) < 1$ *for all small positive* $\delta$, *then* $\hat{\theta}_n \overset{a.s.}{\to} a$ *as* $n \to \infty$.

*(2) If* $\theta = b$ *and there is a positive constant* $R_1 > 1$ *such that* $1(\delta, R_1\delta) < 1$ *for all small positive* $\delta$, *then* $\hat{\theta}_n \overset{a.s.}{\to} b$ *as* $n \to \infty$.

*(3) If* $\theta \in (a, b)$ *and there are positive constants* $R_1 > 1$ *and* $R_2 > 1$ *such that* $\alpha(\delta, R_1, R_2) < 1$ *for all small positive* $\delta$, *then* $\hat{\theta}_n \overset{a.s.}{\to} \theta$ *as* $n \to \infty$.

The following trivial lemma is necessary to prove part (3) of Theorem 1.3.2.

**Lemma 1.3.2** *Let* $\{c_n\}_{n=1}^{\infty}$ *and* $\{d_n\}_{n=1}^{\infty}$ *be sequences of real numbers such that, for some finite constant* $\phi$, $c_n \le \phi \le d_n$ *for all large* $n$ *and* $d_n - c_n = o(1)$ *as* $n \to \infty$. *Then* $d_n = \phi + o(1)$ *and* $c_n = \phi + o(1)$ *as* $n \to \infty$.

**Proof**

Let $\varepsilon > 0$ be arbitrary. Then there exists a positive integer $N(\varepsilon)$ such that for all $n \ge N(\varepsilon), 0 \le d_n - c_n < \varepsilon$. Hence,

$$\phi \le d_n = (d_n - c_n) + c_n \le \varepsilon + \phi,$$

which implies that $d_n \to \phi$ as $n \to \infty$. Similarly, $c_n \to \phi$ as $n \to \infty$.

□

**Proof of Theorem 1.3.2**

Suppose the hypothesis of part (1) of the theorem holds. Let $J_n$ be defined by (1.7) with $\theta = a$. Since $f(\theta) > 0$, $F$ assigns positive probability to every interval $[a, a+\varepsilon]$, $\varepsilon > 0$. So, by (1.5) and the SLLN, $Y_{J_n+s_n}$ and $Y_{J_n-s_n}$ converge almost surely to $a$.

Consider the following events,

$$\begin{aligned}
\Omega_0 &= [\lim_{n\to\infty} Y_{K_n-s_n} = a], \\
\Omega_1 &= \left[\lim_{n\to\infty} \frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{F_n(Y_{K_n+s_n}) - F_n(Y_{K_n-s_n})} = 1\right], \\
\Omega_2 &= \left[\lim_{n\to\infty} \frac{F(Y_{J_n+s_n}) - F(Y_{J_n-s_n})}{F_n(Y_{J_n+s_n}) - F_n(Y_{J_n-s_n})} = 1\right], \\
\Omega_3 &= [\lim_{n\to\infty} Y_{J_n+s_n} = a].
\end{aligned}$$

By Lemma 1.3.1 with $S_n = Y_{K_n-s_n}$ and $T_n = Y_{K_n+s_n}$, we know that $P(\Omega_1) = 1$ and with $S_n = Y_{J_n-s_n}$ and $T_n = Y_{J_n+s_n}$, we have $P(\Omega_2) = 1$. Also, $P(\Omega_3) = 1$. Next, we show that $P(\Omega_0) = 1$.

It suffices to show that $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_0^c = \emptyset$. Suppose that $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_0^c$. Since $\omega \in \Omega_0^c$, there is an $\varepsilon > 0$ and a subsequence $\{n(j)\}$ such that

$$Y_{K_{n(j)}-s_{n(j)}} > a + R_2\varepsilon \text{ for all } j \text{ large enough.} \tag{1.11}$$

From the definition of $r(\cdot,\cdot)$, it follows that there exists a $\delta_0 > 0$ such that $r(\delta_0, R_2\delta_0) < 1$ and $\delta_0 < \varepsilon$. Hence, from (1.11) and the fact that $P(\Omega_1) = P(\Omega_2) = P(\Omega_3) = 1$, we obtain

$$\begin{aligned}
1 &> r(\delta_0, R_2\delta_0) \\
&= \frac{r^+(\delta_0)}{r^-(R_2\delta_0)} \\
&\geq \left\{\frac{r^+(\delta_0)}{r^-(R_2\delta_0)}\right\}\left\{\frac{Y_{J_{n(j)}+s_{n(j)}} - Y_{J_{n(j)}-s_{n(j)}}}{Y_{K_{n(j)}+s_{n(j)}} - Y_{K_{n(j)}-s_{n(j)}}}\right\} \\
&= \frac{\int_{[Y_{J_{n(j)}-s_{n(j)}};Y_{J_{n(j)}+s_{n(j)}}]} r^+(\delta_0)dx}{\int_{[Y_{K_{n(j)}-s_{n(j)}};Y_{K_{n(j)}+s_{n(j)}}]} r^-(R_2\delta_0)dx} \\
&\geq \frac{\int_{[Y_{J_{n(j)}-s_{n(j)}};Y_{J_{n(j)}+s_{n(j)}}]} f(x)dx}{\int_{[Y_{K_{n(j)}-s_{n(j)}};Y_{K_{n(j)}+s_{n(j)}}]} f(x)dx} \\
&= \frac{[F(Y_{J_{n(j)}+s_{n(j)}}) - F(Y_{J_{n(j)}-s_{n(j)}})]n(j)(2s_{n(j)})^{-1}}{[F(Y_{K_{n(j)}+s_{n(j)}}) - F(Y_{K_{n(j)}-s_{n(j)}})]n(j)(2s_{n(j)})^{-1}} \\
&\xrightarrow{a.s.} 1 \text{ as } j \to \infty,
\end{aligned}$$

which leads to a contradiction. Hence, $P(\Omega_0) = 1$.

As in the proof of Theorem 1.3.1, we conclude that

$$Y_{K_n+s_n} - Y_{K_n-s_n} \overset{a.s.}{\to} 0 \text{ as } n \to \infty. \tag{1.12}$$

Thus,

$$Y_{K_n+s_n} = Y_{K_n+s_n} - Y_{K_n-s_n} + Y_{K_n-s_n} \to a,$$

almost surely. Part (1) of the theorem now follows since $Y_{K_n-s_n} \le \hat{\theta}_n \le Y_{K_n+s_n}$ (see (1.4)).

The proof of part (2) is similar to that of part (1).

Now, suppose the hypothesis of part (3) holds. Since

$$\alpha(\delta, R_1, R_2) \ge \max\{r(\delta, R_2\delta), 1(\delta, R_1\delta)\},$$

it follows that for all small positive $\delta$, $r(\delta, R_2\delta) < 1$ and $1(\delta, R_1\delta) < 1$. Thus parts (1) and (2) of the theorem are applicable.

Define $I_n$ and $L_n$ by the following: If $[a, \theta]$ contains at least $2s_n + 1$ observations, let

$$\begin{aligned} &Y_{I_n+s_n} - Y_{I_n-s_n} \\ &\quad = \max(Y_{j+s_n} - Y_{j-s_n}; j = s_n + 1, \ldots, n - s_n; a \le Y_{j-s_n} \le Y_{j+s_n} \le \theta). \end{aligned} \tag{1.13}$$

If $[a, \theta]$ contains fewer than $2s_n + 1$ observations, let $I_n = s_n + 1$. If $[\theta, b]$ contains at least $2s_n + 1$ observations, let

$$\begin{aligned} &Y_{L_n+s_n} - Y_{L_n-s_n} \\ &\quad = \max(Y_{j+s_n} - Y_{j-s_n}; j = s_n + 1, \ldots, n - s_n; \theta \le Y_{j-s_n} \le Y_{j+s_n} \le b). \end{aligned} \tag{1.14}$$

If $[\theta, b]$ contains fewer than $2s_n + 1$ observations, let $L_n = s_n + 1$.

By part (1) of the theorem and (1.12), we have that $Y_{L_n-s_n}$ and $Y_{L_n+s_n}$ converge to $\theta$ almost surely. Similarly (by using part (2) of the theorem) it follows that $Y_{I_n-s_n}$ and $Y_{I_n+s_n}$ converge to $\theta$ almost surely. Let $\{n(i)\}_{i=1}^{\infty}$, $\{m(j)\}_{j=1}^{\infty}$ and $\{l(k)\}_{k=1}^{\infty}$ be subsequences of $\{1, 2, \ldots\}$ such that $\{n(i)\}_{i=1}^{\infty} \cup \{m(j)\}_{j=1}^{\infty} \cup \{l(k)\}_{k=1}^{\infty} = \{1, 2, \ldots\}$ and $Y_{K_{n(i)}+s_{n(i)}} \le \theta$ for each $i$, $Y_{K_{m(j)}-s_{m(j)}} \ge \theta$ for each $j$ and $Y_{K_{l(k)}-s_{l(k)}} \le \theta \le Y_{K_{l(k)}+s_{l(k)}}$ for each $k$.

Since $[Y_{K_{n(i)}-s_{n(i)}}, Y_{K_{n(i)}+s_{n(i)}}] \subset [a, \theta]$ it follows that $Y_{K_{n(i)}-s_{n(i)}} = Y_{I_{n(i)}-s_{n(i)}}$ and $Y_{K_{n(i)}+s_{n(i)}} = Y_{I_{n(i)}+s_{n(i)}}$. Hence, $\hat{\theta}_{n(i)} \stackrel{a.s.}{\to} \theta$ as $i \to \infty$. It follows similarly that $\hat{\theta}_{m(j)} \stackrel{a.s.}{\to} \theta$ as $j \to \infty$. Since $Y_{K_{l(k)}-s_{l(k)}} \leq \theta \leq Y_{K_{l(k)}+s_{l(k)}}$ for each $k$, it follows from Lemma 1.3.2 and (1.12) that $Y_{K_{l(k)}-s_{l(k)}}$ and $Y_{K_{l(k)}+s_{l(k)}}$ both converge to $\theta$ almost surely. This implies that $\hat{\theta}_{l(k)} \stackrel{a.s.}{\to} \theta$ as $k \to \infty$. Hence, $\hat{\theta}_n \stackrel{a.s.}{\to} \theta$ as $n \to \infty$.

□

Let us turn attention to estimation of the minimum $f(\theta)$ of an unknown density $f$. Let $\{r_n\}$ be another nonrandom sequence of positive integers such that $r_n \to \infty$ as $n \to \infty$. We propose estimating $f(\theta)$ by

$$\eta_n = \frac{n^{-1}(2r_n + 1)}{Y_{K_n+r_n} - Y_{K_n-r_n}}, \tag{1.15}$$

where $K_n$ is defined by (1.3).

The definition of $\eta_n$ is motivated heuristically as follows. Let $A_n(x) = \sum_{i=1}^n I(Y_i \leq x)$ and suppose $t \geq 1$ is some integer depending on $n$ ($I(B)$ is the indicator function of the event $B$). Van Ryzin (1973) en Kim and Van Ryzin (1975) proposed and studied the following nonparametric estimator of $f$,

$$\hat{f}_{n,t}(x) = \frac{n^{-1}(2t + 1)}{Y_{A_n(x)+t} - Y_{A_n(x)-t}},$$

for $x \in I_t = [Y_{t+1}, Y_{n-t+1})$. Since we are interested in estimating $f(\theta) = \inf_x f(x)$, consider

$$\begin{aligned} \inf\{\hat{f}_{n,s_n}(x) : x \in I_{s_n}\} &= \frac{n^{-1}(2s_n + 1)}{\max_{s_n+1 \leq j \leq n-s_n}(Y_{j+s_n} - Y_{j-s_n})} \\ &= \frac{n^{-1}(2s_n + 1)}{Y_{K_n+s_n} - Y_{K_n-s_n}} \\ &= \hat{f}_{n,s_n}(Y_{K_n}). \end{aligned}$$

Note that $\eta_n = \hat{f}_{n,r_n}(Y_{K_n})$. The incorporation of the sequence $\{r_n\}$ in the definition of $\eta_n$ allows the estimator to be "more flexible" and some recommendations regarding the choice of $\{r_n\}$ (and $\{s_n\}$) will be made in the theorems and the numerical studies below. If, for example, $\{r_n\}$ and $\{s_n\}$ are chosen such that $r_n < s_n$ for all $n$, then Theorem

1.5.2 shows that, under certain regularity assumptions, $\eta_n$ is asymptotically (as $n \to \infty$) normally distributed!

The following theorem shows that $\eta_n$ is strongly consistent.

**Theorem 1.3.3** *Suppose (1.5) holds, $r_n \leq s_n$ for all $n$, and*

$$\sum_{n=1}^{\infty} n\lambda^{r_n} < \infty \quad \textit{for all } \lambda,\ 0 < \lambda < 1.$$

*Further, suppose the antimode is unique, i.e., the antimodal set (see Definition 1.2.1) $M = \{\theta\}$, and $F$ has a first derivative in some neighborhood of $\theta \in [a, b]$ with $F'$ continuous at $\theta$ and $f(\theta) = F'(\theta)$. Then $\eta_n \overset{a.s.}{\to} f(\theta)$ as $n \to \infty$.*

**Proof**

Using the mean-value theorem, it follows that

$$F(Y_{K_n+r_n}) - F(Y_{K_n-r_n}) = (Y_{K_n+r_n} - Y_{K_n-r_n})F'(\alpha_n),$$

where $Y_{K_n-r_n} \leq \alpha_n \leq Y_{K_n+r_n}$. Now, since $Y_{K_n-s_n} \leq Y_{K_n-r_n} \leq Y_{K_n+r_n} \leq Y_{K_n+s_n}$, we obtain from Corollary 1.3.1 that $Y_{K_n-r_n}$ and $Y_{K_n+r_n}$ converge almost surely to $\theta$. Hence, as $n \to \infty$

$$\frac{F(Y_{K_n+r_n}) - F(Y_{K_n-r_n})}{Y_{K_n+r_n} - Y_{K_n-r_n}} \overset{a.s.}{\to} f(\theta). \tag{1.16}$$

Using Lemma 1.3.1 and (1.16), it follows that

$$\begin{aligned}
\eta_n &= \frac{n^{-1}(2r_n+1)}{Y_{K_n+r_n} - Y_{K_n-r_n}} \\
&= \left\{\frac{F_n(Y_{K_n+r_n}) - F_n(Y_{K_n-r_n})}{Y_{K_n+r_n} - Y_{K_n-r_n}}\right\}\left\{\frac{n^{-1}(2r_n+1)}{n^{-1}(2r_n)}\right\} \\
&= \left\{\frac{F_n(Y_{K_n+r_n}) - F_n(Y_{K_n-r_n})}{F(Y_{K_n+r_n}) - F(Y_{K_n-r_n})}\right\}\left\{\frac{F(Y_{K_n+r_n}) - F(Y_{K_n-r_n})}{Y_{K_n+r_n} - Y_{K_n-r_n}}\right\}\left\{\frac{2r_n+1}{2r_n}\right\} \qquad (1.17) \\
&\overset{a.s.}{\to} f(\theta).
\end{aligned}$$

□

**Remark**

Suppose $r_n = s_n$ for all $n$, and $F$ satisfies the standard conditions on $[a, b]$, with associated nonempty antimodal set $M$. Without any further assumptions on $F$, the strong consistency of $\eta_n$ follows directly from (1.17) by applying Lemma 1.3.1 and the fact that $P(\Omega_4) = 1$ (see (1.8)).

## 1.4 Strong convergence rates

In this section we retain all the background and assumptions of the previous sections, except assumptions (1.1), (1.2) and the third requirement of Definition 1.2.1. Let $r(\delta, R_2\delta)$, $1(\delta, R_1\delta)$ and $\alpha(\delta, R_1, R_2)$ be defined as in Definition 1.3.1.

**Theorem 1.4.1** *Suppose the antimode $\theta \in [a, b]$ is unique and its estimator $\hat{\theta}_n$ is defined as in (1.4). Let $s_n$ be of the form $An^{2k/(1+2k)}$ for some $A > 0$, and set $\delta_n = n^{-1/(1+2k)}(\log n)^{1/k}$, for $k$ specified below.*

*(1) If $\theta = a$ and there are positive constants $R_2 > 1, \rho$ and $k$ such that $r(\delta, R_2\delta) \leq 1 - \rho\delta^k$ for all small positive $\delta$, then $\hat{\theta}_n = a + o(\delta_n)$ almost surely.*

*(2) If $\theta = b$ and there are positive constants $R_1 > 1, \rho$ and $k$ such that $1(\delta, R_1\delta) \leq 1 - \rho\delta^k$ for all small positive $\delta$, then $\hat{\theta}_n = b + o(\delta_n)$ almost surely.*

*(3) If $\theta \in (a, b)$ and there are positive constants $R_1 > 1, R_2 > 1, \rho$ and $k$ such that $\alpha(\delta, R_1, R_2) \leq 1 - \rho\delta^k$ for all small positive $\delta$, then $\hat{\theta}_n = \theta + o(\delta_n)$ almost surely.*

The following lemmas will be needed for the proof of the theorem. Lemma 1.4.1 was proved by Sager (1975).

**Lemma 1.4.1** *Let $S_1, S_2, \ldots$, and $T_1, T_2, \ldots$, be sequences of random variables such that $S_n \leq T_n$ for each $n$ and $[S_n, T_n]$ contains exactly $2l_n+1$ of the observations $X_1, X_2, \ldots, X_n$, where $\{l_n\}$ is a nonrandom sequence of positive integers of the form $An^\alpha$, for some finite constants $A > 0$ and $0 < \alpha < 1$. Then, as $n \to \infty$*

$$\frac{F(T_n) - F(S_n)}{F_n(T_n) - F_n(S_n)} = 1 + o(l_n^{-1/2} \log l_n) \quad \text{a.s.},$$

*where $F_n$ denotes the empirical distribution function of $X_1, X_2, \ldots, X_n$.*

**Lemma 1.4.2** *Suppose the hypothesis of part (1) of Theorem 1.4.1 holds. Let $J_n$ be defined by (1.7) with $\theta = a$. Then, as $n \to \infty$*

$$Y_{J_n+s_n} = a + o(\delta_n) \quad \text{a.s.}$$

**Proof**

Let $\varepsilon > 0$ be arbitrary. Since $f(x) > f(a) > 0$ for each $x \in (a, b]$, we obtain

$$\begin{aligned} F(a+\varepsilon\delta_n) - F(a) &= \int_{(a,a+\varepsilon\delta_n]} f(x)dx \\ &> \int_{(a,a+\varepsilon\delta_n]} f(a)dx \\ &= f(a)\varepsilon\delta_n. \end{aligned}$$

This implies that

$$\liminf_{n\to\infty} \left\{ \frac{F(a+\varepsilon\delta_n) - F(a)}{f(a)\varepsilon\delta_n} \right\} \geq 1. \tag{1.18}$$

By Lemma 1.3.1, we have

$$\{F(Y_{J_n+s_n}) - F(a)\} n(2s_n)^{-1} \xrightarrow{a.s.} 1. \tag{1.19}$$

But $n^{-1}(2s_n)/\{f(a)\varepsilon\delta_n\} \to 0$ as $n \to \infty$, so by (1.19) we have

$$\frac{F(Y_{J_n+s_n}) - F(a)}{f(a)\varepsilon\delta_n} \xrightarrow{a.s.} 0. \tag{1.20}$$

From (1.18) and (1.20) we deduce that

$$\limsup_{n\to\infty} \left\{ \frac{F(Y_{J_n+s_n}) - F(a)}{f(a)\varepsilon\delta_n} \right\} < \liminf_{n\to\infty} \left\{ \frac{F(a+\varepsilon\delta_n) - F(a)}{f(a)\varepsilon\delta_n} \right\},$$

and hence that $F(a+\varepsilon\delta_n) > F(Y_{J_n+s_n})$ for all large $n$, almost surely. Since $\varepsilon$ is arbitrary, this implies $Y_{J_n+s_n} = a + o(\delta_n)$ almost surely.

□

Note that, since $a \leq Y_{J_n-s_n} \leq Y_{J_n+s_n}$, we also have that $Y_{J_n-s_n} = a + o(\delta_n)$ almost surely.

**Lemma 1.4.3** *Suppose that the hypothesis of part (1) of Theorem 1.4.1 holds. Then, as* $n \to \infty$

$$Y_{K_n - s_n} = a + o(\delta_n) \quad \text{a.s.},$$

*where* $K_n$ *is defined in (1.3).*

**Proof**

Consider the following events,

$$\begin{aligned}
\Omega_0 &= [Y_{K_n - s_n} = a + o(\delta_n)], \\
\Omega_1 &= \left[\frac{F(Y_{K_n + s_n}) - F(Y_{K_n - s_n})}{F_n(Y_{K_n + s_n}) - F_n(Y_{K_n - s_n})} = 1 + o(s_n^{-1/2} \log s_n)\right], \\
\Omega_2 &= \left[\frac{F(Y_{J_n + s_n}) - F(Y_{J_n - s_n})}{F_n(Y_{J_n + s_n}) - F_n(Y_{J_n - s_n})} = 1 + o(s_n^{-1/2} \log s_n)\right], \\
\Omega_3 &= [Y_{J_n + s_n} = a + o(\delta_n)].
\end{aligned}$$

By Lemma 1.4.1 with $S_n = Y_{K_n - s_n}$ and $T_n = Y_{K_n + s_n}$, we know that $P(\Omega_1) = 1$ and with $S_n = Y_{J_n - s_n}$ and $T_n = Y_{J_n + s_n}$, we have $P(\Omega_2) = 1$. By Lemma 1.4.2 we know that $P(\Omega_3) = 1$. Next, we show that $P(\Omega_0) = 1$.

It suffices to show that $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_0^c = \emptyset$. Suppose that $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_0^c$. Since $\omega \in \Omega_0^c$, there is an $\varepsilon > 0$ and a subsequence $\{n(j)\}$ such that

$$Y_{K_{n(j)} - s_{n(j)}} > a + R_2 \varepsilon \delta_{n(j)} \text{ for all } j \text{ large enough.} \tag{1.21}$$

Using the hypothesis of (1), Lemma 1.4.2, (1.21) and Lemma 1.4.1, we have

$$\begin{aligned}
1 - \rho(\varepsilon \delta_{n(j)})^k &\geq r(\varepsilon \delta_{n(j)}, R_2 \varepsilon \delta_{n(j)}) \\
&= \frac{r^+(\varepsilon \delta_{n(j)})}{r^-(R_2 \varepsilon \delta_{n(j)})} \\
&\geq \left\{\frac{r^+(\varepsilon \delta_{n(j)})}{r^-(R_2 \varepsilon \delta_{n(j)})}\right\} \left\{\frac{Y_{J_{n(j)} + s_{n(j)}} - Y_{J_{n(j)} - s_{n(j)}}}{Y_{K_{n(j)} + s_{n(j)}} - Y_{K_{n(j)} - s_{n(j)}}}\right\} \\
&= \frac{\int_{[Y_{J_{n(j)} - s_{n(j)}}; Y_{J_{n(j)} + s_{n(j)}}]} r^+(\varepsilon \delta_{n(j)}) dx}{\int_{[Y_{K_{n(j)} - s_{n(j)}}; Y_{K_{n(j)} + s_{n(j)}}]} r^-(R_2 \varepsilon \delta_{n(j)}) dx} \\
&\geq \frac{\int_{[Y_{J_{n(j)} - s_{n(j)}}; Y_{J_{n(j)} + s_{n(j)}}]} f(x) dx}{\int_{[Y_{K_{n(j)} - s_{n(j)}}; Y_{K_{n(j)} + s_{n(j)}}]} f(x) dx}
\end{aligned}$$

$$\begin{aligned} &= \frac{[F(Y_{J_{n(j)}+s_{n(j)}}) - F(Y_{J_{n(j)}-s_{n(j)}})]n(j)(2s_{n(j)})^{-1}}{[F(Y_{K_{n(j)}+s_{n(j)}}) - F(Y_{K_{n(j)}-s_{n(j)}})]n(j)(2s_{n(j)})^{-1}} \\ &= 1 + o(s_{n(j)}^{-1/2} \log s_{n(j)}) \\ &= 1 + o(n(j)^{-k/(1+2k)} \log n(j)). \end{aligned}$$

However, $1 - \rho(\varepsilon\delta_{n(j)})^k = 1 - \rho\varepsilon^k n(j)^{-k/(1+2k)} \log n(j)$, which contradicts the above inequality for large $j$. Hence, $P(\Omega_0) = 1$.

□

The proof of the following trivial lemma is analogous to that of Lemma 1.3.2 and will therefore be omitted.

**Lemma 1.4.4** *Let $\{c_n\}_{n=1}^{\infty}$, $\{d_n\}_{n=1}^{\infty}$ and $\{\lambda_n\}_{n=1}^{\infty}$ be sequences of real numbers such that, for some finite constant $\phi$, $c_n \leq \phi \leq d_n$ for all large $n$ and $d_n - c_n = o(\lambda_n)$ as $n \to \infty$. Then $d_n = \phi + o(\lambda_n)$ and $c_n = \phi + o(\lambda_n)$ as $n \to \infty$.*

**Proof of Theorem 1.4.1**

By (1.9),

$$\begin{aligned} 0 &\leq Y_{K_n+s_n} - Y_{K_n-s_n} \\ &\leq \{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})\}/f(a), \end{aligned}$$

and Lemma 1.4.1 implies that

$$F(Y_{K_n+s_n}) - F(Y_{K_n-s_n}) = n^{-1}(2s_n) + o(n^{-1}s_n) = o(\delta_n).$$

Hence, as $n \to \infty$

$$Y_{K_n+s_n} - Y_{K_n-s_n} = o(\delta_n) \quad \text{a.s.} \tag{1.22}$$

This, together with Lemma 1.4.3, yield

$$Y_{K_n+s_n} = Y_{K_n+s_n} - Y_{K_n-s_n} + Y_{K_n-s_n} = a + o(\delta_n),$$

almost surely. Part (1) of the theorem now follows easily, since

$$0 \leq \hat{\theta}_n - a \leq Y_{K_n+s_n} - a = o(\delta_n),$$

almost surely. The proof of part (2) is similar to that of part (1).

Now, suppose the hypothesis of part (3) holds. Since

$$\alpha(\delta, R_1, R_2) \geq \max\{r(\delta, R_2\delta), 1(\delta, R_1\delta)\},$$

it follows that for all small positive $\delta$, $r(\delta, R_2\delta) \leq 1 - \rho\delta^k$ and $1(\delta, R_1\delta) \leq 1 - \rho\delta^k$. Thus parts (1) and (2) of the theorem are applicable.

Define $I_n$ and $L_n$ as in (1.13) and (1.14). Since $Y_{L_n+s_n} - Y_{L_n-s_n} = o(\delta_n)$ almost surely (which follows as in (1.22)), we obtain from Lemma 1.4.3 that

$$Y_{L_n-s_n} = \theta + o(\delta_n); \;\; Y_{L_n+s_n} = \theta + o(\delta_n), \tag{1.23}$$

almost surely. Similarly, we obtain that

$$Y_{I_n-s_n} = \theta + o(\delta_n); \;\; Y_{I_n+s_n} = \theta + o(\delta_n), \tag{1.24}$$

almost surely.

Similarly, as in the proof of part (3) of Theorem 1.3.2, by using (1.22), (1.23), (1.24) and Lemma 1.4.4, it follows that $\hat{\theta}_n = \theta + o(\delta_n)$ almost surely.

□

We now derive strong convergence rates for $\eta_n$ (defined in (1.15)), the estimator of $f(\theta)$.

**Theorem 1.4.2** *Suppose the assumptions of Theorem 1.4.1 hold and $r_n = s_n$ for all $n$. If $F$ has a bounded second derivative in some neighborhood of $\theta \in [a, b]$ and $f(\theta) = F'(\theta)$, then*

$$\eta_n = \begin{cases} f(\theta) + o(\delta_n), & \text{if } k \geq 1, \\ f(\theta) + o(\delta_n^k), & \text{if } k \leq 1, \end{cases}$$

*almost surely.*

**Proof**

Using Lemma 1.4.1 with $T_n = Y_{K_n+s_n}$ and $S_n = Y_{K_n-s_n}$, it follows that

$$\begin{aligned} \frac{F_n(Y_{K_n+s_n}) - F_n(Y_{K_n-s_n})}{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})} &= 1 + o(s_n^{-1/2} \log s_n) \\ &= 1 + o(\delta_n^k). \end{aligned}$$

Also, using a Taylor expansion,

$$F(Y_{K_n+s_n}) = F(\theta) + (Y_{K_n+s_n} - \theta)F'(\theta) + \tfrac{1}{2}(Y_{K_n+s_n} - \theta)^2 F''(\alpha_n),$$

where $\alpha_n$ is a point between $\theta$ and $Y_{K_n+s_n}$. Similarly,

$$F(Y_{K_n-s_n}) = F(\theta) + (Y_{K_n-s_n} - \theta)F'(\theta) + \tfrac{1}{2}(Y_{K_n-s_n} - \theta)^2 F''(\beta_n),$$

where $\beta_n$ is a point between $\theta$ and $Y_{K_n-s_n}$. Hence,

$$\begin{aligned} F(Y_{K_n+s_n}) - F(Y_{K_n-s_n}) &= (Y_{K_n+s_n} - Y_{K_n-s_n})F'(\theta) + \tfrac{1}{2}(Y_{K_n+s_n} - \theta)^2 F''(\alpha_n) \\ &\quad - \tfrac{1}{2}(Y_{K_n-s_n} - \theta)^2 F''(\beta_n). \end{aligned}$$

Using the definition of $\hat{\theta}_n$ (see (1.4)) and (1.22), we have

$$\begin{aligned} \hat{\theta}_n - \theta &\leq Y_{K_n+s_n} - \theta \\ &= Y_{K_n+s_n} - Y_{K_n-s_n} + Y_{K_n-s_n} - \theta \\ &= o(\delta_n) + Y_{K_n-s_n} - \theta \\ &\leq o(\delta_n) + \hat{\theta}_n - \theta. \end{aligned}$$

This, together with part (3) of Theorem 1.4.1, implies that $Y_{K_n+s_n} = \theta + o(\delta_n)$ almost surely. Also, (1.22) implies that $Y_{K_n-s_n} = \theta + o(\delta_n)$ almost surely. Hence,

$$\frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{Y_{K_n+s_n} - Y_{K_n-s_n}} = f(\theta) + o(\delta_n).$$

It now follows that

$$\begin{aligned} \eta_n &= \frac{n^{-1}(2s_n+1)}{Y_{K_n+s_n} - Y_{K_n-s_n}} \\ &= \left\{\frac{F_n(Y_{K_n+s_n}) - F_n(Y_{K_n-s_n})}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\}\left\{\frac{n^{-1}(2s_n+1)}{n^{-1}(2s_n)}\right\} \\ &= \left\{\frac{F_n(Y_{K_n+s_n}) - F_n(Y_{K_n-s_n})}{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}\right\}\left\{\frac{F(Y_{K_n+s_n}) - F(Y_{K_n-s_n})}{Y_{K_n+s_n} - Y_{K_n-s_n}}\right\}\left\{1 + \frac{1}{2s_n}\right\} \\ &= \{1 + o(\delta_n^k)\}\{f(\theta) + o(\delta_n)\}\{1 + O(\delta_n^{2k}/(\log n)^2)\}. \end{aligned}$$

Consequently,

$$\eta_n = \begin{cases} f(\theta) + o(\delta_n), & \text{if } k \geq 1, \\ f(\theta) + o(\delta_n^k), & \text{if } k \leq 1, \end{cases}$$

almost surely.

□

## 1.5 Asymptotic distributions

In this section the following will be assumed without further statement: For some finite constants $a$ and $b$, $a < b$, $f(x) > 0$ for all $x \in (a, b)$ and $f(x) = 0$ otherwise. There exists a $\theta \in (a, b)$ such that, for all $x \in (a, b)$, $x \neq \theta$, $f(x) > f(\theta) > 0$.

Define $K_n$ as in (1.3), viz.,

$$Y_{K_n+s_n} - Y_{K_n-s_n} = \max_{s_n+1 \leq j \leq n-s_n} (Y_{j+s_n} - Y_{j-s_n}),$$

where $\{s_n\}$ is a nonrandom sequence of positive integers such that $s_n \to \infty$ as $n \to \infty$. Throughout the discussion below we consider

$$\hat{\theta}_n = Y_{K_n}$$

as estimator of $\theta$, and $\eta_n$, the proposed estimator of $f(\theta)$, as defined in (1.15).

In Theorems 1.5.1 and 1.5.2 limiting distributions are derived for $\hat{\theta}_n$ and $\eta_n$. Firstly, we prove two lemmas that are needed for the proof of Theorem 1.5.1.

It is well-known that $F(Y_1), F(Y_2), \ldots, F(Y_n)$ may be thought of as the order statistics of an independent sample from the uniform distribution on $[0, 1]$ and that the vector $(F(Y_1), F(Y_2), \ldots, F(Y_n))$ has the same distribution as

$$\left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\right),$$

where

$$S_i = Z_1 + Z_2 + \cdots + Z_i, \quad i = 1, 2, \ldots, n+1,$$

with $Z_1, Z_2, \ldots, Z_{n+1}$ independent random variables, each with a standard exponential distribution (e.g., see David, 1981). Hence, writing $G = F^{-1}$ (the inverse exists, since $F$ is continuous and strictly increasing), the vector $(Y_1, Y_2, \ldots, Y_n)$ has the same distribution as

$$\left(G\left(\frac{S_1}{S_{n+1}}\right), G\left(\frac{S_2}{S_{n+1}}\right), \ldots, G\left(\frac{S_n}{S_{n+1}}\right)\right).$$

Hence, since we intend deriving limiting distributions for $\hat{\theta}_n$ and $\eta_n$, we can replace $Y_i$, $i = 1, 2, \ldots, n$ in all proofs by $G(S_i/S_{n+1})$, $i = 1, 2, \ldots, n$. Let $\tilde{K}_n$ be defined as $K_n$ above by applying this representation. Then $\tilde{K}_n$ and $K_n$ have the same distribution.

Also, for example, $Y_{K_n}$ and $G(S_{\tilde{K}_n}/S_{n+1})$ have the same distribution, etc. For ease of notation we shall not distinguish between $K_n$ and $\tilde{K}_n$, and if two statistics have the same distribution, it will merely be denoted by an equality sign. The almost sure results obtained for statistics in terms of the $Y_i$'s now hold in probability for these statistics defined in terms of the $S_i$'s.

**Lemma 1.5.1** *Suppose* $n^{-1}s_n \to 0$ *as* $n \to \infty$ *and* $n^{-1}(s_n+1) \le p \le 1 - n^{-1}s_n$. *Then,*

$$\sup_p |S_{[np]+s_n} S_{n+1}^{-1} - p| = n^{-1}s_n + o(n^{-1/2}\log n) \quad \text{a.s.}, \tag{1.25}$$

*and*

$$\sup_p |S_{[np]-s_n} S_{n+1}^{-1} - p| = -n^{-1}s_n + o(n^{-1/2}\log n) \quad \text{a.s.}, \tag{1.26}$$

*where* $[z]$ *denotes the largest integer less than or equal to* $z$.

**Proof**

Write

$$\begin{aligned} \frac{S_{[np]+s_n}}{S_{n+1}} - p &= \frac{S_{[np]+s_n} - [np] - s_n}{S_{n+1}} \\ &\quad + \frac{[np]-np}{S_{n+1}} + p\left\{\frac{n}{S_{n+1}} - 1\right\} + \frac{s_n}{S_{n+1}}. \end{aligned} \tag{1.27}$$

From the law of the iterated logarithm (e.g., see Breiman, 1968), it follows that $S_{n+1} = n + O(n^{1/2}(\log_2 n)^{1/2})$ almost surely. Hence, the second and third terms on the right-hand side of (1.27) are almost surely $o(n^{-1/2}\log n)$ uniformly in $p$, while the last term is $n^{-1}s_n + o(n^{-1/2}\log n)$, almost surely. Consider the first term. We have

$$\begin{aligned} &P\{\sup_p |S_{[np]+s_n} - [np] - s_n| > \epsilon n^{1/2}\log n\} \\ &\le P\{\text{for some } j, s_n+1 \le j \le n - s_n, |S_j - j| > \epsilon n^{1/2}\log n\} \\ &\le \sum_{j=1}^{n} P\{S_j > j + \epsilon n^{1/2}\log n\} + \sum_{j=1}^{n} P\{S_j < j - \epsilon n^{1/2}\log n\}. \end{aligned} \tag{1.28}$$

For any random variable $X$ and any constant $c$,

$$P(X > c) \le E\exp\{t(X-c)\}, \quad t > 0,$$

provided that this expectation exists. Applying this inequality to $S_j$ which has density $x^{j-1}e^{-x}/\Gamma(j)$ for $x \geq 0$ and 0 otherwise, $\Gamma(\cdot)$ being the gamma function, we get

$$P\{S_j > j + \varepsilon n^{1/2}\log n\} \leq (1-t)^{-j}e^{-t(j+\varepsilon n^{1/2}\log n)}, \quad 0 < t < 1.$$

Summing over $j$ between 1 and $n$, we find that the first sum on the right in (1.28) is bounded by $h\lambda^n$ where

$$h = \{1 - (1-t)e^t\}^{-1}, \tag{1.29}$$

and

$$\lambda = (1-t)^{-1}e^{-t(1+\varepsilon n^{-1/2}\log n)}. \tag{1.30}$$

Taking

$$t = \frac{\varepsilon n^{-1/2}\log n}{1 + \varepsilon n^{-1/2}\log n}$$

in (1.29) and (1.30), one finds that

$$h \leq 4/(\varepsilon n^{-1/2}\log n)^2,$$

and

$$\lambda \leq 1 - (\varepsilon n^{-1/2}\log n)^2/4 \leq e^{-(\varepsilon n^{-1/2}\log n)^2/4}.$$

Hence, we obtain an upper bound for the first sum in (1.28), viz.

$$h\lambda^n \leq 4\varepsilon^{-2}n(\log n)^{-2}e^{-\varepsilon^2(\log n)^2/4},$$

so that $\sum_{n=1}^{\infty} h\lambda^n < \infty$. Similarly for the second sum in (1.28). We have proved for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left\{\sup_p \frac{|S_{[np]+s_n} - [np] - s_n|}{n} > \varepsilon n^{-1/2}\log n\right\} < \infty.$$

The Borel-Cantelli Lemma now implies that

$$\frac{|S_{[np]+s_n} - [np] - s_n|}{n^{1/2}\log n} = o(1),$$

uniformly in $p$, almost surely. Thus,

$$\{S_{[np]+s_n} - [np] - s_n\}/n = o(n^{-1/2}\log n),$$

uniformly in $p$, almost surely. Hence, from (1.27), it follows that

$$S_{[np]+s_n} S_{n+1}^{-1} - p = n^{-1} s_n + o(n^{-1/2} \log n),$$

uniformly in $p$, almost surely. This completes the proof of (1.25). The proof of (1.26) follows similarly.

□

**Lemma 1.5.2** *If* $Y_{K_n} \xrightarrow{a.s.} \theta$ *as* $n \to \infty$, *then* $n^{-1} K_n \xrightarrow{a.s.} q = F(\theta)$ *as* $n \to \infty$ *for* $a < \theta < b$.

**Proof**

Let $A = \{Y_{K_n} \to \theta \text{ as } n \to \infty\}$ and $B = \{n^{-1} K_n \to q \text{ as } n \to \infty\}$. Choose and fix an $\omega \in A$. Suppose $\omega \in B^c$. Hence, there exists a subsequence $\{n(i)\}$ of integers such that $n(i)^{-1} K_{n(i)} \to l$ as $i \to \infty$ for some finite constant $l \neq q$, with $0 \leq l \leq 1$.

Choose $\varepsilon > 0$. For all $i$ large enough, if $l \neq 0$ and $l \neq 1$, we have that

$$[n(i)(l-\varepsilon)] < K_{n(i)} < [n(i)(l+\varepsilon)] + 1,$$

which implies that

$$Y_{[n(i)(l-\varepsilon)]} < Y_{K_{n(i)}} < Y_{[n(i)(l+\varepsilon)]+1}.$$

Similarly, if $l = 0$ then $a < Y_{K_{n(i)}} < Y_{[n(i)\varepsilon]+1}$, and if $l = 1$ then $Y_{[n(i)(1-\varepsilon)]} < Y_{K_{n(i)}} < b$ for all $i$ large enough.

Since $F$ is continuous and strictly increasing, $Y_{[n(i)(l-\varepsilon)]} \to F^{-1}(l-\varepsilon)$ as $i \to \infty$, for $0 < l \leq 1$ and $\varepsilon < l$. Also, $Y_{[n(i)(l+\varepsilon)]+1} \to F^{-1}(l+\varepsilon)$ as $i \to \infty$, for $0 \leq l < 1$ and $\varepsilon < 1-l$ (e.g., see Serfling, 1980). Hence, since $\varepsilon$ is arbitrary, we conclude that

$$Y_{K_{n(i)}} \to F^{-1}(l) \neq F^{-1}(q) = \theta \text{ as } i \to \infty,$$

if $0 < l < 1$. Also, since $0 < F(\theta) < 1$ (which is implied by the assumptions imposed on $f$ and $\theta$), we have that

$$\limsup_{i \to \infty} Y_{K_{n(i)}} < F^{-1}(q) = \theta,$$

if $l = 0$, and

$$\liminf_{i \to \infty} Y_{K_{n(i)}} > F^{-1}(q) = \theta,$$

if $l = 1$. Each of these three cases leads to a contradiction. Thus $\omega \in B$, which implies that $A \subset B$. This completes the proof of the lemma.

□

Henceforth, let $\xrightarrow{d}$ denote convergence in distribution as $n \to \infty$.

**Theorem 1.5.1** *Suppose that the following conditions hold:*

*(i) $f$ has a bounded third derivative in some neighborhood of $\theta$, with $f''(\theta) > 0$,*

*(ii) for each open set $U$ containing $\theta$, there exists an $\varepsilon = \varepsilon(U) > 0$ such that $f(x) - \varepsilon \geq f(\theta)$ for each $x \in (a, b) - U$,*

*(iii) $n^{-11} s_n^{13} \to 0$ as $n \to \infty$,*

*(iv) $n^{-4} s_n^5 \to C$, for some constant $C$, $0 < C \leq \infty$.*

*Then, as $n \to \infty$*

$$2^{-1/3}(f''(\theta))^{2/3} f(\theta)^{-1} n^{-1/3} s_n^{2/3}(\hat{\theta}_n - \theta) \xrightarrow{d} T,$$

*where $T$ is a random variable that maximises the process $\{Z(t) - t^2, -\infty < t < \infty\}$, and $\{Z(t)\}$ is a Gaussian process, originating from zero, with expectation 0 and covariance function given by*

$$\mathrm{Cov}\{Z(t), Z(t^*)\} = \tfrac{1}{2}\{\min(|t|, 2B) + \min(|t^*|, 2B) - \min(|t - t^*|, 2B)\},$$

*where*

$$B = C^{1/3} 2^{-1/3} (f''(\theta))^{2/3} f(\theta)^{-2}.$$

*If $C = \infty$, $\{Z(t)\}$ is a two-sided Wiener-Levy process, which is defined as follows: Let $\{W_1(t), t \geq 0\}$ and $\{W_2(t), t \geq 0\}$ be two independent standard Wiener-Levy processes. Then,*

$$Z(t) = \begin{cases} W_1(t), & \text{if } t \geq 0, \\ W_2(-t), & \text{if } t < 0. \end{cases}$$

*In this case the covariance function becomes*

$$\mathrm{Cov}\{Z(t), Z(t^*)\} = \min(|t|, |t^*|)\{I(t \geq 0, t^* \geq 0) + I(t < 0, t^* < 0)\},$$

*where $I(\cdot)$ denotes (as before) the indicator function.*

**Proof**

Let $q = F(\theta)$. Using the definition of $\hat{\theta}_n$, we obtain from the mean-value theorem that

$$\hat{\theta}_n - \theta = G\left(\frac{S_{K_n}}{S_{n+1}}\right) - G(q) = \left(\frac{S_{K_n}}{S_{n+1}} - q\right) G'(\Psi_n), \tag{1.31}$$

with $\Psi_n$ a point between $q$ and $S_{K_n}/S_{n+1}$. Note that, since $f$ is continuous at $\theta$, $F'(\theta)$ exists and $F'(\theta) = f(\theta)$. Hence, from Corollary 1.3.1 (note that Conditions *(iii)* and *(iv)* imply (1.5) and (1.6)), it follows that $Y_{K_n} \xrightarrow{a.s.} \theta$, so that $n^{-1}K_n \to q$ in probability by Lemma 1.5.2. It follows from Lemma 1.5.1 that $S_{K_n+s_n}/S_{n+1} \to q$ and $S_{K_n-s_n}/S_{n+1} \to q$ in probability, and therefore $S_{K_n}/S_{n+1} \to q$ in probability. This implies that $\Psi_n \to q$ in probability, and consequently $G'(\Psi_n) \to f(\theta)^{-1}$ in probability, by using Condition *(i)*. Further

$$\frac{S_{K_n}}{S_{n+1}} - q = \frac{S_{K_n} - K_n}{S_{n+1}} + \frac{K_n - nq}{S_{n+1}} + \frac{q(n - S_{n+1})}{S_{n+1}}, \tag{1.32}$$

$$\frac{S_{K_n} - K_n}{S_{n+1}} = n^{-1/2}\left\{\frac{S_{K_n} - K_n}{K_n^{1/2}}\right\}\left\{\frac{S_{n+1}}{n}\right\}^{-1}\left\{\frac{K_n}{n}\right\}^{1/2}. \tag{1.33}$$

By the fact that $n^{-1}K_n \to q$ in probability and the SLLN, the last two factors in (1.33) converge in probability to $q^{1/2}$. Using the central limit theorem (CLT) for a random number of summands (Blum *et al.*, 1963), the second factor in (1.33) converges in distribution to a $N(0,1)$-distribution. Hence,

$$\frac{S_{K_n} - K_n}{S_{n+1}} = O_p(n^{-1/2}).$$

A similar result holds for the last term in (1.32), since the CLT and the law of the iterated logarithm hold for $S_{n+1}$. Hence

$$\frac{S_{K_n}}{S_{n+1}} - q = O_p(n^{-1/2}) + \frac{K_n - nq}{S_{n+1}}. \tag{1.34}$$

Suppose $\{U_n\}$ is a sequence of positive numbers satisfying

$$n^{1/2}U_n \to \infty \text{ and } U_n \to 0 \text{ as } n \to \infty.$$

Then multiplication of (1.34) by $U_n^{-1}$ and substitution into (1.31) readily show that if $U_n^{-1}(n^{-1}K_n - q)$ has a limiting distribution, then $U_n^{-1}f(\theta)(\hat{\theta}_n - \theta)$ has the same limiting distribution.

Now, define the following random process

$$W_n(t) = (2U_n n)^{-1/2} S_{n+1} f(\theta)\{Y_{[n(q+U_n t)]+s_n} - Y_{[n(q+U_n t)]-s_n} - (Y_{[nq]+s_n} - Y_{[nq]-s_n})\}, \quad (1.35)$$

for $t \in [-t_0, t_0]$, with $t_0$ a finite positive constant, and

$$U_n = 2^{1/3} n^{1/3} s_n^{-2/3} (f''(\theta))^{-2/3} f(\theta)^2. \quad (1.36)$$

We shall prove that the process $W_n$ converges weakly to a limit process $W$ on the space $D[-t_0, t_0]$ of functions on $[-t_0, t_0]$ that are right-continuous and have left-hand limits. The limiting distribution of $U_n^{-1}(n^{-1}K_n - q)$ will follow by applying the continuous mapping theorem.

For $p$ close to $q$, since $f'(\theta) = 0$, we have the following Taylor series expansion,

$$\begin{aligned} G(p) &= G(q) + \frac{1}{f(\theta)}(p-q) - \frac{f''(\theta)}{6f(\theta)^4}(p-q)^3 \\ &\quad + \frac{1}{24} G^{(4)}(\Psi_n)(p-q)^4, \end{aligned} \quad (1.37)$$

where $G^{(4)}$ denotes the fourth derivative of $G$ and $\Psi_n$ is some point between $p$ and $q$. From Lemma 1.5.1 we obtain,

$$\frac{S_{[n(q+U_n t)]+s_n}}{S_{n+1}} = q + U_n t + n^{-1} s_n + o(n^{-1/2} \log n) \quad \text{a.s.}, \quad (1.38)$$

and similar expressions hold for $S_{[n(q+U_n t)]-s_n}/S_{n+1}$, $S_{[nq]+s_n}/S_{n+1}$ and $S_{[nq]-s_n}/S_{n+1}$. Expressing the $Y_i$'s in terms of the $S_i$'s, (1.35) becomes, by using (1.37), (1.38) and its equivalents,

$$W_n(t) = Z_n(t) - t^2 + R_{n1}(t) + R_{n2}(t), \quad (1.39)$$

where

$$\begin{aligned} Z_n(t) &= (2U_n n)^{-1/2}\{(S_{[n(q+U_n t)]+s_n} - S_{[n(q+U_n t)]-s_n}) \\ &\quad -(S_{[nq]+s_n} - S_{[nq]-s_n})\}, \end{aligned} \quad (1.40)$$

$$R_{n1}(t) = -\frac{1}{6}\frac{f''(\theta)}{f(\theta)^4}(2U_n n)^{-1/2} S_{n+1} f(\theta)\left[\left\{\frac{S_{n(q+U_n t)+s_n}}{S_{n+1}} - q\right\}^3\right.$$

$$
\begin{aligned}
&-\left\{\frac{S_{[n(q+U_nt)]-s_n}}{S_{n+1}}-q\right\}^3-\left\{\frac{S_{[nq]+s_n}}{S_{n+1}}-q\right\}^3+\left\{\frac{S_{[nq]-s_n}}{S_{n+1}}-q\right\}^3\Big]+t^2\\
=\;&-\frac{1}{6}\frac{f''(\theta)}{f(\theta)^4}(2U_nn)^{-1/2}S_{n+1}f(\theta)\Big\{[n^{-1}s_n+U_nt+o(n^{-1/2}\log n)]^3\\
&-[-n^{-1}s_n+U_nt+o(n^{-1/2}\log n)]^3-[n^{-1}s_n+o(n^{-1/2}\log n)]^3\\
&+[-n^{-1}s_n+o(n^{-1/2}\log n)]^3\Big\}+t^2, \qquad (1.41)
\end{aligned}
$$

and

$$
\begin{aligned}
R_{n2}(t) =\;&\frac{1}{24}(2U_nn)^{-1/2}S_{n+1}f(\theta)\Big[\left\{\frac{S_{[n(q+U_nt)]+s_n}}{S_{n+1}}-q\right\}^4G^{(4)}(\Psi_{n1})\\
&-\left\{\frac{S_{[n(q+U_nt)]-s_n}}{S_{n+1}}-q\right\}^4G^{(4)}(\Psi_{n2})\\
&-\left\{\frac{S_{[nq]+s_n}}{S_{n+1}}-q\right\}^4G^{(4)}(\Psi_{n3})+\left\{\frac{S_{[nq]-s_n}}{S_{n+1}}-q\right\}^4G^{(4)}(\Psi_{n4})\Big]\\
=\;&\frac{1}{24}(2U_nn)^{-1/2}S_{n+1}f(\theta)\Big\{[n^{-1}s_n+U_nt+o(n^{-1/2}\log n)]^4G^{(4)}(\Psi_{n1})\\
&-[-n^{-1}s_n+U_nt+o(n^{-1/2}\log n)]^4G^{(4)}(\Psi_{n2})\\
&-[n^{-1}s_n+o(n^{-1/2}\log n)]^4G^{(4)}(\Psi_{n3})\\
&+[-n^{-1}s_n+o(n^{-1/2}\log n)]^4G^{(4)}(\Psi_{n4})\Big\},
\end{aligned}
$$

where $\Psi_{n1}$ is a point between $S_{[n(q+U_nt)]+s_n}/S_{n+1}$ and $q$, $\Psi_{n2}$ is a point between $S_{[n(q+U_nt)]-s_n}/S_{n+1}$ and $q$, $\Psi_{n3}$ is a point between $S_{[nq]+s_n}/S_{n+1}$ and $q$, and $\Psi_{n4}$ is a point between $S_{[nq]-s_n}/S_{n+1}$ and $q$.

The leading term in $\{\cdot\}$ of (1.41) is $6n^{-1}s_nU_n^2t^2$. It therefore follows from (1.36) and the fact that $n^{-1}s_n^2\to\infty$ as $n\to\infty$, that

$$\sup_{-t_0\le t\le t_0}|R_{n1}(t)|=o(1),$$

almost surely. By Condition *(iii)*,

$$\sup_{-t_0\le t\le t_0}|R_{n2}(t)|=o(1),$$

almost surely.

Hence, it now suffices to prove the weak convergence of the process $Z_n$ to $Z$. For this it is sufficient to show that the finite-dimensional distributions of $Z_n$ converge to those of $Z$ and that the sequence $\{Z_n\}$ is tight (see, e.g., Billingsley, 1968, Theorem 15.1).

Consider first a single time point $t$; we must prove

$$Z_n(t) \xrightarrow{d} Z(t) \text{ as } n \to \infty. \tag{1.42}$$

From (1.40) it follows that $Z_n(t)$ can be written for large $n$ as

$$Z_n(t) = \begin{cases} (2U_n n)^{-1/2} \left( \sum_{i=1}^{2s_n} Z_i - \sum_{i=2s_n+1}^{4s_n} Z_i \right), & \text{if } |t| > 2B, \\ (2U_n n)^{-1/2} \left( \sum_{i=1}^{l_n(t)} Z_i - \sum_{i=l_n(t)+1}^{2l_n(t)} Z_i \right), & \text{if } |t| \le 2B, \end{cases} \tag{1.43}$$

where $l_n(t) = [nU_n|t|]$. Using Condition *(iv)*, (1.36), (1.43) and the CLT, it follows that $Z_n(t) \xrightarrow{d} N(0, 2B)$ if $|t| > 2B$, and $Z_n(t) \xrightarrow{d} N(0, |t|)$ if $|t| \le 2B$. This proves (1.42). Next, consider two time points $s$ and $t$ with $s < t$. We must now prove that

$$(Z_n(s), Z_n(t)) \xrightarrow{d} (Z(s), Z(t)) \text{ as } n \to \infty. \tag{1.44}$$

The validity of (1.44) will only be illustrated for the case $|s| \le 2B$, $|t| \le 2B$ and $|t - s| \le 2B$. Other cases can be dealt with similarly. Using expressions for $Z_n(s)$ and $Z_n(t)$ analogous to (1.43), it immediately follows, as above, that

$$(Z_n(s), Z_n(t)) \xrightarrow{d} \begin{cases} (s^{1/2}Z_1, s^{1/2}Z_1 + (t-s)^{1/2}Z_2), & \text{if } 0 \le s < t, \\ ((-t)^{1/2}Z_1 + (t-s)^{1/2}Z_2, (-t)^{1/2}Z_1), & \text{if } s < t \le 0, \\ ((-s)^{1/2}Z_1, t^{1/2}Z_2), & \text{if } s < 0 < t, \end{cases}$$

where $Z_1$ and $Z_2$ are two independent $N(0,1)$-distributed random variables. This proves (1.44). A set of three or more time points can be treated in the same way, and hence the finite-dimensional distributions converge properly.

It remains to show that $\{Z_n\}$ is tight. From Theorem 15.6 of Billingsley (1968), a sufficient condition for this is that there exist constants $\gamma \ge 0$ and $\alpha > \frac{1}{2}$ and a nondecreasing, continuous function $H$ on $[-t_0, t_0]$ such that for all $t_1 \le t \le t_2$ and $n \ge 1$,

$$E\{|Z_n(t) - Z_n(t_1)|^\gamma |Z_n(t_2) - Z_n(t)|^\gamma\} \le \{H(t_2) - H(t_1)\}^{2\alpha}. \tag{1.45}$$

Consider the case $|t_1| \le 2B$ and $|t_2| \le 2B$. Using (1.43), (1.45) follows directly by choosing $\gamma = 2$, $\alpha = 1$ and $H(t) = A \cdot t$ for some finite positive constant $A$. Other cases can be dealt with similarly.

At this point it has been shown that (see (1.39)) $W_n \to W$ weakly on $D[-t_0, t_0]$ as $n \to \infty$, where $W(t) = Z(t) - t^2$. Since $t_0$ is arbitrary, it follows from Whitt (1970, 1971) (and the references therein) that the weak convergence result holds for all $t \in (-\infty, \infty)$. Now, for $x \in D(-\infty, \infty)$, let

$$h(x) = \min\left\{t : x(t) = \max_s x(s)\right\}.$$

From the definition of $K_n$ (see (1.3)), (1.35), Theorem 5.1 of Billingsley (1968) and the fact that $F$ is continuous, it now follows that $h(W_n) \xrightarrow{d} h(W)$ as $n \to \infty$, where $h(W_n) = U_n^{-1}(n^{-1}K_n - q)$ and $h(W) = T$, as defined in the statement of the theorem. It was proved by Chernoff (1964) and Groeneboom (1989) that $P(T < \infty) = 1$. This completes the proof of the theorem.

□

**Remarks**

(a) From the proof of Theorem 1.5.1 it is clear that $U_n^{-1}(n^{-1}K_n - q) \xrightarrow{d} T$ as $n \to \infty$, *only* under Conditions *(i)*, *(iii)* and *(iv)*.

(b) Note that if $\{s_n\}$ is selected so that $s_n \sim An^{\alpha}$ and $A > 0$, then Theorem 1.5.1 holds for $\frac{4}{5} \leq \alpha < \frac{11}{13}$.

(c) Suppose Conditions *(ii)* and *(iv)* of Theorem 1.5.1 hold, and instead of *(i)* and *(iii)* we assume

*(i)'* $f$ has a bounded fourth derivative in some neighbourhood of $\theta$, with $f'''(\theta) > 0$,

*(iii)'* $n^{-7}s_n^8 \to c$, for some constant $c$, $0 \leq c < \infty$.

Then, following the same arguments as in the proof of Theorem 1.5.1, we can show that

$$2^{-1/3}(f''(\theta))^{2/3}f(\theta)^{-1}n^{-1/3}s_n^{2/3}(\hat{\theta}_n - \theta)$$

is asymptotically distributed as the variable $T$ which maximises the process $\{Z(t) - t^2 + C^*t, -\infty < t < \infty\}$, where $\{Z(t)\}$ is as before and

$$C^* = \tfrac{1}{3}f'''(\theta)(f(\theta))^{-4}(f''(\theta))^{-1/3}2^{-1/3}c^{1/3}.$$

In this case, if $\{s_n\}$ is selected so that $s_n \sim An^{\alpha}$ and $A > 0$, then the above holds for $\frac{4}{5} \leq \alpha \leq \frac{7}{8}$.

We now derive the limiting distribution of $\eta_n$, the estimator of $f(\theta)$, as defined in (1.15), viz.

$$\eta_n = \frac{n^{-1}(2r_n + 1)}{Y_{K_n+r_n} - Y_{K_n-r_n}}.$$

Recall that $K_n$ (see (1.3)) is defined in terms of the sequence $\{s_n\}$.

**Theorem 1.5.2** *Suppose the following conditions hold:*

*(i) $f$ has a bounded third derivative in a neighbourhood of $\theta$, with $f''(\theta) > 0$,*

*(ii) for each open set $U$ containing $\theta$, there exists an $\varepsilon = \varepsilon(U) > 0$ such that $f(x) - \varepsilon \geq f(\theta)$ for each $x \in (a, b) - U$,*

*(iii) $n^{-11}s_n^{13} \to 0$ as $n \to \infty$,*

*(iv) $n^{-4}s_n^5 \to \infty$ as $n \to \infty$,*

*(v) $n^{-4}r_n^5 \to k$ as $n \to \infty$, for some constant $k$, $0 \leq k < \infty$,*

*(vi) $n^4 s_n^{-2} r_n^{-3} \to 0$ as $n \to \infty$.*

*Then, as $n \to \infty$, we have*

$$(2r_n + 1)^{1/2}\{\eta_n - f(\theta)\} \xrightarrow{d} N\left(\tfrac{1}{3}(k/2)^{1/2} f''(\theta)(f(\theta))^{-2}, (f(\theta))^2\right).$$

**Proof**

Expanding $G(S_{K_n+r_n}/S_{n+1})$ and $G(S_{K_n-r_n}/S_{n+1})$ in a Taylor series around $q = F(\theta)$ to third order terms and using the fact that $f'(\theta) = 0$, we have

$$\begin{aligned}
&(2r_n + 1)^{1/2}\{f(\theta)\eta_n^{-1} - 1\} \\
&= (2r_n + 1)^{1/2}\{f(\theta)n(2r_n + 1)^{-1}[Y_{K_n+r_n} - Y_{K_n-r_n}] - 1\} \\
&= \{(S_{K_n+r_n} - S_{K_n-r_n}) - 2r_n\}(2r_n + 1)^{-1/2} n S_{n+1}^{-1} \\
&\quad + \{(2r_n)(2r_n + 1)^{-1/2} n S_{n+1}^{-1} - (2r_n + 1)^{1/2}\}
\end{aligned}$$

$$
\begin{aligned}
&+n(2r_n+1)^{-1/2}\frac{f''(\theta)}{6(f(\theta))^3}\left\{\left[\frac{S_{K_n-r_n}}{S_{n+1}}-q\right]^3-\left[\frac{S_{K_n+r_n}}{S_{n+1}}-q\right]^3\right\}\\
&+n(2r_n+1)^{-1/2}\frac{G^{(4)}(\xi_{n1})f(\theta)}{24}\left[\frac{S_{K_n+r_n}}{S_{n+1}}-q\right]^4\\
&-n(2r_n+1)^{-1/2}\frac{G^{(4)}(\xi_{n2})f(\theta)}{24}\left[\frac{S_{K_n-r_n}}{S_{n+1}}-q\right]^4\\
=\ &\{(S_{K_n+r_n}-S_{K_n-r_n})-2r_n\}(2r_n+1)^{-1/2}nS_{n+1}^{-1}\\
&+R_{n1}+R_{n2}+R_{n3}+R_{n4}\ \text{(say)},
\end{aligned}
\tag{1.46}
$$

where $\xi_{n1}$ is a point between $S_{K_n+r_n}/S_{n+1}$ and $q$ and $\xi_{n2}$ a point between $S_{K_n-r_n}/S_{n+1}$ and $q$.

From Corollary 1.3.1 and Lemma 1.5.2 it follows immediately that

$$n^{-1}K_n = q + o_p(1). \tag{1.47}$$

Also, using Lemma 1.5.1 and (1.47) we have that $G^{(4)}(\xi_{n1}) = O_p(1)$ and $G^{(4)}(\xi_{n2}) = O_p(1)$.

From (1.47) and the CLT for a random number of summands (Blum *et al.*, 1963) we have

$$n^{-1}(S_{K_n+r_n} - K_n - r_n) = O_p(n^{-1/2}). \tag{1.48}$$

Since $U_n^{-1}(n^{-1}K_n - q)$ (with $U_n$ defined in (1.36)) has a limiting distribution (see Remark (a) above), it follows that

$$n^{-1}K_n = q + O_p(U_n). \tag{1.49}$$

Hence, from (1.48), (1.49) and the CLT we obtain

$$\frac{S_{K_n+r_n}}{S_{n+1}} - q = n^{-1}r_n + O_p(U_n),\ \ \frac{S_{K_n-r_n}}{S_{n+1}} - q = -n^{-1}r_n + O_p(U_n). \tag{1.50}$$

By the conditions of the theorem and (1.50) it readily follows that $R_{n1} = o_p(1)$, $R_{n3} = o_p(1)$, $R_{n4} = o_p(1)$ and $R_{n2} = -\frac{1}{3}(k/2)^{1/2}f''(\theta)(f(\theta))^{-3} + o_p(1)$. Hence, from (1.46) we have

$$
\begin{aligned}
&(2r_n+1)^{1/2}\{f(\theta)\eta_n^{-1} - 1\}\\
&\quad = \{(S_{K_n+r_n} - S_{K_n-r_n}) - 2r_n\}(2r_n+1)^{-1/2}nS_{n+1}^{-1}\\
&\qquad -\tfrac{1}{3}(k/2)^{1/2}f''(\theta)(f(\theta))^{-3} + o_p(1).
\end{aligned}
$$

Since $nS_{n+1}^{-1} \to 1$ almost surely, to complete the proof of the theorem, it suffices to show that

$$(S_{K_n+r_n} - S_{K_n-r_n} - 2r_n)(2r_n)^{-1/2} \xrightarrow{d} N(0,1). \tag{1.51}$$

The left-hand side of (1.51) can be written as

$$\begin{aligned}(2r_n)^{-1/2}\{T_{K_n+r_n} - T_{[nq]+r_n}\} \;&+\; (2r_n)^{-1/2}\{T_{[nq]-r_n} - T_{K_n-r_n}\} \\ &+\; (2r_n)^{-1/2}\{T_{[nq]+r_n} - T_{[nq]-r_n}\},\end{aligned} \tag{1.52}$$

where $T_n = S_n - n$. Let $\delta > 0$ be arbitrary. From (1.49), it follows that there exist finite positive constants $M(\delta)$ and $N(\delta)$ such that for all $n > N(\delta)$,

$$P(|K_n - [nq]| > nM(\delta)U_n) < \delta. \tag{1.53}$$

By Kolmogorov's inequality for sums of independent random variables (e.g., see Breiman, 1968) and (1.53), we have for all $\varepsilon > 0$,

$$\begin{aligned}&P\{(2r_n)^{-1/2}|T_{K_n+r_n} - T_{[nq]+r_n}| > \varepsilon\} \\ &\leq\; P\{|T_{K_n+r_n} - T_{[nq]+r_n}| > \varepsilon(2r_n)^{1/2}, |K_n - [nq]| \leq nM(\delta)U_n\} + \delta \\ &\leq\; 2P\left\{\max_{1\leq k\leq [nM(\delta)U_n]} |T_k| > \varepsilon(2r_n)^{1/2}\right\} + \delta \\ &\leq\; \frac{2[nM(\delta)U_n]}{\varepsilon^2(2r_n)} + \delta.\end{aligned}$$

From this we conclude that the first term in (1.52) is $o_p(1)$, by letting $n \to \infty$ (applying Condition *(vi)*) and then $\delta \to 0$. A similar argument yields that the second term in (1.52) is $o_p(1)$. The third term has the same distribution as $(2r_n)^{-1/2}T_{2r_n}$, which converges to a $N(0,1)$-distribution. From this and Slutsky's theorem, the proof of the theorem is completed.

□

**Remark**

Note that if $\{s_n\}$ is selected so that $s_n \sim A_1 n^\alpha$ and $\{r_n\}$ is selected so that $r_n \sim A_2 n^\beta$, then Theorem 1.5.2 holds for $\frac{4}{5} < \alpha < \frac{11}{13}$ and $\frac{2}{3}(2-\alpha) < \beta \leq \frac{4}{5}$. The bias in the limiting distribution derived above, is non-zero only if $\beta = \frac{4}{5}$.

## 1.6 Relationship with maximal spacings

Let $X_1, X_2, \ldots$, be a sequence of independent and identically distributed random variables on some probability space $(\Omega,\mathcal{F},P)$ with unknown univariate distribution function $F$ on the real line. Suppose $F$ is absolutely continuous (with respect to Lebesgue measure) with density $f$. Denote (as before) the order statistics of $X_1, X_2, \ldots, X_n$ by

$$Y_1 \leq Y_2 \leq \cdots \leq Y_n.$$

Let $\{k_n\}$ be a nonrandom sequence of positive integers. The *maximal $k_n$-spacing* is defined by

$$M_n = \max_{1 \leq j \leq n-k_n} (Y_{j+k_n} - Y_j).$$

A great deal is known about the behaviour of $M_n$ when $k_n \equiv 1$ for all $n$ and the $X_i$'s are uniformly distributed on (0,1). For example, Devroye (1981, 1982) and Deheuvels (1982, 1983) derived laws of the iterated logarithm for $M_n$. If $k_n \to \infty$ as $n \to \infty$ at certain rates, Deheuvels and Devroye (1984) obtained analogous results.

However, few results are available when $F$ is arbitrary. For $k_n \equiv 1$, Deheuvels (1984) derived strong limiting bounds for $M_n$. He pointed out, among others, that if $F$ has a continuous density $f$, the major influence on the behaviour of maximal spacings is exerted by the behaviour of $f$ in the neighbourhood of its minimum. Under the assumption that $Y_1$ and $Y_n$ belong to the domain of attraction of extreme-value distributions and that $k_n \equiv 1$, Deheuvels (1986) showed that the weak limiting behaviour of $Y_1$ and $Y_n$ characterises completely the weak limiting behaviour of $M_n$ and he also obtained the corresponding limiting non-normal distributions. Also, Barbe (1992) proved that $M_n$ (appropriately standardised) converges in distribution to a Gumbel distribution if it is assumed, among other things, that the density $f$ has a positive minimum and $k_n \equiv 1$ for all $n$. The weak limiting behaviour of $M_n$ is related to the minimum of the density function and to the local behaviour of the density function near its minimum, as is the case for the almost sure behaviour of $M_n$ (Barbe, 1992).

In previous sections we studied a modified version of $M_n$. Firstly, we defined a *maximal*

$2s_n$-*spacing* by

$$V_n \equiv Y_{K_n+s_n} - Y_{K_n-s_n} = \max_{s_n+1 \le j \le n-s_n} (Y_{j+s_n} - Y_{j-s_n}),$$

and then modified it to

$$\tilde{V}_n = Y_{K_n+r_n} - Y_{K_n-r_n}.$$

The estimator of $f(\theta)$ was then defined in terms of $\tilde{V}_n$, viz.

$$\eta_n = n^{-1}(2r_n + 1)\tilde{V}_n^{-1}.$$

In deriving the strong and weak limiting properties of $\eta_n$, the corresponding results for the modified statistic $\tilde{V}_n$ were in fact being obtained. A strong law of large numbers and a limiting distribution for $\tilde{V}_n$ can thus be formally stated as follows:

**Theorem 1.6.1** *Under the conditions of Theorem 1.3.3, as* $n \to \infty$

$$f(\theta)n(2r_n + 1)^{-1}\tilde{V}_n \to 1 \quad \text{a.s.}$$

**Theorem 1.6.2** *Under the conditions of Theorem 1.5.2, we have, as* $n \to \infty$

$$(2r_n + 1)^{1/2}\{f(\theta)n(2r_n + 1)^{-1}\tilde{V}_n - 1\} \xrightarrow{d} N(0, 1).$$

The result in Theorem 1.6.2 is surprising, since it is in contrast with the non-normal asymptotic distributions obtained in the literature for the maximal $k_n$-spacing $M_n$. The incorporation of the second sequence $\{r_n\}$ of integers enabled me to derive the limiting normal distribution.

# Chapter 2

# Kernel density estimation

## 2.1 Introduction

The antimode and minimum of a density estimator provide indirect estimators of $\theta$ and $f(\theta)$. In Chapter 4 the small and moderate sample behaviour of my proposed estimators of $\theta$ and $f(\theta)$ are compared with these obvious alternatives. The well-known and popular kernel method introduced by Rosenblatt (1956) is used here, as density estimation technique. The practical application of kernel density estimation is crucially dependent on the choice of the so-called smoothing parameter. The ultimate aim of this chapter is to motivate the specific preferences of smoothing parameters, applied in the numerical studies, from the extensive recent literature on data-based selection of the smoothing parameter in kernel density estimation. To reach this goal, a short background is first provided of kernel density estimation in general and secondly some of the current smoothing methods are discussed in general terms.

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with unknown univariate distribution function $F$ and probability density function $f$. The kernel estimator of $f$ is defined by

$$\hat{f}_h(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{2.1}$$

where $K$ is the kernel function. The value $h = h_n$ is known as the smoothing parameter; also called the window width or bandwidth. The value of $h$ will generally depend on the

sample size $n$, but this dependence will not always be made explicit.

It is interesting to note that Cacoullos (1966) appears to have been the first to call $K$ in (2.1) a **kernel function.** Previously, $K$ was referred to as a **weight function.** The simplest class of kernels consists of probability density functions that satisfy

$$K(u) \geq 0, \quad \int K(u)du = 1. \tag{2.2}$$

If a kernel $K$ from this class is used in (2.1), then $\hat{f}_h$ will always be a bona fide density. Popular choices of univariate kernels include the Gaussian kernel with unbounded support,

$$K(u) = (2\pi)^{-1/2} \exp(-\tfrac{1}{2}u^2), \quad -\infty < u < \infty,$$

and the compactly supported "polynomial" kernels,

$$K(u) = \begin{cases} \kappa_{rs}(1 - |u|^r)^s, & \text{if } -1 \leq u \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$\kappa_{rs} = \frac{r}{2B(s+1, \frac{1}{r})}, \quad r > 0, \ s \geq 0,$$

with $B(\cdot, \cdot)$ denoting the beta-function. The rectangular kernel is obtained if $s = 0$ ($\kappa_{r0} = \frac{1}{2}$); the triangular kernel if $r = 1$, $s = 1$ ($\kappa_{11} = 1$); the Epanechnikov kernel if $r = 2$, $s = 1$ ($\kappa_{21} = \frac{3}{4}$); the biweight kernel if $r = 2$, $s = 2$ ($\kappa_{22} = \frac{15}{16}$) and the triweight kernel if $r = 2$, $s = 3$ ($\kappa_{23} = \frac{35}{32}$). The Gaussian kernel can be obtained if $r = 2$, $s = \infty$, after a suitable rescaling.

Before discussing various properties of the density estimator $\hat{f}_h$, some measures of discrepancy that are often used in establishing the closeness of the estimator $\hat{f}_h$ to the true density $f$ are given.

## 2.2 Measures of discrepancy

Various measures of discrepancy of a density estimator $\hat{f}$ ($\hat{f}$ may be the kernel estimator $\hat{f}_h$, or any other density estimator) from the true density $f$ have been proposed. When

considering estimation at a single point, a natural measure is the mean squared error (MSE), defined by

$$\mathrm{MSE}_x(\hat{f}) = E[\hat{f}(x) - f(x)]^2.$$

By elementary properties of mean and variance, the MSE of $\hat{f}$ can be written as

$$\mathrm{MSE}_x(\hat{f}) = [E\hat{f}(x) - f(x)]^2 + \mathrm{Var}\hat{f}(x), \tag{2.3}$$

i.e., the sum of the squared bias and the variance of $\hat{f}$.

A measure of global accuracy of $\hat{f}$ as an estimator of $f$ is the mean integrated squared error (MISE) defined by

$$\mathrm{MISE}(\hat{f}) = E\int[\hat{f}(x) - f(x)]^2 dx = \int E[\hat{f}(x) - f(x)]^2 dx, \tag{2.4}$$

According to (2.3), $\mathrm{MISE}(\hat{f})$ can be written as

$$\begin{aligned} \mathrm{MISE}(\hat{f}) &= \int \mathrm{MSE}_x(\hat{f}) dx \\ &= \int [E\hat{f}(x) - f(x)]^2 dx + \int \mathrm{Var}\hat{f}(x) dx. \end{aligned} \tag{2.5}$$

Other global measures of deviation such as the mean integrated absolute error (MIAE),

$$E\int |\hat{f}(x) - f(x)| dx,$$

may appear to be more natural under some circumstances (see Devroye & Györfi, 1985).

It can easily be shown from the early work of Whittle (1958) that the mean and variance of the kernel estimator are given by

$$E\hat{f}_h(x) = \frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right) f(y) dy, \tag{2.6}$$

and

$$\mathrm{Var}\hat{f}_h(x) = \frac{1}{nh_n^2}\left\{\int K\left(\frac{x-y}{h_n}\right)^2 f(y) dy - \left[\int K\left(\frac{x-y}{h_n}\right) f(y) dy\right]^2\right\}. \tag{2.7}$$

These expressions may be substituted into (2.3) and (2.5) to obtain exact expressions for the MSE and MISE, but except in very special cases, the calculations become intractable and the expressions obtained have little intuitive meaning. It is more instructive to obtain approximations to (2.6) and (2.7) under suitable conditions, as is done in the following section.

## 2.3 Large-sample properties

There is an enormous literature on the large-sample properties of kernel density estimators. The reader is referred to Rosenblatt (1971), Prakasa Rao (1983) and Rosenblatt (1991) for reviews of the literature. In the first part of this section approximate expressions for the bias and variance are given. In the second part these expressions are used to investigate the behaviour of the MSE and the MISE and consequently to establish the ideal choices for the bandwidth and kernel. The last part includes a few asymptotic results to give an indication of the large-sample behaviour of the kernel density estimator in general. Assume throughout this section that the kernel $K$ is a symmetric probability density function defined on $\Re$ that satisfies the following,

$$\int uK(u)du = 0, \tag{2.8}$$

$$0 < \int u^2 K(u)du < \infty. \tag{2.9}$$

### 2.3.1 Approximate expressions for bias and variance

If the density function $f$ is bounded and continuous at $x$ and $h_n \to 0$ as $n \to \infty$, the bias

$$b_n(x) \equiv E\hat{f}_h(x) - f(x)$$

converges to zero as $n \to \infty$. If $f$ is bounded and

$$nh_n \to \infty, \quad h_n \to 0, \quad \text{as } n \to \infty, \tag{2.10}$$

the variance $\text{Var}\hat{f}_h(x)$ tends to zero as $n \to \infty$ (Rosenblatt, 1991).

If $f$ is continuously differentiable up to second order with bounded derivatives, the bias can be approximated, using (2.8) and (2.9), by

$$b_n(x) \simeq \frac{1}{2}h_n^2 f''(x) \int u^2 K(u)du. \tag{2.11}$$

An approximation for the variance is

$$\text{Var}\hat{f}_h(x) \simeq \frac{1}{nh_n} f(x) \int K(u)^2 du. \tag{2.12}$$

The integrated squared bias and the integrated variance are required in (2.5) for the MISE and are given by

$$\int b_n(x)^2 dx \simeq \frac{1}{4}h_n^4 \int f''(x)^2 dx \left\{\int u^2 K(u) du\right\}^2, \tag{2.13}$$

$$\int \mathrm{Var}\hat{f}_h(x) dx \simeq \frac{1}{nh_n} \int K(u)^2 du. \tag{2.14}$$

Suppose now that one wants to minimise the MISE. One of the fundamental problems of density estimation is represented by the two components of the MISE (or the MSE). To reduce the bias a small value for the bandwidth should be used. In this case, the integrated variance (the second component of the expression for the MISE) will become large. On the other hand, choosing a large value for $h$ will reduce the random variation as quantified by the variance, at the expense of introducing systematic error, or bias, into the estimation. In other words, by adjusting the amount of smoothing, the bias can be reduced at the expense of increasing the variance, and vice versa.

### 2.3.2 Optimal bandwidth and kernel

Using (2.11) and (2.12), the MSE can be approximated by

$$E[\hat{f}_h(x) - f(x)]^2 \simeq \frac{1}{4}h_n^4 f''(x)^2 \left[\int u^2 K(u) du\right]^2 + \frac{1}{nh_n} f(x) \int K(u)^2 du.$$

It is clear that one gets the most rapid rate of decay to zero if

$$h_n = \left[\frac{f(x) \int K(u)^2 du}{(f''(x) \int u^2 K(u) du)^2}\right]^{1/5} n^{-1/5}.$$

Then

$$E[\hat{f}_h(x) - f(x)]^2 \simeq \frac{5}{4}\left[f''(x) \int u^2 K(u) du\right]^{2/5} \left[f(x) \int K(u)^2 du\right]^{4/5} n^{-4/5}.$$

Note that the magnitude of the locally optimal $h_n$ is $O(n^{-1/5})$, which decreaces rather slowly to zero as $n \to \infty$.

From (2.13) and (2.14), it follows that the approximate mean integrated squared error (AMISE) is

$$\frac{1}{4}h_n^4 \left[\int u^2 K(u) du\right]^2 \int f''(x)^2 dx + \frac{1}{nh_n} \int K(u)^2 du. \tag{2.15}$$

The optimal value of $h$, from the point of view of minimising AMISE in (2.15), is $h_*$, where

$$h_* = \left[\frac{\int K(u)^2 du}{(\int f''(x)^2 dx)(\int u^2 K(u) du)^2}\right]^{1/5} n^{-1/5}. \tag{2.16}$$

Substituting the value of $h_*$ back into (2.15) shows that AMISE becomes

$$\frac{5}{4} C(K) \left[\int f''(x)^2 dx\right]^{1/5} n^{-4/5}, \tag{2.17}$$

where the constant $C(K)$ is given by

$$C(K) = \left[\int u^2 K(u) du\right]^{2/5} \left[\int K(u)^2 du\right]^{4/5}.$$

As far as the choice of a kernel is of interest, formula (2.17) shows that one should choose a kernel $K$ with a small value of $C(K)$. The problem of minimising $C(K)$ (see Epanechnikov, 1969) reduces to that of minimising

$$\int K(u)^2 du,$$

subject to the constraints

1. $\int K(u) du = 1$,
2. $K(u) = K(-u)$,
3. $\int u^2 K(u) du = 1$.

The solution to this simple variational problem is the kernel function

$$K_e(u) = \begin{cases} (3/4)5^{-1/2}(1 - u^2/5), & \text{if } |u| \leq 5^{1/2}, \\ 0, & \text{otherwise.} \end{cases}$$

The kernel $K_e(u)$ is often called the *Epanechnikov kernel.*

It is interesting to compare the optimal kernel function $K_e(u)$ with other kernel functions $K(u)$, by computing the ratio

$$r = \frac{\int K^2(u) du}{\int K_e^2(u) du}.$$

For example, it is clear from Table I (Prakasa Rao, 1983:66), Table 3.1 (Silverman, 1986:43) and Table 2.1 (Rosenblatt, 1991:10), that AMISE is quite insensitive to the

shape of the kernel. However, in multidimensional problems, the shape of the kernel may be of greater importance. In the one-dimensional case the choice of the kernel should be based on other considerations, for example the continuity of the kernel function or the degree of differentiability required.

### 2.3.3 Consistency and limiting distribution results

The property of density estimators that received a good deal of attention is consistency in various senses. Before giving any of the results, it will be useful to give an indication of some of the various meanings of the term "consistency".

A sequence $\{\hat{f}_n\}$ of density estimators is **consistent** (in $L_1$) if the integrated absolute error (IAE) tends to zero in probability as $n \to \infty$, that is,

$$\int_{-\infty}^{\infty} |\hat{f}_n(x) - f(x)|dx \xrightarrow{P} 0,$$

where $\xrightarrow{P}$ denotes convergence in probability as $n \to \infty$.

A sequence $\{\hat{f}_n\}$ of density estimators is **strongly consistent** (in $L_1$) if the IAE tends to zero almost surely as $n \to \infty$, that is,

$$\int_{-\infty}^{\infty} |\hat{f}_n(x) - f(x)|dx \xrightarrow{a.s.} 0.$$

A sequence $\{\hat{f}_n\}$ of density estimators is **consistent in quadratic mean** if the MSE tends to zero for every $x$, that is,

$$\lim_{n\to\infty} E[\hat{f}_n(x) - f(x)]^2 = 0,$$

and **uniformly consistent in quadratic mean** when $E[\hat{f}_n(x) - f(x)]^2$ converges to zero uniformly in $x$.

A sequence $\{\hat{f}_n\}$ of density estimators is **integratedly consistent in quadratic mean** if the MISE tends to zero, that is,

$$\lim_{n\to\infty} E\left\{\int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx\right\} = 0,$$

which is also referred to as **integratedly uniformly consistent in quadratic mean.**

A sequence $\{\hat{f}_n\}$ of density estimators is said to be **weakly consistent** if

$$\hat{f}_n(x) \xrightarrow{P} f(x),$$

for every $x$.

A sequence $\{\hat{f}_n\}$ of density estimators is **uniformly weakly consistent** if

$$\sup_x |\hat{f}_n(x) - f(x)| \xrightarrow{P} 0.$$

A sequence $\{\hat{f}_n\}$ of density estimators is **strongly consistent** if

$$\hat{f}_n(x) \xrightarrow{a.s.} f(x),$$

for every $x$.

A sequence $\{\hat{f}_n\}$ of density estimators is **uniformly strongly consistent** if

$$\sup_x |\hat{f}_n(x) - f(x)| \xrightarrow{a.s.} 0.$$

Since unbiased estimators do not exist for $f$ (Rosenblatt (1956) proved this rather disappointing result), the concept of asymptotic unbiasedness was introduced.

A sequence of density estimators $\{\hat{f}_n\}$ is **asymptotically unbiased** if, for every $x$,

$$\lim_{n\to\infty} E\hat{f}_n(x) = f(x).$$

A sequence of density estimators $\{\hat{f}_n\}$ is **uniformly asymptotically unbiased** if

$$\lim_{n\to\infty} \sup_x |E\hat{f}_n(x) - f(x)| = 0.$$

Under certain mild regularity conditions on $f$ and $K$, Parzen (1962) proved that the kernel estimator $\hat{f}_h$ is asymptotically unbiased if $h_n \to 0$ as $n \to \infty$. Under (2.10) and surprisingly mild regularity conditions on $K$ and $f$, Parzen (1962) proved the consistency of $\hat{f}_h$ at a single point $x$. Bertrand-Retali (1978) studied uniform consistency under slightly stronger conditions than those of (2.10), namely

$$nh_n(\log n)^{-1} \to \infty, \quad \text{and} \quad h_n \to 0, \quad \text{as } n \to \infty. \tag{2.18}$$

Devroye (1983), using the $L_1$-approach, proved the remarkably simple result that if $K$ satisfies (2.2), then the kernel estimator is a strongly consistent estimator of $f$ if and only

if (2.10) holds, without any assumptions on $f$. More results concerning consistency in its various senses can be found in Prakasa Rao (1983).

Asymptotic normality of the kernel estimator $\hat{f}_h$ was first proved by Parzen (1962). Rosenblatt (1971) proved that if (2.10) holds and certain regularity conditions on $K$ and $f$ are met, then

$$(nh_n)^{1/2}[\hat{f}_h(x) - E\hat{f}_h(x)]$$

is asymptotically normally distributed with mean zero and variance

$$f(u)\int K(u)^2 du.$$

Finally, it is clear that conditions (2.10) and (2.18) under which consistency was proved are extremely weak. This does not suggest that good estimates of the density can be obtained for a wide range of values of $h_n$. In fact, the rate at which $\hat{f}_h$ converges to $f$ is quite sensitive to the choice of $h_n$. For example, if $h_n$ is chosen optimally, then AMISE tends to zero at the rate $n^{-4/5}$ (see (2.17)). If $h_n$ is not chosen optimally, but still satisfying (2.10), AMISE converges to zero at a slower rate, for instance, if $h_n = n^{-1/2}$, then (see (2.15)) AMISE is $O(n^{-1/2})$.

## 2.4 Data-based bandwidth selection

Practical application of kernel density estimation is crucially dependent on the choice of the smoothing parameter, $h$. (For ease of notation, I use $h$ instead of $h_n$ in this section.) Too small an $h$ gives a curve which is too noisy in that it is quite dependent on the particular realisation of the data at hand, showing features which are not shared by the density $f$. Too large an $h$ creates a bias which can eliminate, by oversmoothing, some interesting features of $f$. The aim of this section is to provide a review of the extensive recent literature on automatic, data-based selection of a global smoothing parameter in univariate kernel density estimation. During the discussion it will become clear that no one method is universally best. However, some conclusions are reached that enable us to point out an overall current preference.

Throughout, the mean integrated squared error (MISE) is used as a measure of discrepancy between $f$ and $\hat{f}_h$ (see (2.4)),

$$\text{MISE}(h) \equiv \text{MISE}(\hat{f}_h) = E\int[\hat{f}_h(x) - f(x)]^2 dx.$$

The global minimiser of MISE (for fixed $n$) is denoted by $h_0$. It is the ideal or optimal bandwidth with respect to MISE, in that it represents the bandwidth of the best possible kernel estimator in these terms. The asymptotic representation for $h_0$ is given by $h_*$ (see (2.16)).

A useful tool for the comparison of smoothing parameter selection techniques (e.g., see Park & Marron, 1990) is the asymptotic rate of convergence to the optimum. Hall and Marron (1991) showed that the best attainable rate of convergence of any data-based procedure to $h_0$ is $n^{-1/2}$, assuming that $f$ is sufficiently smooth. The optimal *root n* rate can be achieved by some of the choices of $h$ which will be discussed below, even without using higher order kernels in some cases.

Other performance criteria which may be used instead of the MISE include the integrated squared error (ISE),

$$\text{ISE}(h) = \int[\hat{f}_h(x) - f(x)]^2 dx, \tag{2.19}$$

and the Kullback-Leibler information loss function,

$$\text{I}(h) = \int f(x)\log\{f(x)/\hat{f}_h(x)\}dx. \tag{2.20}$$

There is a growing body of opinion (e.g., see Mammen, 1990) which maintains that the bandwidth which should be aimed at is not the minimiser of the MISE, but rather the minimiser of the ISE, that is, one should focus on loss rather than risk. In more practical terms, one should try to minimise the integrated squared error for the particular sample at hand, not for some hypothetical average sample. There is disagreement in the literature concerning this statement. These issues are addressed in Hall and Marron (1991) and Jones (1991). Hall and Johnstone (1992) pointed out that $h_0$ can be estimated very accurately, with a relative error of $O_p(n^{-1/2})$ even in a nonparametric setting, yet the minimiser of the ISE can be estimated only poorly with an accuracy of no better

that $n^{-1/10}$ even in simple parametric problems. They gave an explanation why the minimiser of ISE is so difficult to estimate. They quantified the difficulty by providing sharp lower bounds and showed efficient and potentially practical estimators which attain these bounds. It is clear from the extensive discussion on Hall and Johnstone's paper that the final word has not been spoken as to whether the "risk" MISE or the "loss" ISE should be preferred in measuring the accuracy of a nonparametric density estimate. I follow the recommendations of Hall and Marron (1991) and Jones (1991) to prefer the risk function to the loss function.

Another question is whether one should consider $L_1$-error, for example, the mean absolute error (MAE),

$$\text{MAE}(h) = E|\hat{f}_h(x) - f(x)|,$$

and the mean integrated absolute error (MIAE),

$$\text{MIAE}(h) = E\int |\hat{f}_h(x) - f(x)|dx,$$

instead of $L_2$-error. Hall and Wand (1988) compared two asymptotically optimal bandwidths for kernel estimation of a probability density function at a point, by considering the two norms MSE and MAE. In most cases the two results obtained differed by only a few percent. Schucany (1989) showed that the ratio of these two bandwidths is constant (and equals 0.985) for all kernels and density functions that satisfy the usual smoothness conditions. This implies that the $L_1$-error and $L_2$-error criteria do not yield large-sample results that differ by any meaningful amount. Jones *et al.* (1994) mentioned that for most densities they considered, the results concerning comparisons of various bandwidth selectors were much the same when MISE was replaced by a suitable $L_1$-analogue. Cao *et al.* (1994) also concluded that all the selectors they considered in their simulation studies showed a similar performance for the metrics $L_1$ and $L_2$. For now, I continue with $L_2$-error, specifically MISE, because it is technically simpler to deal with.

### 2.4.1 Plug-in methods

A natural method for choosing the smoothing parameter is to plot out several curves and to choose the estimate that is most in accordance with one's prior ideas about the density.

This approach is appropriate and sufficient if the purpose of the density estimation is to explore the data in order to suggest possible models and hypotheses.

Another very easy and natural parametric approach is to use a standard family of distributions to assign a value to the term $\int f''(x)^2 dx$ in the expression for some optimal bandwidth, for example $h_*$ (see (2.16)). If the Gaussian distribution (with variance $\sigma^2$) is used as reference distribution,

$$\int f''(x)^2 dx \approx 0.212\sigma^{-5},$$

and the expression for the optimal bandwidth becomes

$$\sigma \left[ \frac{\int K(u)^2 du}{0.212(\int u^2 K(u) du)^2} \right]^{1/5} n^{-1/5}.$$

A quick way of choosing the smoothing parameter would be to estimate the scale $\sigma$ from the data to obtain an estimated value for $h_*$. Silverman (1986) considered estimating the scale $\sigma$ by $s$, the sample standard deviation. This method works well if $f$ is a normal density. However, it may oversmooth if $f$ is multimodal or heavily skewed. He also considered a more robust measure of spread, viz. the sample interquartile range (IQR) normalised by the theoretical $N(0,1)$ interquartile range, $\mathrm{IQR}/(\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4}))$. Using this estimate of $\sigma$, better results are achieved for long-tailed and skew distributions, but the problem of oversmoothing becomes even worse for bimodal distributions. Silverman (1986) claimed that better results can be achieved by using the smaller bandwidth, that is, by estimating the scale by

$$\hat{\sigma}_{ROT1} = \min\{s, \mathrm{IQR}/(\Phi^{-1}(\tfrac{3}{4}) - \Phi^{-1}(\tfrac{1}{4}))\}.$$

This resulted in Silverman's first rule-of-thumb (denoted by $\hat{h}_{ROT1}$). However, this choice of bandwidth frequently tends to be too large. Hence Silverman (1986) proposed a reduction of the bandwidth, by an arbitrary factor of 90%, which resulted in his second rule-of-thumb (denoted by $\hat{h}_{ROT2}$).

In this context, Janssen *et al.* (1995) proposed bandwidth selectors based on alternative measures of scale, which are more local in character. Firstly, they considered the

collection of differences between order statistics that are a fixed percentage of observations apart, and then took the minimum of these differences in order to make the scale measure as focussed as possible. This measure of scale results in density estimates that are superior to density estimates in which $\hat{h}_{ROT1}$ and $\hat{h}_{ROT2}$ are used. However, they are still biased towards oversmoothing. The authors addressed this problem heuristically for an important special case of a bimodal density. This led to their second proposal in which they provided an improvement of the scale estimator (denoted by $D3$) by incorporating an estimate of the proportion of data that lie in the more prominent peak of the density. This choice of bandwidth performed better in the bimodal case. However, it was pointed out that it could perform poorly in situations where the density has several high peaks near each other. To solve this problem, they followed Silverman's idea and defined their "super scale" measure,

$$SS = \min(s, D3).$$

The bandwidth selector based on $SS$ is denoted by $\hat{h}_{SS}$.

Janssen *et al.* (1995) considered a variety of densities, including some very challenging, in their simulation studies. The selector $\hat{h}_{SS}$ performed much better than the other selectors (sometimes with the exception of the selector $\hat{h}_{ROT2}$), except for densities close to the Gaussian. They pointed out that $\hat{h}_{SS}$ has very good mean properties, but is much more variable than the other selectors. Janssen *et al.* (1995) (see also Jones *et al.*, 1994) classified the methods discussed above as "quick and dirty" smoothing methods. The practical implementation of $\hat{h}_{SS}$ depends heavily on the choice of an unknown parameter $\beta$ and, of course, the choice of the reference density to evaluate $\int f''(x)^2 dx$. Therefore, this bandwidth selector will not be applied in the Monte Carlo studies of Chapter 4.

If (2.16) is used to derive data-driven bandwidth selectors, this method is known in the literature as the "plug-in" approach. Alternative plug-in selectors will be discussed in Section 2.4.3, but a different approach, viz. cross-validation, which has received much attention in the literature, will be discussed first.

### 2.4.2 Cross-validation

The basic algorithm of cross-validation involves removing a single value, say $X_i$, from the sample, computing the appropriate density estimate at that $X_i$ from the remaining $n-1$ sample values,

$$\hat{f}_{h,-i}(X_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right), \tag{2.21}$$

and then choosing $h$ to optimise some given criterion involving all values of $\hat{f}_{h,-i}(X_i)$, $i = 1, 2, \ldots, n$. Two different versions of cross-validation have been used in density estimation, least-squares cross-validation and maximum likelihood cross-validation, which pre-dated the former.

#### Maximum likelihood cross-validation

The basic idea of this method is that the smoothing parameter $h$ is considered as a parameter which can be estimated by the maximum likelihood procedure. The mean log-likelihood is

$$LC(h) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_{h,-i}(X_i).$$

The value of $h$ maximising the function $LC(h)$ for the given data is the maximum likelihood cross-validation choice of $h$, denoted by $\hat{h}_{LC}$.

The function $LC(h)$ was suggested by Habbema *et al.* (1974) and Duin (1976). It has strong intuitive appeal and does not present serious computational difficulties. However, the performance of $LC(h)$ is very sensitive to outliers (see Scott & Factor, 1981). It is known that when using compactly supported kernels such as the compactly supported "polynomial" kernel (see Section 2.1), maximum likelihood cross-validation produces strongly consistent (in $L_1$) estimates of compactly supported densities (Chow *et al.*, 1983), but does not necessarily do so for estimating infinitely supported densities (Schuster & Gregory, 1981). Devroye and Györfi (1985, Chapter 6, Theorem 4) proved that the $h$ chosen by maximum likelihood cross-validation provides *universal consistency* (in $L_1$) with the only condition of compact support of $f$. No continuity or differentiability assumption is imposed on $f$. They also made some suggestions in the case of densities

with noncompact support (Devroye & Györfi, 1985:153) and thereby showed the invariance of the $L_1$-metric under monotone transformations. This ensures the consistency of this method *for all densities* $f$ which can be viewed as a property of robustness.

Hall (1987) proved that $\hat{h}_{LC}$ asymptotically minimises the Kullback-Leibler discrepancy between $\hat{f}_h$ and $f$ (see (2.20)). He studied the complex influence that the tails of both $K$ and $f$ have on maximum likelihood cross-validation in terms of the Kullback-Leibler loss. Its poor behaviour in the heavy-tailed case is a major drawback of the maximum likelihood procedure (Bowman, 1984, 1985, and Marron, 1985). In connection with this, Broniatowski *et al.* (1989) related this problem to the stability of the extreme order statistics. They proved that

$$\int |\hat{f}_{\hat{h}_{LC}}(x) - f(x)|dx \to 0,$$

almost surely (in probability) if and only if the extreme order statistics $Y_1 = \min(X_1, \ldots, X_n)$ and $Y_n = \max(X_1, \ldots, X_n)$ are strongly (weakly) stable. Since many common distributions (including those with heavier-than-exponential tails) do not satisfy this condition, the above result could be considered as a nearly decisive argument against the maximum likelihood method. Finally, Chow *et al.* (1983) and the simulation studies by Scott and Factor (1981) indicated that, depending upon the type of kernel employed, maximum likelihood cross-validation could lead to either a severely undersmoothed or oversmoothed density estimate.

### Least-squares cross-validation

The method of least-squares cross-validation was suggested independently by Rudemo (1982) and Bowman (1984). See also Bowman *et al.* (1984) for further discussion.

As mentioned earlier, an optimal choice of $h$ (not depending on the location $x$) is obtained by minimising the MISE as a function of $h$. Since the answer depends on the unknown $f$, it is useless. However, a plausible alternative might be to minimise some estimate of MISE. To derive such an estimate, write ISE($h$) (see (2.19)) as

$$\text{ISE}(h) = \int [\hat{f}_h(x) - f(x)]^2 dx = \int \hat{f}_h(x)^2 dx - 2\int \hat{f}_h(x) f(x) dx + \int f(x)^2 dx.$$

Since only the first two integrals depend on $h$, consider

$$R(\hat{f}_h) = \int \hat{f}_h(x)^2 dx - 2 \int \hat{f}_h(x) f(x) dx. \tag{2.22}$$

The second integral in (2.22) depends on the unknown $f$ and should be estimated. Define $\hat{f}_{h,-i}$ as in (2.21), and define

$$CV(h) = \int \hat{f}_h(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,-i}(X_i). \tag{2.23}$$

It can easily be shown that

$$E[CV(h)] = E[R(\hat{f}_h)],$$

so that $CV(h)$ is an unbiased estimator of MISE $- \int f(x)^2 dx$. The idea of least-squares cross-validation is to minimise $CV(h)$ with respect to $h$. Denote this minimiser by $\hat{h}_{CV}$.

Least-squares cross-validation does not seem to display the peculiar behaviour exhibited by maximum likelihood cross-validation. Indeed, very mild tail conditions on $f$ and $K$ are needed to prove asymptotic optimality results. For example, Hall (1983) and Stone (1984) showed that least-squares cross-validation achieves asymptotically the best possible choice of smoothing parameter in the sense of minimising the integrated squared error. Bowman (1984) also showed, via a limited simulation study, that least-squares cross-validation achieves satisfactory results for long-tailed $f$.

However, the general performance of the method is far from satisfactory, despite its practical popularity. Firstly, Hall and Marron (1987b) and Scott and Terrell (1987) showed that the relative rate of convergence to the optimum is of the extremely slow order of $n^{-1/10}$, i.e.,

$$\frac{\hat{h}_{CV}}{h_0} = 1 + O_p(n^{-1/10}), \tag{2.24}$$

as $n \to \infty$. Secondly, the bandwidth selected by least-squares cross-validation has a large variability, i.e., different data sets from the same distribution will often give results which are very different. See in this regard the results in Hall and Marron (1987c) and the numerical results in Scott and Terrell (1987).

### 2.4.3 Other smoothing methods

In spite of some attractive asymptotic properties, the performance of least-squares cross-validation has been often disappointing in simulations and applications. Many simulation studies in the literature have shown that this selector tends to choose smaller smoothing parameters more frequently than predicted by asymptotic theorems (e.g., see Scott & Terrell, 1987). A density estimate with a smaller smoothing parameter tends to show too many false features (structures) of the data. Moreover, as previously mentioned, $\hat{h}_{CV}$ exhibits too much variability. Chiu (1991) derived an approximate expression for $\hat{h}_{CV}$ which clearly points out the source of variation. The major source of variability of $\hat{h}_{CV}$ is contributed by $|\tilde{\phi}(\lambda)|^2$ at high frequencies which do not contain much information about $f(x)$; $\tilde{\phi}(\lambda)$ being the sample characteristic function,

$$\tilde{\phi}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} \exp(i\lambda X_j), \quad -\infty < \lambda < \infty. \tag{2.25}$$

The terminology "frequency" for $\lambda$ is borrowed from time series analysis (e.g., see Brillinger, 1975:8), where the parameter $\lambda$ is called the **frequency** of the sinusoid $\exp(i\lambda t)$, $\lambda$ real, and the index $t$ is the time of recording of the measurements, $t = 0, \pm 1, \pm 2, \ldots$.

Because of the unacceptably large dependence of least-squares cross-validation on sampling fluctuations, there has been serious investigation into other, more stable, methods of bandwidth selection.

#### Further plug-in methods

The only unknown quantity in the expression for AMISE and hence in the asymptotic optimal bandwidth, $h_*$, (see (2.15) and (2.16)) is $\int f''(x)^2 dx$. As mentioned in Section 2.4.1, the plug-in approach attempts to estimate $\int f''(x)^2 dx$ in order to obtain a bandwidth selector. Estimating the functional $\int f''(x)^2 dx$ is a different smoothing problem from estimating $f$ itself. The value of the bandwidth most appropriate for kernel estimation of this functional, denoted by $g$, will differ from $h$. In the earliest of the plug-in methods, no distinction was made between $g$ and $h$ (Woodroofe, 1970).

The plug-in method has an apparent advantage in that it does not need an opti-

misation program. The motivation of these plug-in smoothing parameters is based on asymptotic arguments. This can be considered as a drawback from the point of view of practitioners. Care must be taken as to which estimates are plugged in for the above functional.

Park and Marron (1990) studied a plug-in selector which can be described as follows. Estimate the curvature $\int f''(x)^2 dx$ by $\int \hat{f}_g''(x)^2 dx$, where $\hat{f}_g$ is a kernel estimate based on a bandwidth $g$ and a kernel $L$, not necessarily equal to $h$ and $K$ respectively. The authors then proposed a slight modification of this curvature estimate, viz.

$$n^{-2}g^{-6}\sum_{i\neq j}\int L''\left(\frac{x-X_i}{g}\right)L''\left(\frac{x-X_j}{g}\right)dx.$$

The fact that the auxiliary bandwidth $g$ is used instead of $h$ is a crucial difference between this approach and biased cross-validation described below.

Let $g_1$ be the asymptotically optimal (with respect to MISE) bandwidth for estimating $\int f''(x)^2 dx$. Park and Marron (1990) derived an exact expression for $g_1$ by using the work of Hall and Marron (1987a). Again, $g_1$ depends on an unknown functional of $f$. However, at this second stage, it turns out to be sufficiently good to estimate this functional less well, for example, by using a known scale model for $f$, with the scale estimated robustly. This gives an estimate $\hat{g}_1$ of $g_1$ and Park and Marron (1990) suggested using $g = \hat{g}_1$. Their proposed selector of $h$ has a relative error rate of convergence of order $n^{-4/13}$.

Sheather and Jones (1991) proposed an improved version of the plug-in bandwidth of Park and Marron (1990). It consists of re-incorporating the diagonal terms in the estimation of the curvature and using the auxiliary bandwidth $g$ to approximately cancel the positive bias due to the diagonal terms with the negative leading smoothing bias term. The plug-in bandwidth is defined by

$$\hat{h}_{SJ} = \left[\frac{\int K(u)^2 du}{\hat{S}(g)(\int u^2 K(u) du)^2}\right]^{1/5} n^{-1/5},$$

where $\hat{S}(g)$ is a kernel estimate of the curvature, now given by

$$\hat{S}(g) = n^{-2}g^{-6}\sum_i\sum_j\int L''\left(\frac{x-X_i}{g}\right)L''\left(\frac{x-X_j}{g}\right)dx.$$

A data-dependent choice of $g$ is made by Sheather and Jones (1991) by using similar techniques as those discussed by Hall and Marron (1987a) and Jones and Sheather (1991). The bandwidth, $\hat{h}_{SJ}$, has a relative error rate of convergence of the order $n^{-5/14}$.

Jones *et al.* (1994) reported the results of a major Monte Carlo study. They included some densities that are very hard to estimate. Overall, $\hat{h}_{SJ}$ delivered accurate density estimates. The careful choice of the pilot bandwidth $g$ is, however, very important. A wrong choice of $g$ can give rise to a very poor density estimate.

Hall *et al.* (1991) proved that the optimum convergence rate $n^{-1/2}$ can be achieved by using plug-in bandwidths. This approach lacks appeal because it requires the use of higher-order kernels in the estimation of preliminary quantities and it uses an asymptotic expansion of MISE which needs to be carried out to two terms in the bias. Of course, using higher-order kernels can lead to density estimates that take on negative values.

Chiu (1991) proposed a simple plug-in bandwidth estimate. Since it involves some of the ideas of his stabilised methods, it is discussed below.

### Biased cross-validation

Biased cross-validation was proposed by Scott and Terrell (1987). This method is actually a hybrid of cross-validation and plug-in methods in that a score function is minimised as for least-squares cross-validation (see (2.23)), but the score function makes use of some plug-in ideas. The score function is given by

$$BC(h) = \frac{1}{nh}\int K(u)^2 du + \frac{1}{2n^2 h}\left[\int u^2 K(u) du\right]^2 \sum_{i<j} \varphi\left(\frac{X_i - X_j}{h}\right),$$

where $\varphi(x) = \int K''(u)K''(u+x)du$.

Unlike $CV(h)$ given in (2.23), which is an unbiased estimator of MISE $- \int f(x)^2 dx$, $BC(h)$ is based on the AMISE. In fact (see (2.15)),

$$\begin{aligned} E[BC(h)] &= \frac{1}{4}h^4\left[\int u^2 K(u) du\right]^2 \int f''(x)^2 dx + \frac{1}{nh}\int K(u)^2 du + O(n^{-1}) \\ &= \text{AMISE} + O(n^{-1}). \end{aligned}$$

The effect of this method is to provide a data-based bandwidth with substantially less sample variability than ordinary cross-validation at the price of including some bias.

Scott and Terrell (1987) compared $\hat{h}_{CV}$ with the minimiser of $BC(h)$, denoted by $\hat{h}_{BC}$. Theoretically they showed that $\hat{h}_{BC}$ shares with $\hat{h}_{CV}$ (see (2.24)) the rate property

$$\frac{\hat{h}_{BC}}{h_0} = 1 + O_p(n^{-1/10}).$$

Cao *et al.* (1994) showed that $BC(h)$ has no finite global minimum. Plots of the function $BC(h)$ typically showed a local minimum which they used in their simulation results. The smoothing parameter chosen by this method has a high efficiency for symmetric thin-tailed as well as heavy-tailed densities, but a very low efficiency for asymmetric densities with thin or medium tails. A comparison of least-squares cross-validation, plug-in rules and biased cross-validation can be found in Park and Marron (1990).

### Smoothed cross-validation

The method of smoothed cross-validation, proposed by Hall *et al.* (1992), has the feature of obtaining $n^{-1/2}$ convergence to the optimum $h_0$. This is achieved without using higher order kernels. The idea of smoothed cross-validation has also been developed by Jones *et al.* (1991) with a slight technical difference.

Consider the following estimate of MISE,

$$SC(h) = \frac{1}{nh}\int K(u)^2 du + \hat{B}_g(h),$$

where

$$\hat{B}_g(h) = \int \left[\int h^{-1} K\left(\frac{x-t}{h}\right) \hat{f}_g(t)dt - \hat{f}_g(x)\right]^2 dx.$$

$\hat{f}_g$ is an auxiliary (pilot) estimator of $f$, i.e., $\hat{f}_g$ is a kernel estimator with bandwidth $g$ and kernel $L$, allowed to be different from $K$. The above-mentioned authors assumed that $g$ is of the form $g = Cn^p h^m$, for constants $C$, $p$ and $m$. The smoothed cross-validation selector is defined as the value $h$ which minimises $SC(h)$. The selector is denoted by $\hat{h}_{SC}$.

Motivation for using $SC(h)$ is that the first term is a good approximation of the integrated variance of $\hat{f}_h$, while $\hat{B}_g(h)$ provides an estimate of the true integrated squared bias. This method has therefore a simple, intuitive nonasymptotic motivation. (See in this regard (2.5).) This approach also provides a more direct estimate of the integrated

squared bias than the usual asymptotic bias approximations, for example those used in biased cross-validation and the usual plug-in methods discussed above. Jones *et al.* (1991) (and similarly Hall *et al.*, 1992) showed that

$$\frac{\hat{h}_{SC}}{h_0} = 1 + O_p(n^{-1/2}),$$

with $m = -2$, $p = -\frac{23}{45}$ and $C$ depending on $f$ and its first four derivatives in a complicated way.

The smoothed cross-validation selector showed a fairly good behaviour in the simulation results of Cao *et al.* (1994). However, specification of $C$ seems to be a serious problem, and therefore the smoother $\hat{h}_{SC}$ cannot be recommended for practical use.

### Trimmed cross-validation

The proposal of Feluch and Koronacki (1992) is based on the idea of excluding the pairs $(X_i, X_j)$ of observations that are *too close* in the cross-validation criterium $CV(h)$ (see (2.23)). As already mentioned, the bandwidth chosen by least-squares cross-validation $\hat{h}_{CV}$ has high variability when sample sizes are too small. Usually (see Section 2.4.3 of Silverman, 1986 and also Devroye, 1989) the reason that it assumes "too small" values is that some differences $|X_i - X_j|$ in (2.23) happen to be "very small". Therefore, in order to prevent the tendency of a resulting cross-validated bandwidth to be "too small", it seems natural to disregard in (2.23) components with small differences $|X_i - X_j|$. Hence, trimmed cross-validation is done by minimising with respect to $h$, the trimmed version of (2.23),

$$\int \hat{f}_h(x)^2 dx - \frac{2}{n(n-1)h} \sum_{i \neq j} K\left(\frac{X_i - X_j}{h}\right) I(|X_i - X_j| > c_n), \tag{2.26}$$

where $I(\cdot)$ denotes the indicator function and $\{c_n\}$ is a sequence of positive constants, $c_n/h \to 0$ as $n \to \infty$. Feluch and Koronacki (1992) showed that (2.26) can be written as the $U$-statistic

$$TC(h) = \frac{1}{nh} \int K(u)^2 du + \frac{1}{n(n-1)h} \sum_{i \neq j} \varphi_n\left(\frac{X_i - X_j}{h}\right),$$

where $\varphi_n(c) = \int K(u)K(u+c)du - 2K(c)I(|c| > c_n/h)$.

The authors obtained conditions under which $TC(h)$ is asymptotically equivalent to its original untrimmed version. They showed via simulation that some amount of trimming leads to an apparent improvement on $\hat{h}_{CV}$ for small samples. However, no indication is given as to how the trimming constants $c_n$ should be chosen, and consequently this smoothing procedure was not implemented in the Monte Carlo studies in Chapter 4.

### Stabilised smoothing method (with variations)

Chiu (1991) introduced three simple bandwidth selectors which are much more stable than that given by least-squares cross-validation. To reduce the variability in the cross-validation bandwidth estimator, he suggested modifying the sample characteristic function beyond some cut-off frequency (henceforth, denoted by $\Lambda$) in estimating the MISE and proposed a procedure for estimating the cut-off frequency.

As mentioned in the first paragraph of Section 2.4.3, Chiu (1991) pointed out that the major source of variability in the bandwidth estimator selected by least-squares cross-validation, is contributed by $|\tilde{\phi}(\lambda)|^2$ at high frequencies, which do not contain much information about $f$ ($\tilde{\phi}(\lambda)$ is the sample characteristic function given in (2.25)).

Formally, his procedure to reduce this influence of $|\tilde{\phi}(\lambda)|^2$ at high frequencies is as follows. Find the random variable $\hat{\Lambda}$ which is the first $\lambda$ such that $|\tilde{\phi}(\lambda)|^2 \leq c/n$ for some constant $c > 1$. The selection of $c$ is not important when $f$ is sufficiently smooth, as is proved in Theorem 4 of Chiu (1991), and also confirmed by his empirical results. For most practical purposes, setting $-\log_e(0.15) \leq c \leq -\log_e(0.05)$ should yield good results. The bandwidth estimate $\hat{h}_S$, called the **stabilised bandwidth estimate**, is obtained by minimising (with respect to $h$),

$$S(h) = \int_0^{\hat{\Lambda}} \{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{\hat{k}(h\lambda)^2 - 2\hat{k}(h\lambda)\}d\lambda + \frac{\pi}{nh}\int K(u)^2 du, \qquad (2.27)$$

where $\hat{k}(\lambda) = \int \exp(i\lambda u)K(u)du$ is the characteristic function of $K(u)$. $S(h)$ uses the first part of (2.27) to estimate the bias term in $\pi\text{MISE}(h)$, where

$$\pi\text{MISE}(h) = \int_0^{\infty} |\phi(\lambda)|^2\{1 - \hat{k}(h\lambda)\}^2 d\lambda + n^{-1}\int_0^{\infty} \hat{k}(h\lambda)^2\{1 - |\phi(\lambda)|^2\}d\lambda,$$

(see Lemma 1 and the discussion thereafter, Chiu (1991)).

Under commonly assumed smoothness conditions, the convergence rate of the stabilised bandwidth selector is faster than the convergence rate of the cross-validation estimate. For sufficiently smooth $f$, the stabilised bandwidth estimator can attain the convergence rate of $n^{-1/2}$, instead of the rate $n^{-1/10}$ of the cross-validation estimator. Chiu (1991) also proved the asymptotic normality and asymptotic unbiasedness of $\hat{h}_S$.

His simulation results verified that the proposed selector performed much better than cross-validation for finite samples. The stabilised procedure is adaptive to the smoothness of $f$. When the true density is not smooth enough, the stabilised procedure is more biased towards oversmoothing than least-squares cross-validation. The procedure performs excellently when $|\phi(\lambda)|$ decays nicely (this can be seen from the plot of $|\tilde{\phi}(\lambda)|^2$), where $\phi(\lambda)$ is the characteristic function of $f(x)$,

$$\phi(\lambda) = \int \exp(i\lambda x) f(x) dx.$$

However, if $|\phi(\lambda)|^2$ does not decay monotonically, the procedure may ignore the sidelobes and, consequently, over-estimate the optimal bandwidth. In all the examples I considered (see Section 4.2), $|\tilde{\phi}(\lambda)|^2$ did decay monotonically.

Note that the bandwidth selectors proposed by Park and Marron (1990), Sheather and Jones (1991), Hall *et al.* (1991) and Jones *et al.* (1991) are derived from score functions that can be written in forms similar to (2.27) (Chiu, 1991).

More recently, Chiu (1992) proposed two new procedures for selecting the cut-off frequency $\Lambda$. Firstly, replace $\hat{\Lambda}$ in (2.27) by $\hat{\Lambda}_\infty$, where $\hat{\Lambda}_\infty$ is the minimiser (with respect to $\Lambda$) of the cross-validation score function

$$CV^\infty(\Lambda) = -\frac{1}{2\pi} \int_{|\lambda|<\Lambda} |\tilde{\phi}(\lambda)|^2 d\lambda + \frac{2\Lambda}{\pi n}.$$

The minimiser of (2.27) with $\hat{\Lambda}$ replaced by $\hat{\Lambda}_\infty$, is the proposed bandwidth selector, $\hat{h}_\infty$.

However, $CV^\infty(\Lambda)$ occasionally selects an unduly large $\Lambda$. In order to reduce the chance of selecting big $\Lambda$, Chiu (1992) proposed a second modification. The modified estimate $\hat{\Lambda}_m$ is the global minimiser of

$$CV^m(\Lambda) = CV^\infty(\Lambda) + 1.65 \left\{ 2 \max(0, \Lambda - \hat{\Lambda}_1) \int f(x)^2 dx/\pi \right\}^{1/2} /n,$$

where $\hat{\Lambda}_1$ is the first local minimiser of $CV^{\infty}(\Lambda)$. The unknown quantity $\int f(x)^2 dx$ is estimated by

$$\frac{1}{\pi}\int_0^{\hat{\Lambda}_{\infty}} \{|\tilde{\phi}(\lambda)|^2 - 1/n\} d\lambda.$$

The minimiser with respect to $h$ of $S(h)$ with $\hat{\Lambda}$ replaced by $\hat{\Lambda}_m$ in (2.27), is the second proposed selector, and is denoted by $\hat{h}_m$.

Chiu (1992) carried out extensive simulation studies to check the performance of the bandwidth selectors for finite samples. He included the usual least-squares cross-validation selector $\hat{h}_{CV}$, the plug-in selector of Sheather and Jones (1991) $\hat{h}_{SJ}$, his own proposed stabilised selector $\hat{h}_S$, as well as the two improvements on the stabilised selector, $\hat{h}_{\infty}$ and $\hat{h}_m$, in the simulation studies. The bandwidth estimator of Sheather and Jones (1991) was included, because according to Chiu (1992) it is "the most promising selector of the procedures proposed by Hall *et al.* (1991), Jones *et al.* (1991), Park and Marron (1990) and Sheather and Jones (1991)".

Chiu (1992) concluded that $\hat{h}_S$ should be replaced by $\hat{h}_{\infty}$ or $\hat{h}_m$ in practice. For all the cases he considered, the performance of $\hat{h}_m$ was either the best or comparable to the best one, which was often $\hat{h}_{SJ}$. However, the latter was substantially biased towards over-smoothing in some of the cases. In general, Chiu (1992) recommended $\hat{h}_m$ when one prefers a more stable estimate.

As mentioned earlier, Chiu (1991) also proposed a simple plug-in bandwidth estimate. Since

$$\int f''(x)^2 dx = (2\pi)^{-1} \int \lambda^4 |\phi(\lambda)|^2 d\lambda,$$

he proposed to estimate $\int f''(x)^2 dx$ by

$$\pi^{-1} \int_0^{\Lambda} \lambda^4 \{|\tilde{\phi}(\lambda)|^2 - 1/n\} d\lambda,$$

where $\Lambda$ is the cut-off frequency mentioned above. The plug-in bandwidth estimate $\hat{h}_{P1}$, is given by

$$\hat{h}_{P1} = \left[ \frac{\int K(u)^2 du}{(\pi^{-1} \int_0^{\hat{\Lambda}} \lambda^4 \{|\tilde{\phi}(\lambda)|^2 - 1/n\} d\lambda)(\int u^2 K(u) du)^2} \right]^{1/5} n^{-1/5}. \qquad (2.28)$$

In his simulation studies, Chiu (1991) concluded that, for small sample sizes or rougher densities, the estimate $\hat{h}_{P1}$ outperformed $\hat{h}_S$. He also introduced an adjusted plug-in estimate that is asymptotically equivalent to the stabilised estimate.

Surprisingly enough, the modified choices of the cut-off frequency $\Lambda$ were not applied by Chiu (1992) to the plug-in bandwidth estimate $\hat{h}_{P1}$. Replacing the value $\hat{\Lambda}$ by $\hat{\Lambda}_m$ in (2.28) results in a modified plug-in bandwidth estimate, which we denote by $\hat{h}_{P2}$. Tables 4.4 and 4.5 display the results of a comparison between $\hat{h}_{P2}$ and $\hat{h}_m$, showing that the latter performs slightly better. However, $\hat{h}_{P2}$ is much faster to compute.

### Bootstrap-based procedures

The reader is referred to Chapter 3 for a brief review of bootstrap methodology. This section deals with bootstrap-based procedures to select the smoothing parameter in kernel density estimation. The basic idea is to calculate a bootstrap estimate of MISE, and then to minimise it with respect to $h$.

Let $X_1^*, X_2^*, \ldots, X_n^*$ be a bootstrap sample from the empirical distribution function $F_n$ of $X_1, X_2, \ldots, X_n$. Suppose $\hat{f}_h^*(x)$ is the bootstrap kernel estimator

$$\hat{f}_h^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right). \tag{2.29}$$

The bootstrap estimate of MISE is given by

$$\mathrm{MISE}_*(h) = E_* \int [\hat{f}_h^*(x) - \hat{f}_h(x)]^2 dx,$$

where $E_*$ denotes expectation with respect to $X_1^*, X_2^*, \ldots, X_n^*$, and $\hat{f}_h$ is (as before) the kernel estimator based on $X_1, X_2, \ldots, X_n$. However, $\mathrm{MISE}_*(h)$ is in fact not suitable for bandwidth selection because

$$E_*[\hat{f}_h^*(x)] = E_*\left[h^{-1} K\left(\frac{x - X_1^*}{h}\right)\right] = \hat{f}_h(x).$$

This shows that there is no bias in this bootstrap world, which is disastrous because as shown by the $\mathrm{MISE}(h)$ analysis in Section 2.3, bias constitutes one of the two essential quantities to be balanced in bandwidth selection.

This motivated researchers to estimate $\mathrm{MISE}(h)$ differently, and consequently obtaining different estimates of $h$. Some of these methods will now be discussed.

### (a) Smoothed bootstrap

Taylor (1989) suggested applying the smoothed bootstrap (see Section 3.3 of Chapter 3). Let $X_1^*, X_2^*, \ldots, X_n^*$ be independent random variables with common density function $\hat{f}_h$. $\mathrm{MISE}_*(h)$ is defined as above, using the smoothed bootstrap data. Taylor (1989) derived an exact expression for $\mathrm{MISE}_*(h)$, which is given if a Gaussian kernel is used, by

$$\begin{aligned}\mathrm{MISE}_*(h) &= \frac{1}{2n^2h(2\pi)^{1/2}}\Big[(1+n^{-1})\sum_{i,j}\exp\Big\{-\frac{(X_j-X_i)^2}{8h^2}\Big\} \\ &\quad -\frac{4}{3^{1/2}}\sum_{i,j}\exp\Big\{-\frac{(X_j-X_i)^2}{6h^2}\Big\} \\ &\quad +2^{1/2}\sum_{i,j}\exp\Big\{-\frac{(X_j-X_i)^2}{4h^2}\Big\}+n2^{1/2}\Big].\end{aligned}$$

Hence, no resampling is required in order to calculate $\mathrm{MISE}_*(h)$. Numerical methods can be used to minimise $\mathrm{MISE}_*(h)$ with respect to $h$. Denote this data-based bandwidth by $\hat{h}_{BT}$. In Remark 3.6 of Hall *et al.* (1992), it is shown that

$$\frac{\hat{h}_{BT}}{h_0} = 1 + O_p(n^{-1/10}).$$

This very slow rate of convergence is the same as that obtained by least-squares cross-validation and biased cross-validation. For this reason the smoothed bootstrap, using a pilot bandwidth, was introduced.

### (b) Smoothed bootstrap with pilot bandwidth

Faraway and Jhun (1990) proposed a smoothed bootstrap-based choice of the smoothing parameter $h$ by using a pilot bandwidth $g$. Let $X_1^*, X_2^*, \ldots, X_n^*$ be a bootstrap random sample from $\hat{f}_g$, the kernel estimator based on a bandwidth $g$ instead of $h$ and a kernel $L$ which may differ from $K$. The bootstrap estimate of MISE is now defined by

$$\mathrm{MISE}_*(h) = E_* \int [\hat{f}_h^*(x) - \hat{f}_g(x)]^2 dx,$$

with $\hat{f}_h^*(x)$ calculated as in (2.29) by using the bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{f}_g$.

Once again, $\mathrm{MISE}_*(h)$ can be expressed directly in terms of the data $X_1, X_2, \ldots, X_n$, so that no resampling is needed. For example, if $K = L = \phi$, the standard normal

density, we have

$$\begin{aligned}\text{MISE}_*(h) \;=\; & n^{-2}\Big[(1+n^{-1})\{4\pi(h^2+g^2)\}^{-1/2}\sum_{i,j}\exp\Big\{-\frac{(X_i-X_j)^2}{4(h^2+g^2)}\Big\} \\ & -2\{2\pi(h^2+2g^2)\}^{-1/2}\sum_{i,j}\exp\Big\{-\frac{(X_i-X_j)^2}{2(h^2+2g^2)}\Big\} \\ & +\{4\pi g^2\}^{-1/2}\sum_{i,j}\exp\Big\{-\frac{(X_i-X_j)^2}{4g^2}\Big\}+n\{4\pi h^2\}^{-1/2}\Big].\end{aligned}$$

Faraway and Jhun (1990) chose $g$ by least-squares cross-validation, but good insight into the problem of how to choose $g$ was provided by Jones *et al.* (1991) and Hall *et al.* (1992). These authors proved that, under certain conditions,

$$\frac{\hat{h}_{BFJ}}{h_0} = 1 + O_p(n^{-1/2}),$$

where $\hat{h}_{BFJ}$ is the minimiser of $\text{MISE}_*(h)$. This result is not surprising, since $\text{MISE}_*(h)$ is almost identical to $SC(h)$, the smoothed cross-validation criterion discussed earlier. Consequently, $\hat{h}_{BFJ}$ and $\hat{h}_{SC}$ behave similarly for small samples, and $\hat{h}_{BFJ}$ therefore suffers the same drawbacks as $\hat{h}_{SC}$, specifically with respect to the choice of $g$.

### (c) Modified bootstrap

Hall (1990) proposed the following way to estimate the bandwidth $h$, by applying the modified bootstrap (see Section 3.5 of Chapter 3). Let $X_1^*, X_2^*, \ldots, X_m^*$ be a bootstrap random sample of size $m < n$ from the empirical distribution $F_n$. Let $h = g(\frac{m}{n})^{1/5}$. He suggested estimating MISE by

$$\text{MISE}_*(g) = E_* \int [\hat{f}_h^*(x) - \hat{f}_g(x)]^2 dx,$$

with $\hat{f}_h^*(x)$ calculated as in (2.29). An estimate of $h$ is obtained by minimising $\text{MISE}_*(g)$. Hall (1990) showed that, under certain conditions, this bandwidth estimate has a relative error of $n^{-1/10}$. However, this procedure cannot be implemented in practice, since the choice of $m$ is unknown.

**Main conclusion**

Jones *et al.* (1994) performed a comprehensive Monte Carlo study to compare various data-driven bandwidth selectors. They included the densities of Marron and Wand (1992) that ranged from very easy to very hard to estimate. Their simulation studies led them to commend the use of $\hat{h}_{SJ}$ for general purposes. Unfortunately, Jones *et al.* (1994) did not include the bandwidth selector proposed by Chiu (1992), because "this fine contribution arose too late to be included in our simulation study". In his simulation studies, Chiu (1992) included, among others, $\hat{h}_{SJ}$ and came to the conclusion that "for all cases, the performance of $\hat{h}_m$ is either the best or comparable to the best one". In view of this and the discussions above, the selectors $\hat{h}_m$ and $\hat{h}_{P2}$ are preferred, and will therefore be applied in the Monte Carlo studies of Chapter 4.

## 2.5 Estimating functionals of the density

There are a number of statistical problems where it is necessary to obtain estimates of functionals of the density like $\int f(x)^2 dx$, $\int f(x) \log f(x) dx$, $\int f''(x)^2 dx$, etc. Another important example is estimation of the mode or the antimode of a density. An obvious way of constructing these estimates is to find an estimate $\hat{f}$ of $f$ and then to use this estimate in the functional of $f$. Sometimes it is possible to express these estimators in closed form. For example, if $\hat{f}$ is the kernel estimator $\hat{f}_h$, then $\int \hat{f}_h(x)^2 dx$ can be expressed explicitly using the formula

$$\int \hat{f}_h(x)^2 dx = \frac{1}{n^2 h_n} \sum_i \sum_j K_{(2)} \left( \frac{X_i - X_j}{h} \right),$$

where $K_{(2)} = K * K$, i.e., the convolution of $K$ with itself.

The problem of estimation of the mode of a density has received some attention in this regard. Parzen (1962) proposed using the location of the maximum of the kernel density estimate to estimate the mode of $f$. Assume that $f$ is unimodal in the sense that there exists a unique real number $\phi$ such that

$$f(\phi) = \max_{-\infty < x < \infty} f(x).$$

Suppose $f(x)$ is uniformly continuous and suppose $K(\cdot)$ is a continuous function such that $K(x) \to 0$ as $x \to \pm\infty$. Then $\hat{f}_h(x)$ is continuous and there is a random variable $\phi_n$ such that

$$\hat{f}_h(\phi_n) = \max_{-\infty<x<\infty} \hat{f}_h(x).$$

$\phi_n$ is called the *sample mode.*

The uniform strong convergence property of $\hat{f}_h$ to $f$ gives the convergence of $\phi_n$ to $\phi$ in the corresponding sense. Parzen (1962, Theorem 3A) gave conditions under which $\phi_n$ is a strongly consistent estimator of $\phi$. Nadaraya (1965) and Van Ryzin (1969) derived stronger consistency results. Parzen (1962, Theorem 5A) gave conditions under which $\phi_n$ (appropriately normalised) has an asymptotic normal distribution.

In general, the mode of $f$ may not be unique, in which case, let

$$\phi = M(f) = \inf\left\{x | f(x) = \sup_y f(y)\right\},$$

and $\phi_n = M(\hat{f}_h)$. Eddy (1980, Corollary 2.2) proved the asymptotic normality of $\phi_n$ under less stringent conditions than those of Parzen.

In Chapter 4 the small and moderate sample behaviour of the kernel-based estimate of the antimode and the kernel-based estimate of the minimum of a density are studied. The large-sample behaviour of these estimators has not been addressed in the literature.

## 2.6 Incorporating support constraints

So far in the discussion of kernel density estimation, it has been assumed implicitly either that one is interested in estimating over all of $\Re$, or if the density has a compact support, that estimation is carried out at a point well away from the boundary points. If the point of estimation is at the boundary or too close to that, serious inconsistency problems develop. In fact, if $f$ has support $[a, b]$, $a$ and $b$ finite, and $K(x) = K(-x)$ for all $x \in \Re$, then under some mild regularity conditions, we have

(i) $\hat{f}_h(x) \overset{a.s.}{\to} f(x)$ if $a < x < b$,

(ii) $\hat{f}_h(a) \overset{a.s.}{\to} f(a)/2$,

(iii) $\hat{f}_h(b) \xrightarrow{a.s.} f(b)/2$.

A heuristic proof of the statements is the following. By using the strong law of large numbers and a Taylor expansion,

$$\begin{aligned}
\hat{f}_h(x) &= \frac{1}{h}\Big[\frac{1}{n}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h}\Big)\Big] \\
&\xrightarrow{a.s.} \frac{1}{h}EK\Big(\frac{x-X_1}{h}\Big) \qquad \text{(for } h \text{ fixed)} \\
&= \frac{1}{h}\int_a^b K\Big(\frac{x-y}{h}\Big)f(y)dy \\
&= \int_{(x-b)/h}^{(x-a)/h} K(z)f(x-hz)dz \\
&= \int_{(x-b)/h}^{(x-a)/h} K(z)\{f(x) - hzf'(x) + \cdots\}dz \\
&= f(x)\int_{(x-b)/h}^{(x-a)/h} K(z)dz - hf'(x)\int_{(x-b)/h}^{(x-a)/h} zK(z)dz + \cdots \\
&= f(x)\int_{(x-b)/h}^{(x-a)/h} K(z)dz + O(h).
\end{aligned}$$

Let $a < x < b$. For this case, as $n \to \infty$ and $h \to 0$,

$$\begin{aligned}
\hat{f}_h(x) &\xrightarrow{a.s.} f(x)\int_{-\infty}^{\infty} K(z)dz \\
&= f(x).
\end{aligned}$$

Let $x = a$, then as $n \to \infty$ and $h \to 0$,

$$\begin{aligned}
\hat{f}_h(a) &\xrightarrow{a.s.} f(a)\int_{-\infty}^{0} K(z)dz \\
&= \frac{f(a)}{2}.
\end{aligned}$$

Case (iii) follows similarly.

□

According to Marron and Ruppert (1994), the boundary region is often 20-50% (sometimes even more) of the support of $f$, so that boundary bias can be a serious problem.

Several methods have been proposed to modify kernel estimators to handle boundary effects. One view is that one can no longer make use of a kernel function symmetric

about zero if the point of estimation is at the boundary or too close to it. This suggests that one ought to investigate the existence of nonsymmetric kernel functions that will allow one to obtain the same asymptotic properties as in the case of symmetric kernel functions. Such kernel functions will be useful in estimation at the boundary or close to it. Rosenblatt (1991) considered a simple family of kernel functions that are useful in the estimation of density functions which are continuously differentiable up to second order. Let

$$p(t,y) = \begin{cases} [1-(t-y)^2](\alpha+\beta\{t-y\}), & \text{if } -1+y \le t \le 1+y, \\ 0, & \text{otherwise,} \end{cases}$$

with the properties that

$$\int_{-1+y}^{1+y} p(t,y)dt = 1, \quad \int_{-1+y}^{1+y} tp(t,y)dt = 0.$$

Solving these two equations for $\alpha$ and $\beta$ on the interval $[-1+y, 1+y]$, results in

$$p(t,y) = [1-(t-y)^2]\left(\tfrac{3}{4} - \tfrac{15}{4}y\{t-y\}\right).$$

Rosenblatt (1991) suggested applying the kernel estimator where the kernel $K(\cdot)$ is now defined by

$$K(\cdot) = \begin{cases} p(\cdot,0), & \text{if } a+h \le x \le b-h, \\ p(\cdot,(a-x+h)/h), & \text{if } a \le x < a+h, \\ p(\cdot,(b-x-h)/h), & \text{if } b-h < x \le b. \end{cases}$$

Note that the kernel function can take on negative values. This boundary kernel approach is computationally complicated since each point in the boundary region requires its unique kernel.

Another way of incorporating support constraints into kernel density estimators was suggested by Schuster (1985). He proposed a "reflection about the boundaries" adjustment of the kernel estimator $\hat{f}_h$, viz.

$$\tilde{f}_h(x) = \begin{cases} \hat{f}_h(2a-x) + \hat{f}_h(x) + \hat{f}_h(2b-x), & \text{if } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

The basic problem with the usual kernel estimator is that $\hat{f}_h(x) > 0$ for some $x$'s with $x < a$ and $x > b$, but $f(x) = 0$ for $x < a$ and $x > b$. If $X_i$ is close to $a$, then part

of the contribution of $X_i$ to $\hat{f}_h(x)$ flows over to $(-\infty, a)$. Similarly, if $X_i$ is close to $b$, then part of the contribution of $X_i$ to $\hat{f}_h(x)$ flows over to $(b, \infty)$. The estimator $\tilde{f}_h(x)$ tries to incorporate the overflow to $(-\infty, a)$ and $(b, \infty)$ back into $[a, b]$. Schuster (1985) limited consideration to symmetric kernels with compact support which can be taken to be $[-1, 1]$ and choices of the smoothing parameter $h$ were limited to $0 < h < (b - a)/2$.

Figure 2.1 illustrates the "reflection about the boundaries" technique for two different densities, both with compact support $[0, 1]$. The densities, Case 11 and Case 15, as described in Section 4.2, and the bandwidth $\hat{h} = \hat{h}_m$ were used.

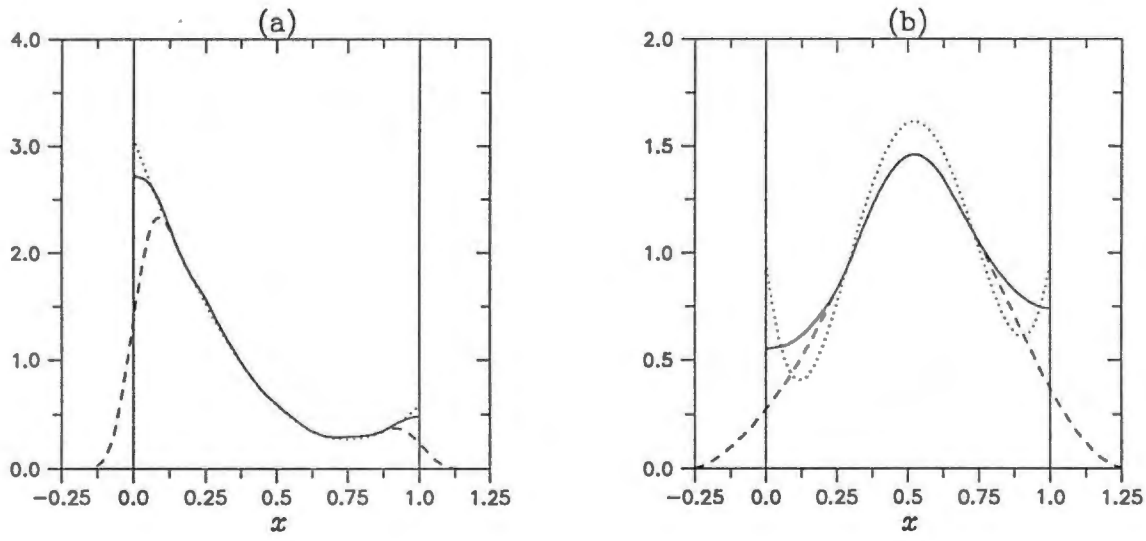


Figure 2.1: (a) *Case 11 and* (b) *Case 15 - The averaged simulated densities* $\tilde{f}_{\hat{h}}(x)$ *and* $\hat{f}_{\hat{h}}(x)$ *are represented by the solid and dashed curves, respectively, and the theoretical densities* $f(x)$ *by the dotted curves.*

Some convergence properties are given by the next two theorems (Schuster, 1985).

**Theorem 2.6.1** *Suppose the kernel* $K$ *is a bounded symmetric probability density function and vanishes off* $[-1, 1]$. *Let* $\{h_n\}$ *be a sequence of positive constants with* $h_n \to 0$ *and* $nh_n \to \infty$. *If* $F' = f$ *is uniformly continuous on* $[a, b]$, *then, for every* $x$,

1. $\lim_{n\to\infty} E\tilde{f}_h(x) = f(x)$,

2. $\lim_{n\to\infty} nh_n \mathrm{Var}\tilde{f}_h(x) = f(x)\int_{-\infty}^{\infty} K(u)^2 du \equiv \sigma^2$,

3. $\tilde{f}_h(x)$ *is strongly consistent for* $f(x)$,

4. $(nh_n)^{1/2}(\tilde{f}_h(x) - f(x))$ *is asymptotically normally distributed with mean zero and variance* $\sigma^2$, *when* $x \in [a, b]$.

**Theorem 2.6.2** *Suppose the kernel* $K$ *is a bounded symmetric probability density function and vanishes off* $[-1, 1]$. *Let* $\{h_n\}$ *be a sequence of positive constants with* $\sum_{n=1}^{\infty} \exp(\alpha n h_n^2) < \infty$ *for all* $\alpha > 0$. *Then*

$$\lim_{n\to\infty} \sup_x |\tilde{f}_h(x) - f(x)| = 0,$$

*almost surely, i.e.,* $\tilde{f}_h$ *is uniformly strongly consistent.*

Schuster's (1985) "tied down" technique which incorporates knowledge of $f(a)$ and/or $f(b)$ into $\hat{f}_h$ will not be discussed.

Marron and Ruppert (1994) showed that Schuster's reflection technique reduces the bias, but that the bias at $a$, for example, after reflection is approximately proportional to $h_n f'(a)$ as $n \to \infty$. In contrast, the bias in the interior is approximately proportional to $h_n^2 f''(x)/2$. Thus, unless $f$ has a derivative equal to zero at both boundaries and $h_n$ is sufficiently small, the bias at the boundaries will be larger than in the interior. They proposed to transform the data to a density that has first derivative equal to zero at both boundaries, estimate the density of the transformed data and obtain an estimate of the density of the original data by a change of variables. Under certain conditions, Marron and Ruppert (1994) proved that their density estimator so obtained has boundary bias $O(h_n^2)$. Their estimator depends on two bandwidths. Unfortunately, no indication is given as to how they can be estimated from the data. However, some insights of how this can be done were recently given by Cheng (1994).

# Chapter 3

# Bootstrap methodology

## 3.1 Introduction

Efron (1979) introduced a very general resampling procedure for estimating the distributions of statistics based on independent observations. This resampling procedure is called the bootstrap. In contrast with the popular Quenouille-Tukey jackknife method, the bootstrap is more widely applicable and perhaps has a sounder theoretical basis. With the advancement of computer technology, the bootstrap (which is a computer-based method) becomes increasingly more feasible and acceptable.

The bootstrap has some attractive properties for the statistical practitioner. It requires few assumptions, little modelling or analysis, and can be applied in an automatic way in a wide variety of situations. The use of the bootstrap either relieves the analyst from having to do complex mathematical calculations, or in some instances provides an answer where no analytical answer can be obtained. In Chapter 4 the bootstrap technique will be applied to derive data-based smoothing parameters. Furthermore, the bootstrap will be used to construct confidence intervals.

The bootstrap can be used either nonparametrically, or parametrically. In the nonparametric situation, it avoids restrictive and sometimes unrealistic assumptions about the form of the underlying population distribution. Since the bootstrap is used most often in the nonparametric set-up, the discussion below will be limited to this approach,

emphasising aspects which are relevant for our purposes.

## 3.2 Formal description

The review of the bootstrap method will be started by giving a formal description of the procedure. Let $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ be a sample of independent, identically distributed random variables with common *unknown* distribution function $F$. Suppose $T_n(\boldsymbol{X}_n; F)$ is some specified random variable of interest, possibly depending upon the unknown $F$. Let $F_n$ denote the empirical distribution function of $\boldsymbol{X}_n$, defined to be the discrete distribution that puts probability $1/n$ on each value $X_i$, $i = 1, 2, \ldots, n$. The bootstrap method consists of approximating the sampling distribution of $T_n(\boldsymbol{X}_n; F)$ under $F$ by the bootstrap distribution of $T_n(\boldsymbol{X}_n^*; F_n)$ under $F_n$, where $\boldsymbol{X}_n^* = (X_1^*, X_2^*, \ldots, X_n^*)$ denotes a random sample of size $n$ drawn with replacement from $F_n$.

The actual calculation of the bootstrap distribution of $T_n(\boldsymbol{X}_n^*; F_n)$ is usually difficult. However, Efron (1979) suggested a Monte Carlo method that provides an accurate approximation of the bootstrap distribution. Generate $B$ bootstrap random samples of size $n$ from $F_n$,

$$\boldsymbol{X}_n^*(1), \boldsymbol{X}_n^*(2), \ldots, \boldsymbol{X}_n^*(B).$$

The empirical distribution of the corresponding values

$$T_n(\boldsymbol{X}_n^*(1); F_n), T_n(\boldsymbol{X}_n^*(2); F_n), \ldots, T_n(\boldsymbol{X}_n^*(B); F_n)$$

is taken as an approximation of the actual bootstrap distribution of $T_n(\boldsymbol{X}_n^*; F_n)$. This approximation can be made arbitrarily accurate by taking $B$ sufficiently large. Monte Carlo approximation is not always required, because direct theoretical calculation of the bootstrap distribution is possible in some cases. Some nontrivial examples are discussed by Efron (1979).

The mechanism of the bootstrap procedure is illustrated by the following simple example concerning accuracy estimates. The bootstrap was initially introduced as a computer-based method for estimating the standard error of $\hat{\tau}$, an estimate of a parameter of interest

$\tau$. The bootstrap estimate of the standard error requires no theoretical calculations, and is available no matter how mathematically complicated the estimator $\hat{\tau}$ may be.

**Example**

In this example the bootstrap estimate of the standard error for one-sample situations are considered. Let $\hat{\tau} = \hat{\tau}(X_1, X_2, \ldots, X_n)$ be a real-valued statistic with standard error $\sigma(F)$, i.e.,

$$\sigma(F) = \{\mathrm{Var}_F(\hat{\tau})\}^{1/2}.$$

Of course, $\sigma(F)$ is also a function of the sample size $n$ and the form of the statistic $\hat{\tau}$, but since both of these are known, these dependencies will not be made explicitly in the notation. Since $F$ is unknown, $\sigma(F)$ is also unknown. The bootstrap estimate of $\sigma(F)$ is simply $\sigma(F)$ evaluated at some estimate of $F$, usually the empirical distribution $F_n$. Hence, the bootstrap estimate of standard error is simply given by

$$\hat{\sigma} = \sigma(F_n) = \{\mathrm{Var}_*(\hat{\tau}^*)\}^{1/2}, \tag{3.1}$$

where $\hat{\tau}^* = \hat{\tau}(X_1^*, X_2^*, \ldots, X_n^*)$ and $(X_1^*, X_2^*, \ldots, X_n^*)$ is a bootstrap sample from $F_n$. $\mathrm{Var}_*$ denotes variance over the conditional law of $\boldsymbol{X}_n^*$ given $\boldsymbol{X}_n$.

The argument leading to the bootstrap estimate of the standard error in (3.1) is a direct example of the so-called plug-in principle. In this case, the plug-in estimate of the parameter, which is some functional of the distribution function $F$, is simply given by the same functional of the empirical distribution $F_n$.

Sometimes it is possible to derive an explicit expression for (3.1). Suppose $\hat{\tau} = \bar{X}_n$, in which case $\sigma(F) = \{\mu_2(F)/n\}^{1/2}$, where

$$\mu_2(F) = \int (x - E_F(X_1))^2 dF(x).$$

Note that $\hat{\sigma} = \{s_n^2/n\}^{1/2}$, where $s_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, so that (3.1) can be written in closed form in this case.

Since there is no simple expression for the functional $\sigma(F)$ in most cases, the Monte Carlo algorithm (Efron, 1979) is now implemented as follows.

1. Construct $F_n$, putting mass $n^{-1}$ at each point $X_1, X_2, \ldots, X_n$.

2. Using a random number generator, draw a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $F_n$ with replacement and calculate $\hat{\tau}^*(1) = \hat{\tau}(X_1^*, X_2^*, \ldots, X_n^*)$.

3. Independently repeat step 2 some number $B$ times to obtain bootstrap replications $\hat{\tau}^*(1), \hat{\tau}^*(2), \ldots, \hat{\tau}^*(B)$.

4. Calculate

$$\hat{\sigma}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} [\hat{\tau}^*(b) - \hat{\tau}^*(\cdot)]^2 \right\}^{1/2},$$

where $\hat{\tau}^*(\cdot) = B^{-1} \sum_{b=1}^{B} \hat{\tau}^*(b)$.

Note that $\hat{\sigma}_B$ converges to $\hat{\sigma} = \sigma(F_n)$ as $B \to \infty$. In most cases a value $B$ in the range of 50 to 200 is adequate for estimating standard errors (Efron, 1981). Larger values of $B$ are required for bootstrap confidence intervals (see Section 3.4).

Other measures of statistical error, or accuracy, such as bias or prediction error, can also be easily assessed using the bootstrap. For example, consider estimation of the bias. For a given estimator $\hat{\tau}(\boldsymbol{X}_n)$ of a parameter $\tau(F)$, let

$$T_n(\boldsymbol{X}_n; F) = \hat{\tau}(\boldsymbol{X}_n) - \tau(F).$$

The bias of $\hat{\tau}(\boldsymbol{X}_n)$ for estimating $\tau(F)$ is

$$\beta(F) = E_F\{T_n(\boldsymbol{X}_n; F)\} = E_F\{\hat{\tau}(\boldsymbol{X}_n)\} - \tau(F).$$

The bootstrap estimate of bias is

$$\hat{\beta} = \beta(F_n) = E_*\{\hat{\tau}(\boldsymbol{X}_n^*)\} - \tau(F_n),$$

where $E_*$ denotes expectation over the conditional law of $\boldsymbol{X}_n^*$ given $\boldsymbol{X}_n$. The only change in the numerical evaluation of $\hat{\beta}$ occurs in Step 4 of the algorithm given above,

$$\hat{\beta}_B = \hat{\tau}^*(\cdot) - \tau(F_n).$$

Once again, we have that $\hat{\beta}_B \to \hat{\beta}$ as $B \to \infty$.

**Remark**

If the parameter $\tau$ cannot suitably be written as a functional of $F$, one can (as is often being done in the literature) replace $\tau(F_n)$ in the definition of $\hat{\beta}$ by any accurate estimator of $\tau$.

## 3.3 The smoothed bootstrap

So far, the empirical distribution function $F_n$ was used in the bootstrap calculations. $F_n$ is not a bad choice, since it is *asymptotically minimax* and is the *nonparametric maximum likelihood estimator* of $F$ (Beran, 1984). However, since $F_n$ is a discrete distribution, samples constructed from $F_n$ in the bootstrap simulation will have some rather peculiar properties. All the values taken by the members of the bootstrap samples will be drawn from the original sample values. Therefore nearly every sample will contain repeated values. This seems unacceptable when dealing with continuous variables. The smoothed bootstrap is a modification done to the usual (sometimes referred to as the "naive") bootstrap procedure to avoid bootstrap samples with this property. The basic idea is to perform the repeated sampling not from $F_n$ itself, but from a smoothed version of $F_n$. According to Swanepoel (1990) a smoothed version of $F_n$ is, for example, the kernel estimator of $F$, which is defined by

$$\hat{F}_h(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h_n}\right)$$

where $h = h_n$ is a sequence of smoothing parameters such that $h_n \to 0$ as $n \to \infty$. The kernel $H$ is a known continuous distribution function symmetric around zero. Azzalini (1981) obtained second-order results that show an asymptotic improvement in the estimation of $F$ by using $\hat{F}_h$ instead of $F_n$, provided certain regularity conditions on $F$ are met and that the sequence $h_n$ converges to zero at a certain rate. An algorithm to construct a bootstrap sample from $\hat{F}_h$ is the following.

1. Generate $X_1^*, X_2^*, \ldots, X_n^*$ which are independent random variables with common cumulative distribution function $F_n$.

2. Generate $Z_1, Z_2, \ldots, Z_n$ which are independent random variables with common cumulative distribution function $H$.

3. If both these samples are independent, then we take $Y_i^* = X_i^* + hZ_i$, $i = 1, 2, \ldots, n$, as a bootstrap sample from $\hat{F}_h$.

No definite answer exists whether the smoothed bootstrap is superior to the unsmoothed bootstrap, especially for small samples. The difficult part in applying the smoothed bootstrap comprises the choice of the smoothing parameter $h$. Silverman and Young (1987) studied the problem of choosing the smoothing parameter but unfortunately their result is not very useful in practice, since it does not specify the value of the smoothing parameter $h$ for which the smoothed bootstrap is superior. Hall *et al.* (1989) showed that in estimating the variance of a sample quantile, the rate of convergence of the relative error can be improved by using a smoothed bootstrap instead of the usual bootstrap. In Section 2.4 the smoothed bootstrap was considered as a possible bandwidth selection procedure. Since the direct application of the bootstrap fails in trying to estimate the mean integrated squared error of the kernel density estimator, this proves to be a situation where smoothing is not only desirable but necessary.

## 3.4 Confidence intervals

One of the main purposes in obtaining $\hat{\sigma}$, the estimated standard error of an estimator $\hat{\tau}$ of an unknown parameter $\tau$, is to assign approximate $100(1-\alpha)\%$-confidence intervals to $\tau$. Typically these intervals are of the form

$$\hat{\tau} \pm z(\alpha/2)\hat{\sigma},$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)$ percentile point of the standard normal distribution. However, these intervals may often be very inaccurate, in the sense that the nominal coverage probability is not attained. The bootstrap can be applied very successfully to obtain nonparametric confidence intervals. Various bootstrap methods for constructing confidence intervals exist in the literature. Comprehensive surveys are given by DiCiccio and Romano (1988) and Hall (1988b).

A brief discussion of the following nonparametric procedures will now be given:

- bootstrap-$t$,
- percentile intervals.

### Bootstrap-$t$ intervals

Let $c = c(F)$ be a constant satisfying

$$P(|\hat{\tau} - \tau|/\hat{\sigma} \leq c) \simeq 1 - \alpha.$$

The quantity $(\hat{\tau} - \tau)/\hat{\sigma}$ is called an approximate pivot. This means that its distribution is approximately the same for each value of $\tau$. If the constant $c(F)$ were known, a $100(1-\alpha)\%$-confidence interval for $\tau$ would be

$$\hat{\tau} \pm c\hat{\sigma}.$$

However, replacing $c(F)$ by the bootstrap estimate $c(F_n)$, an approximate $100(1-\alpha)\%$-confidence interval for $\tau$ is

$$\hat{\tau} \pm c(F_n)\hat{\sigma}.$$

The bootstrap estimate $c(F_n)$ is defined by

$$P^*(|\hat{\tau}^* - \hat{\tau}|/\hat{\sigma}^* \leq c(F_n)) \simeq 1 - \alpha,$$

where $\hat{\tau}^*$, $\hat{\sigma}^*$ are the estimates $\hat{\tau}$ and $\hat{\sigma}$ based on the bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $F_n$. Note that $P^*$ denotes the conditional probability law of $(X_1^*, X_2^*, \ldots, X_n^*)$ given $(X_1, X_2, \ldots, X_n)$. The random variable $c(F_n)$ can now be approximated by means of the following Monte Carlo algorithm.

1. Draw a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ of size $n$ with replacement from $F_n$ and calculate $\hat{\tau}^*$, $\hat{\sigma}^*$ and

$$T_n^*(1) = |\hat{\tau}^* - \hat{\tau}|/\hat{\sigma}^*.$$

2. Independently repeat Step 1 some number $B$ times, obtaining bootstrap replications $T_n^*(1), T_n^*(2), \ldots, T_n^*(B)$.

3. Let $D_{(i)}$, $i = 1, 2, \ldots, B$, denote the order statistics of $T_n^*(i)$, $i = 1, 2, \ldots, B$.

4. The random variable $c(F_n)$ is now approximated by $D_{(k)}$, where $k = [B(1-\alpha)]$ with $[z]$ denoting the largest integer less than or equal to $z$.

**Percentile intervals**

Let $\hat{G}$ denote the cumulative distribution function of $\hat{\tau}^* = \hat{\tau}(X_1^*, X_2^*, \ldots, X_n^*)$, i.e.,

$$\hat{G}(t) = P^*(\hat{\tau}^* \leq t).$$

The **percentile** $100(1-\alpha)\%$-confidence intervals are given by

$$\tau \in [\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1-\alpha/2)], \tag{3.2}$$

which can be approximated as follows. Draw $B$ independent bootstrap random samples of size $n$ and calculate $\hat{\tau}_1^*, \hat{\tau}_2^*, \ldots, \hat{\tau}_B^*$. If $\hat{\tau}_{(1)}^* \leq \hat{\tau}_{(2)}^* \leq \cdots \leq \hat{\tau}_{(B)}^*$ denote the corresponding order statistics, the percentile interval can be approximated by

$$\tau \in [\hat{\tau}_{(r)}^*, \hat{\tau}_{(s)}^*],$$

where the integers $r$ and $s$ are defined by $r = [B\alpha/2]$ and $s = [B(1-\alpha/2)]$.

The **bias-corrected percentile** $100(1-\alpha)\%$-confidence intervals are given by

$$\tau \in [\hat{G}^{-1}\{\Phi(2z_o - z(\alpha/2))\}, \hat{G}^{-1}\{\Phi(2z_o + z(\alpha/2))\}]. \tag{3.3}$$

Here $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\Phi(z(\alpha/2)) = 1 - \alpha/2$ and $z_o = \Phi^{-1}(\hat{G}(\hat{\tau}))$.

These intervals improve on the percentile intervals in the sense that they attempt to eliminate the effects of bias of the bootstrap distribution of $\hat{\tau}^*$. Note that in the median unbiased case, i.e., $\hat{G}(\hat{\tau}) = 0.5$, $z_o = 0$ and (3.3) reduces to (3.2).

The bias-corrected percentile intervals can be calculated exactly as above, except that we now define

$$r = [B\Phi(2z_o - z(\alpha/2))],$$

$$s = [B\Phi(2z_o + z(\alpha/2))],$$

and

$$\hat{G}(\hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\tau}_b^* < \hat{\tau}),$$

where $I(\cdot)$ denotes the indicator function.

Efron (1982) gave a detailed discussion of these two types of confidence intervals. Efron (1987) introduced an improved version of the bias-corrected percentile method, called the **accelerated bias-corrected percentile** method. This method incorporates both a bias and skewness correction. The $100(1-\alpha)\%$-confidence intervals are defined by

$$\tau \in [\hat{G}^{-1}\{\Phi(2z_o - b(\alpha/2))\}, \hat{G}^{-1}\{\Phi(2z_o + c(\alpha/2))\}], \tag{3.4}$$

where

$$b(\alpha/2) = \frac{z(\alpha/2) - z_o}{1 - a(z_o - z(\alpha/2))} + z_o,$$

$$c(\alpha/2) = \frac{z(\alpha/2) + z_o}{1 - a(z_o + z(\alpha/2))} - z_o,$$

and $a$ being some constant depending on the unknown $F$. Note that if $a = 0$ then (3.4) reduces to (3.3). Now, suppose $\tau = t(F)$, for some smooth functional $t$, and $\hat{\tau} = t(F_n)$. There are various ways to compute $a$. Efron (1987) suggested that $a$ be estimated by

$$\hat{a} = \frac{\sum_{i=1}^{n} U_i^3}{6(\sum_{i=1}^{n} U_i^2)^{3/2}}, \tag{3.5}$$

where $U_i$ is the so-called empirical influence function of $\hat{\tau}$ evaluated at $X_i$, $i = 1, 2, \ldots, n$, i.e.,

$$U_i = \lim_{\varepsilon \to 0}\left\{\frac{t((1-\varepsilon)F_n + \varepsilon\delta_i) - t(F_n)}{\varepsilon}\right\},$$

with $\delta_i$ the degenerate distribution function at the point $X_i$. Instead of letting $\varepsilon \to 0$ to compute $U_i$, some small value of $\varepsilon$, for example $\varepsilon = 1/(n+1)$, is usually selected. (See Efron, 1982:41 for a detailed discussion.) Alternatively, we may use the *jackknife influence function* for $\hat{\tau}$, viz.

$$U_i = (n-1)(\hat{\tau}_{(\cdot)} - \hat{\tau}_{(i)}), \tag{3.6}$$

where

$$\hat{\tau}_{(i)} = \hat{\tau}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n),$$

and

$$\hat{\tau}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_{(i)}.$$

This avoids having to explicitly define $\hat{\tau}$ as a functional statistic.

Note that the quantity $\hat{a}$ is called the *acceleration* because it refers to the rate of change of the standard error of $\hat{\tau}$ with respect to the true parameter value $\tau$ (see (3.5), (3.6) and Efron & Tibshirani, 1993:186).

### Discussion

The bootstrap-$t$ procedure is a useful and interesting generalisation of the usual Student's $t$-method. It is particularly applicable to location statistics like the sample mean, the median, the trimmed mean, or a sample percentile. However, the bootstrap-$t$ has some drawbacks. Bootstrap-$t$ intervals are difficult to compute when $\hat{\sigma}$, the bootstrap estimate of the standard error of $\hat{\tau}$, needs also to be calculated. The entire computation is then a double-nested bootstrap. The bootstrap-$t$ method is not invariant under reparameterisation. Beran (1987) suggested a different pivotal construction, which mimics the probability integral approach. The theoretical and numerical properties of his pivotal iteration confidence intervals are comparable to those of bootstrap-$t$ intervals. Finally, the bootstrap-$t$ intervals have good theoretical coverage probabilities, but tend to be erratic in practice.

The percentile intervals are reasonably stable in practice, but have less satisfactory coverage probabilities. They have a tendency to undercover. The three percentile bootstrap confidence intervals discussed above are in order of increasing generality. Like the standard $t$, or normal intervals, they are invariant under reparameterisation. Of these three techniques, the accelerated bias-corrected intervals generally perform best. Their coverage accuracy can, however, still be erratic for small sample sizes. The coverage probabilities of both the accelerated bias-corrected percentile interval and the bootstrap-$t$ interval differ from the nominal level (i.e., the $(1-\alpha)$-level) by only $O(n^{-1})$ instead of $O(n^{-1/2})$ for normal intervals. In other words, the accelerated bias-corrected percentile and the bootstrap-$t$ intervals are second-order accurate while the standard and percentile

methods are only first-order accurate. This was proved for the mean by Singh (1981) and for more general statistics by Hartigan (1986), Bickel (1987) and Hall (1988b). If the standard error estimate $\hat{\sigma}$ is chosen correctly, the bootstrap-$t$ can perform better than the accelerated bias-corrected procedure.

The main disadvantage of the bias-corrected accelerated percentile method is the large number of bootstrap replications required. The discussion in Section 19.3 of Efron and Tibshirani (1993) shows that at least $B = 1000$ replications are needed in order to sufficiently reduce the Monte Carlo sampling error. To assess this disadvantage, we mention another method of obtaining bootstrap intervals. DiCiccio and Efron (1992) discussed the *approximate bootstrap confidence interval.* It is a method of approximating the bias-corrected accelerated percentile interval endpoints analytically, without using any Monte Carlo replications at all. The approximation is usually quite good as can be seen from Table 14.2 of Efron and Tibshirani (1993:183). Most importantly, the approximate bootstrap interval endpoints require substantially less computational effort than is required for the bias-corrected accelerated percentile intervals. The approximate bootstrap interval endpoints are invariant under reparameterisation and are second-order accurate.

Most of the published work on bootstrap confidence intervals has concentrated on *equal-tailed* two-sided intervals. An equal-tailed $(1-\alpha)$-level confidence interval for $\tau$ is of the form

$$[\hat{\tau} - \hat{a}, \hat{\tau} + \hat{b}],$$

where $\hat{a}$ and $\hat{b}$ are chosen so that

$$P(\tau < \hat{\tau} - \hat{a}) \doteq P(\tau > \hat{\tau} + \hat{b}) \doteq \frac{\alpha}{2}.$$

Hall (1988a) discussed *symmetric* two-sided bootstrap-$t$ intervals of the form

$$[\hat{\tau} - \hat{c}, \hat{\tau} + \hat{c}],$$

where $\hat{c}$ is chosen so that

$$P(|\hat{\tau} - \tau| > \hat{c}) \doteq \alpha.$$

Hall(1988a) showed that in quite general circumstances equal-tailed bootstrap-$t$ intervals have coverage error $O(n^{-1})$ and that symmetric bootstrap-$t$ intervals have coverage error $O(n^{-2})$. Beran's (1987) pivotal iteration method reduces the error to a very low $O(n^{-3})$. It was also proved that symmetric intervals are not necessarily any longer than equal-tailed intervals.

## 3.5 The modified bootstrap

Suppose $T_n(\boldsymbol{X}_n; F)$ is some specified variable of interest, depending upon the underlying cumulative distribution function $F$. The bootstrap method consists of approximating the sampling distribution of $T_n(\boldsymbol{X}_n; F)$ under $F$ by the bootstrap distribution of $T_n(\boldsymbol{X}_n^*; F_n)$ under $F_n$, where $\boldsymbol{X}_n^* = (X_1^*, X_2^*, \ldots, X_n^*)$ denotes a random sample of size $n$ from $F_n$, i.e.,

$$P(T_n(\boldsymbol{X}_n; F) \in A) \simeq P^*(T_n(\boldsymbol{X}_n^*; F_n) \in A), \tag{3.7}$$

for any Borel set $A$. $P^*$ denotes the conditional probability law of $T_n(\boldsymbol{X}_n^*; F_n)$ given $\boldsymbol{X}_n$.

Singh (1981) and Bickel and Freedman (1981) showed that approximation (3.7) is asymptotically valid in a large number of situations, including $t$-statistics, the empirical and quantile processes and von Mises functionals. However, they also provided some counter-examples to show where the bootstrap apparently fails. These examples are for pivotal quantities based on a $U$-statistic, extreme order statistics and spacings of the observations $\boldsymbol{X}_n$.

The basic idea of the modified bootstrap procedure, suggested by Swanepoel (1986), is to replace $T_n(\boldsymbol{X}_n^*; F_n)$ in (3.7) by $T_m(\boldsymbol{X}_m^*; F_n)$ for some suitable choice of the bootstrap sample size $m$ in terms of $n$. Swanepoel (1990) showed how the counter-examples to Efron's (1979) bootstrap method, given by Bickel and Freedman (1981), can be mended by the modified bootstrap procedure. Hall (1990) applied the modified bootstrap procedure in the context of automatic bandwidth selection (see Section 2.4).

# Chapter 4

# Numerical studies

## 4.1 Introduction

In this chapter the results of extended simulation studies are reported. Data were generated from a variety of populations. The different underlying probability density functions are discussed in Section 4.2. Sections 4.3-4.7 deal with the estimation of the minimum of a density, $f(\theta)$, in various aspects. Consideration of the estimation of the antimode $\theta$, is postponed to Section 4.8 for reasons to be explained later. Monte Carlo estimates of the bias and the mean squared error of the estimator proposed in Chapter 1, $\eta_n$, are reported in Section 4.3. This is done in order to investigate the role played by the smoothing parameters, $s_n$ and $r_n$, in the behaviour of $\eta_n$. Data-driven methods to determine the smoothing parameters are discussed in Section 4.4. In Section 4.5 possible alternative estimators of $f(\theta)$ are discussed. Section 4.6 consists of the results of simulation studies in which the small and moderate sample behaviour of all the estimators is investigated and compared. Confidence intervals are constructed by using, among others, the double bootstrap in Section 4.7. Section 4.8 deals with the estimation of $\theta$ for selected cases; alternative estimators, possible smoothing methods and some simulation results are discussed. Interesting real data applications from the field of Astrophysics are considered in Section 4.9.

## 4.2 Target densities

In this chapter the results of a series of Monte Carlo studies which were performed to investigate the accuracy of the proposed and possible alternative estimators for $\theta$ and $f(\theta)$ are reported. For convenience, densities with compact support $[0,1]$ were considered. Data were generated from fifteen different populations. The first nine cases were populations with different von Mises probability density functions according to different choices of parameters. The last six cases were constructed in order to include nonperiodical and more complex densities.

A random variable $X$ is said to have a von Mises distribution if its probability density function is given by

$$f(x) = \frac{1}{2\pi I_o(\kappa)} e^{\kappa \cos(x-\mu)}, \quad 0 < x \leq 2\pi, \ \kappa > 0, \ 0 \leq \mu < 2\pi,$$

where $I_o(\kappa)$ is the modified Bessel function of the first kind and order zero, i.e.,

$$I_o(\kappa) = \sum_{r=0}^{\infty} \frac{1}{(r!)^2} \left(\frac{1}{2}\kappa\right)^{2r}.$$

The parameter $\mu$ is the mean direction while the parameter $\kappa$ is described as the concentration parameter. Note that $f(0) = f(2\pi)$.

This distribution plays a dominant role in statistical inference on the circle and its importance there is almost the same as that of the normal distribution on the line. The distribution is unimodal and is symmetrical about $x = \mu$. The mode is at $x = \mu$ and the antimode $\theta$, is at $x = \mu - \pi$. As $\kappa \to 0$, the distribution converges to the uniform distribution; as $\kappa \to \infty$, the distribution tends to the point distribution concentrated in the direction $\mu$. The reader is referred to Mardia (1972) and Fisher (1993) for a discussion on the various properties of the von Mises distribution as well as its relation with other distributions.

Since consideration is limited to the interval $[0,1]$, the von Mises density defined by the following, was used,

$$f(x) = \begin{cases} \frac{1}{I_o(\kappa)} e^{\kappa \cos[2\pi(x-\mu)]}, & 0 \leq x \leq 1, \ \kappa > 0, \ 0 \leq \mu < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Nine von Mises densities were considered, according to different choices for $\kappa$ and $\theta = \mu - \frac{1}{2}$. These are displayed in the following table.

Table 4.1: *Von Mises densities.*

| Case | $\theta$ | $\kappa$ | $f(\theta)$ |
|---|---|---|---|
| 1 | 0.125 | 0.25 | 0.767 |
| 2 | 0.125 | 0.50 | 0.570 |
| 3 | 0.125 | 1.00 | 0.291 |
| 4 | 0.250 | 0.25 | 0.767 |
| 5 | 0.250 | 0.50 | 0.570 |
| 6 | 0.250 | 1.00 | 0.291 |
| 7 | 0.500 | 0.25 | 0.767 |
| 8 | 0.500 | 0.50 | 0.570 |
| 9 | 0.500 | 1.00 | 0.291 |

In order to include, among others, nonperiodical densities with $f(0) \neq f(1)$, six additional density functions were constructed.

The densities are given by:

**Case 10:** $f(x) = 5x^2 - 5x + \frac{11}{6}$,

**Case 11:** $f(x) = 5x^2 - \frac{15}{2}x + \frac{37}{12}$,

**Case 12:** $f(x) = x^2 - x + \frac{7}{6}$,

**Case 13:** $f(x) = x^2 - \frac{3}{2}x + \frac{17}{12}$,

**Case 14:** $f(x) = g(x)/\{\int g(x)dx\}$, where

$$g(x) = c_1x^4 + c_2x^3 + c_3x^2 + c_4x + c_5,$$

with

$$c_1 = 54.025,\ c_2 = -111.1,\ c_3 = 69.126,\ c_4 = -12.052,\ c_5 = 1.1,$$

**Case 15:** $f(x) = g(x)/\{\int g(x)dx\}$, where

$$g(x) = c_1x^6 + c_2x^5 + c_3x^4 + c_4x^3 + c_5x^2 + c_6x + c_7,$$

with

$$c_1 = -1712.006,\ c_2 = 5136.018,\ c_3 = -5755.922,\ c_4 = 2951.814,$$

$$c_5 = -672.680,\ c_6 = 52.776,\ c_7 = 0.8.$$

The constants $c_1, c_2, \ldots, c_5$ (Case 14) and $c_1, c_2, \ldots, c_7$ (Case 15) were numerically obtained by fitting polynomials to specified values for the modes and maxima, and the antimodes and minima of the densities. Table 4.2 displays the values of $\theta$ and $f(\theta)$ for Cases 10-15.

Table 4.2: *Nonperiodical densities.*

| Case | $\theta$ | $f(\theta)$ |
|---|---|---|
| 10 | 0.50 | 0.583 |
| 11 | 0.75 | 0.271 |
| 12 | 0.50 | 0.917 |
| 13 | 0.75 | 0.854 |
| 14 | 0.12 | 0.409 |
| 15 | 0.25, 0.75 | 0.161 |

Cases 1-11 are displayed in Figure 4.1. Cases 12-15 are viewed as being more difficult as far as the estimation of $f(\theta)$ is concerned (see Figure 4.2).

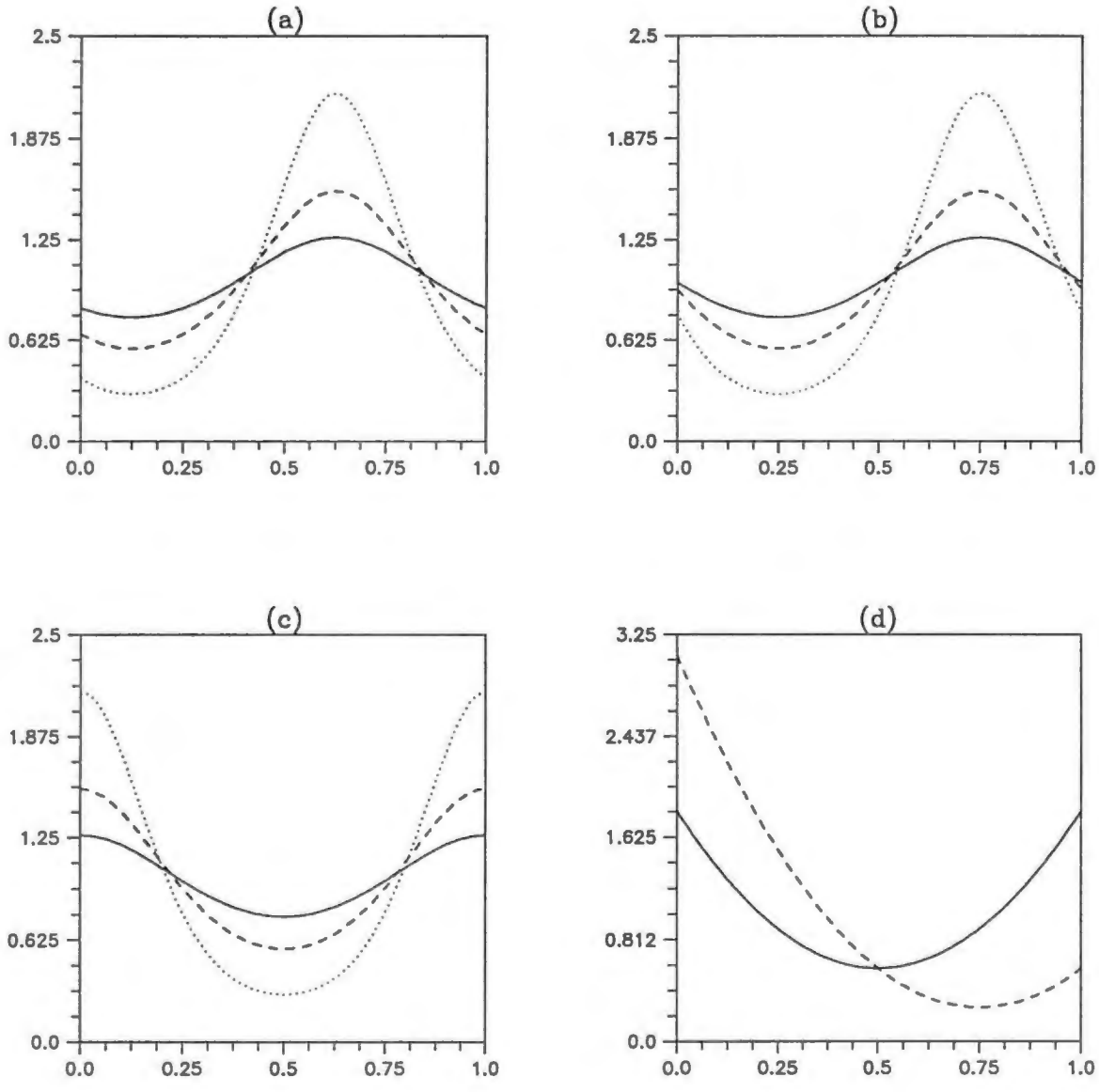


Figure 4.1: (a) *Case 1 is represented by the solid curve, Case 2 by the dashed curve and Case 3 by the dotted curve.* (b) *Case 4 is represented by the solid curve, Case 5 by the dashed curve and Case 6 by the dotted curve.* (c) *Case 7 is represented by the solid curve, Case 8 by the dashed curve and Case 9 by the dotted curve.* (d) *Case 10 is represented by the solid curve and Case 11 by the dashed curve.*

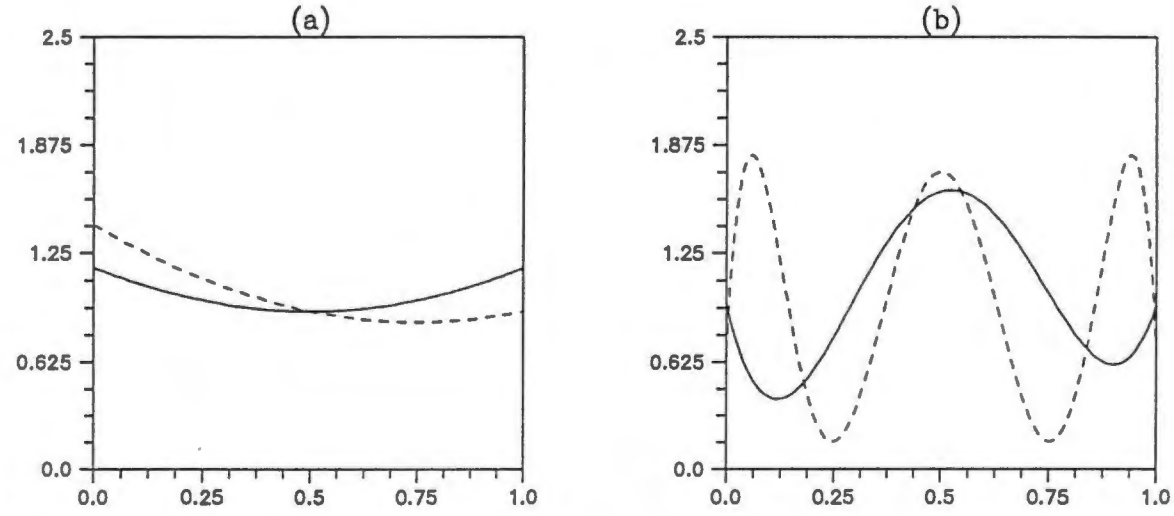


Figure 4.2: (a) *Case 12 is represented by the solid curve and Case 13 by the dashed curve.* (b) *Case 14 is represented by the solid curve and Case 15 by the dashed curve.*

## 4.3 Optimal choice of smoothing parameter

The estimator of $f(\theta)$, $\eta_n$, (see (1.15)), is defined in terms of two smoothing parameters, $s = s_n$ and $r = r_n$. A simulation study was first conducted to investigate the influence of different choices of $s$ and $r$ on the MSE of $\eta_n$, viz.

$$\text{MSE}(\eta_n) = E\left(\eta_n - \min_{0\le x\le 1} f(x)\right)^2 = E(\eta_n - f(\theta))^2. \tag{4.1}$$

Monte Carlo estimates (based on 2000 independent trials) were obtained for $\text{MSE}(\eta_n)$ for all pairs $(r, s)$, where both $r$ and $s$ were allowed to vary over all integers between 1 and $[(n-1)/2]$, under the restriction $r \le s$. All fifteen densities were considered and sample sizes $n = 100$ and $n = 400$ were used. It turned out that in each case $\text{MSE}(\eta_n)$ attained its minimum at some unique pair $(r, s)$ with $r = s$. No theoretical explanation for this behaviour can be given at this point. Table 4.3 displays the minimum values of $\text{MSE}(\eta_n)$, and the corresponding optimal values $r = s$. Estimates of $E\eta_n$ based on these optimal values are also included and comparing them with the $f(\theta)$-values, show that $\eta_n$ is in all cases (except Case 15, for $n = 100$) almost unbiased.

Table 4.3: *Monte Carlo estimates of bias and minimum MSE.*

| | | $n = 100$ | | | $n = 400$ | | |
|---|---|---|---|---|---|---|---|
| Case | $f(\theta)$ | $s_n$ | $E\eta_n$ | min(MSE) | $s_n$ | $E\eta_n$ | min(MSE) |
| 1 | 0.767 | 13 | 0.782 | 0.006 | 46 | 0.773 | 0.003 |
| 2 | 0.570 | 6 | 0.564 | 0.007 | 28 | 0.578 | 0.003 |
| 3 | 0.291 | 2 | 0.301 | 0.005 | 11 | 0.293 | 0.002 |
| 4 | 0.767 | 16 | 0.775 | 0.005 | 57 | 0.776 | 0.002 |
| 5 | 0.570 | 9 | 0.581 | 0.006 | 33 | 0.578 | 0.002 |
| 6 | 0.291 | 3 | 0.299 | 0.004 | 13 | 0.294 | 0.001 |
| 7 | 0.767 | 19 | 0.782 | 0.004 | 57 | 0.771 | 0.002 |
| 8 | 0.570 | 10 | 0.585 | 0.005 | 34 | 0.580 | 0.002 |
| 9 | 0.291 | 3 | 0.294 | 0.004 | 13 | 0.295 | 0.001 |
| 10 | 0.583 | 11 | 0.598 | 0.005 | 37 | 0.593 | 0.002 |
| 11 | 0.271 | 3 | 0.276 | 0.003 | 13 | 0.276 | 0.001 |
| 12 | 0.917 | 35 | 0.929 | 0.002 | 125 | 0.926 | 0.001 |
| 13 | 0.854 | 26 | 0.865 | 0.003 | 96 | 0.862 | 0.001 |
| 14 | 0.409 | 3 | 0.432 | 0.006 | 11 | 0.418 | 0.003 |
| 15 | 0.161 | 1 | 0.233 | 0.008 | 2 | 0.158 | 0.001 |

Consequently, the following estimator of $f(\theta)$ was considered during the rest of the numerical studies,

$$\eta_n = \frac{n^{-1}(2s_n + 1)}{Y_{K_n+s_n} - Y_{K_n-s_n}}, \tag{4.2}$$

with $K_n$ as defined in (1.3).

Figures 4.3-4.17 show the behaviour of $E\eta_n$ and MSE as functions of $s_n$. Two sample sizes, $n = 100$ and $n = 400$, were considered. The number of trials was 2000. For each case it is clear that there exists a unique choice of $s_n$ where the MSE is minimised. This is indicated by a vertical dotted line. Note that in some cases the MSE has values close to its minimum for a range of $s_n$-values. (See Figures 4.6, 4.9, 4.12, 4.14 and 4.15.) This implies (as will be seen later) that the estimator $\eta_n$ will be quite accurate for any of these values of $s_n$. This is in contrast with the cases in Figures 4.5, 4.8, 4.11, 4.13, 4.16 and 4.17.

Hence, from the figures it is clear that a "correct" choice of smoothing $s_n$ is possible. In Section 4.4 data-driven choices of $s_n$ will be discussed. The performance of $\eta_n$ based on these data-based smoothers will be empirically evaluated in Section 4.6.

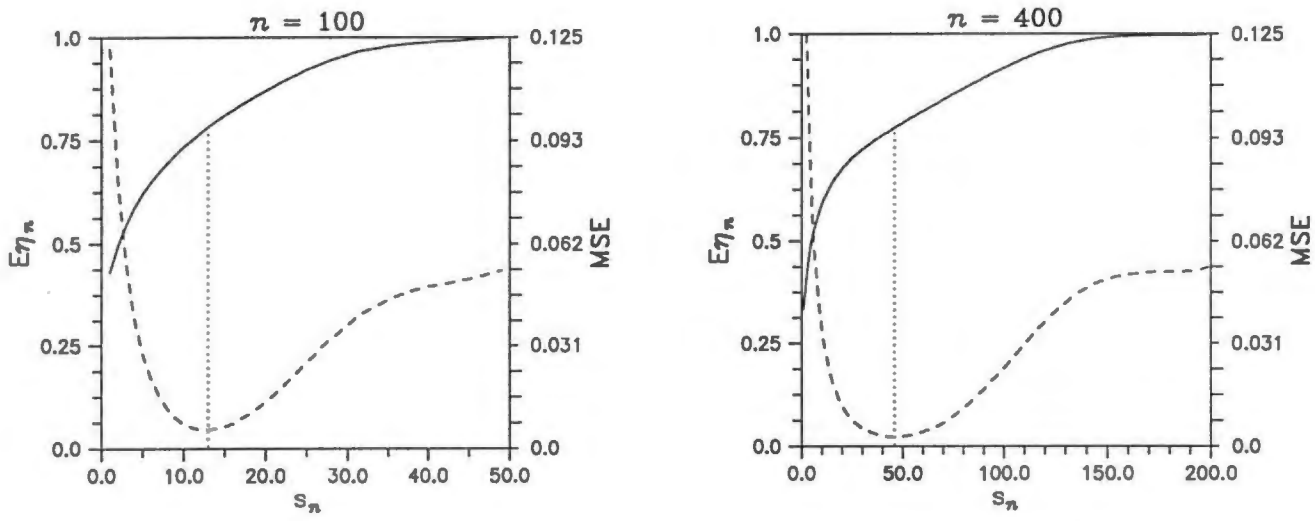


Figure 4.3: *Case 1 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

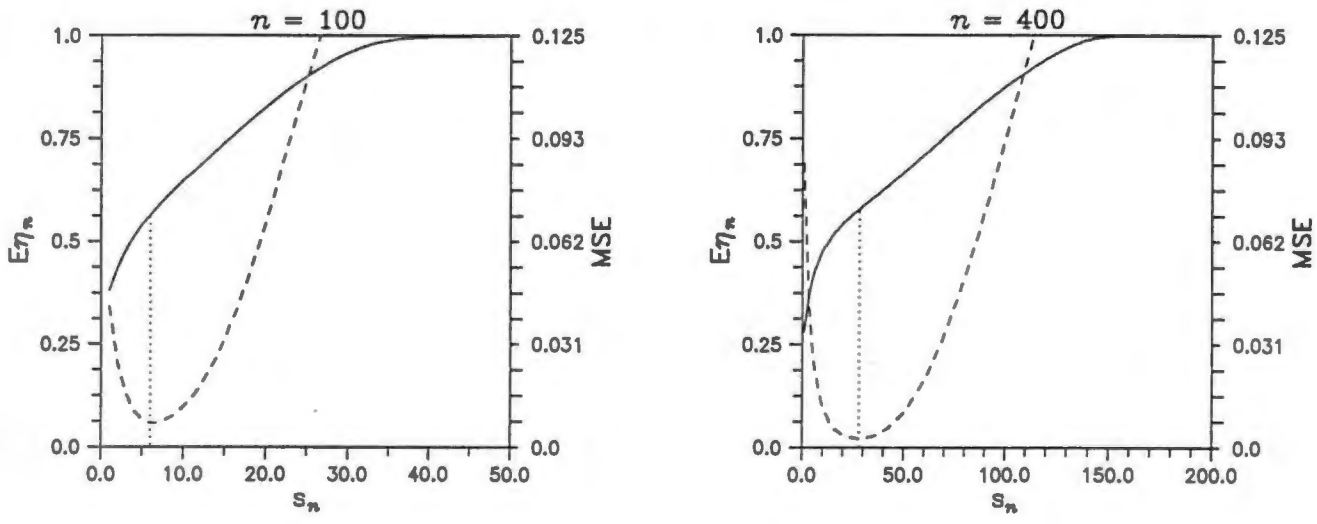


Figure 4.4: *Case 2 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

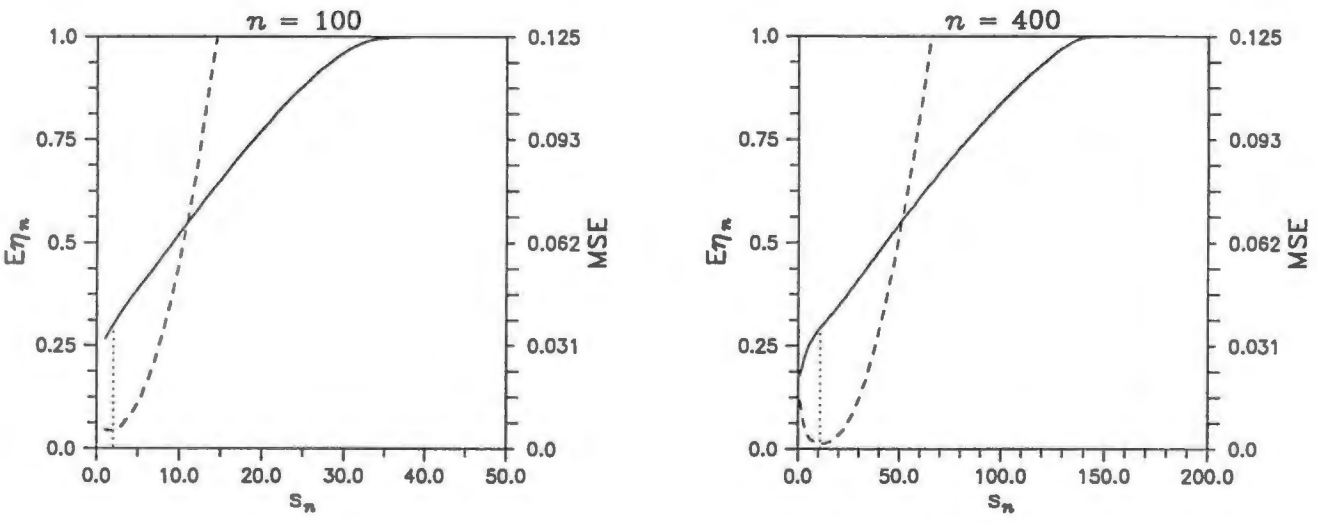


Figure 4.5: *Case 3 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

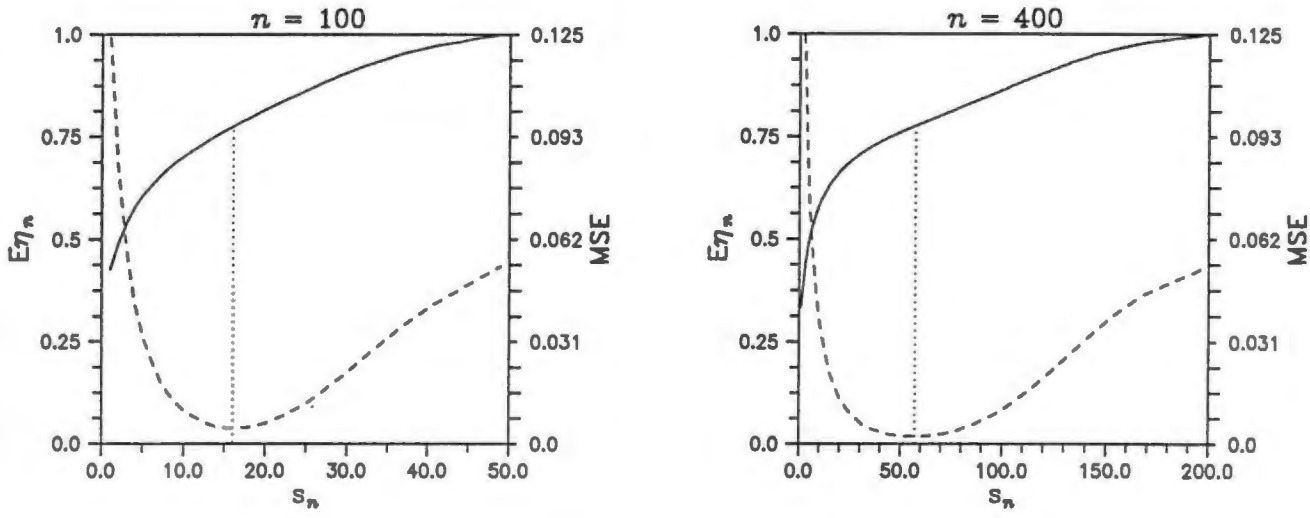


Figure 4.6: *Case 4 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

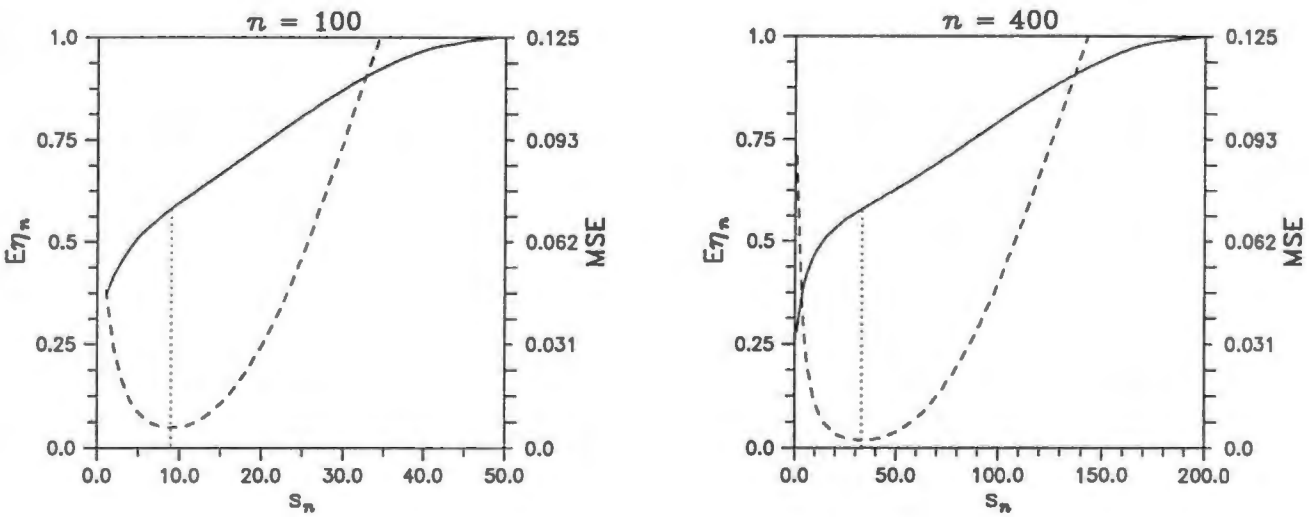


Figure 4.7: *Case 5 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

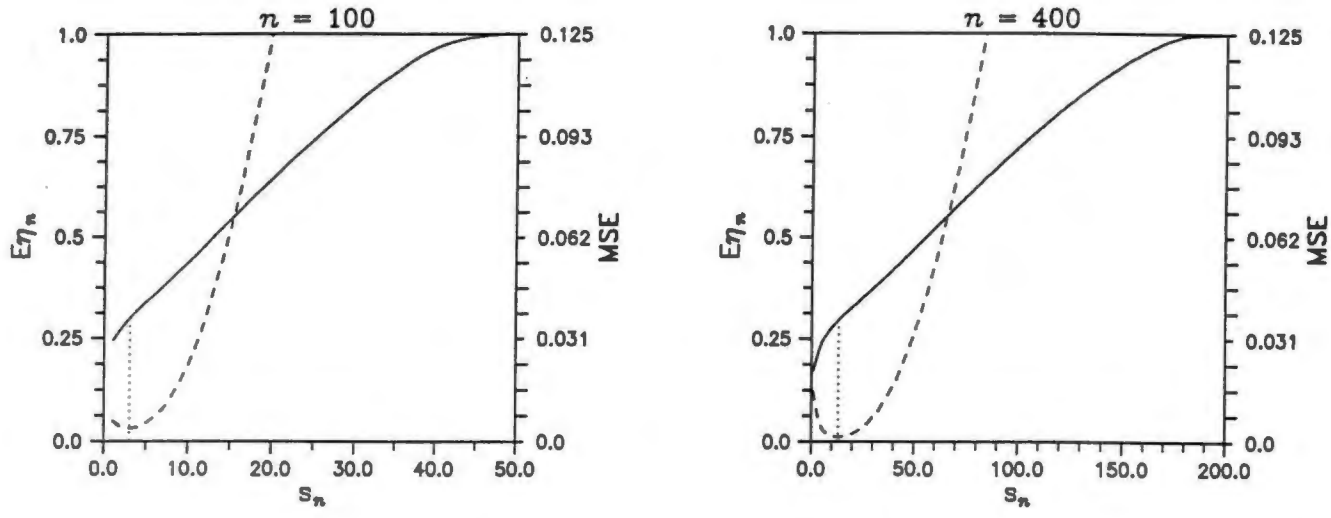


Figure 4.8: *Case 6 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

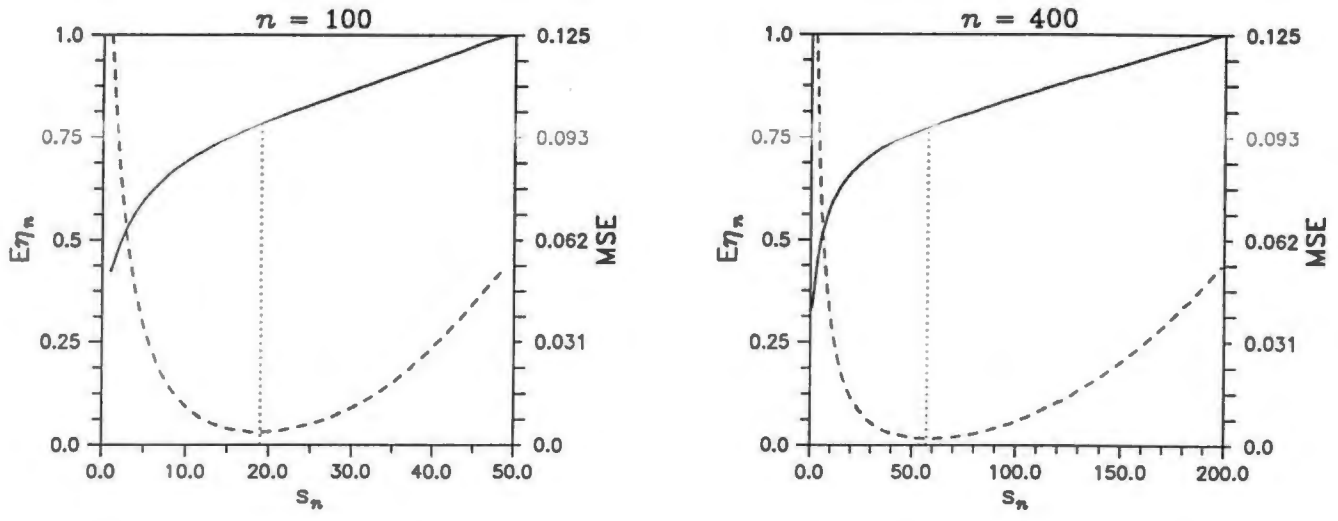


Figure 4.9: *Case 7 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

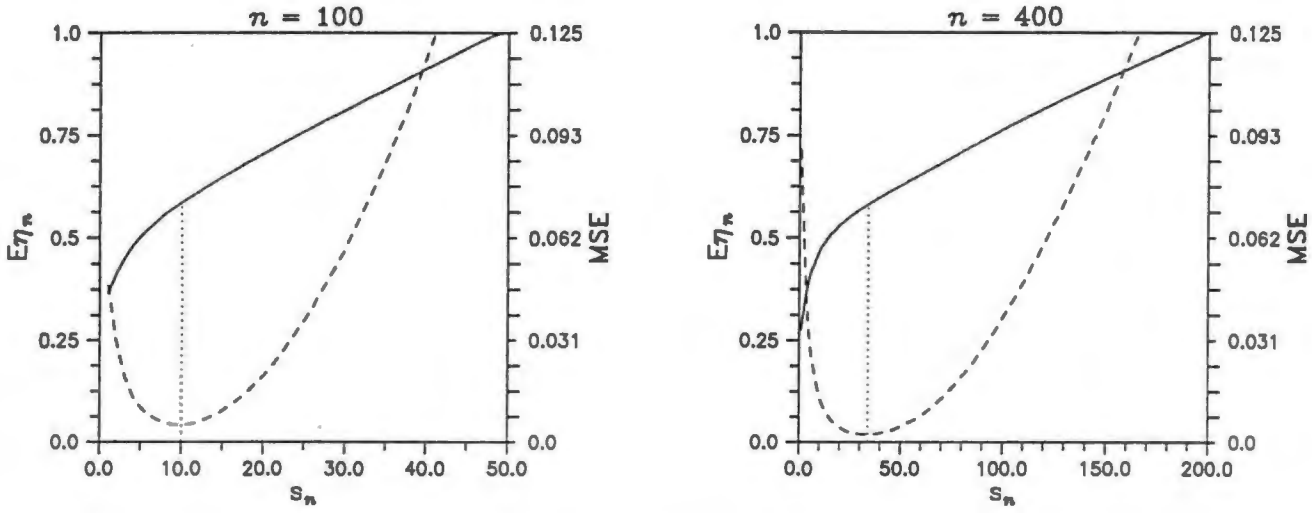


Figure 4.10: *Case 8 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

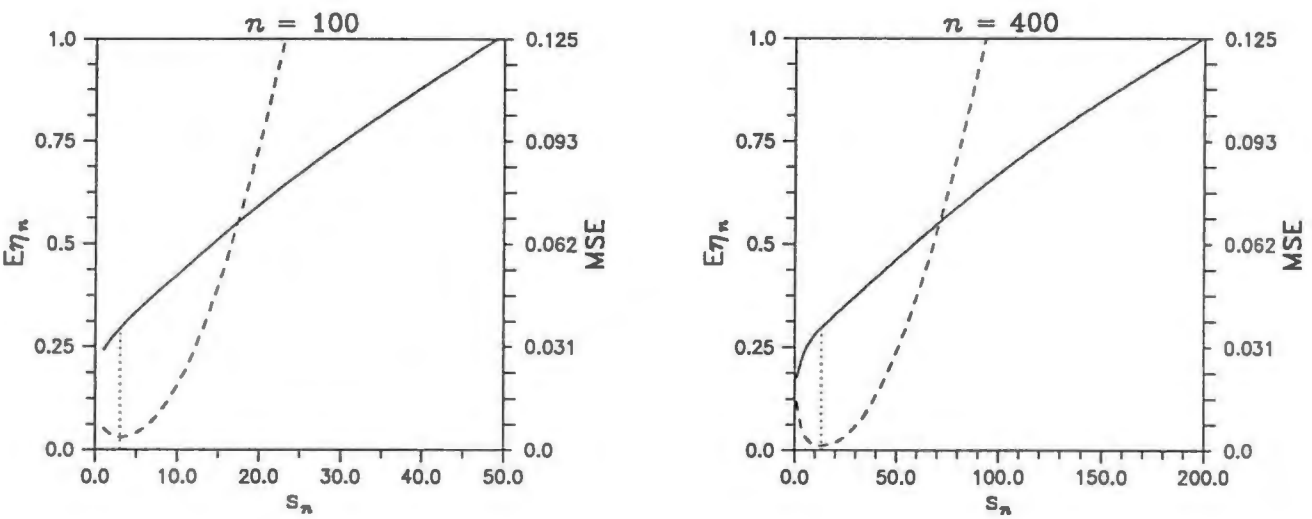


Figure 4.11: *Case 9 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

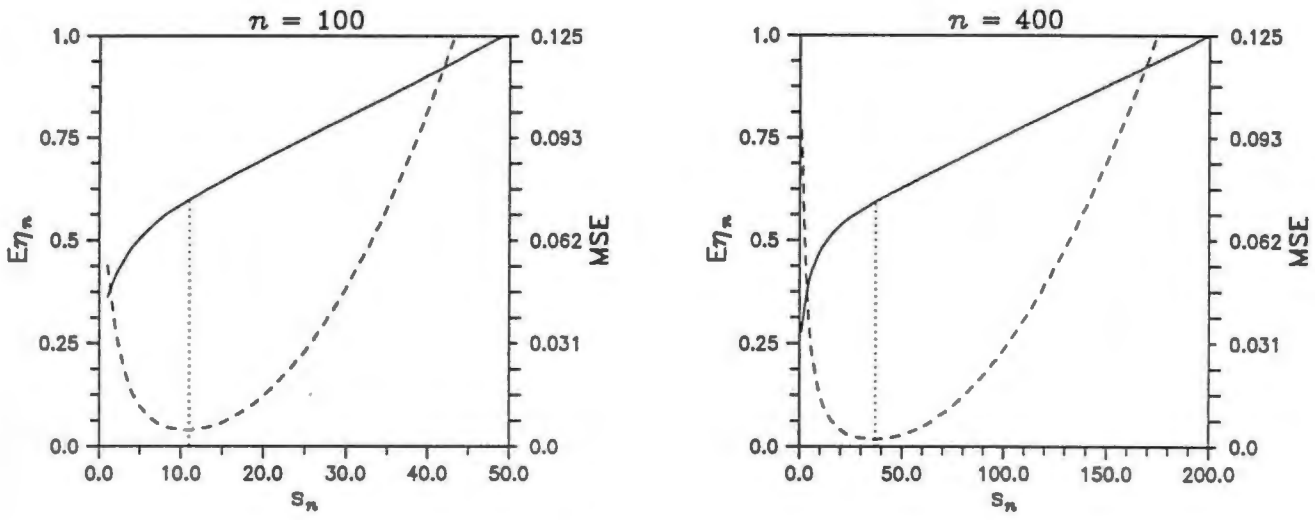


Figure 4.12: *Case 10 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

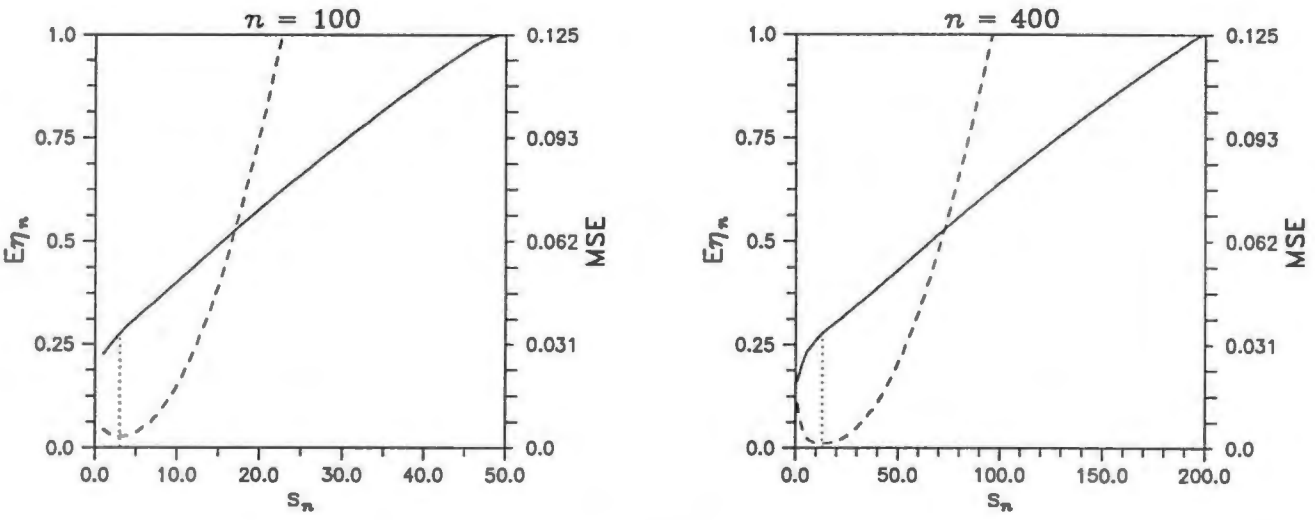


Figure 4.13: *Case 11 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

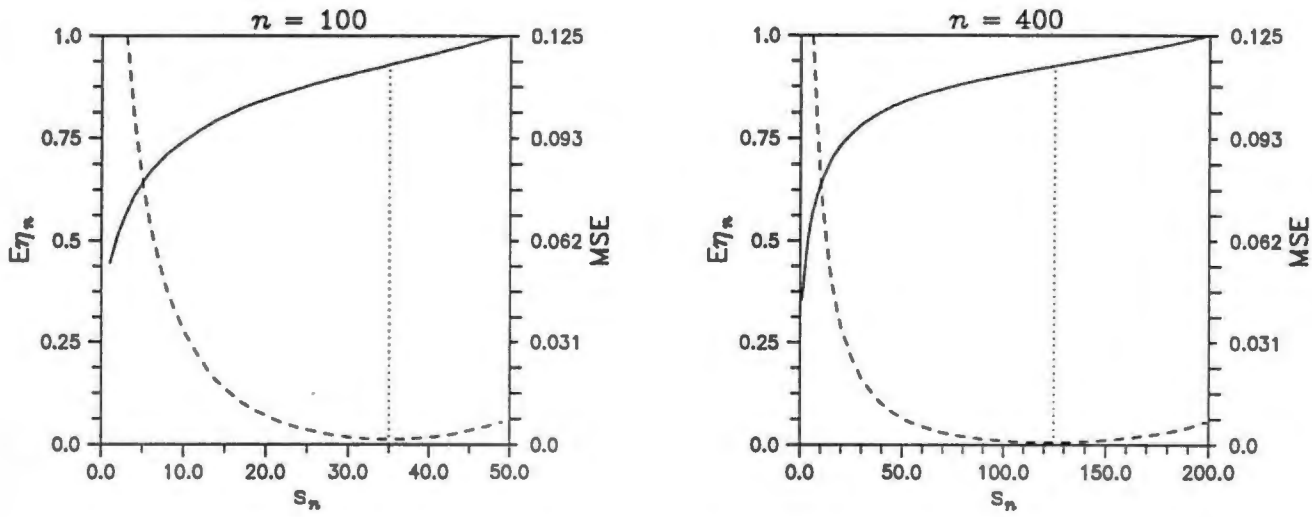


Figure 4.14: *Case 12 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

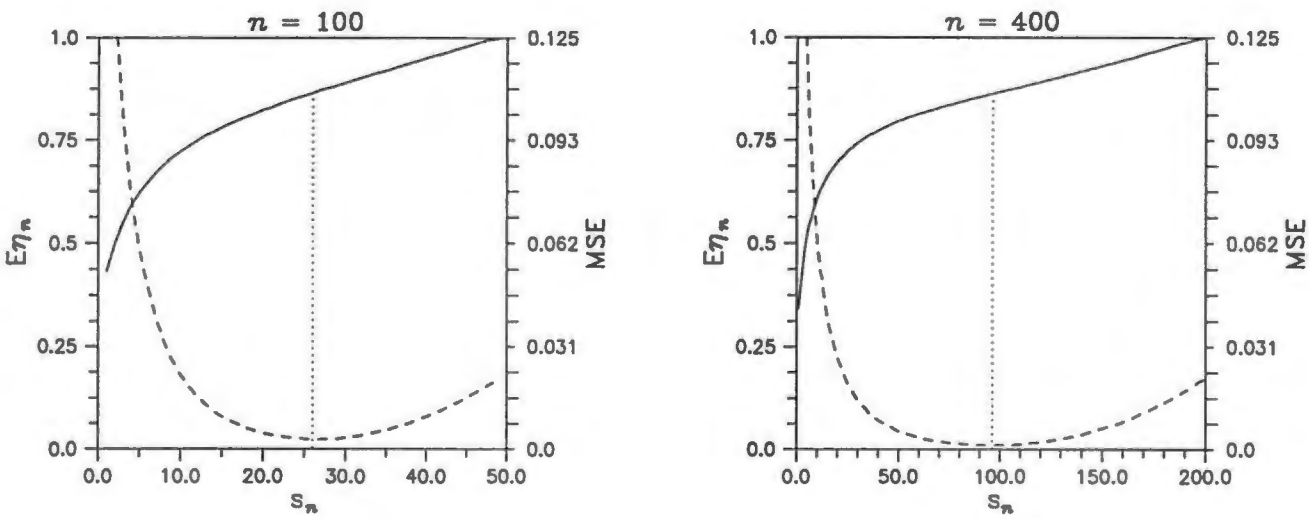


Figure 4.15: *Case 13 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

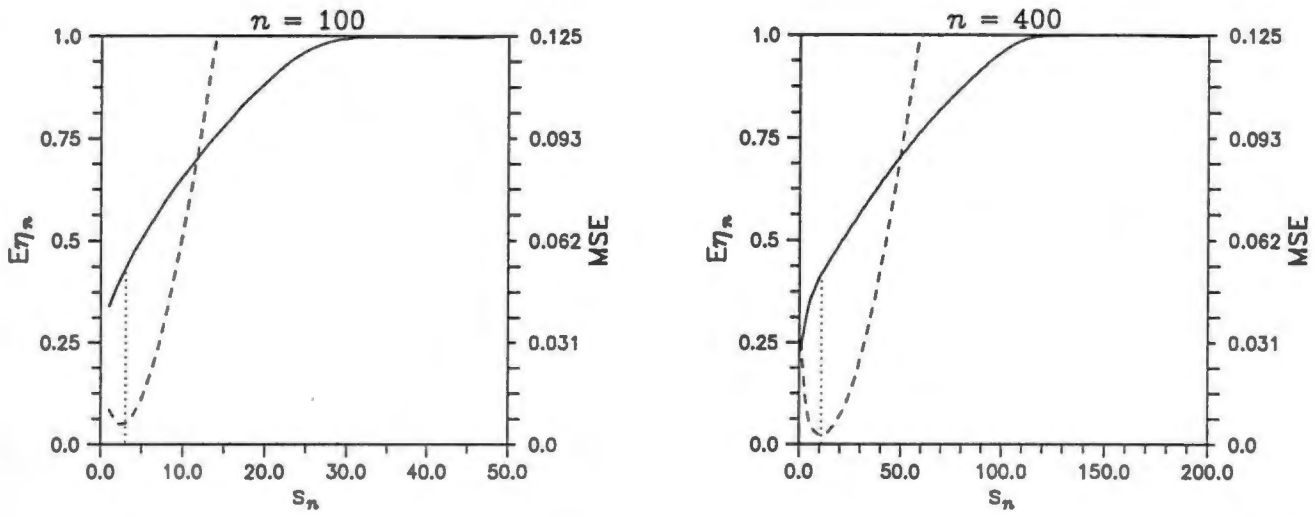


Figure 4.16: *Case 14 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

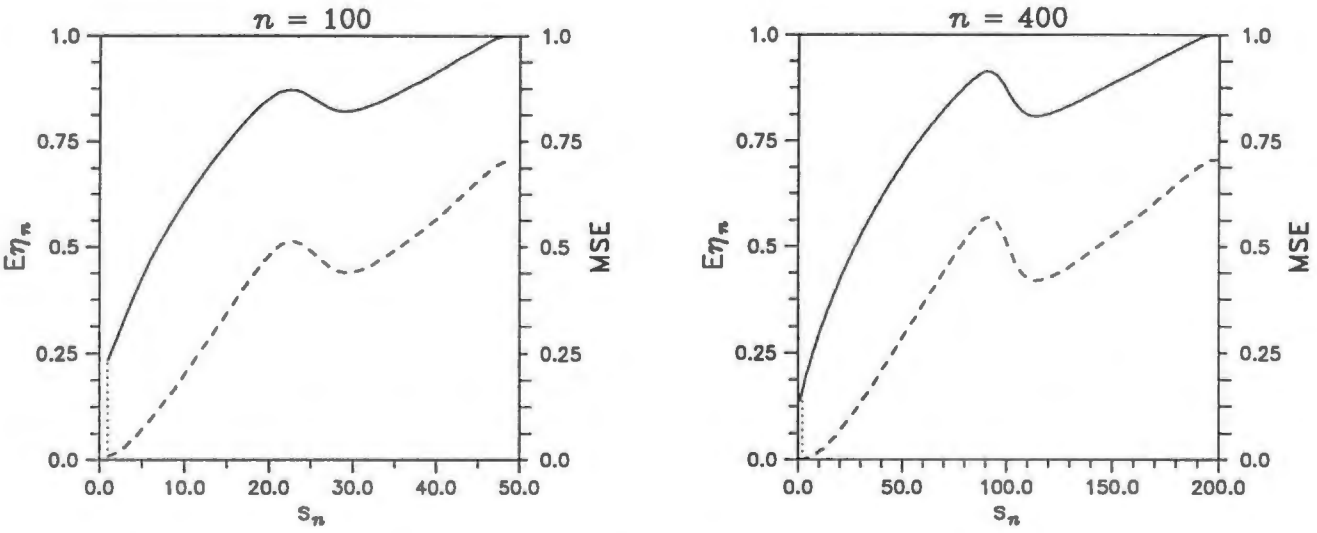


Figure 4.17: *Case 15 - Monte Carlo estimates of* $E\eta_n$ *(solid curve) and* $E(\eta_n - f(\theta))^2$ *(dashed curve).*

## 4.4 Data-based choices of the smoothing parameter

The importance of a data-dependent method of selecting the smoothing parameter $s_n$ was discussed in Section 4.3. In this section some smoothing methods are discussed.

A first possibility is to choose $s_n$ as the minimiser of an estimator of $\text{MSE}(\eta_n)$ (see (4.1)). The naive bootstrap estimate of $\text{MSE}(\eta_n)$ is $E_*(\eta_n^* - \eta_n)^2$, where $\eta_n^* = \eta_n(X_1^*, X_2^*, \ldots, X_n^*)$ and $X_1^*, X_2^*, \ldots, X_n^*$ is the bootstrap random sample from $F_n$, the empirical distribution function of $X_1, X_2, \ldots, X_n$. $E_*$ denotes the expectation over the conditional law of $X_1^*, X_2^*, \ldots, X_n^*$ given $X_1, X_2, \ldots, X_n$. In all the simulation studies, it was found that this estimate adequately estimated the variance component of the MSE, but it estimated the bias component as almost zero. This failure of the naive bootstrap is also experienced in bootstrap-based procedures determining the bandwidth in kernel density estimation. (See the discussion in Section 2.4.)

However, an alternative bootstrap estimate of $\text{MSE}(\eta_n)$ is (see the remark at the end of Section 3.2),

$$\text{MSE}_*(\eta_n) = E_*\left(\eta_n^* - \min_{0\le x\le 1} \hat{f}_n(x)\right)^2, \tag{4.3}$$

where $\hat{f}_n$ is some density estimate. Simulation studies proved that in using this definition of $\text{MSE}_*(\eta_n)$, the bias of $\eta_n$ is estimated adequately. The first proposed data-based choice of $s_n$ is the minimiser (with respect to $s_n$) of $\text{MSE}_*(\eta_n)$ defined by (4.3).

Let (see (4.2))

$$V_n = Y_{K_n+s_n} - Y_{K_n-s_n} = \max_{s_n+1\le j\le n-s_n} (Y_{j+s_n} - Y_{j-s_n}),$$

the stochastic component of $\eta_n$. Since the smallest concentration of observations occurs in a neighbourhood of the antimode (which implies that $\text{Var}(V_n)$ will typically be large), a second data-based choice of $s_n$ is suggested as the maximiser of $\text{Var}(V_n)$. In case of local maxima, the $s_n$ corresponding to the first local maximum is chosen. Figures 4.18-4.24 display the behaviour of $\text{Var}(V_n)$ as a function of $s_n$. Additional motivation for this method is found by comparing the optimal $s_n$-values displayed in Table 4.3 with the $s_n$-values that maximise $\text{Var}(V_n)$. The bias and MSE of $\eta_n$ based on these $s_n$-values can be obtained from Figures 4.3-4.17, and compared with the optimal values of bias and MSE.

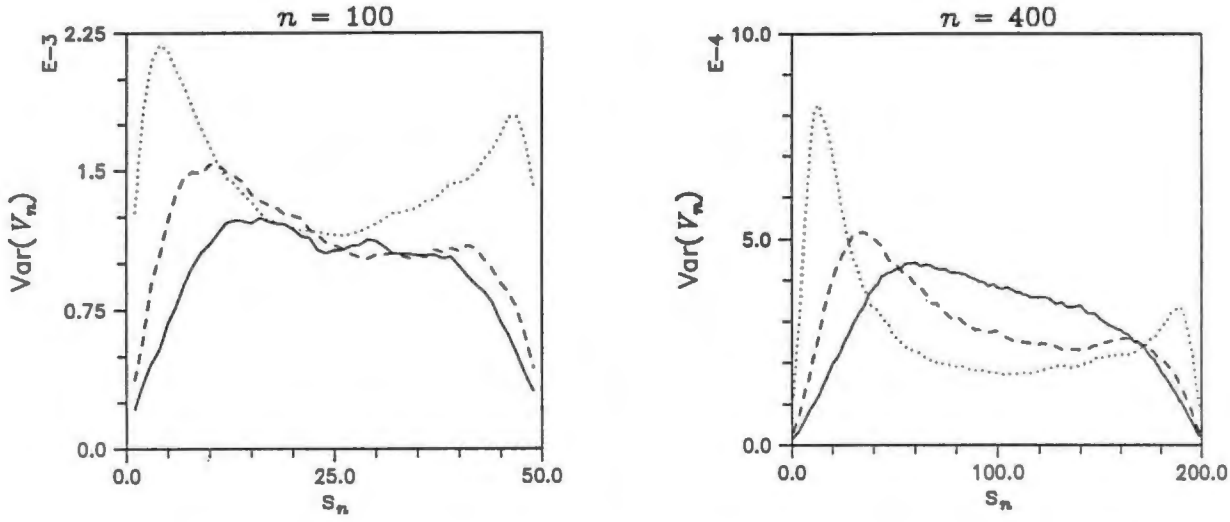


Figure 4.18: *Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials. Case 1 is represented by the solid curve, Case 2 by the dashed curve and Case 3 by the dotted curve.*

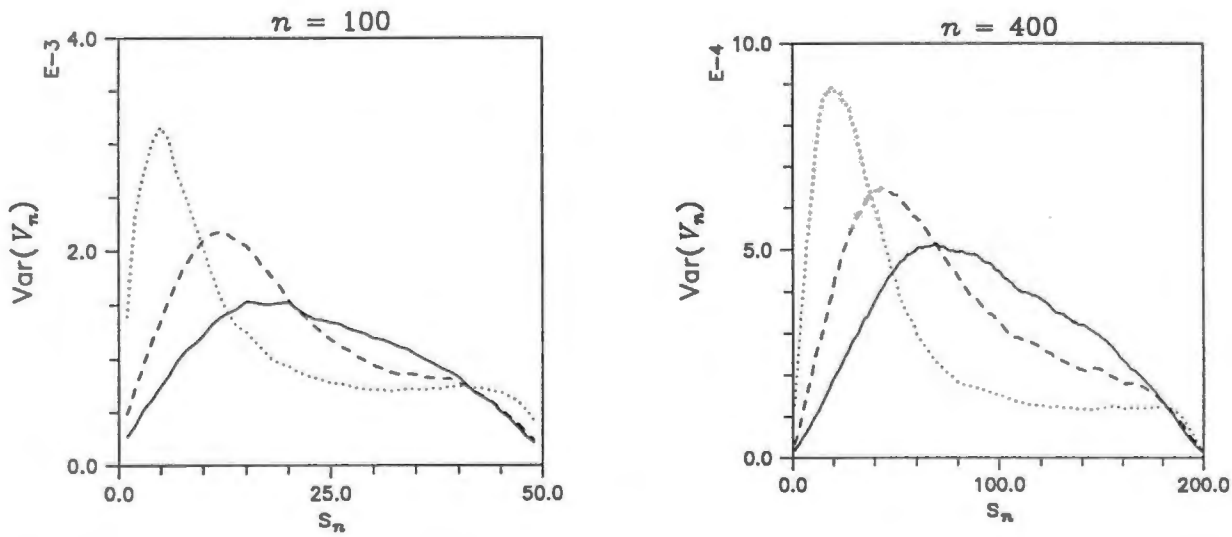


Figure 4.19: *Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials. Case 4 is represented by the solid curve, Case 5 by the dashed curve and Case 6 by the dotted curve.*

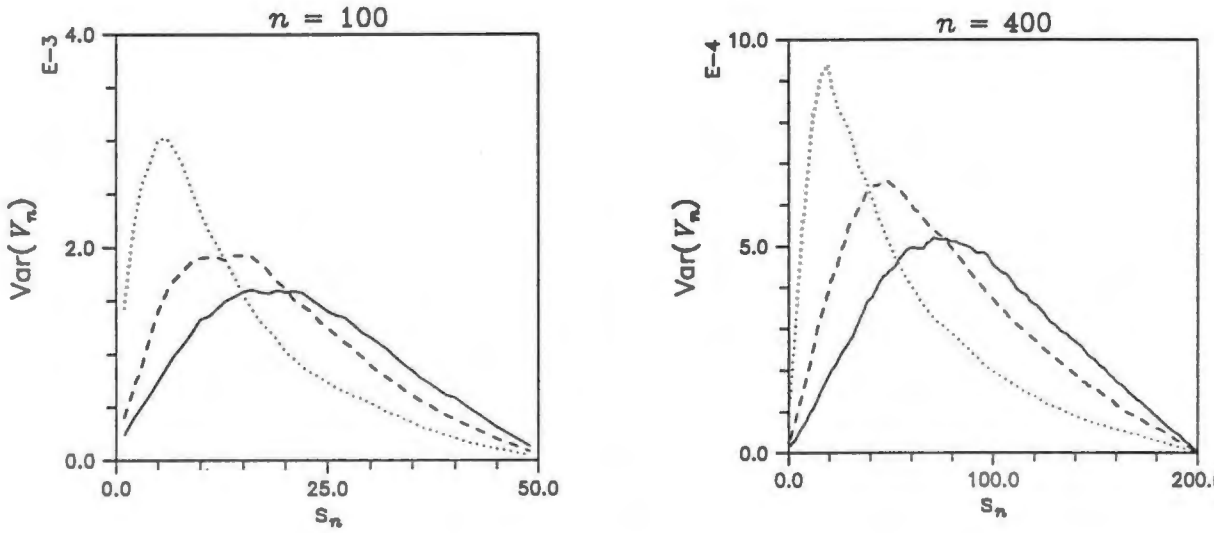


Figure 4.20: *Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials. Case 7 is represented by the solid curve, Case 8 by the dashed curve and Case 9 by the dotted curve.*

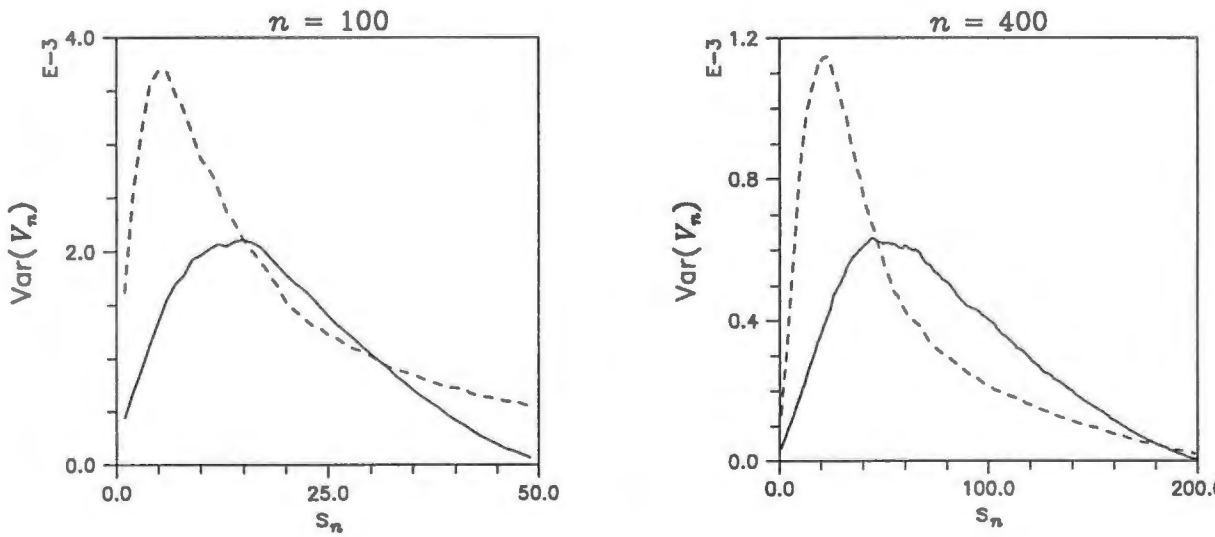


Figure 4.21: *Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials. Case 10 is represented by the solid curve and Case 11 by the dashed curve.*

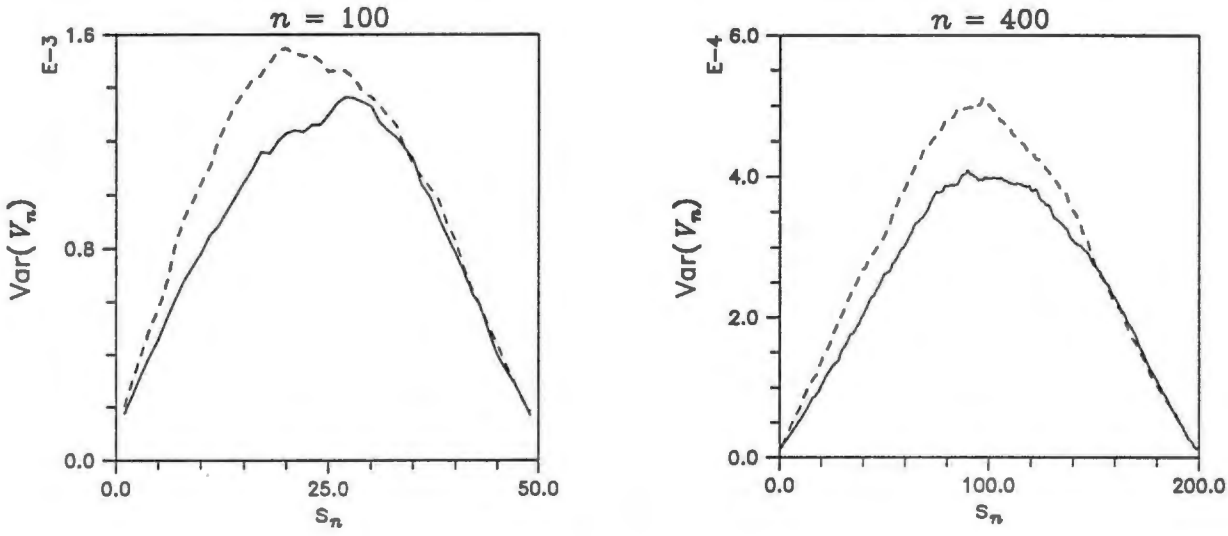


Figure 4.22: *Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials. Case 12 is represented by the solid curve and Case 13 by the dashed curve.*

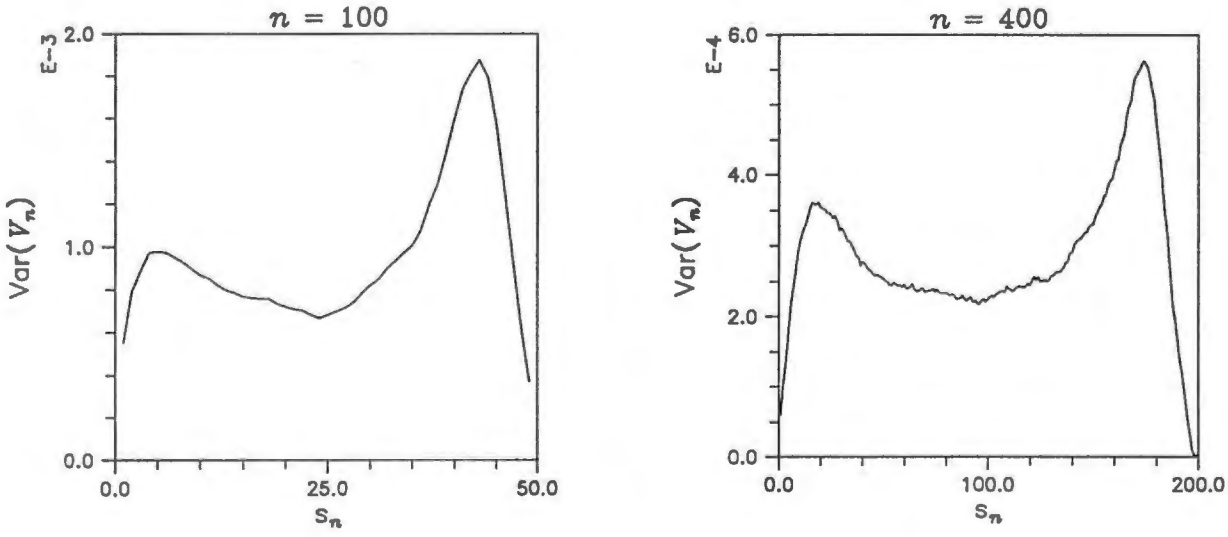


Figure 4.23: *Case 14: Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials.*

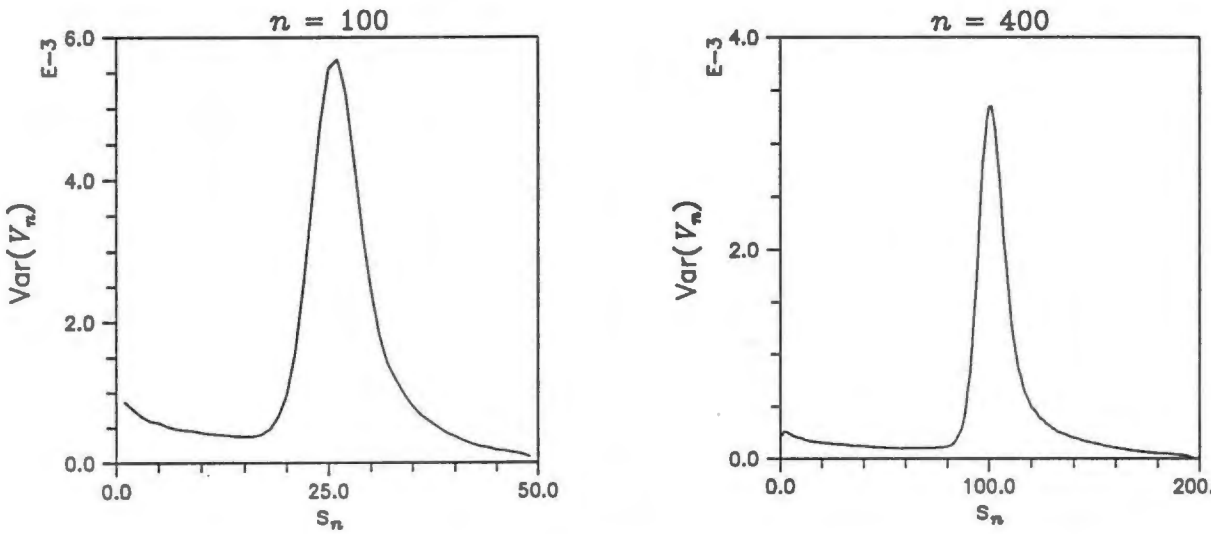


Figure 4.24: *Case 15: Monte Carlo estimates of* $\mathrm{Var}(V_n)$ *based on 2000 independent trials.*

## 4.5 Alternative estimators

Deheuvels (1984) derived strong limit theorems for maximal spacings. Under certain conditions, he proved that $\hat{\zeta}_n \xrightarrow{a.s.} f(\theta)$ as $n \to \infty$, where

$$\hat{\zeta}_n = \frac{n^{-1} \log n}{\max_{1 \le i \le n-1}(Y_{i+1} - Y_i)}.$$

All the simulation studies done showed that $\hat{\zeta}_n$ performed poorly in comparison with the proposed estimator $\eta_n$. A possible explanation for this is that $\hat{\zeta}_n$ does not depend on any smoothing. I therefore do not include $\hat{\zeta}_n$ in what follows.

A kernel-based estimator of $f(\theta)$ is the following,

$$\hat{\xi}_n = \min_{0 \le x \le 1} \tilde{f}_{\hat{h}}(x), \tag{4.4}$$

where $\tilde{f}_{\hat{h}}(x)$ is the corrected kernel estimator suggested by Schuster (1985) (see Section 2.6), viz.

$$\tilde{f}_{\hat{h}}(x) = \hat{f}_{\hat{h}}(-x) + \hat{f}_{\hat{h}}(x) + \hat{f}_{\hat{h}}(2 - x), \tag{4.5}$$

where $\hat{f}_{\hat{h}}(x)$ is the kernel estimator based on some data-based bandwidth $\hat{h}$. Two good choices of $\hat{h}$ are $\hat{h} = \hat{h}_m$ (modified stabilised smoother) and $\hat{h} = \hat{h}_{P2}$ (modified plug-in

smoother), as was pointed out in Section 2.4. Since the density $f$ is assumed to have compact support, we take $K$ as the Epanechnikov kernel,

$$K(x) = \tfrac{3}{4}(1 - x^2), \quad -1 \leq x \leq 1.$$

Firstly, a Monte Carlo study was conducted to investigate the influence of $\hat{h}_m$ and $\hat{h}_{P2}$ on the behaviour op $\hat{\xi}_n$. The sample size was $n = 100$ and the number of independent trials 2000. Tables 4.4 and 4.5 display estimates of $E\hat{h}$, $E\hat{\xi}_n$, RMSE($\hat{\xi}_n$) (root mean squared error of $\hat{\xi}_n$) and RATIO, the ratio (expressed as a percentage) of the RMSE's with the RMSE of $\hat{\xi}_n$ based on $\hat{h}_m$ as nominator. The estimated standard errors of the averages were found to be negligibly small, and were therefore omitted from the tables. It is evident that there is little to choose between the two selectors, although $\hat{h}_m$ delivered slightly better results for almost all cases. Note, however, that $\hat{h}_{P2}$ requires far fewer calculations and is consequently much faster to compute. Similar results were also obtained for other sample sizes.

Another alternative estimator of the minimum of a density, $f(\theta)$, is the following. Let, for $j = s_n + 1, s_n + 2, \ldots, n - s_n$,

$$S^2_{s_n}(j) = \frac{1}{2s_n + 1} \sum_{i=j-s_n}^{j+s_n} (Y_i - \bar{Y}(j))^2,$$

where

$$\bar{Y}(j) = \frac{1}{2s_n + 1} \sum_{i=j-s_n}^{j+s_n} Y_i.$$

Let $J_n$ be an integer-valued random variable defined by

$$S^2_{s_n}(J_n) = \max_{s_n+1 \leq j \leq n-s_n} S^2_{s_n}(j). \tag{4.6}$$

The estimator of $f(\theta)$ which I propose is (compare with (4.2)),

$$\gamma_n = \frac{n^{-1}(2s_n + 1)}{Y_{J_n+s_n} - Y_{J_n-s_n}}. \tag{4.7}$$

Results regarding the strong consistency, strong rates of consistency and asymptotic distribution of $\gamma_n$ can be proved along the same lines as those of $\eta_n$ (see Sections 1.3-1.5), and will therefore be omitted.

Table 4.4: *Monte Carlo estimates of bias and root mean squared error for Cases 1-8.*

| Case | $f(\theta)$ | $\hat{h}$ | $E\hat{h}$ | $E\hat{\xi}_n$ | RMSE($\hat{\xi}_n$) | RATIO |
|---|---|---|---|---|---|---|
| 1 | 0.767 | $\hat{h}_m$ | 0.288 | 0.713 | 0.138 | 95.2 |
| | | $\hat{h}_{P2}$ | 0.275 | 0.705 | 0.145 | |
| 2 | 0.570 | $\hat{h}_m$ | 0.247 | 0.552 | 0.135 | 98.5 |
| | | $\hat{h}_{P2}$ | 0.235 | 0.543 | 0.137 | |
| 3 | 0.291 | $\hat{h}_m$ | 0.183 | 0.264 | 0.111 | 99.1 |
| | | $\hat{h}_{P2}$ | 0.173 | 0.256 | 0.112 | |
| 4 | 0.767 | $\hat{h}_m$ | 0.260 | 0.718 | 0.138 | 94.5 |
| | | $\hat{h}_{P2}$ | 0.246 | 0.708 | 0.146 | |
| 5 | 0.570 | $\hat{h}_m$ | 0.209 | 0.549 | 0.124 | 96.9 |
| | | $\hat{h}_{P2}$ | 0.193 | 0.537 | 0.128 | |
| 6 | 0.291 | $\hat{h}_m$ | 0.165 | 0.276 | 0.096 | 98.0 |
| | | $\hat{h}_{P2}$ | 0.156 | 0.267 | 0.098 | |
| 7 | 0.767 | $\hat{h}_m$ | 0.209 | 0.680 | 0.178 | 93.2 |
| | | $\hat{h}_{P2}$ | 0.197 | 0.666 | 0.191 | |
| 8 | 0.570 | $\hat{h}_m$ | 0.147 | 0.490 | 0.161 | 93.1 |
| | | $\hat{h}_{P2}$ | 0.137 | 0.474 | 0.173 | |

Table 4.5: *Monte Carlo estimates of bias and root mean squared error for Cases 9-15.*

| Case | $f(\theta)$ | $\hat{h}$ | $E\hat{h}$ | $E\hat{\xi}_n$ | RMSE($\hat{\xi}_n$) | RATIO |
|---|---|---|---|---|---|---|
| 9 | 0.291 | $\hat{h}_m$ | 0.102 | 0.199 | 0.139 | 94.6 |
| | | $\hat{h}_{P2}$ | 0.095 | 0.186 | 0.147 | |
| 10 | 0.583 | $\hat{h}_m$ | 0.140 | 0.472 | 0.185 | 93.0 |
| | | $\hat{h}_{P2}$ | 0.130 | 0.454 | 0.199 | |
| 11 | 0.271 | $\hat{h}_m$ | 0.122 | 0.193 | 0.123 | 94.6 |
| | | $\hat{h}_{P2}$ | 0.114 | 0.182 | 0.130 | |
| 12 | 0.917 | $\hat{h}_m$ | 0.269 | 0.775 | 0.201 | 93.9 |
| | | $\hat{h}_{P2}$ | 0.257 | 0.764 | 0.214 | |
| 13 | 0.854 | $\hat{h}_m$ | 0.257 | 0.721 | 0.195 | 94.2 |
| | | $\hat{h}_{P2}$ | 0.244 | 0.710 | 0.207 | |
| 14 | 0.409 | $\hat{h}_m$ | 0.253 | 0.505 | 0.164 | 100.6 |
| | | $\hat{h}_{P2}$ | 0.242 | 0.499 | 0.163 | |
| 15 | 0.161 | $\hat{h}_m$ | 0.091 | 0.176 | 0.099 | 103.1 |
| | | $\hat{h}_{P2}$ | 0.086 | 0.162 | 0.096 | |

Monte Carlo estimates (based on 2000 independent trials) were obtained for $\text{MSE}(\gamma_n)$, viz.

$$\text{MSE}(\gamma_n) = E(\gamma_n - f(\theta))^2,$$

for $s_n = 1, 2, \ldots, [(n-1)/2]$. All fifteen densities were considered and sample sizes $n = 100$ and $n = 400$ were used. Table 4.6 displays the minimum values of $\text{MSE}(\gamma_n)$ and the corresponding optimal $s_n$-values. Estimates of $E\gamma_n$ based on these optimal values are also included and comparing them with the $f(\theta)$-values, shows that $\gamma_n$ is in all cases (except Case 15, for $n = 100$) almost unbiased.

Graphs of $\text{MSE}(\gamma_n)$ and $E\gamma_n$ similar to that of $\text{MSE}(\eta_n)$ and $E\eta_n$ (see Figures 4.3-4.17) were compiled. They revealed the same tendencies and are therefore omitted. However, from these graphs it was once again clear that a "correct" choice of smoothing $s_n$ is possible. The performance of $\gamma_n$ based on a data-based bandwidth (discussed in Section 4.4), will be empirically evaluated in Section 4.6.

## 4.6 Comparison of the estimators

The small and moderate sample behaviour of the following estimators of $f(\theta)$ was compared mutually.

- MSB1:

$$\hat{\eta}_n = \frac{n^{-1}(2\hat{s}_n + 1)}{Y_{\hat{K}_n + \hat{s}_n} - Y_{\hat{K}_n - \hat{s}_n}}, \tag{4.8}$$

  where $\hat{s}_n$ is the minimiser of (see (4.3) and (4.5)),

$$\text{MSE}_*(\eta_n) = E_* \left( \eta_n^* - \min_{0 \le x \le 1} \tilde{f}_{\hat{h}}(x) \right)^2,$$

  with $\hat{h} = \hat{h}_m$ (modified stabilised smoother) and $\hat{K}_n$ is defined by

$$Y_{\hat{K}_n + \hat{s}_n} - Y_{\hat{K}_n - \hat{s}_n} = \max_{\hat{s}_n + 1 \le j \le n - \hat{s}_n} (Y_{j+\hat{s}_n} - Y_{j-\hat{s}_n}).$$

- MVS:

$$\hat{\gamma}_n = \frac{n^{-1}(2\hat{s}_n + 1)}{Y_{\hat{J}_n + \hat{s}_n} - Y_{\hat{J}_n - \hat{s}_n}},$$

Table 4.6: *Monte Carlo estimates of bias and minimum MSE.*

| | | $n = 100$ | | | $n = 400$ | | |
|---|---|---|---|---|---|---|---|
| Case | $f(\theta)$ | $s_n$ | $E\gamma_n$ | min(MSE) | $s_n$ | $E\gamma_n$ | min(MSE) |
| 1 | 0.767 | 11 | 0.789 | 0.009 | 41 | 0.786 | 0.004 |
| 2 | 0.570 | 6 | 0.593 | 0.010 | 19 | 0.566 | 0.004 |
| 3 | 0.291 | 1 | 0.265 | 0.006 | 9 | 0.296 | 0.002 |
| 4 | 0.767 | 12 | 0.771 | 0.008 | 45 | 0.783 | 0.004 |
| 5 | 0.570 | 7 | 0.583 | 0.009 | 23 | 0.574 | 0.004 |
| 6 | 0.291 | 2 | 0.286 | 0.005 | 9 | 0.289 | 0.002 |
| 7 | 0.767 | 16 | 0.799 | 0.007 | 43 | 0.778 | 0.003 |
| 8 | 0.570 | 7 | 0.574 | 0.008 | 23 | 0.573 | 0.004 |
| 9 | 0.291 | 2 | 0.283 | 0.005 | 10 | 0.296 | 0.002 |
| 10 | 0.583 | 8 | 0.595 | 0.008 | 27 | 0.592 | 0.003 |
| 11 | 0.271 | 2 | 0.262 | 0.004 | 10 | 0.277 | 0.002 |
| 12 | 0.917 | 31 | 0.941 | 0.003 | 100 | 0.932 | 0.002 |
| 13 | 0.854 | 21 | 0.869 | 0.005 | 78 | 0.867 | 0.002 |
| 14 | 0.409 | 2 | 0.405 | 0.007 | 8 | 0.412 | 0.004 |
| 15 | 0.161 | 1 | 0.233 | 0.008 | 2 | 0.164 | 0.001 |

where $\hat{s}_n$ is the minimiser of

$$\mathrm{MSE}_*(\gamma_n) = E_*\left(\gamma_n^* - \min_{0\le x\le 1} \tilde{f}_{\hat{h}}(x)\right)^2,$$

with $\hat{h} = \hat{h}_m$, $\gamma_n^* = \gamma_n(X_1^*, X_2^*, \ldots, X_n^*)$ (see (4.7)), and $\hat{J}_n$ is defined by (4.6) with $s_n$ replaced by $\hat{s}_n$.

- MSB2:

  $\hat{\eta}_n$, as defined in (4.8), where $\hat{s}_n$ is now the value where $\mathrm{Var}_*(V_{s_n}^*)$ attains its first local maximum (see Section 4.4), with

$$V_{s_n}^* = Y_{K_n^*+s_n}^* - Y_{K_n^*-s_n}^* = \max_{s_n+1\le j\le n-s_n}\left(Y_{j+s_n}^* - Y_{j-s_n}^*\right).$$

  $\mathrm{Var}_*$ denotes the variance over the conditional law of $X_1^*, X_2^*, \ldots, X_n^*$ given $X_1, X_2, \ldots, X_n$. For our purposes, $\hat{s}_n$ was calculated as follows:

$$\hat{s}_n = \begin{cases} \text{the first integer } s_n \text{ as } s_n \text{ increases from 1 upwards such that} \\ s_n \le N - [0.05n] \text{ and } \mathrm{Var}_*(V_{s_n}^*) \ge \max_{s_n+1\le i\le s_n+[0.05n]}\mathrm{Var}_*(V_i^*), \\ N - [0.05n] + 1, \text{ if no such } s_n \text{ exists,} \end{cases} \tag{4.9}$$

  where $N = [(n-1)/2]$ and $[z]$ denotes the largest integer less than or equal to $z$.

- MKE:

  $\hat{\xi}_n$, as defined in (4.4), with $\hat{h} = \hat{h}_m$.

Monte Carlo simulations were performed for sample sizes $n = 100$ and $n = 400$, using all fifteen densities. All estimates were based on 1000 independent random samples for $n = 100$, and 200 independent samples for $n = 400$. In order to calculate bootstrap estimates (see Section 3.2), $B = 500$ independent bootstrap replications were generated for each trial. Tables 4.7-4.14 display Monte Carlo estimates of $E\hat{s}_n$, $E\widehat{f(\theta)}$ and $\mathrm{RMSE}(\widehat{f(\theta)})$ (root mean squared error), where $\widehat{f(\theta)}$ refers to the estimators MSB1, MVS, MSB2 and MKE. To facilitate comparison, the following ratios were calculated

$$\mathrm{RATIO} = 100 \times \frac{\mathrm{RMSE}(\widehat{f(\theta)})}{\mathrm{RMSE}(\mathrm{MKE})}\%,$$

for $\widehat{f(\theta)}$ = MSB1, MVS and MSB2. Estimated standard errors of the sample averages appear in parentheses.

From the tables it is evident that MSB1 and MVS performed very similarly for both sample sizes (compare the RMSE's), although the former performed slightly better in most cases. MSB1 has smaller bias than MKE, except for Case 14 (both sample sizes) and Case 15 ($n = 100$). A comparison of the RMSE's of MSB1 and MKE, reveals that the new proposed estimator (MSB1) outperformed MKE substantially (the improvement in RMSE was up to 35%), except in the above-mentioned cases.

The behaviour of MSB2 is very interesting. From the tables it seems that it has exceptionally good performance for larger $f(\theta)$-values or for densities with nonpronounced minima. If one therefore has prior information that this is indeed the case, the use of MSB2 is recommended. However, for smaller $f(\theta)$-values or densities with pronounced minima, MSB2 can perform poorly.

Table 4.7: *Monte Carlo estimates of bias and RMSE for Case 1.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 1 | 0.767 | 100 | MSB1 | 13.8 (0.3) | 0.749 (0.004) | 0.138 | 98.6 |
| | | | MVS | 11.3 (0.2) | 0.750 (0.005) | 0.143 | 102.1 |
| | | | MSB2 | 12.7 (0.2) | 0.750 (0.004) | 0.140 | 100.0 |
| | | | MKE | | 0.713 (0.004) | 0.140 | |
| | | 400 | MSB1 | 40.4 (1.2) | 0.739 (0.006) | 0.088 | 89.8 |
| | | | MVS | 31.6 (1.0) | 0.741 (0.006) | 0.093 | 94.9 |
| | | | MSB2 | 53.5 (1.3) | 0.785 (0.006) | 0.088 | 89.8 |
| | | | MKE | | 0.714 (0.006) | 0.098 | |

Table 4.8: *Monte Carlo estimates of bias and RMSE for Case 2 and Case 3.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 2 | 0.570 | 100 | MSB1 | 8.2 (0.1) | 0.583 (0.004) | 0.136 | 103.0 |
| | | | MVS | 6.7 (0.1) | 0.585 (0.004) | 0.141 | 106.8 |
| | | | MSB2 | 9.7 (0.1) | 0.621 (0.005) | 0.156 | 118.2 |
| | | | MKE | | 0.556 (0.004) | 0.132 | |
| | | 400 | MSB1 | 29.3 (0.7) | 0.571 (0.005) | 0.073 | 101.4 |
| | | | MVS | 24.5 (0.7) | 0.572 (0.005) | 0.075 | 104.2 |
| | | | MSB2 | 35.2 (0.9) | 0.598 (0.006) | 0.095 | 131.9 |
| | | | MKE | | 0.555 (0.005) | 0.072 | |
| 3 | 0.291 | 100 | MSB1 | 2.5 (0.1) | 0.305 (0.003) | 0.093 | 86.9 |
| | | | MVS | 2.1 (0.0) | 0.303 (0.003) | 0.094 | 87.9 |
| | | | MSB2 | 5.9 (0.1) | 0.406 (0.004) | 0.182 | 170.1 |
| | | | MKE | | 0.260 (0.003) | 0.107 | |
| | | 400 | MSB1 | 12.2 (0.3) | 0.291 (0.004) | 0.060 | 100.0 |
| | | | MVS | 9.4 (0.3) | 0.291 (0.004) | 0.062 | 103.3 |
| | | | MSB2 | 16.8 (0.5) | 0.325 (0.005) | 0.080 | 133.3 |
| | | | MKE | | 0.279 (0.004) | 0.060 | |

Table 4.9: *Monte Carlo estimates of bias and RMSE for Case 4 and Case 5.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 4 | 0.767 | 100 | MSB1 | 16.7 (0.3) | 0.746 (0.004) | 0.137 | 93.8 |
| | | | MVS | 13.5 (0.3) | 0.748 (0.004) | 0.142 | 97.3 |
| | | | MSB2 | 14.0 (0.2) | 0.735 (0.004) | 0.140 | 96.0 |
| | | | MKE | | 0.712 (0.004) | 0.146 | |
| | | 400 | MSB1 | 49.4 (1.3) | 0.741 (0.005) | 0.081 | 87.1 |
| | | | MVS | 37.0 (1.0) | 0.741 (0.006) | 0.086 | 92.5 |
| | | | MSB2 | 62.4 (1.3) | 0.778 (0.005) | 0.077 | 82.8 |
| | | | MKE | | 0.716 (0.006) | 0.093 | |
| 5 | 0.570 | 100 | MSB1 | 10.1 (0.2) | 0.573 (0.004) | 0.125 | 100.8 |
| | | | MVS | 7.8 (0.1) | 0.573 (0.004) | 0.129 | 104.0 |
| | | | MSB2 | 10.2 (0.1) | 0.582 (0.004) | 0.130 | 104.8 |
| | | | MKE | | 0.545 (0.004) | 0.124 | |
| | | 400 | MSB1 | 34.2 (0.9) | 0.562 (0.006) | 0.079 | 96.3 |
| | | | MVS | 24.9 (0.6) | 0.561 (0.006) | 0.083 | 101.2 |
| | | | MSB2 | 41.2 (0.9) | 0.589 (0.006) | 0.080 | 97.6 |
| | | | MKE | | 0.541 (0.005) | 0.082 | |

Table 4.10: *Monte Carlo estimates of bias and RMSE for Case 6 and Case 7.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 6 | 0.291 | 100 | MSB1 | 3.7 (0.1) | 0.301 (0.003) | 0.092 | 98.9 |
| | | | MVS | 2.8 (0.0) | 0.300 (0.003) | 0.094 | 101.1 |
| | | | MSB2 | 5.1 (0.1) | 0.334 (0.003) | 0.110 | 118.3 |
| | | | MKE | | 0.276 (0.003) | 0.093 | |
| | | 400 | MSB1 | 15.0 (0.4) | 0.294 (0.004) | 0.057 | 100.0 |
| | | | MVS | 11.1 (0.3) | 0.295 (0.004) | 0.059 | 103.5 |
| | | | MSB2 | 20.0 (0.6) | 0.320 (0.004) | 0.069 | 121.1 |
| | | | MKE | | 0.280 (0.004) | 0.057 | |
| 7 | 0.767 | 100 | MSB1 | 16.7 (0.3) | 0.716 (0.005) | 0.157 | 89.2 |
| | | | MVS | 13.4 (0.3) | 0.715 (0.005) | 0.163 | 92.6 |
| | | | MSB2 | 15.2 (0.2) | 0.726 (0.004) | 0.132 | 75.0 |
| | | | MKE | | 0.677 (0.005) | 0.176 | |
| | | 400 | MSB1 | 33.8 (1.2) | 0.688 (0.006) | 0.119 | 83.2 |
| | | | MVS | 24.2 (0.8) | 0.687 (0.007) | 0.123 | 86.0 |
| | | | MSB2 | 66.8 (1.4) | 0.784 (0.005) | 0.072 | 50.3 |
| | | | MKE | | 0.653 (0.006) | 0.143 | |

Table 4.11: *Monte Carlo estimates of bias and RMSE for Case 8 and Case 9.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 8 | 0.570 | 100 | MSB1 | 8.1 (0.2) | 0.524 (0.004) | 0.144 | 89.4 |
| | | | MVS | 6.1 (0.1) | 0.524 (0.004) | 0.148 | 91.9 |
| | | | MSB2 | 11.2 (0.2) | 0.580 (0.004) | 0.126 | 78.3 |
| | | | MKE | | 0.484 (0.004) | 0.161 | |
| | | 400 | MSB1 | 20.5 (0.7) | 0.512 (0.006) | 0.107 | 87.0 |
| | | | MVS | 14.7 (0.5) | 0.510 (0.007) | 0.111 | 90.2 |
| | | | MSB2 | 44.7 (1.0) | 0.600 (0.006) | 0.086 | 69.9 |
| | | | MKE | | 0.482 (0.006) | 0.123 | |
| 9 | 0.291 | 100 | MSB1 | 2.0 (0.1) | 0.264 (0.003) | 0.090 | 64.3 |
| | | | MVS | 1.6 (0.0) | 0.265 (0.003) | 0.093 | 66.4 |
| | | | MSB2 | 5.6 (0.1) | 0.338 (0.003) | 0.111 | 79.3 |
| | | | MKE | | 0.199 (0.003) | 0.140 | |
| | | 400 | MSB1 | 5.4 (0.2) | 0.231 (0.005) | 0.090 | 85.7 |
| | | | MVS | 4.0 (0.2) | 0.231 (0.005) | 0.094 | 89.5 |
| | | | MSB2 | 21.2 (0.6) | 0.328 (0.004) | 0.072 | 68.6 |
| | | | MKE | | 0.207 (0.005) | 0.105 | |

Table 4.12: *Monte Carlo estimates of bias and RMSE for Case 10 and Case 11.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 10 | 0.583 | 100 | MSB1 | 8.3 (0.2) | 0.529 (0.005) | 0.157 | 87.2 |
| | | | MVS | 6.3 (0.2) | 0.528 (0.005) | 0.162 | 90.0 |
| | | | MSB2 | 12.3 (0.2) | 0.602 (0.004) | 0.119 | 66.1 |
| | | | MKE | | 0.485 (0.005) | 0.180 | |
| | | 400 | MSB1 | 18.2 (0.7) | 0.505 (0.007) | 0.124 | 86.1 |
| | | | MVS | 13.0 (0.5) | 0.502 (0.007) | 0.128 | 88.9 |
| | | | MSB2 | 48.4 (1.0) | 0.609 (0.005) | 0.070 | 48.6 |
| | | | MKE | | 0.471 (0.006) | 0.144 | |
| 11 | 0.271 | 100 | MSB1 | 2.1 (0.1) | 0.245 (0.003) | 0.084 | 64.6 |
| | | | MVS | 1.7 (0.0) | 0.245 (0.003) | 0.085 | 65.4 |
| | | | MSB2 | 5.9 (0.1) | 0.326 (0.003) | 0.113 | 86.9 |
| | | | MKE | | 0.187 (0.003) | 0.130 | |
| | | 400 | MSB1 | 6.1 (0.3) | 0.222 (0.004) | 0.077 | 84.6 |
| | | | MVS | 4.4 (0.2) | 0.222 (0.004) | 0.078 | 85.7 |
| | | | MSB2 | 21.8 (0.7) | 0.305 (0.004) | 0.072 | 79.1 |
| | | | MKE | | 0.199 (0.004) | 0.091 | |

Table 4.13: *Monte Carlo estimates of bias and RMSE for Case 12 and Case 13.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 12 | 0.917 | 100 | MSB1 | 22.5 (0.4) | 0.819 (0.004) | 0.164 | 85.4 |
| | | | MVS | 18.8 (0.3) | 0.823 (0.004) | 0.167 | 87.0 |
| | | | MSB2 | 15.9 (0.2) | 0.781 (0.004) | 0.182 | 94.8 |
| | | | MKE | | 0.782 (0.004) | 0.192 | |
| | | 400 | MSB1 | 50.3 (1.7) | 0.806 (0.006) | 0.142 | 81.1 |
| | | | MVS | 36.2 (1.2) | 0.803 (0.007) | 0.148 | 84.6 |
| | | | MSB2 | 79.1 (1.7) | 0.868 (0.004) | 0.078 | 44.6 |
| | | | MKE | | 0.765 (0.006) | 0.175 | |
| 13 | 0.854 | 100 | MSB1 | 18.3 (0.3) | 0.767 (0.004) | 0.160 | 87.0 |
| | | | MVS | 14.8 (0.3) | 0.771 (0.004) | 0.163 | 88.6 |
| | | | MSB2 | 15.7 (0.2) | 0.761 (0.004) | 0.151 | 82.1 |
| | | | MKE | | 0.730 (0.004) | 0.184 | |
| | | 400 | MSB1 | 42.1 (1.3) | 0.753 (0.006) | 0.131 | 79.9 |
| | | | MVS | 30.2 (1.0) | 0.755 (0.006) | 0.133 | 81.1 |
| | | | MSB2 | 73.0 (1.7) | 0.816 (0.005) | 0.081 | 49.4 |
| | | | MKE | | 0.714 (0.006) | 0.164 | |

Table 4.14: *Monte Carlo estimates of bias and RMSE for Case 14 and Case 15.*

| Case | $f(\theta)$ | $n$ | $\widehat{f(\theta)}$ | $E\hat{s}_n$ | $E\widehat{f(\theta)}$ | RMSE($\widehat{f(\theta)}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 14 | 0.409 | 100 | MSB1 | 6.5 (0.1) | 0.531 (0.004) | 0.183 | 111.6 |
| | | | MVS | 5.5 (0.1) | 0.528 (0.004) | 0.181 | 110.4 |
| | | | MSB2 | 6.5 (0.1) | 0.534 (0.005) | 0.209 | 127.4 |
| | | | MKE | | 0.506 (0.004) | 0.164 | |
| | | 400 | MSB1 | 25.7 (0.5) | 0.527 (0.006) | 0.142 | 108.4 |
| | | | MVS | 22.1 (0.5) | 0.524 (0.005) | 0.138 | 105.3 |
| | | | MSB2 | 20.5 (0.7) | 0.485 (0.008) | 0.133 | 101.5 |
| | | | MKE | | 0.516 (0.005) | 0.131 | |
| 15 | 0.161 | 100 | MSB1 | 1.5 (0.0) | 0.250 (0.003) | 0.119 | 114.4 |
| | | | MVS | 1.3 (0.0) | 0.249 (0.003) | 0.120 | 115.4 |
| | | | MSB2 | 3.2 (0.1) | 0.329 (0.005) | 0.220 | 211.5 |
| | | | MKE | | 0.179 (0.003) | 0.104 | |
| | | 400 | MSB1 | 2.9 (0.1) | 0.167 (0.004) | 0.054 | 98.2 |
| | | | MVS | 2.3 (0.1) | 0.168 (0.004) | 0.059 | 107.3 |
| | | | MSB2 | 4.8 (0.4) | 0.203 (0.006) | 0.095 | 172.7 |
| | | | MKE | | 0.147 (0.004) | 0.055 | |

## 4.7 Confidence intervals

In this section the behaviour of the proposed estimator $\hat{\eta}_n$ with regard to confidence intervals for $f(\theta)$ is briefly studied. Firstly, a standard interval, based on the standard normal distribution quantiles, is studied. To do this, an estimate of the standard error of $\hat{\eta}_n$ is needed. Since $\eta_n$ in Chapter 1 was defined in terms of two different sequences $\{r_n\}$ and $\{s_n\}$ of smoothing parameters, the result of Theorem 1.5.2 cannot be used. However, it is claimed that for large $n$,

$$\mathrm{Var}(\eta_n) \cong \frac{f(\theta)^2}{2s_n + 1}, \tag{4.10}$$

which can be proved heuristically as follows.

Let $G = F^{-1}$, and define $S_i$, $i = 1, 2, \ldots, n$, as in the introduction of Section 1.5. Now, since $\eta_n$ is (under certain conditions) a strongly consistent estimator of $f(\theta)$, we have that $\eta_n \cong \eta_n^{-1} f(\theta)^2$, so that

$$\mathrm{Var}(\eta_n) \cong f(\theta)^4 \mathrm{Var}(\eta_n^{-1}).$$

Using the fact that $K_n \cong nq$, $q = F(\theta)$, (which was proved in Chapter 1), it follows that

$$\begin{aligned}
\eta_n^{-1} &= \frac{Y_{K_n+s_n} - Y_{K_n-s_n}}{n^{-1}(2s_n+1)} \\
&\cong \frac{Y_{nq+s_n} - Y_{nq-s_n}}{n^{-1}(2s_n+1)} \\
&\stackrel{d}{=} n(2s_n+1)^{-1}\Big\{G\Big(\frac{S_{nq+s_n}}{S_{n+1}}\Big) - G\Big(\frac{S_{nq-s_n}}{S_{n+1}}\Big)\Big\} \\
&= n(2s_n+1)^{-1}\Big\{G(q) + G'(q)\Big(\frac{S_{nq+s_n}}{S_{n+1}} - q\Big) + \ldots \\
&\quad - \Big[G(q) + G'(q)\Big(\frac{S_{nq-s_n}}{S_{n+1}} - q\Big) + \ldots\Big]\Big\} \\
&\cong n(2s_n+1)^{-1} G'(q)\Big(\frac{S_{nq+s_n} - S_{nq-s_n}}{S_{n+1}}\Big) \\
&= n(2s_n+1)^{-1} f(\theta)^{-1}\Big(\frac{S_{nq+s_n} - S_{nq-s_n}}{S_{n+1}}\Big) \\
&\stackrel{d}{=} n(2s_n+1)^{-1}(nf(\theta))^{-1} \sum_{i=1}^{2s_n} Z_i,
\end{aligned}$$

where the $Z_i$'s are independent random variables, each with a standard exponential distribution. Hence,

$$\mathrm{Var}(\eta_n^{-1}) \cong \{f(\theta)^2(2s_n+1)\}^{-1},$$

and

$$\mathrm{Var}(\eta_n) \cong \frac{f(\theta)^2}{2s_n + 1}.$$

□

Hence, I suggest estimating the standard error of $\hat{\eta}_n$ by $(2\hat{s}_n + 1)^{-1/2}\hat{\eta}_n$. A standard $100(1-\alpha)\%$-confidence interval for $f(\theta)$ is therefore

$$\left[\hat{\eta}_n \left(1 - z(\alpha/2)/(2\hat{s}_n + 1)^{1/2}\right), \hat{\eta}_n \left(1 + z(\alpha/2)/(2\hat{s}_n + 1)^{1/2}\right)\right], \tag{4.11}$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)$ percentile point of the standard normal distribution.

If the normal quantiles in (4.11) are under suspicion, the bootstrap can be used to construct nonparametric confidence intervals (see Section 3.4 of Chapter 3). This is illustrated by applying the percentile method discussed in Section 3.4. From (3.2) it follows that a percentile $100(1-\alpha)\%$-confidence interval for $f(\theta)$ is given by

$$\left[\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1-\alpha/2)\right], \tag{4.12}$$

where $\hat{G}(t) = P^*(\hat{\eta}_n^* \leq t)$, and $\hat{\eta}_n^*$ is defined in (4.8) with $X_1, X_2, \ldots, X_n$ replaced by a bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$. Note that the calculation of $\hat{s}_n^* = \hat{s}_n(X_1^*, X_2^*, \ldots, X_n^*)$ requires second-level bootstrapping, i.e., sampling from the empirical distribution of $X_1^*, X_2^*, \ldots, X_n^*$ (often also referred to as the **double bootstrap**). To calculate (4.12), one can apply the Monte Carlo approximation method described in Section 3.4.

A simulation study was conducted to investigate the performance of the confidence intervals given in (4.11) and (4.12) with respect to coverage probability and expected length. I chose $n = 100$, $\alpha = 0.05$, $\alpha = 0.10$ and only considered Cases 1, 2, 5 and 6. The number of Monte Carlo trials was 200, and the numbers of first-level and second-level bootstrap replications were 200 and 100 respectively. It should be emphasised that the calculation of the bootstrap interval (4.12) is extremely computer-time expensive. It was therefore necessary to use $\hat{h} = \hat{h}_{P2}$ (modified plug-in smoother) instead of $\hat{h} = \hat{h}_m$ in the definition of MSB1 given in (4.8). Estimated standard errors of the sample averages appear in parentheses.

From Tables 4.15-4.18 it is clear that all the intervals behaved satisfactorily, in the sense that the estimated coverage probabilities were close to the prescribed confidence

levels. The satisfactory coverage probabilities attained by the standard intervals can be an indication of rapid weak convergence of $\hat{\eta}_n$ (properly standardised) to a standard normal distribution. However, it should be noted that the standard intervals based on MSB1 and MSB2 tend to be conservative. The percentile interval based on MSB1 is slightly anti-conservative, and the percentile interval based on MSB2 is conservative, but less than the standard intervals. As far as the expected length is concerned, the intervals behaved almost identically. Nonparametric confidence intervals can also be constructed by using the techniques discussed in Section 3.4, but due to the good performance of the percentile interval they were not included in the study. Moreover, it would require an immense amount of additional computer time.

Table 4.15: $100(1-\alpha)\%$ *standard intervals, (4.11), using* MSB1.

| Case | $f(\theta)$ | $E\hat{\eta}_n$ | $\alpha$ | Coverage | $E$(Length) |
|---|---|---|---|---|---|
| 1 | 0.77 | 0.76 (0.01) | 0.05 | 0.98 (0.01) | 0.61 (0.01) |
| | | | 0.10 | 0.92 (0.02) | 0.51 (0.01) |
| 2 | 0.57 | 0.56 (0.01) | 0.05 | 0.97 (0.01) | 0.62 (0.01) |
| | | | 0.10 | 0.94 (0.02) | 0.52 (0.01) |
| 5 | 0.57 | 0.57 (0.01) | 0.05 | 0.96 (0.02) | 0.53 (0.01) |
| | | | 0.10 | 0.92 (0.02) | 0.44 (0.01) |
| 6 | 0.29 | 0.30 (0.01) | 0.05 | 0.98 (0.01) | 0.47 (0.01) |
| | | | 0.10 | 0.95 (0.02) | 0.39 (0.01) |

Table 4.16: $100(1-\alpha)\%$ *percentile intervals, (4.12),* MSB1.

| Case | $f(\theta)$ | $E\hat{\eta}_n$ | $\alpha$ | Coverage | $E$(Length) |
|---|---|---|---|---|---|
| 1 | 0.77 | 0.76 (0.01) | 0.05 | 0.93 (0.02) | 0.63 (0.01) |
| | | | 0.10 | 0.88 (0.02) | 0.56 (0.01) |
| 2 | 0.57 | 0.56 (0.01) | 0.05 | 0.92 (0.02) | 0.52 (0.01) |
| | | | 0.10 | 0.87 (0.02) | 0.45 (0.01) |
| 5 | 0.57 | 0.57 (0.01) | 0.05 | 0.92 (0.02) | 0.52 (0.01) |
| | | | 0.10 | 0.87 (0.02) | 0.44 (0.01) |
| 6 | 0.29 | 0.30 (0.01) | 0.05 | 0.91 (0.02) | 0.29 (0.01) |
| | | | 0.10 | 0.85 (0.03) | 0.25 (0.01) |

Table 4.17: $100(1-\alpha)\%$ *standard intervals, (4.11), using* MSB2.

| Case | $f(\theta)$ | $E\hat{\eta}_n$ | $\alpha$ | Coverage | $E$(Length) |
|---|---|---|---|---|---|
| 1 | 0.77 | 0.76 (0.01) | 0.05 | 0.98 (0.01) | 0.61 (0.01) |
| | | | 0.10 | 0.94 (0.02) | 0.51 (0.01) |
| 2 | 0.57 | 0.63 (0.01) | 0.05 | 0.98 (0.01) | 0.61 (0.01) |
| | | | 0.10 | 0.94 (0.02) | 0.51 (0.01) |
| 5 | 0.57 | 0.56 (0.01) | 0.05 | 0.97 (0.01) | 0.52 (0.01) |
| | | | 0.10 | 0.92 (0.02) | 0.44 (0.01) |
| 6 | 0.29 | 0.33 (0.01) | 0.05 | 0.95 (0.02) | 0.43 (0.01) |
| | | | 0.10 | 0.92 (0.02) | 0.36 (0.01) |

Table 4.18: $100(1-\alpha)\%$ *percentile intervals, (4.12), using* MSB2.

| Case | $f(\theta)$ | $E\hat{\eta}_n$ | $\alpha$ | Coverage | $E$(Length) |
|---|---|---|---|---|---|
| 1 | 0.77 | 0.76 (0.01) | 0.05 | 0.95 (0.02) | 0.59 (0.01) |
| | | | 0.10 | 0.93 (0.02) | 0.51 (0.01) |
| 2 | 0.57 | 0.63 (0.01) | 0.05 | 0.95 (0.02) | 0.59 (0.01) |
| | | | 0.10 | 0.93 (0.02) | 0.51 (0.01) |
| 5 | 0.57 | 0.56 (0.01) | 0.05 | 0.95 (0.02) | 0.50 (0.01) |
| | | | 0.10 | 0.92 (0.02) | 0.42 (0.01) |
| 6 | 0.29 | 0.33 (0.01) | 0.05 | 0.97 (0.01) | 0.40 (0.01) |
| | | | 0.10 | 0.93 (0.01) | 0.34 (0.01) |

## 4.8 Estimation of the antimode

In Section 1.3, it was proposed to estimate the antimode $\theta$ by $\hat{\theta}_n$, where $\hat{\theta}_n$ is any statistic satisfying

$$Y_{K_n-s_n} \leq \hat{\theta}_n \leq Y_{K_n+s_n},$$

and $K_n$ is defined in (1.3). The small and moderate sample behaviour of $\hat{\theta}_n = Y_{K_n}$ will now be discussed briefly. Because of this choice, attention will be limited to densities with antimodes not too close to the boundaries. Clearly, one should choose $\hat{\theta}_n$ differently when this is not the case, for example, $\hat{\theta}_n = Y_{K_n-s_n}$ or $\hat{\theta}_n = Y_{K_n+s_n}$. Since $Y_{K_n}$ depends on the smoothing sequence $\{s_n\}$, two data-based smoothing methods will be considered. The following estimators of $\theta$ were compared mutually.

- YKB1:

  $Y_{\hat{K}_n}$, where $\hat{s}_n$ is the minimiser of

$$\mathrm{MSE}_*(Y_{K_n}) = E_*\left(Y^*_{K^*_n} - \arg\min_{0\leq x\leq 1} \tilde{f}_{\hat{h}}(x)\right)^2,$$

with $\hat{h} = \hat{h}_m$ and $Y^*_{K^*_n}$ is the bootstrap version of $Y_{K_n}$. As before, $\hat{K}_n$ is defined by

$$Y_{\hat{K}_n+\hat{s}_n} - Y_{\hat{K}_n-\hat{s}_n} = \max_{\hat{s}_n+1\le j\le n-\hat{s}_n} (Y_{j+\hat{s}_n} - Y_{j-\hat{s}_n}).$$

- YJB:

  $Y_{\hat{J}_n}$, where $\hat{s}_n$ is the minimiser of

  $$\mathrm{MSE}_*(Y_{J_n}) = E_*\left(Y^*_{J^*_n} - \arg\min_{0\le x\le 1} \tilde{f}_{\hat{h}}(x)\right)^2,$$

  with $\hat{h} = \hat{h}_m$ and $Y^*_{J^*_n}$ is the bootstrap version of $Y_{J_n}$ (see (4.6)). As before, $\hat{J}_n$ is defined by (4.6) with $s_n$ replaced by $\hat{s}_n$.

- YKB2:

  $Y_{\hat{K}_n}$, where $\hat{s}_n$ is now the value where $\mathrm{Var}_*(V^*_n)$ attains its first local maximum (see Section 4.4), with

  $$V^*_n = Y^*_{K^*_n+s_n} - Y^*_{K^*_n-s_n} = \max_{s_n+1\le j\le n-s_n} \left(Y^*_{j+s_n} - Y^*_{j-s_n}\right).$$

- AKE:

  $\arg\min_{0\le x\le 1} \tilde{f}_{\hat{h}}(x)$, with $\hat{h} = \hat{h}_m$.

Monte Carlo simulations were performed for sample sizes $n = 100$ and $n = 400$, using densities, Cases 4-13. All estimates were based on 1000 independent random samples for $n = 100$, and 200 independent samples for $n = 400$. In order to calculate bootstrap estimates (see Section 3.2), I generated $B = 500$ independent bootstrap replications for each trial. Tables 4.19-4.23 display Monte Carlo estimates of $E\hat{s}_n$, $E\hat{\theta}$ and $\mathrm{RMSE}(\hat{\theta})$ (root mean squared error), where $\hat{\theta}$ refers to the estimators YKB1, YJB, YKB2 and AKE. To facilitate comparison, the following ratios were calculated

$$\mathrm{RATIO} = 100 \times \frac{\mathrm{RMSE}(\hat{\theta})}{\mathrm{RMSE}(\mathrm{AKE})}\%,$$

for $\hat{\theta}$ = YKB1, YJB and YKB2. Estimated standard errors of the sample averages appear in parentheses.

Table 4.19: *Monte Carlo estimates of bias and RMSE for Case 4 and Case 5.*

| Case | $\theta$ | $n$ | $\hat{\theta}$ | $E\hat{s}_n$ | $E\hat{\theta}$ | RMSE($\hat{\theta}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 4 | 0.25 | 100 | YKB1 | 15.7 (0.6) | 0.311 (0.005) | 0.157 | 60.4 |
| | | | YJB | 18.9 (0.7) | 0.315 (0.005) | 0.157 | 60.4 |
| | | | YKB2 | 14.0 (0.2) | 0.330 (0.004) | 0.144 | 55.4 |
| | | | AKE | | 0.269 (0.008) | 0.260 | |
| | | 400 | YKB1 | 60.9 (5.1) | 0.270 (0.007) | 0.104 | 73.2 |
| | | | YJB | 75.0 (6.0) | 0.271 (0.007) | 0.101 | 71.1 |
| | | | YKB2 | 62.4 (1.3) | 0.287 (0.004) | 0.071 | 50.0 |
| | | | AKE | | 0.247 (0.010) | 0.142 | |
| 5 | 0.25 | 100 | YKB1 | 9.1 (0.4) | 0.257 (0.003) | 0.092 | 60.1 |
| | | | YJB | 11.9 (0.5) | 0.260 (0.003) | 0.089 | 58.2 |
| | | | YKB2 | 10.2 (0.1) | 0.287 (0.002) | 0.074 | 48.4 |
| | | | AKE | | 0.218 (0.005) | 0.153 | |
| | | 400 | YKB1 | 41.1 (3.5) | 0.253 (0.004) | 0.063 | 75.0 |
| | | | YJB | 53.1 (4.2) | 0.252 (0.004) | 0.063 | 75.0 |
| | | | YKB2 | 41.2 (0.9) | 0.269 (0.003) | 0.050 | 59.5 |
| | | | AKE | | 0.241 (0.006) | 0.084 | |

Table 4.20: *Monte Carlo estimates of bias and RMSE for Case 6 and Case 7.*

| Case | $\theta$ | $n$ | $\hat{\theta}$ | $E\hat{s}_n$ | $E\hat{\theta}$ | RMSE($\hat{\theta}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 6 | 0.25 | 100 | YKB1 | 3.8 (0.2) | 0.255 (0.002) | 0.075 | 80.6 |
| | | | YJB | 5.5 (0.2) | 0.252 (0.002) | 0.070 | 75.3 |
| | | | YKB2 | 5.1 (0.1) | 0.276 (0.002) | 0.070 | 75.3 |
| | | | AKE | | 0.239 (0.003) | 0.093 | |
| | | 400 | YKB1 | 19.6 (1.8) | 0.255 (0.004) | 0.053 | 96.4 |
| | | | YJB | 28.1 (2.3) | 0.252 (0.004) | 0.050 | 90.9 |
| | | | YKB2 | 20.0 (0.6) | 0.255 (0.003) | 0.046 | 83.6 |
| | | | AKE | | 0.252 (0.004) | 0.055 | |
| 7 | 0.50 | 100 | YKB1 | 26.2 (1.0) | 0.504 (0.004) | 0.123 | 65.1 |
| | | | YJB | 28.9 (1.0) | 0.504 (0.004) | 0.123 | 65.1 |
| | | | YKB2 | 15.2 (0.2) | 0.497 (0.004) | 0.114 | 60.3 |
| | | | AKE | | 0.506 (0.006) | 0.189 | |
| | | 400 | YKB1 | 130.7 (10.5) | 0.500 (0.006) | 0.081 | 79.4 |
| | | | YJB | 139.1 (10.8) | 0.498 (0.005) | 0.075 | 73.5 |
| | | | YKB2 | 66.8 (1.4) | 0.494 (0.005) | 0.072 | 70.6 |
| | | | AKE | | 0.493 (0.007) | 0.102 | |

Table 4.21: *Monte Carlo estimates of bias and RMSE for Case 8 and Case 9.*

| Case | $\theta$ | $n$ | $\hat{\theta}$ | $E\hat{s}_n$ | $E\hat{\theta}$ | RMSE($\hat{\theta}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 8 | 0.5 | 100 | YKB1 | 19.0 (0.8) | 0.499 (0.003) | 0.088 | 85.4 |
| | | | YJB | 23.8 (0.9) | 0.498 (0.003) | 0.079 | 76.7 |
| | | | YKB2 | 11.2 (0.2) | 0.497 (0.003) | 0.085 | 82.5 |
| | | | AKE | | 0.498 (0.003) | 0.103 | |
| | | 400 | YKB1 | 86.4 (8.2) | 0.497 (0.005) | 0.065 | 90.3 |
| | | | YJB | 105.3 (8.9) | 0.501 (0.004) | 0.053 | 73.6 |
| | | | YKB2 | 44.7 (1.0) | 0.499 (0.004) | 0.055 | 76.4 |
| | | | AKE | | 0.502 (0.005) | 0.072 | |
| 9 | 0.5 | 100 | YKB1 | 7.0 (0.4) | 0.500 (0.003) | 0.087 | 102.4 |
| | | | YJB | 14.1 (0.6) | 0.499 (0.002) | 0.071 | 83.5 |
| | | | YKB2 | 5.6 (0.1) | 0.502 (0.002) | 0.074 | 87.1 |
| | | | AKE | | 0.498 (0.003) | 0.085 | |
| | | 400 | YKB1 | 29.2 (4.1) | 0.499 (0.004) | 0.061 | 95.3 |
| | | | YJB | 62.2 (6.1) | 0.503 (0.003) | 0.046 | 71.9 |
| | | | YKB2 | 21.2 (0.6) | 0.500 (0.003) | 0.048 | 75.0 |
| | | | AKE | | 0.503 (0.005) | 0.064 | |

Table 4.22: *Monte Carlo estimates of bias and RMSE for Case 10 and Case 11.*

| Case | $\theta$ | $n$ | $\hat{\theta}$ | $E\hat{s}_n$ | $E\hat{\theta}$ | RMSE($\hat{\theta}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 10 | 0.50 | 100 | YKB1 | 20.6 (0.9) | 0.500 (0.003) | 0.097 | 85.1 |
| | | | YJB | 25.5 (1.0) | 0.497 (0.003) | 0.085 | 74.6 |
| | | | YKB2 | 12.3 (0.2) | 0.500 (0.003) | 0.087 | 76.3 |
| | | | AKE | | 0.496 (0.004) | 0.114 | |
| | | 400 | YKB1 | 102.6 (9.1) | 0.503 (0.005) | 0.067 | 74.4 |
| | | | YJB | 121.1 (9.9) | 0.500 (0.004) | 0.060 | 66.7 |
| | | | YKB2 | 48.4 (1.0) | 0.500 (0.004) | 0.054 | 60.0 |
| | | | AKE | | 0.503 (0.006) | 0.090 | |
| 11 | 0.75 | 100 | YKB1 | 3.8 (0.2) | 0.737 (0.003) | 0.096 | 84.2 |
| | | | YJB | 5.0 (0.2) | 0.740 (0.003) | 0.087 | 76.3 |
| | | | YKB2 | 5.9 (0.1) | 0.690 (0.002) | 0.098 | 86.0 |
| | | | AKE | | 0.760 (0.004) | 0.114 | |
| | | 400 | YKB1 | 19.1 (1.8) | 0.745 (0.005) | 0.071 | 89.9 |
| | | | YJB | 23.8 (2.0) | 0.746 (0.005) | 0.067 | 84.8 |
| | | | YKB2 | 21.8 (0.7) | 0.732 (0.004) | 0.060 | 75.9 |
| | | | AKE | | 0.750 (0.006)) | 0.079 | |

Table 4.23: *Monte Carlo estimates of bias and RMSE for Case 12 and Case 13.*

| Case | $\theta$ | $n$ | $\hat{\theta}$ | $E\hat{s}_n$ | $E\hat{\theta}$ | RMSE($\hat{\theta}$) | RATIO |
|---|---|---|---|---|---|---|---|
| 12 | 0.50 | 100 | YKB1 | 24.5 (0.9) | 0.510 (0.006) | 0.179 | 59.1 |
| | | | YJB | 27.1 (1.0) | 0.510 (0.006) | 0.181 | 59.7 |
| | | | YKB2 | 15.9 (0.2) | 0.497 (0.005) | 0.163 | 53.8 |
| | | | AKE | | 0.518 (0.010) | 0.303 | |
| | | 400 | YKB1 | 120.9 (9.8) | 0.521 (0.010) | 0.149 | 66.8 |
| | | | YJB | 134.6 (10.5) | 0.525 (0.010) | 0.147 | 65.9 |
| | | | YKB2 | 79.1 (0.7) | 0.503 (0.009) | 0.132 | 59.2 |
| | | | AKE | | 0.534 (0.016) | 0.223 | |
| 13 | 0.75 | 100 | YKB1 | 18.9 (0.8) | 0.660 (0.005) | 0.177 | 76.6 |
| | | | YJB | 21.1 (0.8) | 0.655 (0.005) | 0.181 | 78.4 |
| | | | YKB2 | 15.7 (0.2) | 0.608 (0.004) | 0.197 | 85.3 |
| | | | AKE | | 0.747 (0.007) | 0.231 | |
| | | 400 | YKB1 | 80.9 (7.0) | 0.693 (0.010) | 0.147 | 79.0 |
| | | | YJB | 88.2 (7.6) | 0.694 (0.010) | 0.149 | 80.1 |
| | | | YKB2 | 73.0 (1.7) | 0.665 (0.007) | 0.128 | 68.8 |
| | | | AKE | | 0.748 (0.013) | 0.186 | |

From Tables 4.19-4.23 it is clear that the newly proposed estimators (YKB1, YJB and YKB2) are far superior (in terms of RMSE) to AKE, the standard method of estimating the antimode $\theta$. The estimators YKB1, YJB and YKB2 have small bias, except for Cases 4 and 13. However, it seems that AKE is almost unbiased in all cases, implying that (in view of the large RMSE) it has large variability.

## 4.9 Application to real data

One of the aims in $\gamma$-ray, $X$-ray and optic ray astronomy is to identify these rays from rotating bodies called pulsars. Pulsars are objects in the universe with masses more or less that of our sun's mass and radii of about 10 km. They are fast rotating objects with periods ranging between 1 millisecond and a few hundred seconds. Some of them radiate $\gamma$-rays with the same period as the spin period of the pulsar. These $\gamma$-rays are detected on earth via satellites or ground-based telescopes.

However, aperiodic cosmic rays (background radiation or noise) are also detected. Thus, a typical data set consists of a sequence of arrival times $t_i$, each arrival time representing either noise or pulsed radiation. In the pre-analysis, the $t_i$'s are folded modulo 1 with pulsar period $q$,

$$X_i = \frac{t_i}{q} - \left[\frac{t_i}{q}\right], \quad i = 1, 2, \ldots .$$

The pulsar's signal period $q$ can be accurately determined from observations of its radio pulses.

The unknown periodic density function (or light curve) $f(x)$ of the folded (modulo 1) arrival times can be represented as

$$f(x) = 1 - p + p f_s(x) ,$$

where $p$, $0 \leq p \leq 1$, is the unknown strength of the periodic signal and the unknown source function $f_s(x)$ gives the relative radiation intensity as a function of $x$. An important problem is to estimate the strength of the pulsed signal, $p$, in a series of high-energy photon arrival times.

Tests for the following hypothesis have been developed (e.g., see Protheroe, 1985, and Swanepoel & De Beer, 1990),

$$H_0 : p = 0 \text{ versus } H_A : p > 0.$$

The interpretation of the null hypothesis is that the pulsar doesn't radiate $\gamma$-rays. If this hypothesis is rejected and under the reasonable assumption that

$$\min_x f_s(x) = 0,$$

we have

$$\min_x f(x) = 1 - p,$$

so that estimation of $p$ can be reduced to the estimation of the minimum of $f$. The estimation of a unique antimode is not important in this set-up, as will become clear from the discussion below.

In order to estimate $p$ from the data, it is usually assumed in the literature (e.g., see De Jager *et al.*, 1989) that $f_s(x)$ has some known parametric form, such as the cardioid or von Mises density functions. Estimators of $p$ can then be derived by using standard statistical methods, for example, maximum likelihood and method-of-moments. However, these estimators can perform poorly (i.e., having large bias and/or large variance) if the assumed parametric form of $f_s(x)$ deviates from its true form. A more realistic approach is to estimate $p$ nonparametrically. To the best of my knowledge, this problem has not yet been formally addressed in the literature.

However, the following *subjective* technique to estimate $p$ (for $f_s(x)$ unknown) is often applied in practice. Denote the sample of folded (modulo 1) arrival times by $X_i$ ($0 < X_i < 1$), $i = 1, 2, \ldots, n$, which are usually referred to as *sample phases*. Construct a histogram of the $X_i$'s, using a bin width $c$. Use *visual* inspection to determine an interval $J$ of phases corresponding to that part of the histogram characterised by only random fluctuations around its lowest level. The subjective estimator of $p$ is then

$$\tilde{p} = \max\left\{0\ ,\ 1 - \frac{(\text{number of } X_i\text{'s in } J)}{n \times \text{length}(J)}\right\}. \tag{4.13}$$

The performance of $\tilde{p}$ can be seriously affected by wrong choices of $c$ and $J$. Moreover, $J$ is a random variable and quantities defined in terms of $J$ have totally unknown distributions. Hence, a standard error and bias cannot be attached to $\tilde{p}$. Typically, if $J$ is chosen such that length($J$) is "too large", then $\tilde{p}$ will have large bias (and small variability) and vice versa if length($J$) is "too small".

The estimator for $p$ involving $\hat{\eta}_n$ (see (4.8)) has the same form as $\tilde{p}$ in (4.13). However, the important exception is that $J$ is replaced by an interval $I$ which is chosen *objectively*, in the sense that it can be calculated automatically from the data, without having to construct a histogram or any other density estimate.

Suppose the sample phases $X_1, X_2, \ldots, X_n$ are arranged in ascending order of magnitude and then denoted by $Y_1, Y_2, \ldots, Y_n$. Let $\hat{K}_n$ be the integer defined as before by

$$Y_{\hat{K}_n+\hat{s}_n} - Y_{\hat{K}_n-\hat{s}_n} = \max_{\hat{s}_n+1\le j\le n-\hat{s}_n} (Y_{j+\hat{s}_n} - Y_{j-\hat{s}_n}),$$

and let $I$ be the interval of phases with lower bound $Y_{\hat{K}_n-\hat{s}_n}$ and upper bound $Y_{\hat{K}n+\hat{s}_n}$, that is,

$$I = [Y_{\hat{K}_n-\hat{s}_n}, Y_{\hat{K}_n+\hat{s}_n}].$$

From the expression for $\hat{\eta}_n$ in (4.8), the proposed estimator of $p$ is now given by (compare with (4.13)),

$$\begin{aligned} \hat{p} &= \max\{0\,,\, 1-\hat{\eta}_n\} \\ &= \max\left\{0\,,\, 1 - \frac{(\text{number of } X_i\text{'s in } I)}{n \times \text{length}(I)}\right\}. \end{aligned}$$

The standard error of $\hat{p}$ (henceforth denoted by SE($\hat{p}$)) is given by (see (4.10))

$$\text{SE}(\hat{p}) = \frac{(2\hat{s}_n+1)^{1/2}}{n(Y_{\hat{K}_n+\hat{s}_n} - Y_{\hat{K}_n-\hat{s}_n})}.$$

It is important to bear in mind that the sample phases represent either noise or pulsed radiation. In other words, background radiation is always present and this implies that the minimum of $f$ is never small. Furthermore, by nature of the source function, the minimum of $f$ is typically nonpronounced. For these two reasons, $\hat{s}_n$ was calculated as for MSB2, that is, by using (4.9).

The signal strength $p$ was estimated for five sets of data. The first data set, AE Aquarii, consists of 7712 sample phases. AE Aquarii is a 33 second period accreting white dwarf emitting periodic X-rays with energies between 0.1–4 keV, as discovered by Patterson *et al.* (1980). The arrival times as described by De Jager (1991) were folded with the recent ephemeris of De Jager *et al.* (1994). The next four data sets were obtained by considering all pulsar phases above 50 MeV for Geminga (Mayer-Hasselwander *et al.*, 1994), Vela (Kanbach *et al.*, 1994) and Crab (Nolan *et al.*, 1993), but above 300 MeV for PSR1706-44 (Thompson *et al.*, 1992). The choice of 300 MeV for PSR1706-44 follows from the fact that the signal is weak below 300 MeV, as was pointed out by Thompson *et al.* (1993). The number of sample phases for these four data sets are 5018, 4691, 1470 and 477, respectively. The data were extracted from the public domain Phase I of the EGRET experiment on *Compton Gamma Ray Observatory* (*CGRO*).

For each of the data sets, $\hat{s}_n$, $I$, $\hat{p}$ and $\mathrm{SE}(\hat{p})$ were calculated (see Table 4.24). The values of $(2/n)^{1/2}$, the "standard conventional" measure of error in estimating the signal strength (De Jager *et al.*, 1989), are also included. It is evident that $\hat{p}$ estimates $p$ very accurately, since its standard error $\mathrm{SE}(\hat{p})$ is very small in each case. A striking feature of the results in the table is the remarkable correspondence between $\mathrm{SE}(\hat{p})$ and $(2/n)^{1/2}$.

Table 4.24: *Signal strength estimates and standard errors.*

| Data Set | $n$ | $\hat{s}_n$ | $I$ | $\hat{p}$ | $\mathrm{SE}(\hat{p})$ | $(2/n)^{1/2}$ |
|---|---|---|---|---|---|---|
| AE Aquarii | 7712 | 1540 | [0.229, 0.679] | 0.112 | 0.016 | 0.016 |
| Geminga | 5018 | 306 | [0.257, 0.486] | 0.468 | 0.022 | 0.020 |
| Vela | 4691 | 310 | [0.612, 0.996] | 0.656 | 0.014 | 0.021 |
| Crab | 1470 | 119 | [0.652, 0.947] | 0.450 | 0.036 | 0.037 |
| PSR1706-44 | 477 | 64 | [0.178, 0.559] | 0.290 | 0.063 | 0.065 |

Each data set is represented by binning the data into a phase histogram, as is typically done in the astrophysical literature. This is done merely to give a visual representation of the data. In the figures, I included the estimated background level $1-\hat{p}$, $1-\hat{p}\pm\mathrm{SE}(1-\hat{p})$

(as is also typically done in the astrophysical literature), and the phase interval $I$. The histograms displayed in Figure 4.25 are based on 50 phase bins each, and the histogram in Figure 4.26 on 20 phase bins.

All numerical computations were performed using FORTRAN programs together with IMSL (Version 1.0) and *exponent* GRAPHICS (Version 1.0) on an IBM RS6000 computer.

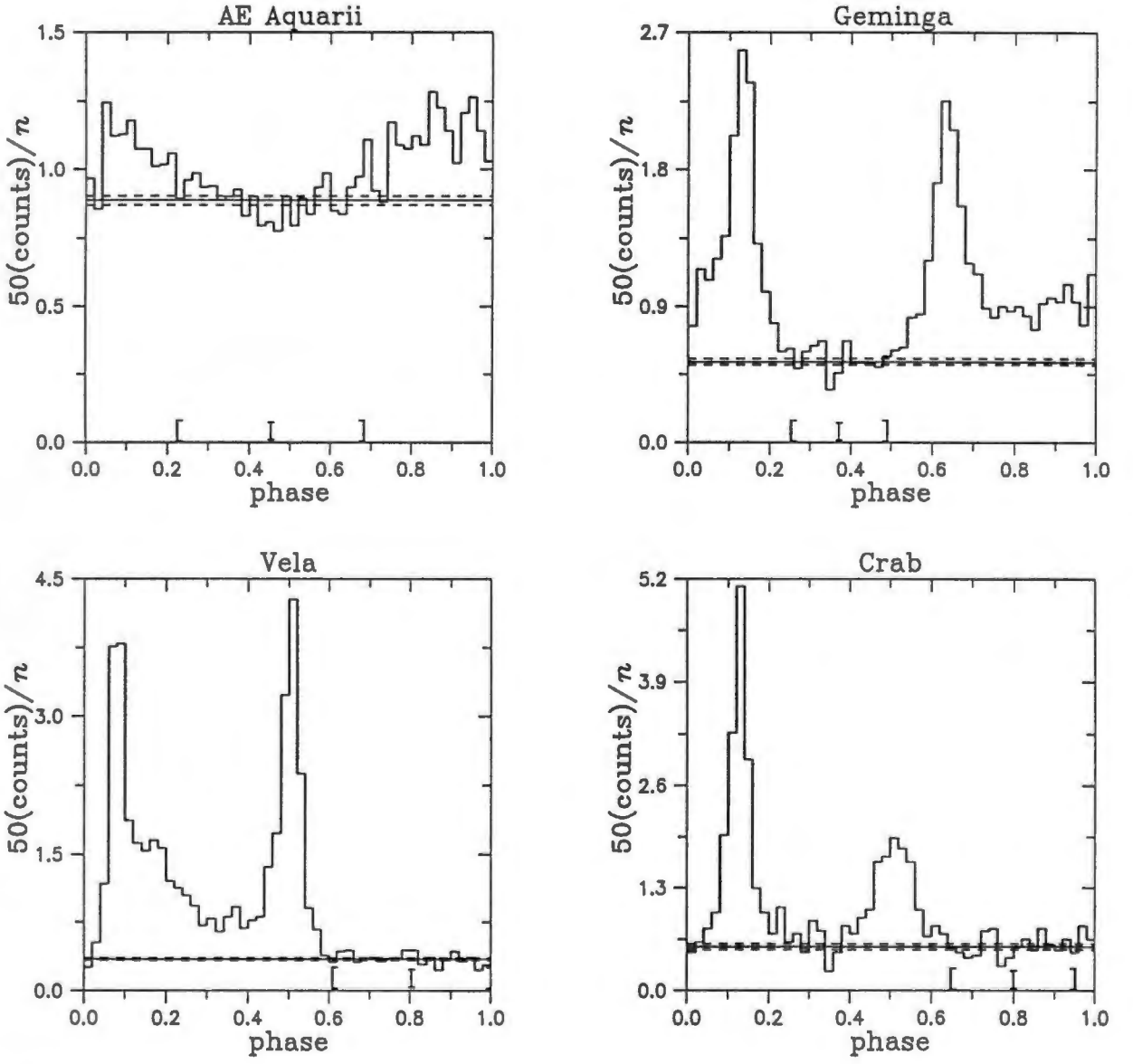


Figure 4.25: *Phase histograms, each with estimated background level* $1 - \hat{p}$ *(solid line),* $1 - \hat{p} \pm \mathrm{SE}(1 - \hat{p})$ *(dashed lines), and phase interval* $I$.

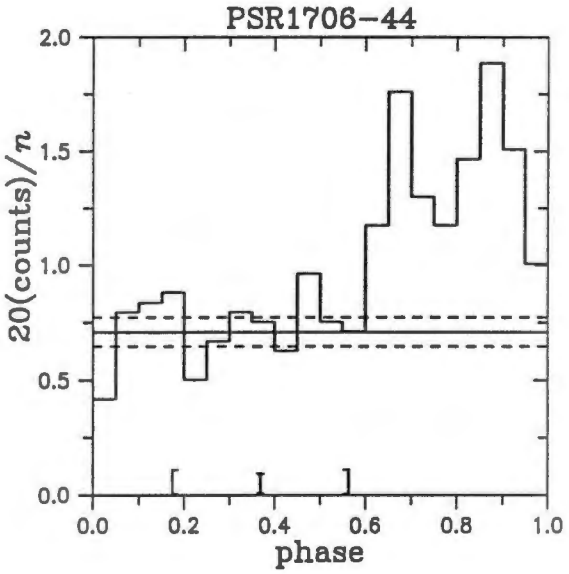


Figure 4.26: *Phase histogram, with estimated background level* $1 - \hat{p}$ *(solid line),* $1 - \hat{p} \pm \mathrm{SE}(1 - \hat{p})$ *(dashed lines), and phase interval* $I$.

# Bibliography


AZZALINI, A. 1981. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68: 326-328.

BARBE, P. 1992. Limiting distribution of the maximal spacing when the density function admits a positive minimum. *Statistics and probability letters*, 14: 53-60.

BERTRAND-RETALI, M. 1978. Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Revue roumaine de mathematiques pures et appliques*, 23: 361-385.

BERAN, R. 1984. Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker Vereinigung*, 86: 14-30.

BERAN, R. 1987. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74: 457-468.

BICKEL, P.J. 1987. Discussion of paper by B. Efron. *Journal of the American Statistical Association*, 82: 191.

BICKEL, P.J. & FREEDMAN, D.A. 1981. Some asymptotic theory for the bootstrap. *Annals of statistics*, 9: 1196-1217.

BILLINGSLEY, P. 1968. *Convergence of probability measures.* New York: Wiley. 253p.

BLUM, J.R., HANSON, D.L. & ROSENBLATT, J.I. 1963. On the central limit theorem for the sum of a random number of independent random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 1: 389-393.

BOWMAN, A.W. 1984. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71: 353-360.

BOWMAN, A.W. 1985. A comparative study of some kernel-based non-parametric density estimators. *Journal of statistical computation and simulation*, 21: 313-327.

BOWMAN, A.W., HALL, P. & TITTERINGTON, D.M. 1984. Cross-validation in non-parametric estimation of probabilities and probability densities. *Biometrika*, 71: 341-351.

BREIMAN, L. 1968. *Probability.* Reading, Massachusetts: Addison-Wesley. 421p.

BRILLINGER, D.R. 1975. *Time series: data analysis and theory.* New York: Holt, Rinehart and Winston. 500p.

BRONIATOWSKI, M., DEHEUVELS, P. & DEVROYE, L. 1989. On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate. *Annals of statistics*, 17: 1070-1086.

CACOULLOS, T. 1966. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18: 179-189.

CAO, R., CUEVAS, A. & GONZÁLEZ-MANTEIGA, W. 1994. A comparative study of several smoothing methods in density estimation. *Computational statistics and data analysis*, 17: 153-176.

CHERNOFF, H. 1964. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16: 31-41.

CHENG, M.-Y. 1994. A bandwidth selector for local linear density estimators. *Manuscript.*

CHIU, S.-T. 1991. Bandwidth selection for kernel density estimation. *Annals of statistics*, 19: 1883-1905.

CHIU, S.-T. 1992. An automatic bandwidth selector for kernel density estimation. *Biometrika*, 79: 771-782.

CHOW, Y.-S., GEMAN, S. & WU, L.-D. 1983. Consistent cross-validated density estimation. *Annals of statistics*, 11: 25-38.

DAVID, H.A. 1981. *Order statistics.* New York: Wiley. 360p.

DEHEUVELS, P. 1982. Strong limiting bounds for maximal uniform spacings. *Annals of probability*, 10: 1058-1065.

DEHEUVELS, P. 1983. Upper bounds for $k$-th maximal spacings. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62: 465-474.

DEHEUVELS, P. 1984. Strong limit theorems for maximal spacings from a general univariate distribution. *Annals of probability*, 12: 1181-1193.

DEHEUVELS, P. 1986 . On the influence of the extremes of an i.i.d. sequence on the maximal spacings. *Annals of probability*, 14: 194-208.

DEHEUVELS, P. & DEVROYE, L. 1984. Strong laws for the maximal $k$-spacing when $k \leq c \log n$. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66: 315-334.

DE JAGER, O.C. 1991. The unusual X-ray pulse timing of AE Aqaurii. *Astrophysical journal*, 378: 286-292.

DE JAGER, O.C., RAUBENHEIMER, B.C. & SWANEPOEL, J.W.H. 1989. The Rayleigh statistic in the case of weak signals - applications and pitfalls. (*In* Di Gesu, V., Scarsi, L., Crane, P., Friedman, J.H., Levialdi, S. & Maccarone, M.C. (eds.), Data Analysis in Astronomy III. London: Plenum Publishing Corporation. p.21-30.)

DE JAGER, O.C., MEINTJES, P.J., O'DONOGHUE, D. & ROBINSON, E.L. 1994. The discovery of a brake on the white dwarf in AE Aquarii. *Monthly notices of the Royal Astronomical Society*, 267: 577-588.

DEVROYE, L. 1981. Laws of the iterated logarithm for order statistics of uniform spacings. *Annals of probability*, 9: 860-867.

DEVROYE, L. 1982. A log log law for maximal uniform spacings. *Annals of probability*, 10: 863-868.

DEVROYE, L. 1983. The equivalence of weak, strong, and complete convergence in $L_1$ for kernel density estimates. *Annals of statistics*, 11: 896-904.

DEVROYE, L. 1989. On the non-consistency of the $L_2$-cross-validated kernel density estimate. *Statistics and probability letters*, 8: 425-433.

DEVROYE, L. & GYÖRFI, L. 1985. *Nonparametric density estimation: The $L_1$-view.* New York: Wiley. 356p.

DICICCIO, T. & EFRON, B. 1992. More accurate confidence intervals in exponential families. *Biometrika*, 79: 231-245.

DICICCIO, T.J. & ROMANO, J.P. 1988. A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society, B*, 50: 338-354.

DUIN, R.P.W. 1976. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on computers*, C-25: 1175-1179.

EDDY, W.F. 1980. Optimum kernel estimators of the mode. *Annals of statistics*, 8: 870-882.

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of statistics*, 7: 1-26.

EFRON, B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68: 589-599.

EFRON, B. 1982. The jackknife, the bootstrap and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics. 92p.

EFRON, B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82: 171-185.

EFRON, B. & TIBSHIRANI, R.J. 1993. *An introduction to the bootstrap.* New York: Chapman and Hall. 436p.

EPANECHNIKOV, V.A. 1969. Nonparametric estimation of a multivariate probability density. *Theory of probability and its applications*, 14: 153-158.

FARAWAY, J.J. & JHUN, M. 1990. Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association*, 85: 1119-1122.

FELUCH, W. & KORONACKI, J. 1992. A note on modified cross-validation in density estimation. *Computational statistics and data analysis*, 13: 143-151.

FISHER, N.I. 1993. *Statistical analysis of circular data.* Cambridge: Cambridge University Press. 277p.

GRENANDER, U. 1965. Some direct estimates of the mode. *Annals of mathematical statistics*, 36: 131-138.

GROENEBOOM, P. 1989. Brownian motion with a parabolic drift and airy functions. *Probability theory and related fields*, 81: 79-109.

HABBEMA, J.D.F., HERMANS, J. & VAN DEN BROEK, K. 1974. A stepwise discriminant analysis program using density estimation. (*In* Bruckmann, G., Ferschl, F. & Schmetterer, L. (eds.), Compstat 1974: Proceedings in Computational Statistics. Vienna: Physica Verlag. p.101-110.)

HALL, P. 1983. Large sample optimality of least squares cross-validation in density estimation. *Annals of statistics*, 11: 1156-1174.

HALL, P. 1987. On Kullback-Leibler loss and density estimation. *Annals of statistics*, 15: 1491-1519.

HALL, P. 1988a. On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society, B*, 50: 35-45.

HALL, P. 1988b. Theoretical comparison of bootstrap confidence intervals. *Annals of statistics*, 16: 927-953.

HALL, P. 1990. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of multivariate analysis*, 32: 177-203.

HALL, P., DICICCIO, T.J. & ROMANO, J.P. 1989. On smoothing and the bootstrap. *Annals of statistics*, 17: 692-704.

HALL, P. & JOHNSTONE, I. 1992. Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society, B*, 54: 475-530.

HALL, P. & MARRON, J.S. 1987a. Estimation of integrated squared density derivatives. *Statistics and probability letters*, 6: 109-115.

HALL, P. & MARRON, J.S. 1987b. Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability theory and related fields*, 74: 567-581.

HALL, P. & MARRON, J.S. 1987c. On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Annals of statistics*, 15: 163-181.

HALL, P. & MARRON, J.S. 1991. Lower bounds for bandwidth selection in density estimation. *Probability theory and related fields*, 90: 149-173.

HALL, P., MARRON, J.S. & PARK, B.U. 1992. Smoothed cross-validation. *Probability theory and related fields*, 92: 1-20.

HALL, P., SHEATHER, S.J., JONES, M.C. & MARRON, J.S. 1991. On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78: 263-269.

HALL, P. & WAND, M.P. 1988. On the minimization of absolute distance in kernel density estimation. *Statistics and probability letters*, 6: 311-314.

HARTIGAN, J.A. 1986. Discussion of the "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy" by B. Efron & R. Tibshirani. *Statistical science*, 1: 75-77.

IZENMAN, A.J. 1991. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86: 205-224.

JANSSEN, P., MARRON, J.S., VERAVERBEKE, N. & SARLE, W. 1995. Scale measures for bandwidth selection. *Manuscript.*

JONES, M.C. 1991. The roles of ISE and MISE in density estimation. *Statistics and probability letters*, 12: 51-56.

JONES, M.C., MARRON, J.S. & PARK, B.U. 1991. A simple root $n$ bandwidth selector. *Annals of statistics*, 19: 1919-1932.

JONES, M.C., MARRON, J.S. & SHEATHER, S.J. 1994. Progress in data-based bandwidth selection for kernel density estimation. *Technical report.*

JONES, M.C. & SHEATHER, S.J. 1991. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and probability letters*, 11: 511-514.

KANBACH, G., ARZOUMANIAN, Z., BERTSCH, D.L., BRAZIER, K.T.S., CHIANG, J., FICHTEL, C.E., FIERRO, J.M., HARTMAN, R.C., HUNTER, S.D., KNIFFEN, D.A., LIN, Y.C., MATTOX, J.R., MAYER-HASSELWANDER, H.A., MICHELSON, P.F., VON MONTIGNY, C., NEL, H.I., NICE, D., NOLAN, P.L., PINKAU, K., ROTHERMEL, H., SCHNEID, E., SOMMER, M., SREEKUMAR, P., TAYLOR, J.H. & THOMPSON, D.J. 1994. EGRET observations of the Vela pulsar, PSR0833-45. *Astronomy and Astrophysics*, 289: 855-867.

KIM, B.K. & VAN RYZIN, J. 1975. Uniform consistency of a histogram density estimator and modal estimation. *Communications in statistics*, 4: 303-315.

MAMMEN, E. 1990. A short note on optimal bandwidth selection for kernel estimators. *Statistics and probability letters*, 9: 23-25.

MARDIA, K.V. 1972. *Statistics of directional data.* London: Academic Press. 357p.

MARRON, J.S. 1985. An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Annals of statistics*, **13**, 1011-1023.

MARRON, J.S. & RUPPERT, D. 1994. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society, B*, 56: 653-671.

MARRON, J.S. & WAND, M.P. 1992. Exact mean integrated squared error. *Annals of statistics*, 20: 712-736.

MAYER-HASSELWANDER, H.A., BERTSCH, D.L., BRAZIER, K.T.S., CHIANG, J., FICHTEL, C.E., FIERRO, J.M., HARTMAN, R.C., HUNTER, S.D., KANBACH, G., KWOK, P.W., KNIFFEN, D.A., LIN, Y.C., MATTOX, J.R., MICHELSON, P.F., NOLAN, P.L., PINKAU, K., ROTHERMEL, H., SCHNEID, E.J., SOMMER, M., SREEKUMAR, P., THOMPSON, D.J. & VON MONTIGNY, C. 1994. High-energy gamma radiation from Geminga observed by EGRET. *Astrophysical journal*, 421: 276-283.

NADARAYA, É.A. 1965. On nonparametric estimates of density functions and regression curves. *Theory of probability and its applications*, 10: 186-190.

NARAYANAN, A. & SAGER, T.W. 1989. Table for the asymptotic distribution of univariate mode estimators. *Journal of statistical computation and simulation*, 33: 37-51.

NOLAN, P.L., ARZOUMANIAN, Z., BERTSCH, D.L., CHIANG, J., FICHTEL, C.E., FIERRO, J.M., HARTMAN, R.C., HUNTER, S.D., KANBACH, G., KNIFFEN, D.A., KWOK, P.W., LIN, Y.C., MATTOX, J.R., MAYER-HASSELWANDER,

H.A., MICHELSON, P.F., VON MONTIGNY, C., NEL, H.I., NICE, D., PINKAU, K., ROTHERMEL, H., SCHNEID, E., SOMMER, M., SREEKUMAR, P., TAYLOR, J.H. & THOMPSON, D.J. 1993. Observations of the Crab pulsar and nebula by the EGRET telescope on the *Compton Gamma-Ray Observatory. Astrophysical journal*, 409: 697-704.

PARK, B.U. & MARRON, J.S. 1990. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85: 66-72.

PARZEN, E. 1962. On estimation of a probability density function and mode. *Annals of mathematical statistics*, 33: 1065-1076.

PATTERSON, J., BRANCH, D., CHINCARINI, G. & ROBINSON, E.L. 1980. 33 second X-ray pulsations in AE Aquarii. *Astrophysical journal*, 240: L133-L136.

PRAKASA RAO, B.L.S. 1983. *Nonparametric functional estimation.* Orlando: Academic Press. 522p.

PROTHEROE, R.J. 1985. A new statistic for the analysis of circular data in gamma-ray astronomy. (*In* Jones, F.C., Adams, J. & Mason, G.M. (eds.), Proceedings: 19th International Cosmic Ray Conference (La Jolla). Washington DC: NASA Scientific and Technical Information Branch. p.485-488.)

ROMANO, J.P. 1988. Bootstrapping the mode. *Annals of the Institute of Statistical Mathematics*, 40: 565-586.

ROSENBLATT, M. 1956. Remarks on some nonparametric estimates of a density function. *Annals of mathematical statistics*, 27: 832-837.

ROSENBLATT, M. 1971. Curve estimates. *Annals of mathematical statistics*, 42: 1815-1842.

ROSENBLATT, M. 1991. *Stochastic curve estimation.* Regional conference series in probability and statistics. Vol. 3. California: Institute of Mathematical Statistics. 93p.

RUDEMO, M. 1982. Empirical choice of histograms and kernel density estimators. *Scandinavian journal of statistics*, 9: 65-78.

SAGER, T.W. 1975. Consistency in nonparametric estimation of the mode. *Annals of statistics*, 3: 698-706.

SAGER, T.W. 1978. Estimation of a multivariate mode. *Annals of statistics*, 6: 802-812.

SCHUCANY, W.R. 1989. Locally optimal window widths for kernel density estimation with large samples. *Statistics and probability letters*, 7: 401-405.

SCHUSTER, E.F. 1985. Incorporating support constraints into nonparametric estimators of densities. *Communications in statistics: Theory and methods*, 14: 1123-1136.

SCHUSTER, E.F. & GREGORY, C.G. 1981. On the nonconsistency of maximum likelihood nonparametric density estimators. (*In* Eddy, W.F. (ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. New York: Springer-Verlag. p.295-298.)

SCOTT, D.W. & FACTOR, L.E. 1981. Monte Carlo study of three data-based nonparametric density estimators. *Journal of the American Statistical Association*, 76: 9-15.

SCOTT, D.W. & TERRELL, G.R. 1987. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82: 1131-1146.

SERFLING, R.J. 1980. *Approximation theorems of mathematical statistics.* New York: Wiley. 371p.

SHEATHER, S.J. & JONES, M.C. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, B*, 53: 683-690.

SILVERMAN, B.W. 1986. *Density estimation for statistics and data analysis.* London: Chapman & Hall. 175p.

SILVERMAN, B.W. & YOUNG, G.A. 1987. The bootstrap: To smooth or not to smooth? *Biometrika*, 74: 469-479.

SINGH, K. 1981. On the asymptotic accuracy of Efron's bootstrap. *Annals of statistics*, 9: 1187-1195.

STONE, C.J. 1984. An asymptotically optimal window selection rule for kernel density estimates. *Annals of statistics*, 12: 1285-1297.

SWANEPOEL, J.W.H. 1986. A note on proving that the (modified) bootstrap works. *Communications in statistics: Theory and methods*, 15: 3193-3203.

SWANEPOEL, J.W.H. 1990. A review of bootstrap methods. *South African Statistical Journal*, 24: 1-34.

SWANEPOEL, J.W.H. & DE BEER, C.F. 1990. A new powerful test for periodic pulsed emission of high-energy photons. *Astrophysical journal*, 350: 754-757.

TAYLOR, C.C. 1989. Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, 76: 705-712.

THOMPSON, D.J., ARZOUMANIAN, Z., BERTSCH, D.L., BRAZIER, K.T.S., D' AMICO, N., FICHTEL, C.E., FIERRO, J.M., HARTMAN, R.C., HUNTER, S.D., JOHNSTON, S., KANBACH, G., KASPI, V.M., KNIFFEN, D.A., LIN, Y.C., LYNE, A.G., MANCHESTER, R.N., MATTOX, J.R., MAYER-HASSELWANDER, H.A., MICHELSON, P.F., VON MONTIGNY, C., NEL, H.I., NICE, D., NOLAN, P.L., PINKAU, K., ROTHERMEL, H., SCHNEID, E.J., SOMMER, M., SREEKUMAR, P. & TAYLOR, J.H. 1992. Pulsed high-energy $\gamma$-rays from the radio pulsar PSR1706-44. *Nature*, 359: 615-616.

THOMPSON, D.J., BERTSCH, D.L., FICHTEL, C.E., HARTMAN, R.C., HOFSTADTER, R., HUGHES, E.B., HUNTER, S.D., HUGHLOCK, B.W., KANBACH, G., KNIF-

FEN, D.A., LIN, Y.C., MATTOX, J.R., MAYER-HASSELWANDER, H.A., VON MONTIGNY, C., NOLAN, P.L., NEL, H.I., PINKAU, K., ROTHERMEL, H., SCHNEID, E.J., SOMMER, M., SREEKUMAR, P., TIEGER, D. & WALKER, A.H. 1993. Calibration of the Energetic Gamma-Ray Experiment Telescope (EGRET) for the *Compton Gamma-Ray Observatory. Astrophysical journal supplement series*, 86: 629-656.

VAN RYZIN, J. 1969. On strong consistency of density estimates. *Annals of mathematical statistics*, 40: 1765-1772.

VAN RYZIN, J. 1973. A histogram method of density estimation. *Communications in statistics*, 2: 493-506.

VENTER, J.H. 1967. On estimation of the mode. *Annals of mathematical statistics*, 38: 1446-1455.

WEGMAN, E.J. 1972. Nonparametric probability density estimation: I. A summary of available methods. *Technometrics*, 14: 533-546.

WHITT, W. 1970. Weak convergence of probability measures on the function space $C[0,\infty)$. *Annals of mathematical statistics*, 41: 939-944.

WHITT, W. 1971. Weak convergence of first passage time processes. *Journal of applied probability*, 8: 417-422.

WHITTLE, P. 1958. On the smoothing of probability density functions. *Journal of the Royal Statistical Society, B*, 20: 334-343.

WOODROOFE, M. 1970. On choosing a delta-sequence. *Annals of mathematical statistics*, 41: 1665-1671.