Some problems in kernel curve estimation

by

Ann Margaret Cowling

A thesis submitted for the degree of

Doctor of Philosophy of The Australian National University

May 1995
Corrigenda

In the text below, the numbers in brackets refer to the page and line number (page, line).

- (4,8) Add the following sentence in brackets at the end of the first paragraph on page 4: "(Marron and Wand (1992, p729ff) give a number of further objections to higher order kernels.)"

- (5,10) Change $h_{oT}$ to $h_{AOPT}$. Add the following immediately after equation (1.4). "Note that $h_{AOPT}$ minimises the asymptotic MISE rather than the actual MISE; see Marron and Wand (1992) for some comparisons of the minimiser of the actual MISE and asymptotic MISE which show that the difference may be large in some practical situations."


- (5,-14) Change the sentence “This produces $h_2$, then $j(r+2)$, $h_3$ and finally $j(r)$.” to “This produces $h_2$, then $f^{(r+2)}$, $h_3$ and finally $f^{(r)}$; see for example Wand and Jones (1995, p72).”

- (7, Section 1.3) Move the 5th paragraph of this Section, beginning “A point $t$ ...” to immediately before Corollary 1.1.

- (7, Corollary 1.1) Change the second sentence to the following: “For any boundary point $x = qh$ where $q \in [0, 1)$, provided $h \to 0$ and $nh \to \infty$ as $n \to \infty$

$$\text{bias } \left\{ \hat{f}(x) \right\} = -f(x) \int_q^1 K(u) \, du + O(h).$$

- (8,16) Add the reference Müller (1993).

- (18,-8) Add at end of sentence, “where $\hat{\tau} = \tau(\hat{\theta})$ is the estimate of the scale parameter $\tau$ based on the original sample $X$.

- (20,9) Change “the convergence of $U(x)$” to “the convergence of $\sup_T T(x)$”.

- (28,5) Replace the first sentence of the paragraph with “Local polynomial kernel estimators gained renewed attention following the papers of Stone (1977) and Cleveland (1979). Recent work of Fan (1992, 1993) has shown that they have a number of advantageous properties, including minimax optimality properties and desirable MSE properties.”

- (28,10) Add the reference Wei and Chu (1994).

- (30,8) Add the following as a new paragraph after the first paragraph on page 30 “Note that while $G$ is monotone, a polynomial interpolant may be non-monotone. In addition, a polynomial fit may behave erratically in places. These matters are discussed further in Section 2.4.1.”

- (30,9) Change “Note” to “Note also”.

- (33,-10) Change the sentence beginning “Let $h \sim \ldots$” to “Let $h \sim \text{const.} n^{-1/(2r+1)}$, and let $m$ be of larger order than $n^{2r/(2r+1)}$, and of smaller order than $n^{(2r/(2r+1)) + \eta}$, where $0 \leq \eta \leq 1/(2r + 1)$.”


- (35,1) Move the headings for Section 2.4 and 2.4.1 to the top of the page, above Figure 2.1.

- (43,7) Change the sentence beginning “Particularly at the ...” to “The left-hand end shows the trend in intensity at the beginning of the period when the accident rate was very high, and the right-hand end shows how current safety policies are working. In both cases it would be of interest to know whether the intensity is increasing or decreasing, but reflection cannot provide an answer.”

- (43,-11) Add at the end of the second paragraph of Section 2.4.3 “The cross-validation score function used was

\[ CV(h) = \int_0^1 \hat{\lambda}(x)^2 \, dx + 2 \sum_i \hat{\lambda}_{-i}(X_i) \]

where \( \hat{\lambda}(x) \) is the estimate of the intensity constructed using both the real data and the pseudodata, and \( \sum_i \hat{\lambda}_{-i}(X_i) \) is the sum over the sample points of the intensity estimates at each sample point, \( X_i \), where each estimate is constructed using the all the sample points except the \( r \)th, and the pseudodata.”

- (59,1) Replace the sentence beginning “The manner in ...” with “The manner in which our univariate algorithms should be modified is straightforward.”

- (61,-7) Change \( h_{opt} \) to \( h_{AOPT} \).

- (74, Section 3.4.1) Change all occurrences of \( h_{opt} \) to \( h_{AOPT} \).

- (74,11) Before “The roughness ratios ...” add “The chosen sample sizes are typical of the sample sizes which arise in many practical settings and were chosen so that the conclusions drawn from the simulations would be applicable in such contexts.”

- (76,-3) Change the sentence beginning “Moreover, the peaks ...” to “While the confidence band is consistent with a monotonically decreasing intensity, examination of a number of bootstrap simulations suggests that the peaks around \( x = 0.2 \) and \( x = 0.78 \) may be real features of the curve; see Cowling et al (1995)”.

- (78–82, captions) Change each occurrence of \( h_{opt} \) to \( h_{AOPT} \).

- (125, References) Add the following references:

**Additional References**


Declaration

I hereby declare that this thesis describes my own work, supervised by Professor Peter Hall. A portion of it has been submitted for publication in papers jointly with him and Dr M. J. Phillips from the University of Leicester, who visited the Centre for Mathematics and its Applications in 1994.

Ann Cowling
Acknowledgements

I am deeply grateful to Professor Peter Hall for his supervision. Although he was extremely busy, his door was always open and he gave generously of his time. His help has been invaluable.

I would like to thank the School of Mathematical Sciences and the ANU for their financial support.

My thanks go also to the other postgraduate students, members and visitors of the Statistical Sciences Program in the Centre for Mathematics and its Applications for their cheerful company, friendship and support.
Related Publications

The following papers have been submitted for publication from the work in this thesis:


Abstract

In this thesis, we present solutions to two problems in kernel curve estimation. First we give a new method for reducing the boundary bias of kernel density and intensity estimators. We apply this method in our second contribution, the development of bootstrap confidence bands for the kernel estimator of the intensity of an inhomogeneous Poisson point process.

The thesis is divided into three chapters. In the first we give an introduction to kernel density and intensity estimation, review the traditional methods of boundary bias correction, and briefly present bootstrap methods and their use in constructing confidence bands for kernel curve estimates. This chapter provides background and motivation for our later work.

In Chapter 2, we describe our new method for reducing the boundary bias of kernel estimators of arbitrary order. The proposed technique involves generating pseudodata beyond the support of the density \( f \), then using the pseudodata as well as the original data in the usual kernel estimator, thereby boundary-correcting the estimator. The pseudodata are simple to generate: they are formed using linear combinations of the order statistics of the sample, with coefficients that vary according to the order of the kernel. We show that the kernel estimators produced in this way have optimal orders of bias and variance throughout the estimation interval. Our method is applied to simulated data, and optimal pseudodata generation rules are proposed. We compare the performance of our method with that of traditional methods, and find that our method outperforms the traditional optimal boundary kernel method when \( f'(0) > 0 \). Finally, we compare pseudodata, reflection and boundary kernel intensity estimates for the classic coal mining disaster data set.
given in Jarrett (1979).

In Chapter 3 we develop theory and methods for calculating confidence bands for the intensity function \( \lambda(x) \) of an inhomogeneous Poisson process. Unlike previous researchers, we treat the Poisson process as genuinely non-stationary rather than as a Cox process. This requires us to derive the asymptotic distribution of \( \hat{\lambda}(x) - E\hat{\lambda}(x) \). Our work complements previous work on point estimation, as well as contributing new results on confidence regions.

There are numerous ways of applying the bootstrap to develop confidence bands for \( E\hat{\lambda} \). We describe methods based on several different bootstrap resampling algorithms and a number of approaches to pivoting. We remove edge effects using our pseudodata estimator, and correct the confidence intervals for bias, producing intervals for \( \lambda \). We demonstrate the effectiveness of these different approaches both theoretically and numerically using simulated data. Lastly, we discuss the intensity of the coal mining disaster data in the light of our confidence bands for \( \lambda \).
Contents

Declaration iii

Acknowledgements iv

Related Publications v

Abstract vi

1 Concepts of kernel curve estimation 1
   1.1 Introduction ........................................... 1
   1.2 Basic results in kernel density estimation .......... 2
   1.3 Boundary effects in kernel density estimation ..... 7
       1.3.1 Boundary bias reduction techniques ............ 8
       1.3.2 Proofs of corollaries ......................... 13
   1.4 Intensity function estimation ....................... 15
       1.4.1 Poisson processes ................................ 15
       1.4.2 Kernel estimator of intensity function ....... 16
   1.5 Bootstrap confidence bands for kernel curve estimates 17
   1.6 Motivation and summary of thesis .................. 21

2 A pseudodata method of removing boundary bias in kernel density estimation 27
   2.1 Introduction ........................................... 27
   2.2 Methodology .......................................... 29
   2.3 Asymptotic theory .................................... 32
   2.4 Numerical results .................................... 35
       2.4.1 Location of pseudodata ......................... 35
2.4.2 Simulation study ............................................. 37
2.4.3 Analysis of coal mining data ................................. 43
2.5 Proof of Theorem 2.1 ............................................. 45

Appendices 49
2A.1 Lemmas ............................................................ 49
2A.2 Tables: simulation results ....................................... 51

3 Bootstrap confidence regions for the intensity of a Poisson point process 57
3.1 Introduction .......................................................... 57
3.2 Methodology ........................................................ 60
  3.2.1 Kernel estimation of a Poisson intensity function .......... 60
  3.2.2 Bootstrap resampling methods ................................ 62
  3.2.3 Confidence bands before bias correction .................. 64
  3.2.4 Bias correction ............................................... 66
  3.2.5 Corrections for edge effects ................................. 67
3.3 Asymptotic theory ................................................. 68
3.4 Numerical results .................................................. 74
  3.4.1 Simulation study ............................................. 74
  3.4.2 Analysis of coal mining data ............................... 75
3.5 Proofs ............................................................... 83
  3.5.1 Proof of Theorem 3.2 ..................................... 83
  3.5.2 Proof of Theorem 3.3 ..................................... 95

Appendices 96
3A.1 Lemmas bounding $\mathbb{E} \{ \sup_x |Q_1(x)|^k \} $ .......................... 97
3A.2 Martingale results and applications ............................ 100
3A.3 Bounds for sums using integral approximations ............... 106
3A.4 A lemma on the closeness of two random processes .......... 110
3A.5 Hölder continuity of $\hat{\mu}'$ ................................ 116
3A.6 Technical Lemmas .............................................. 123

References 125
List of Tables

1.1 Parallels between order statistics and Poisson processes ............. 26
2.1 Boundary slopes of densities used in simulation study .......... .... 38
2.2 Comparison of two-point rules for $f'(0) < 0$ ................. .... 51
2.3 Comparison of two-point rules for $f'(0) > 0$ ................. .... 52
2.4 Comparison of three-point rules for $f'(0) < 0$ ................. .... 53
2.5 Comparison of three-point rules for $f'(0) > 0$ ................. .... 54
2.6 Comparison of boundary correction methods for $f'(0) < 0$ ......... 55
2.7 Comparison of boundary correction methods for $f'(0) > 0$ ......... 56
3.1 Comparison of coverage probabilities of confidence bands for $\lambda$ using smoothed bootstrap with different bandwidths ................. .... 80
3.2 Observed coverage probabilities of confidence bands for $\lambda$: bias correction by undersmoothing with recommended bandwidth choices .... 81
3.3 Observed coverage probabilities of confidence bands for $\lambda$: bias correction by undersmoothing with suboptimal bandwidth choices .... 82
List of Figures

1.1 Usual kernel density estimate: fixed kernel construction ............. 3
1.2 Usual kernel density estimate: moving kernel construction .......... 6
1.3 Boundary kernels for $x = 0, 0.2h, 0.4h, 0.6h, 0.8h, h$ .......... 9
1.4 Boundary kernel density estimate ..................................... 10
1.5 Reflection kernel density estimate .................................... 11
1.6 Cut-and-rescale density estimate ..................................... 12
1.7 Cumulative number of coal mining disasters: 1851–1962 ............ 22
2.1 Location of pseudodata: factors influencing sign of $X_{(i)}$ .......... 35
2.2 Empirical quantile function $\tilde{G}$ and extensions at boundaries ... 36
2.3 Densities used in simulation study ..................................... 39
2.4 Comparison of boundary correction methods for $f'(0) > 0$ ........... 41
2.5 Comparison of boundary correction methods for $f'(0) < 0$ ........... 42
2.6 Intensity of coal mining disasters: comparison of boundary kernel, reflection and pseudodata estimates .............................. 44
3.1 Bootstrap confidence band for intensity of coal mining disasters ... 76
3.2 Comparison of confidence band construction methods: bias correction by undersmoothing with recommended bandwidth choices .... 78
3.3 Comparison of confidence band construction methods: bias correction by undersmoothing with suboptimal bandwidth choices .......... 79
Chapter 1

Concepts of kernel curve estimation

1.1 Introduction

Non-parametric curve estimation techniques make no assumptions about the functional form or distribution of the data. The one assumption concerns the smoothness of the curve: typically, that it has at least two bounded derivatives.

In kernel curve estimation, the estimate at a given point is constructed using only local information: it is constructed from data values that lie in a neighbourhood of the point. To ensure that the estimate is consistent, as the sample size increases the neighbourhood must shrink in such a way that the points become dense in the neighbourhood. As only local information is used in the construction of an estimate, the convergence rate of these estimators is quite slow.

In applications, the support of the curve to be estimated is often known to be bounded on one or both sides. The smoothness assumption renders the kernel estimator unable to reproduce a discontinuity such as may occur at a boundary, and so the usual kernel estimator is much more biased near the boundaries than in the interior. For small and moderate sample sizes, a substantial portion of the support of the curve can be adversely affected by boundary bias. A number of different methods of reducing this bias have been suggested.
Non-parametric smoothing is an especially useful tool for exploratory analysis, as it may reveal structure that would be missed by imposing a parametric model. With the addition of a confidence band around the smooth curve, an unexpected structure can be assessed as either likely to be a real feature of the underlying curve or only an artifact of the sampling process.

In this thesis, we will be concerned with kernel estimation of density and Poisson intensity functions. However, the same principles apply, perhaps with modifications, to the estimation of regression curves.

The remainder of this chapter reviews in more detail the relevant aspects of kernel curve estimation. In Section 1.2 we present some basic concepts of kernel density estimation which are used throughout the thesis. The principal traditional methods of reducing boundary bias are given in Section 1.3. We review Poisson processes and define the kernel estimate of the intensity function in Section 1.4. Section 1.5 introduces the bootstrap and describes how it is used to construct confidence bands for kernel curve estimates. In the final section, Section 1.6, we motivate and outline the research detailed in Chapters 2 and 3.

### 1.2 Basic results in kernel density estimation

Kernel density estimation has attracted much research interest since the publication of the pioneering papers of Rosenblatt (1956) and Parzen (1962). Some recent books include Devroeye and Györfi (1985) developing the $L_1$ theory of density estimation; Silverman (1986), the classical introductory work; Scott (1992) on multivariate density estimation; and the new introductory text of Wand and Jones (1995) covering fundamental aspects of kernel estimation and many of the areas of research interest.

The traditional (see Silverman (1986, p15)) and conceptually natural view of the construction of the kernel density estimate is as follows. Suppose that $X_1, \ldots, X_n$ is a random sample from an unknown continuous distribution with density $f$. A kernel function $K$ spreads equal probability mass through neighbourhoods centred at each $X_i$, with width depending on the bandwidth $h$. The estimate is the sum of the heights of these kernels.
Figure 1.1: Usual kernel density estimate (dashed) for data from a truncated Normal density (solid). The estimate is the sum of the kernels (dotted) centered at the data points. Bandwidth $h = 1.65$.

This is illustrated in Figure 1.1 for a small simulated data set consisting of seven points from a truncated Normal distribution. This data set is used throughout Chapter 1 to demonstrate various kernel estimates and their construction. In practical applications, many more observations would be used: we use only seven observations here for clarity in the figures.

More formally, the usual kernel estimator of $f$ is

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K \{(x - X_i)/h\}, \quad (1.1)$$

where $h$ is the bandwidth, and $K$ is an $r$th order kernel satisfying

$$\int u^i K(u) du = \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{if } 1 \leq i \leq r - 1, \\ \kappa \neq 0 & \text{if } i = r. \end{cases} \quad (1.2)$$

It is obvious from (1.1) that $\hat{f}$ has the same differentiability properties as $K$, so that the order of the kernel $r$ is the number of bounded uniformly continuous derivatives that must be assumed for $f$.

The bias and variance of $\hat{f}$ are given in the following theorem, which is easily proved using Taylor expansion; see for example Silverman (1986, p39).
Theorem 1.1 If \( f \) has \( r \) bounded uniformly continuous derivatives and \( K \) satisfies the conditions in (1.2), then provided \( h \to 0 \) and \( nh \to \infty \) as \( n \to \infty \),

\[
\text{bias}\{\hat{f}(x)\} = (-1)^r \frac{1}{r!} \kappa h^r f^{(r)}(x) + o(h^r) \\
\text{var}\{\hat{f}(x)\} = (nh)^{-1} f(x) \int K^2(u)du + o\{(nh)^{-1}\}.
\]  

(1.3)

Clearly the bias of \( \hat{f} \) can be reduced by using a high-order kernel. However, kernels of order greater than two have negative side lobes, which may result in negative density estimates in places where data are sparse. Therefore, to ensure that \( \hat{f}(x) \geq 0 \) for all \( x \), second order kernels are generally used in practice.

The bias of the estimate can also be reduced by using a smaller bandwidth \( h \), but this leads to a noisy estimate \( \hat{f} \) with local detail masking global features of the sample (\( \hat{f} \) has high variance). If \( h \) is large, \( \hat{f} \) is smoother but the global features are dampened (\( \hat{f} \) has high bias and low variance). The bias, then, can only be reduced at the expense of variance and vice versa, with the bandwidth \( h \) determining the ratio of (squared) bias to variance.

The choice of the bandwidth \( h \) is perhaps the central issue in kernel estimation. We shall discuss bandwidth choice briefly here as it is not addressed elsewhere in this thesis.

Although in some circumstances it may suffice to choose the bandwidth by eye, automatic methods of choosing a bandwidth are generally preferred for their objectivity and relative time efficiency. The automatic methods minimise an error criterion, usually integrated squared error (ISE) or mean integrated squared error (MISE), where \( \text{ISE} = \int (\hat{f} - f)^2 \) and \( \text{MISE} = \mathbb{E} (\hat{f} - f)^2 \). See Jones (1991) for a discussion of the relative merits of these criteria.

Least-squares cross-validation, independently proposed for use in density estimation by Rudemo (1982) and Bowman (1984), minimises the ISE. More precisely, the cross-validatory bandwidth is that value of \( h \) which minimises

\[
\int \hat{f}^2 - 2n^{-1} \sum_i \hat{f}_{(-i)},
\]

an unbiased estimate of \( (\text{ISE} - \int f^2) \), where \( \hat{f}_{(-i)} = \{(n-1)h\}^{-1} \sum_{j \neq i} K\{(x - X_j)/h\} \), the kernel estimate of \( f \) constructed using all the data points except \( X_i \). The chosen
bandwidth is consistent but Hall and Marron (1987, 1991) have shown that the relative rate of convergence of methods minimising the ISE cannot be less than $n^{-1/10}$ whereas the relative error of methods based on minimising the MISE can be reduced to $n^{-1/2}$.

As

$$\text{MISE} = \int \mathbb{E} \left\{ \hat{f}(x) - f(x) \right\}^2 dx$$

$$= \int [\text{bias} \{ \hat{f}(x) \}]^2 dx + \int \text{var} \{ \hat{f}(x) \} dx$$

$$\sim (r!)^{-2} \kappa^2 h^{2r} \int f(r)(x)^2 dx + (nh)^{-1} \int K^2(u) du,$$

it is minimised by

$$h_{opt} = n^{-1/(2r+1)} \left[ \kappa^{-2} \left( \int f(r)(x)^2 dx \right)^{-1} (r!)^2 \left( \int K^2(u) du \right) \right]^{1/(2r+1)}. \quad (1.4)$$

Perhaps the most popular of the bandwidth selection methods targeting MISE are plug-in methods, which are based on using an estimate of $f(r)$ in (1.4); see for example Scott et al (1977), Park and Marron (1990) and Sheather and Jones (1991). The determination of $\hat{f}(r)$ involves a chain-like procedure: typically, an initial bandwidth $h_1$ is chosen and used to estimate $f^{(r+4)}$. This produces $h_2$, then $\hat{f}^{(r+2)}$, $h_3$ and finally $\hat{f}^{(r)}$. An obvious question is how the number of iterations should be chosen.

Empirical comparisons of various bandwidth selection methods using simulation are given in Park and Marron (1990), and Park and Turlach (1992). Comparisons based on real data sets are given in Sheather (1992). As yet no definitive recommendations for an optimal bandwidth selection method can be made.

The bandwidth need not be constant throughout the estimation interval. Variable bandwidths depending on the estimation point $x$, the data points $X_i$ or both have been proposed and studied; recent works include Jones (1990) and Hall (1990, 1992b).

Just as an "optimal" bandwidth can be found by minimising the MISE with respect to $h$, an "optimal" kernel can be found by minimising the MISE with respect to $K$, subject to the constraints (2.2). The optimal second order kernel, for example,
Figure 1.2: Usual kernel density estimate (dashed) for data from a truncated Normal density (solid). The “moving kernel” construction of the estimates at $x = 0$ and $x = 1.1$ is illustrated: a kernel (dotted) is centred at $x$, and the estimate is the sum of the ordinates of the kernel at abscissae $X_i$, $i = 1, \ldots, n$. Bandwidth $h = 1.65$.

is the Epanechnikov kernel

$$K(u) = \frac{3}{4}(1 - u^2)I_{[-1,1]}(u).$$

It can be shown that many other symmetric kernels have a high efficiency compared to the Epanechnikov kernel, and the quartic kernel

$$K(u) = \frac{15}{16}(1 - u^2)^2I_{[-1,1]}(u),$$

which has an efficiency of 0.994 relative to the Epanechnikov, is often preferred as it is differentiable at $u = \pm 1$ producing a smoother estimate. Wand and Jones (1995) tabulate the efficiencies of several kernels relative to the Epanechnikov.

We note here that there is a second view of the construction of a kernel density estimate. In the interpretation given at the beginning of this section, the kernels are regarded as fixed, and centred at the data points $X_i$. In the second, the kernel “moves”, and is centred at the estimation point $x$. This second view is related to the usual explanation of the construction and use of kernel weights in kernel regression; see for example Härdle (1991, p24ff). It is also related to moving averages in time series. We remark here on the different interpretations as the asymmetric kernels used for reducing boundary bias are moving kernels depending on $x$. 
The moving kernel construction of the density estimate at the point $x$ is as follows: centre a scaled kernel at $x$, and take as the estimate the sum of the ordinates of the kernel at abscissae $X_i$. Whilst this construction can be clearly illustrated for an isolated point $x$, see Figure 1.2, it does not allow such clear illustration of the construction of the estimator over the full range of $x$ values as does the fixed kernel construction.

1.3 Boundary effects in kernel density estimation

When the support of $f$ is bounded, the usual kernel estimator (1.1) performs poorly near the boundary. The variance is the same size, $o\{(nh)^{-1}\}$, in the boundary regions as in the interior, but the bias is much greater. More specifically, as $x$ approaches a boundary $b$, $\hat{f}(x)$ tends increasingly to underestimate $f(x)$, until at the boundary under the usual consistency conditions, $E\hat{f}(b) \sim \frac{1}{2}f(b)$, as can be seen from Corollary 1.1 to Theorem 1.1.

The underestimation at the boundary is illustrated in Figure 1.1 on page 3. This figure also shows that kernels centred at points in the boundary region extend past the boundary making $\hat{f}(x) > 0$ beyond the support of $f$. Thus when $f$ has bounded support, $\hat{f}$ is not a density since $\int_{\text{supp}(f)}\hat{f} \neq 1$.

**Corollary 1.1** In addition to the conditions of Theorem 1.1, assume that $K$ and $f$ are supported on $[-1, 1]$ and $[0, 1]$ respectively. Then, provided $h \to 0$ and $nh \to \infty$ as $n \to \infty$,

$$\text{bias } \{\hat{f}(x)\} = -f(x) \int_{q}^{1} K(u)du + O(h).$$

Because the bias in the interior is $O(h^r)$, edge effects dominate the global asymptotic behaviour, influencing bandwidth choice in cross-validatory or plug in methods for example.

We will review the main traditional methods which have been proposed for boundary bias reduction, and in the remainder of this chapter and Chapter 2, will
assume without loss of generality that $f$ has compact support $[0, 1]$. We will further assume that $K$ has compact support $[-1, 1]$ as otherwise the whole estimation interval $[0, 1]$ would be subject to boundary effects.

A point $t$ whose location in $[0, 1]$ does not depend on $n$ is not informative for studying boundary behaviour since as $n \to \infty$, $h \to 0$ and so $t$ will eventually become an interior point. Therefore, for studying boundary behaviour we consider a sequence of points $x(n)$ which have the same relative location in a boundary interval for all $n$: $x(n) = qh$ or $x(n) = 1 - qh$, with $q \in [0, 1)$. Only the correction methods at the left boundary will be described; corrections at the right boundary can be obtained analogously.

### 1.3.1 Boundary bias reduction techniques

#### Boundary kernels

Gasser and Müller (1979) proposed using different kernels, called boundary kernels, in the boundary regions for kernel regression. They developed the theory in Gasser et al (1985), and suggested that boundary kernels could also be employed for correcting edge effects in density estimation.

We must take the moving perspective of the construction of the estimator when discussing the boundary kernel method as boundary kernels vary according to the estimation point $x$. For a boundary point $x = qh$, only the interval $[-1, q)$ of the support of $K$ is mapped into $[0, 1]$. The boundary kernel at $x$, $K_q$, is defined as a modification of the interior kernel $K$, which has support $[-1, q)$ and satisfies the moment conditions of the interior (1.2). It is desirable that $K_q$ should depend continuously on $q$ and that $K_q \to K$ as $q \to 1$.

The boundary kernel estimator is

$$
\hat{f}_{BK}(x) = \begin{cases} 
(nh)^{-1}\sum_{i=1}^{n}K_q\{(x - X_i)/h\} & \text{if } x \in [0, h); \\
(nh)^{-1}\sum_{i=1}^{n}K\{(x - X_i)/h\} & \text{otherwise.}
\end{cases}
$$

The definition of boundary kernels ensures that the bias of $\hat{f}_{BK}$ is of the same order in the boundaries as at interior points.
Suppose that the Epanichnikov kernel is used in the interior. For the left boundary region, Gasser et al (1985) propose the 2nd order polynomials

\[ K_q(u) = (c_0 + c_1 u + c_2 u^2)I_{[-1,q]}(u) \]  

(1.5)

where \( c_0, c_1 \) and \( c_2 \) are constants depending on \( q \) such that

\[ \int_{-1}^{q} K_q(u)du = 1, \quad \int_{-1}^{q} uK_q(u)du = 0, \quad \text{and} \quad K_q(-1) = 0. \]

Figure 1.3 shows these kernels for various values of \( q \).

Using the quartic kernel in the interior, Hall and Wehrly (1991) employ 5th order polynomial boundary kernels based on the quartic kernel:

\[ K_q(u) = (c_0 + c_1 u)^{\frac{15}{16}}(1 - u^2)^2I_{[-1,q]}(u). \]

The coefficients \( c_0 \) and \( c_1 \) are obtained by solving

\[ \int_{-1}^{q} K_q(u)du = 1 \quad \text{and} \quad \int_{-1}^{q} uK_q(u)du = 0. \]

These kernels are defined so that \( K_q(-1) = 0 \). Although similar in general outline to the Gasser, Müller and Mammitsch kernels, the Hall and Wehrly kernels are smoother as they are differentiable at \( u = \pm 1 \). The Hall and Wehrly boundary kernel estimate for the simulated data set is shown in Figure 1.4. Note that \( \int_{\text{supp}(f)} f_{BK} \neq 1 \).
Figure 1.4: Boundary kernel density estimate (dashed) for data from a truncated Normal density (solid). The construction of the estimates at $x = 0$ and $x = 1.1$ is illustrated: a kernel (dotted) is centred at $x$, and the estimate is the sum of the ordinates of the kernel at abscissae $X_i$, $i = 1, \ldots, n$. Bandwidth $h = 1.65$.

Although boundary kernels reduce boundary bias to the appropriate order, they are not strictly positive; see Figure 1.3. This means that the boundary kernel method cannot be guaranteed to give a positive density estimate especially if data are sparse in boundary regions. This occurred with the simulated data set as shown in Figure 1.4. The other drawback of the boundary kernel approach is the more complex programming involved: a different kernel must be constructed for each estimation point lying in the boundary region.

Reflection

The reflection technique is described in Boneva et al (1970). Schuster (1985) gives conditions under which the reflection estimator is asymptotically unbiased, consistent, asymptotically normal and strongly uniformly consistent.

The data $X_1, \ldots, X_n$ are reflected in the boundary, giving rise to pseudodata $X_{-1}, \ldots, X_{-n}$ on $[-1, 0]$, where $X_{-i} = -X_i$, $i = 1, \ldots, n$. The usual estimator is applied to the augmented data set consisting of both the original data and the
Figure 1.5: Reflection density estimate (dashed) for data from a truncated Normal density (solid). The estimate is the sum of the kernels (dotted) centered at the data points and pseudodata points (□). Bandwidth $h = 1.65$.

pseudodata. The reflection estimator is

$$\hat{f}_R(x) = (nh)^{-1} \left[ \sum_{i=1}^{n} K \{(x - X_i)/h\} + \sum_{i=1}^{n} K \{(x - X_{-i})/h\} \right];$$

the estimate for the simulated data set is shown in Figure 1.5. Reflection is the only method of boundary correction considered in this chapter for which $f_{\text{supp}}(f) \hat{f} = 1$.

The technique is particularly simple to implement, because the same estimator is used throughout the estimation interval. The drawback of the method is that it does not adequately correct the boundary bias as indicated by the following Corollary:

**Corollary 1.2** In addition to the conditions of Theorem 1.1, assume that $K$ and $f$ are supported on $[-1,1]$ and $[0,1]$ respectively. Then, provided $h \to 0$ and $nh \to \infty$ as $n \to \infty$,

$$\text{bias} \left\{ \hat{f}_R(x) \right\} = 2hf'(x) \int_{q-1}^{1} (u - q)K(u)du + o(h).$$

**Cut-and-rescale**

Another long-used technique for boundary correction is the cut-and-rescale method; see for example Gasser and Müller (1979), Diggle (1985). In the boundary regions a different kernel is constructed at each estimation point $x = qh$: the usual kernel
Figure 1.6: Cut-and-rescale density estimate (dashed) for data from a truncated Normal density (solid). The construction of the estimates at \( x = 0 \) and \( x = 1.1 \) is illustrated: a kernel (dotted) is centred at \( x \), and the estimate is the sum of the ordinates of the kernel at abscissae \( X_i, \ i = 1, \ldots, n \). Bandwidth \( h = 1.65 \).

is truncated at the boundary, and the portion lying in the estimation interval \([0, 1]\) is rescaled to integrate to one:

\[
K_{qe}(u) = \frac{K(u)}{\int_{-1}^{1} K(u)du} I_{[0, 1]}(u).
\]

The cut-and-rescale estimator is

\[
\hat{f}_{CR}(x) = \begin{cases} 
(nh)^{-1} \sum_{i=1}^{n} K_{qe} \{(x - X_i)/h\} & \text{if } x \in [0, h); \\
(nh)^{-1} \sum_{i=1}^{n} K \{(x - X_i)/h\} & \text{otherwise}.
\end{cases}
\]

Figure 1.6 shows this estimator applied to the simulated data set. The figure clearly shows that \( \int_{\text{supp}(f)} \hat{f} \neq 1 \).

At the boundary, the cut-and-rescale estimate is exactly equal to the reflection estimate, and so these methods have the same bias there.

**Corollary 1.3** In addition to the conditions of Theorem 1.1, assume that \( K \) and \( f \) are supported on \([-1, 1]\) and \([0, 1]\) respectively. Then, provided \( h \to 0 \) and \( nh \to \infty \) as \( n \to \infty \),

\[
\text{bias} \{\hat{f}_{CR}(x)\} = \frac{hf'(x)}{\int_{-1}^{1} K(u)du} \frac{\int_{-1}^{1} uK(u)du}{\int_{-1}^{1} K(u)du} + o(h).
\]
Transformation

Transformation may be useful for removing boundary effects. A transformation $t$ is applied to the sample $X_1, \ldots, X_n$, giving a new data set $Y_1, \ldots, Y_n$ where $Y_i = t(X_i)$. Suppose that the transformed data have density $g$. Then standard theory gives $f(x) = g(t(x))t'(x)$, and hence

$$f_T(x) = (nh)^{-1}t'(x) \sum_{i=1}^{n} K \left[ \{t(x) - t(X_i)/h \} \right].$$

Transformations have been used to reduce boundary bias in density estimation for some time, see for example Copas and Fryer (1980). However, the theory of transformations in kernel density estimation is a relatively new field for research. Marron and Ruppert (1994) propose transforming from a density with bounded support to a density with first derivative of approximately zero at the boundaries. The new density is estimated using reflection for boundary correction, and backtransformed to produce the estimate of the original density. However, the choice of transformation involves estimation of several parameters and the method is not quick and easy to implement.

1.3.2 Proofs of corollaries

Proof of Corollary 1.1  Provided $h \to 0$ as $n \to \infty$, for a point $x = qh$ where $q \in [0, 1)$,

$$\text{bias} \left\{ \hat{f}(x) \right\} = -f(x) + f(x) \int_{-1}^{q} K(u)du - h f'(x) \int_{-1}^{q} uK(u)du + o(h).$$

Therefore,

$$\text{bias} \left\{ \hat{f}(x) \right\} = -f(x) \int_{q}^{1} K(u)du + O(h).$$
Proof of Corollary 1.2  Provided $h \to 0$ as $n \to \infty$, for a point $x = qh$ where $q \in [0, 1)$,

$$E \{ f_R(x) \} = h^{-1} \left( E \left[ K\{(x - Y)/h\} \right] + E \left[ K\{(x + Y)/h\} \right] \right)$$

$$= h^{-1} \left[ \int_0^{x+h} K\{(x - y)/h\} f(y) dy + \int_0^{-x+h} K\{(x + y)/h\} f(y) dy \right]$$

$$= \int_{-1}^{q} K(u) f(x - hu) du + \int_{2q-1}^{q} K(2q - u) f(x - hu) du$$

$$= f(x) \left\{ \int_{-1}^{q} K(u) du + \int_{2q-1}^{q} K(2q - u) du \right\}$$

$$- hf'(x) \left\{ \int_{-1}^{q} uK(u) du + \int_{2q-1}^{q} uK(2q - u) du \right\} + o(h)$$

Therefore,

$$\text{bias} \{ f_R(x) \} = 2hf'(x) \int_{q}^{1} (u - q)K(u) du + o(h).$$

Proof of Corollary 1.3  Provided $h \to 0$ as $n \to \infty$, for a point $x = qh$ where $q \in [0, 1)$,

$$E \{ f_{CR}(x) \} = h^{-1} \int_0^{x+h} K_{CR}\{(x - y)/h\} f(y) dy$$

$$= \frac{\int_{-1}^{q} K(u) f(x - hu) du}{\int_{-1}^{q} K(u) du}$$

$$= f(x) - hf'(x) \frac{\int_{-1}^{q} uK(u) du}{\int_{-1}^{q} K(u) du} + o(h).$$

Therefore,

$$\text{bias} \{ f_{CR}(x) \} = hf'(x) \frac{\int_{-1}^{q} uK(u) du}{\int_{-1}^{q} K(u) du} + o(h).$$
1.4 Intensity function estimation

1.4.1 Poisson processes

A Poisson process is a mathematical model describing the location of point events occurring completely at random in $\mathbb{R}^d$, for example, the times of emissions from a radioactive source or the location of raisins in a cake. The basic theory and properties are well known, see for example Feller (1968).

The three characteristic features of a homogeneous Poisson process on $[0, \infty)$ with rate or intensity parameter $\lambda$ are:

1. The number of points in each finite interval $[u_i, v_i)$ has a Poisson distribution with mean $\lambda(v_i - u_i)$.

2. The numbers of points in disjoint intervals are independent random variables.

3. The distributions are stationary: they depend only on the lengths of the intervals.

An important property of the homogeneous Poisson process is that conditional on the number of events $N$ in the interval of interest, say $[0, T)$, the times of the events are independently uniformly distributed over $[0, T)$.

A Poisson process can be generated from sums of independent exponential random variables: suppose that $\{Z_1, Z_2, \ldots\}$ are independent exponential random variables with parameter $\lambda$ and mean $1/\lambda$. Let $Y_i = \sum_{j=1}^i Z_j$. Then $\{Y_1, Y_2, \ldots\}$ is a Poisson process with parameter $\lambda$. Note that $Y_i \sim \text{Gamma}(\lambda, i)$ with density

$$f_{\lambda,i}(y) = \lambda (\lambda y)^{i-1} e^{-\lambda y} / \Gamma(i) \quad y > 0,$$

so that $\mathbb{E}(Y_i) = i/\lambda$ and $\text{var}(Y_i) = i/\lambda^2$.

Random events need not occur uniformly in $\mathbb{R}^d$: events may occur randomly with some locations being more likely than others, for example, the locations of burglaries in a city, or the times of arrivals at a university cafeteria.

A non-homogeneous Poisson processes on $[0, \infty)$ is determined by a rate or intensity function $\lambda(t) \geq 0$. The process is defined by the following properties:
1. The number of points in each interval \((u_i, v_i]\) has a Poisson distribution with mean \(\int_{u_i}^{v_i} \lambda(t) dt\).

2. The numbers of points in disjoint intervals are independent random variables.

A non-homogeneous Poisson process can be simply related to a homogeneous Poisson process: if \(\{Y_1, Y_2, \ldots\}\) is a Poisson process with unit mean, then writing \(\Lambda(t) = \int_0^t \lambda(x) dx\) and \(X_i = \Lambda^{-1}(Y_i), \{X_1, X_2, \ldots\}\) is a non-homogeneous Poisson process with intensity function \(\lambda(t)\).

For the non-homogeneous Poisson process, conditional on \(N\), the times of the events are independently distributed over \([0, T)\) with density \(\lambda(t)/\Lambda(T)\). Thus \(\lambda(t)\) can be considered as a density of sorts with the difference that \(\Lambda(T) = \int_0^T \lambda(t) dt\) is unlikely to equal 1.

There are a number of ways of simulating a realisation of a non-homogeneous Poisson process on \([0, 1]\), see for example Devroye (1986). A simple (but not always very efficient) method, based on the analogy between intensities and densities, is to first generate a Poisson random variable \(N\) with mean \(\Lambda(1)\), and then to generate \(\{X_i\}\) as the order statistics of a sample of size \(N\) from the distribution with density \(\lambda(t)/\Lambda(1)\). Another method uses inversion of the integrated rate function and the exponential spacings characteristic.

In our simulations in Chapter 3, we used the thinning method, which is similar to the rejection method of generating points from a specified density. In the thinning method, points are generated from a Poisson process on \([0, 1]\) with rate \(\mu\) where \(\mu \geq \lambda(t)\). A prospective event occurring at time \(t \in [0, 1]\) is accepted with probability \(\lambda(t)/\mu\). The "thinned" process, consisting of the accepted events, is a realisation of the non-homogeneous process with intensity \(\lambda(t)\).

### 1.4.2 Kernel estimator of intensity function

Leadbetter and Wold (1983) point out that a common feature of density and intensity estimation is that the function to be estimated is the derivative of a function with a natural discontinuous estimator. The target estimator can therefore be obtained by smoothing the steps using a differentiable function, and differentiating.
The obvious estimators of the distribution function and the integrated rate function are

\[ \tilde{F}(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x) \quad \text{and} \quad \tilde{\Lambda}(x) = \sum_{i=1}^{N} I(X_i \leq x), \]

where \( N \) is the number of events of the Poisson process in the interval \([0, T]\). Smooth versions of these can be obtained by replacing \( I(X_i \leq x) \) with a differentiable function \( \delta\{((x - X_i)/h)\} \). The required estimators, \( \tilde{f}(x) \) and \( \tilde{\lambda}(x) \), are therefore given by

\[ \tilde{f}(x) = (nh)^{-1} \sum_{i=1}^{n} \delta\{(x - X_i)/h\} \quad \text{and} \quad \tilde{\lambda}(x) = h^{-1} \sum_{i=1}^{N} \delta\{((x - X_i)/h)\}. \]

This derivation focuses on the similarity between the distribution function and the integrated rate function, and leads to another interpretation of the kernel function \( K \) — as the derivative of a smoothed indicator function. The kernel estimator of the intensity function is

\[ \hat{\lambda}(x) = h^{-1} \sum_{i=1}^{N} K\{(x - X_i)/h\}. \]

We shall derive the distribution of this estimator in Chapter 3, and therefore delay discussion of its bias and variance until then.

### 1.5 Bootstrap confidence bands for kernel curve estimates

In a non-parametric setting, the bootstrap is a Monte Carlo based method of obtaining properties of the distribution of a particular statistic without making assumptions about its distributional form. Much of the recent research into the bootstrap is aimed at finding reliable, automatic, data-based methods of calculating confidence intervals, but bootstrap methods can also be used to determine other measures of statistical accuracy such as bias and prediction error.

Traditional confidence intervals for a parameter \( \theta \) are often based on the distribution of the asymptotically pivotal statistic \( T = (\hat{\theta} - \theta)/\tau(\hat{\theta}) \), where \( \tau(\hat{\theta}) \) is an appropriate sample-based scale estimate, or on the non-pivotal statistic \( S = \hat{\theta} - \theta \).
Bootstrap confidence intervals based on $T$, known as percentile-$t$ intervals, can be shown to be more accurate than intervals based on $S$, known as percentile intervals; see Beran (1987) and Hall (1992a). We shall derive below the symmetric percentile-$t$ bootstrap confidence interval for a single parameter $\theta$, but shall first introduce some necessary concepts and notation.

We regard the parameter $\theta$ as a functional of the population distribution function $F$, and denote it $\theta(F)$. The bootstrap estimate of $\theta(F)$ is the same functional of the empirical distribution function $\hat{F}$: $\hat{\theta} = \theta(\hat{F})$. For example, the bootstrap estimate of the population mean $\mu = \int x dF(x)$ is the sample mean $\bar{X} = \int x \hat{F}(x)$. Let $X^* = (X_1^*, \ldots, X_n^*)$ be a resample drawn from the original sample $X = (X_1, \ldots, X_n)$, let $\hat{F}^*$ be the empirical distribution function of the resample, and let $\hat{\theta}^* = \theta(\hat{F}^*)$ be the same functional of the bootstrap resample. Then $T^*$, the bootstrap estimate of $T$, is obtained by replacing $\hat{\theta}$ by $\hat{\theta}^*$ and $\theta$ by $\hat{\theta}$ in the expression for $T$, so that $T^* = (\hat{\theta}^* - \hat{\theta})/\tau(\hat{\theta}^*)$. The bootstrap "works" because by judicious choice of the sampling mechanism, the distribution of $T^*$ conditional on $X$, known as the bootstrap distribution of $T^*$, is the same as that of $T$.

A symmetric $\alpha\%$ confidence interval for $T$ is given by $(-t, t)$, where $P(-t \leq T \leq t) = \alpha$. Without making assumptions about the distribution of $T$, we estimate the quantile $t$ by the bootstrap as follows. A number, $B$, of resamples are drawn from $X$, $T^*$ is calculated for each resample, and our estimate $\hat{t}$ of $t$ is the $\alpha$-level quantile of the empirical distribution of $|T^*|$. The symmetric percentile-$t$ bootstrap confidence interval for $\theta$ is then simply $(\hat{\theta} - \hat{t}, \hat{\theta} + \hat{t})$.

In conventional problems with a fixed number of parameters and unknown distribution function $F$, bootstrap confidence intervals for $\theta$ are more accurate than traditional Normal approximation intervals. The reason, which is revealed by an Edgeworth expansion argument (see for example Singh (1981), and Hall (1988, 1992a)), is that the bootstrap correctly estimates first order departures from normality which are due to skewness, whereas traditional Normal theory does not. The essence of the argument (see Hall (1992a, p83)) goes as follows: the pivotal statistic $T$ is generally
asymptotically Normal, and so its distribution function can be expanded as

\[ G(x) = \Pr(T \leq x) = \Phi(x) + n^{-1/2}q(x)\phi(x) + O(n^{-1}), \] (1.6)

where \( q \) is an even quadratic polynomial and \( \Phi, \phi \) are respectively the standard Normal distribution and density functions. The bootstrap estimate of \( G \) can similarly be expanded as

\[ \hat{G}(x) = \Pr(T^* \leq x|X) = \Phi(x) + n^{-1/2}\hat{q}(x)\phi(x) + O_p(n^{-1}), \]

where the polynomial \( \hat{q} \) is obtained from \( q \) by replacing unknowns such as skewness by their bootstrap estimates. Since \( \hat{q} - q = O_p(n^{-1/2}) \),

\[ \Pr(T^* \leq x|X) - \Pr(T \leq x) = O_p(n^{-1}); \]

that is, the bootstrap approximation to \( G \) is in error by only \( n^{-1} \). This is an improvement over the usual Normal approximation, \( G \approx \Phi \), which is in error by \( n^{-1/2} \) as shown in (1.6).

In kernel curve estimation, only the data points within a neighbourhood of width \( h \) of \( x \) contribute to the estimate at \( x \), and so the effective sample size is \( nh \) rather than \( n \). The smaller effective sample size leads to a worsening of the convergence rate, but bootstrap methods still perform better than Gaussian theory methods, as we shall see below.

The effective sample size of \( nh \) can be interpreted as meaning that our procedure is equivalent to estimating \( h^{-1} \sim n^{1/5} \) parameters, a number which increases with \( n \). Thus curve estimation problems are sometimes referred to as infinite (number of) parameter problems.

A confidence band for a curve \( \gamma(x) \) on an interval \( I \) is the region between two curves, \( LB(x) \) and \( UB(x) \), \( x \in I \), where \( LB(x) \) and \( UB(x) \) are chosen to ensure that the true curve is covered by the band \([LB(x), UB(x)]\) with probability \( \alpha \); that is,

\[ \Pr(LB(x) \leq \gamma(x) \leq UB(x) \text{ for all } x \in I) = \alpha. \]

When determining a confidence band on \( I \) or a simultaneous confidence interval for a finite set of points in \( I \), rather than a confidence interval for a single point, our
interest moves from $T$ or $|T|$ to $\sup_T T(x)$ or $\sup_T |T(x)|$. The asymptotically pivotal statistic $T(x) = \{\hat{\gamma}(x) - \gamma(x)\}/\tau\{\hat{\gamma}(x)\}$ is a stochastic process in $x$ which can be shown to converge to a stationary Gaussian process $U(x)$ with known covariance structure.

Confidence bands can be constructed based on the distribution of $\sup_T U(x)$. Extreme value theory for Gaussian processes has been well studied; see for example Hall (1979), Leadbetter et al (1983). The extreme value method of constructing confidence bands for kernel curve estimates was first proposed in Bickel and Rosenblatt (1973) for kernel density estimates. However, the convergence of $U(x)$ to $\sup_T U(x)$ is so slow, typically $(\log n)^{-1}$, see Hall (1991), that extreme value confidence bands are of no use in practice, being much too wide, unless the sample size is simply enormous.

The bootstrap, then, is an especially useful tool for the construction of confidence limits or bands for kernel curve estimates, allowing us to empirically estimate the distribution of $\sup_T T(x)$. The convergence of the bootstrap approximation to the distribution of $\sup_T T(x)$ can be shown to be much faster than $(\log n)^{-1}$; for example, Hall (1991) derives the rate of $(nh)^{-1/2}(\log n)^2$ in the case of the kernel density estimator. In Chapter 3, we derive a similar comparison for the kernel intensity estimator.

In the case of non-parametric curve estimation, the algorithm described above produces a confidence interval for $E \hat{\gamma}$ rather than for $\gamma$ itself. Recall that the bootstrap approximates the distribution of $(\hat{\gamma} - \gamma)/\tau(\hat{\gamma})$ by that of $(\hat{\gamma}^* - \hat{\gamma})/\tau(\hat{\gamma}^*)$. However, as $\hat{\gamma}$ is a “linear” function of the data in the sense that $\hat{\gamma} = \sum_{i=1}^n f(X_i)$ for some function $f$, $\hat{\gamma}^* = \sum_{i=1}^n f(X_i^*)$ has the property that $E(\hat{\gamma}^*|\mathcal{X}) = \hat{\gamma}$, so that bias $(\hat{\gamma}^*|\mathcal{X}) = 0$. Note that $\hat{\gamma} - \gamma = \hat{\gamma} - E\hat{\gamma} + \text{bias}(\hat{\gamma})$. Then in estimating the distribution of $\hat{\gamma} - \gamma$ by that of $\hat{\gamma}^* - \hat{\gamma}$, we are estimating bias $(\hat{\gamma})$ by bias $(\hat{\gamma}^*)$. But bias $(\hat{\gamma}^*) = 0$ so that we are effectively setting bias $(\hat{\gamma})$ to zero and estimating $(\hat{\gamma} - E\hat{\gamma})/\tau(\hat{\gamma})$ not $(\hat{\gamma} - \gamma)/\tau(\hat{\gamma})$.

In the remainder of this thesis, we will explicitly recognise that the bootstrap algorithm produces confidence intervals for $E \hat{\gamma}$ by expressing statistics such as $T$ and $T^*$ in the form $T(x) = \{\hat{\gamma}(x) - E\hat{\gamma}(x)\}/\tau\{\hat{\gamma}(x)\}$ and $T^*(x) = \{\hat{\gamma}^*(x) -
E \gamma^*(x) / \tau\{\gamma^*(x)\}. Confidence intervals for E \gamma can be corrected for bias to obtain the required confidence intervals for \gamma.

The simplest method of bias correction is to undersmooth \gamma to such an extent that its bias becomes negligible. Our confidence interval can then be regarded as one for \gamma as well as for E \gamma. We will use this approach in Chapter 3. Another frequently used method is to explicitly correct the confidence interval for bias using an estimate of the asymptotic bias (for example \(-1) \cdot \frac{1}{\gamma} \kappa h^\gamma \hat{f}(x) in the case of density estimation). Bootstrap iteration could also potentially be used to increase the accuracy of the bootstrap estimate in this context; see for example Hall (1992a). However, it is possible that adding another level of bootstrap sampling and calculation of coverage probabilities would increase the computing time above practical limits in the case of confidence bands.

This completes our general overview of the bootstrap, but we will now add some more specific details. In general, the bootstrap resamples \mathcal{X}^* are drawn from \mathcal{X}, and the bootstrap estimate of \theta(F) is \theta(\hat{F}), where \hat{F} is the empirical distribution function. However, it has also been proposed to smooth the sample empirical distribution function and resample from that, thereby avoiding repeated values in the resample \mathcal{X}^*; see for example Silverman and Young (1987). Although Silverman and Young show that in some cases the smoothed bootstrap does improve performance, Hall et al (1989) show that smoothing is really only beneficial when the quantities under study depend in some way on local properties of the underlying distribution \mathcal{F}. The unknown factor in this is how much to smooth. Hall (1992a) suggests that we should not smooth too much; oversmoothing causes a sharp deterioration in performance. In Chapter 3, we use both the smoothed bootstrap and conventional bootstrap resampling.

1.6 Motivation and summary of thesis

In industrial settings, observations may be collected in the form of point events occurring randomly in space or time; such events might be for example the time of accidents, the failure times of machines, or the location of point faults along a
railway track. The analysis of industrial accidents usually aims to determine possible trends in the rate of occurrence of these events: an increase in the occurrence rate would lead to action to introduce new preventative measures. Similarly the analysis of failure times of industrial equipment concentrates on trends in the failure rate. In both cases, the main interest may be in either the global trend of a series of data, or in early detection of an undesirable trend in the most recent data.

In this thesis we shall analyse the intensity of coal mining disasters using the classic data set given by Jarrett (1979) and previously studied by (among others) Maguire et al (1952), Barnard (1953), Cox and Lewis (1966) and Boneva et al (1970). The data set gives the time interval in days between 190 successive severe explosions in mines for the period 15 March 1851 to 22 March 1962. The cumulative number of disasters is plotted against $t$ in Figure 1.7.

A major issue in previous analysis of these data has been the detection of trends in the accident rate. For example, both Barnard (1953) and Cox and Lewis (1966) fit exponential models to the intensity function, thereby showing that it is decreasing. However, in our view, kernel estimation of this intensity, as used in Diggle (1985), Diggle and Marron (1988) and this thesis, is a more flexible tool for preliminary analysis as it shows not only the general decreasing trend, but also reveals peaks at about 8,000 days and 32,000 days (see Figure 2.6, panels C and D, on page 44).
In Chapter 3, we develop methods for calculating confidence bands for the intensity function of an inhomogeneous Poisson process, and apply them to the coal mining disaster data. Using this confidence band, we can not only address the same issue of global trend as is answered by parametric analysis, but also determine whether the additional information about the data revealed by the non-parametric analysis, the apparent peaks, is likely to be real or merely a chimera of the non-parametric methodology.

It is especially important in the case of intensity estimation to adequately adjust for edge effects because the process continues unobserved outside the observation interval and more data might become available at any time. The boundary bias would ideally be no larger than the bias in the interior so that if more recent data became available, the estimate based on the smaller old data set would co-incide with that based on the new enlarged data set. We also note that if early detection of undesirable trends in the most recent data is the main purpose of the analysis, it is essential that the boundary bias be properly adjusted.

In Section 1.3, we described the traditional boundary kernel, reflection and cut-and-rescale methods of reducing boundary bias. Clearly the boundary kernel method is the most effective, reducing the boundary bias from $O(1)$ to $O(h^r)$, the size of the interior bias, whereas the reflection and cut-and-rescale methods reduce the boundary bias to only $O(h)$.

However, the speed and ease of application of an estimator are very important considerations for the applied statistician—complex methods tend not to be used if there are simple methods that are nearly as good. The boundary kernel estimator shares a major drawback with the cut-and-rescale estimator, namely, its implementation is complicated in that a different kernel is constructed at each estimation point in the boundary region. In addition, the computation of that kernel requires solving a system of equations. On the other hand, the reflection estimator is extremely simple to apply: the same kernel is used throughout the estimation interval and that kernel is simple to implement. Clearly then, there is a place for a new estimator which combines the best features of the boundary kernel and reflection estimators: boundary bias of order $O(h^r)$ and applicability over the whole estimation interval.
In Chapter 2, we describe a new method for reducing the boundary bias of kernel estimators of arbitrary order which has these best features. The proposed technique involves generating pseudodata beyond the support of the density $f$, then using the pseudodata as well as the original data in the usual kernel estimator, thereby boundary-correcting the estimator. The pseudodata are simple to generate: they are formed using linear combinations of the order statistics of the sample, with coefficients that vary according to the order of the kernel. We show that the kernel estimators produced in this way have optimal orders of bias and variance throughout the estimation interval. Our method is applied to simulated data and optimal pseudodata generation rules are proposed. We compare the performance of our method with the traditional methods, and find that our method outperforms the traditional optimal boundary kernel method when $f'(0) > 0$. Finally, we compare pseudodata, reflection and boundary kernel intensity estimators on the coal mining disaster data.

In Chapter 3 we develop theory and methods for calculating confidence bands for the intensity function of an inhomogeneous Poisson process. Unlike Diggle (1985) and Diggle and Marron (1988), we treat the Poisson process as genuinely non-stationary rather than as a Cox process. This requires us to derive the asymptotic distribution of $\hat{\lambda}(x) - E\hat{\lambda}(x)$. Our work complements that of Diggle and of Diggle and Marron on point estimation, as well as contributing new results on confidence regions.

Several different bootstrap resampling algorithms are suggested, ranging from resampling from a Poisson process with intensity equal to that estimated non-parametrically from the data, to resampling the data points themselves in much the same way one would use the bootstrap in problems involving independent and identically distributed observations. For each different resampling method a variety of percentile-$t$ ways of constructing confidence bands is described, producing bands whose width varies in proportion to standard deviation, or is approximately constant, depending on taste. Confidence intervals for $E\hat{\lambda}$ are corrected for bias, producing intervals for $\lambda$. Edge effects are removed using our pseudodata estimator.
The effectiveness of our different approaches is demonstrated theoretically, numerically for simulated data, and applied to the coal mining data.

As well as the link formed between Chapters 2 and 3 by applying the methodology developed in Chapter 2 in Chapter 3, there is a theoretical connection through the relationship between order statistics and Poisson processes; see for example Karlin (1966). Table 1.1 shows the parallels between them which are used in this thesis.

The correspondence allows a similar technique to be used in the proofs of Theorems 2.1 and 3.2. We derive the respective distributions of the density (and intensity) estimators by putting the Taylor expansion for \( X(i) \) (and \( X_i \)) around \( F^{-1}(i/n) \) (and \( M^{-1}(i/l) \)) into the definition of \( \hat{f}(x) \) (and \( \hat{\lambda}(x) \)). A second Taylor expansion gives us a deterministic term \( (nh)^{-1} \sum_i K[(x - F^{-1}(i/n))/h] \) (and \( h^{-1} \sum_i K[(x - M^{-1}(i/l))/h] \)) plus stochastic terms. We derive the asymptotic distribution of these statistics.
### Order statistics

**Density** $f(x)$

\[
\text{Distribution function: } U = F(X)
\]

<table>
<thead>
<tr>
<th>$i/n$</th>
<th>$U(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(i)$</td>
<td>$F^{-1}(i/n)$</td>
</tr>
</tbody>
</table>

$U(i)$ has a Beta$(i, n - i + 1)$ distribution

- $E(U(i)) = i/(n + 1) \sim i/n$
- $E(X(i)) \sim F^{-1}(i/n)$

### Poisson processes

**Intensity** $\lambda(x)$

\[
\text{Integrated rate function: } Y = \Lambda(X)
\]

- $Y_i$ has a Gamma$(1, i)$ distribution
- $E(Y_i / i) = i/i$
- $E(X_i) \sim M^{-1}(i/i)$

<table>
<thead>
<tr>
<th>$i/l$</th>
<th>$Y_i/l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>$M^{-1}(i/l)$</td>
</tr>
</tbody>
</table>

### Table 1.1: Parallels between order statistics and Poisson processes
Chapter 2

A pseudodata method of removing boundary bias in kernel density estimation

2.1 Introduction

The classical approach to correcting edge effects in kernel density estimation is to use "boundary" or "edge" kernels towards the ends of the interval of estimation; see Gasser and Müller (1979), Gasser et al (1985) and Müller (1984, 1988, 1991). However, to the extent that the boundary kernel approach requires a different kernel in the boundary regions, it is more complex than the usual method. In this chapter, we suggest a new technique for removing edge effects in kernel estimators of arbitrary order. It is particularly simple to implement, and permits a minor variant of the usual estimator (1.1), to be constructed right up to the ends of the interval.

The reflection procedure for reducing boundary bias in kernel density estimation has been discussed by Schuster (1985), Silverman (1986, p29ff) and Cline and Hart (1991). However, it is not as adaptable as the techniques considered here. In particular, straight data reflection does not adequately correct for a discontinuity in the density and its derivatives such as may occur at the boundaries. Therefore, straight reflection cannot sufficiently adjust the boundary bias in kernel estimators.
of order 2 or more. Our techniques also have significant advantages in related estimation problems where straight reflection is often advocated; see for example the work of Diggle and Marron (1988) on estimating the intensity of an inhomogeneous point process.

Fan (1992) adds kernel weights to the locally linear least squares regression concepts of Stone (1977) and Cleveland (1979). Fan and Gijbels (1992) show that locally linear kernel regression estimators automatically reduce boundary bias to the same order as the interior. Local linear kernel fits can also be applied in density estimation by binning the data then regarding the binned density as a kind of “response” which is equal to the true density plus some random error; see Jones (1989) and Hall and Wand (1994). Cheng (1994) shows that the mean squared error of a local linear kernel density estimator attains the minimax lower bound at the boundaries, and hence no estimator can be better in a minimax sense.


In Section 2.2 we introduce our rules for generating pseudodata outside the interval of estimation. These data are linear functions of the order statistics of the original sample, with weights varying with the order of the kernel. The pseudodata are incorporated into the usual density estimator to produce a boundary-corrected version. Section 2.3 describes the theoretical properties of our technique. In particular, we show that corrected kernel estimators constructed in this way enjoy optimal orders of bias and variance right up to the endpoints of the interval of estimation. The bias reduction property of our method is the key to the applicability of the usual bandwidth selection methods in the present setting. However, we do not develop explicit theory for bandwidth selection, since in view of the results in Section 2.3, that theory closely parallels classical work in the area (see for example Silverman (1986, p48ff) on cross-validation and Sheather and Jones (1991) for plug-in methods), and so is readily anticipated. The numerical properties of our procedure are discussed in Section 2.4. We give guidelines for constructing pseudodata generation
rules, and compare our estimator to traditional boundary bias correction methods. In a simulation study, for densities on $[0, 1]$ with positive slope at $x = 0$, our method was found to have lower mean squared error at $x = 0$ than the boundary kernel method, especially for small sample sizes.

2.2 Methodology

Suppose that a random sample, $X_1, \ldots, X_n$, is drawn from a distribution $F$ with density $f$ on the interval $I = [0, 1]$. Let us assume that $f$ has $r$ bounded, uniformly continuous derivatives on $I$. In such circumstances the density estimator $\hat{f}$ defined in (1.1) would typically use an $r$th order kernel satisfying the conditions in (1.2). In this setting, we shall describe the generation of pseudodata on $(-\infty, 0)$, and the construction of a variant of $\hat{f}$ corrected for edge effects at the origin. Density estimation at the right boundary of $I$ may be treated analogously.

Let $X_{(1)} < \ldots < X_{(n)}$ be the order statistics of the sample $X_1, \ldots, X_n$, and put $X_{(0)} = 0$. For any real number $t \in (0, n)$, define $X_{(t)}$ by linear interpolation among $X_{(0)}, X_{(1)}, \ldots, X_{(n)}$. Let $G = F^{-1}$ be the quantile function of the $X_i$'s. Our pseudodata generation method is based on the relationship (proved in Lemma 2.1) $E X_{(t)} = G(t/n) + O(n^{-1+\delta})$ for all $\delta > 0$, as follows. We estimate an extension of $G$ on $(-\infty, 0)$, denoted $\tilde{G}$, with the property that $\tilde{G}^{(k)}(0-) = G^{(k)}(0+)$ for $0 \leq k \leq r + 1$. We then generate pseudodata, denoted $X_{(-i)}$, using $X_{(-i)} = \tilde{G}(-i/n)$.

We want our boundary correction method to be simple. We therefore estimate the extension of $G$ by fitting a polynomial of degree $d$ with intercept zero to the points on the empirical quantile function $(\frac{A_1}{n}, X_{(A_1)}), \ldots, (\frac{A_s}{n}, X_{(A_s)})$, where $s \geq d$ and $A_1, \ldots, A_s$ are positive constants. This gives

$$X_{(A_1)} = c_1 \frac{A_1}{n} + c_2 (\frac{A_1}{n})^2 + \cdots + c_d (\frac{A_1}{n})^d + \epsilon_1$$
$$\vdots$$
$$X_{(A_s)} = c_1 \frac{A_s}{n} + c_2 (\frac{A_s}{n})^2 + \cdots + c_d (\frac{A_s}{n})^d + \epsilon_s.$$
Now put \( X = [X_{(A^i)}]_n \), \( c = [c_1, \ldots, c_d]^T \), \( \epsilon = [\epsilon_1, \ldots, \epsilon_d]^T \),

\[
A = \begin{bmatrix}
\frac{A_{1i}}{n} & (\frac{A_{1i}}{n})^2 & \ldots & (\frac{A_{1i}}{n})^d \\
\vdots & \vdots & \ddots & \vdots \\
\frac{A_{ni}}{n} & (\frac{A_{ni}}{n})^2 & \ldots & (\frac{A_{ni}}{n})^d 
\end{bmatrix},
\]

\( m = [-\frac{i}{d}, (-\frac{i}{d})^2, \ldots, (-\frac{i}{d})^d]^T \), and \( a = [a_1, \ldots, a_s]^T \). Writing \( X = Ac + \epsilon \) and using ordinary least-squares to estimate \( c \), we get \( \hat{c} = (A^T A)^{-1} A^T X \). We generate the pseudodata \( X_{(-i)} \) using

\[
X_{(-i)} = \hat{c}_1 \left(-\frac{i}{d}\right) + \hat{c}_2 \left(-\frac{i}{d}\right)^2 + \cdots + \hat{c}_d \left(-\frac{i}{d}\right)^d = \hat{c}^T m \\
= X^T A (A^T A)^{-1} m = X^T a = \sum_{j=1}^s a_j X_{(A_j)},
\]

where \( a = A (A^T A)^{-1} m \). Thus, \( A^T a = m \), or \( \sum_{j=1}^s a_j A^T_j = (-1)^i, 1 \leq i \leq d \).

Note that the reflection technique for generating pseudodata is the pseudodata rule with \( d = 1, s = 1 \) and \( A_1 = 1 \). We expect that pseudodata generation rules using higher values of \( d \) and \( s \) will better capture the shape of the quantile function, thereby giving better estimates of \( f \). In general, we will use \( d = r \), and in Section 2.3 we give the distribution of our estimator for both \( d = r \) and \( d = r + 1 \).

More formally, the pseudodata \( X_{(-i)} \) are given by

\[
X_{(-i)} = \sum_{j=1}^s a_j X_{(A_j)}, \quad 1 \leq i \leq [n/\max(A_i)],
\]

(2.1)

where \( a_1, \ldots, a_n \) are real numbers and \( A_1, \ldots, A_n \) are positive real numbers such that

\[
\sum_{j=1}^s a_j A^T_j = (-1)^i, \quad 1 \leq i \leq r.
\]

(2.2)

Obviously, these equations have solutions in \( a_i \) and \( A_i \) if \( s \geq r \).

The case \( r = 2 \) is of greatest practical interest, and there we suggest taking \( s = 2 \) or 3. We give below two examples of (2.1) in this context:

\[
X_{(-i)} = 3(c - 1) X_{(i)} + (1 - 3c) X_{(2i)} + c X_{(3i)}, \quad (2.3)
\]

\[
X_{(-i)} = 3(c - 5) X_{(i/3)} + 3(2 - c) X_{(2i/3)} + c X_{(i)}, \quad (2.4)
\]

for an arbitrary constant \( c \). Setting \( c = 0 \) in (2.3) we obtain a simple two-point pseudodata generation rule,

\[
X_{(-i)} = X_{(2i)} - 3X_{(i)}.
\]
Putting \( c = \frac{2}{3} \) in (2.3), and \( c = \frac{10}{3} \) in (2.4), we obtain three-point rules with coefficients whose absolute values are close to one another:

\[
X_{(-i)} = -X_{(i)} - X_{(2i)} + \frac{2}{3} X_{(3i)} \\
X_{(-i)} = -5X_{(i/3)} - 4X_{(2i/3)} + \frac{10}{3} X_{(i)} .
\]

Our generalized, edge-corrected kernel estimator \( \hat{f} \) is given by

\[
\hat{f}(x) = (nh)^{-1} \left[ \sum_{i=1}^{n} K\{(x - X_{i})/h\} + \sum_{i=1}^{m} K\{(x - X_{(-i)})/h\} \right] , \quad x > 0 ,
\]

where \( m \) denotes an integer of larger order than \( nh \) but smaller order than \( n \). As we shall show in Section 2.3, if \( x > 0 \) is fixed, then for all sufficiently large \( n \), \( \hat{f} = \hat{f} \) where the latter is the classical, uncorrected kernel estimator defined at (1.1). Thus, our edge correction has very little impact on \( \hat{f} \) away from the origin. However, for \( x \) close to zero, the pseudodata \( X_{(-i)} \) contribute to \( \hat{f} \) in a nontrivial way, rendering the orders of its bias and variance equal to \( h^r \) and \( (nh)^{-1} \) respectively — exactly their sizes away from zero. Hence, the estimator \( \hat{f} \) corrects appropriately for edge effects near the origin.

Moreover we shall show that if condition (2.2) extends to \( j = r + 1 \) then not only does \( \hat{f} \) have bias of order \( h^r \), but the bias admits the classical asymptotic formula

\[
(-1)^r \frac{1}{r!} \kappa h^r f^{(r)}(x) \text{ uniformly on } [0, 1] .
\]

As vast numbers of pseudodata generation rules can be produced, in our simulation study we shall examine several factors involved in the construction of pseudodata generation rules. These factors are discussed below, following the eight rules selected for the simulation study:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_{(-i)} = -\frac{5}{3} X_{(i)} + \frac{1}{3} X_{(4i)} ), ( i = 1, \ldots, [n/4] )</td>
<td>2PR1</td>
</tr>
<tr>
<td>( X_{(-i)} = -3X_{(i)} + X_{(2i)} ), ( i = 1, \ldots, [n/2] )</td>
<td>2PR2</td>
</tr>
<tr>
<td>( X_{(-i)} = -9X_{(i/3)} + 2X_{(i)} ), ( i = 1, \ldots, n )</td>
<td>2PR3</td>
</tr>
<tr>
<td>( X_{(-i)} = -24X_{(i/4)} + 10X_{(i/2)} ), ( i = 1, \ldots, n )</td>
<td>2PR4</td>
</tr>
<tr>
<td>( X_{(-i)} = -X_{(i)} - X_{(2i)} + \frac{2}{9} X_{(3i)} ), ( i = 1, \ldots, [n/3] )</td>
<td>3PR1</td>
</tr>
<tr>
<td>( X_{(-i)} = -6X_{(i)} + 4X_{(2i)} - X_{(3i)} ), ( i = 1, \ldots, [n/3] )</td>
<td>3PR2</td>
</tr>
<tr>
<td>( X_{(-i)} = -20X_{(i/2)} + 15X_{(i)} - 4X_{(3i/2)} ), ( i = 1, \ldots, [2n/3] )</td>
<td>3PR3</td>
</tr>
<tr>
<td>( X_{(-i)} = -5X_{(i/3)} - 4X_{(2i/3)} + \frac{10}{3} X_{(i)} ), ( i = 1, \ldots, n )</td>
<td>3PR4</td>
</tr>
</tbody>
</table>
1. The spacing of the $A_j$. Considerations from experimental design lead us to prefer using equally spaced $A_j$. All rules except 2PR1 and 2PR3 use equally spaced $A_j$.

2. The size of max $A_j$. The extension of $G$ should ideally be based only on the portion of $\hat{G}$ that is close to the boundary. Therefore, we should use small rather than large values for the $A_j$. In our simulations, we will compare rules 2PR3, 2PR4 and 3PR3 and 3PR4, which use points close to the boundary to generate $X(-i)$, with the other rules which take data from further from the boundary.

3. The size of min $A_j$. Rules with small $A_j$ have large coefficients $a_j$: rules 2PR4 and 3PR3 have particularly large $a_j$. While the asymptotic bias of the pseudodata estimator does not involve the coefficients of the pseudodata generation rule, the asymptotic variance of the estimator increases with $A_j$ and $a_j$ and we expect to find that the estimator based on these rules will have high variance at the boundaries.

4. Extension of condition (2.2) to $j = r + 1$. All the rules above satisfy the equations (2.2) for $r = 2$. However, 3PR2 and 3PR3 also satisfy these equations for $j = r + 1 = 3$, and therefore have the classical asymptotic bias (1.3) right up to the endpoints of the estimation interval.

### 2.3 Asymptotic theory

We first address the extent to which the pseudodata $X(-i)$ are negative and decreasing in $i$. For small values of $i/n$, Corollary 2.1 gives

$$E X(-i) \sim -(i/n) f(0)^{-1}.$$

Thus on average, and for small $i/n$, the pseudodata lie to the left of the origin and become successively more negative as $i$ increases. Stochastic variability can mean that this result fails to be strictly correct for $X(-i)$, rather than for the average value of $X(i)$, although the numerical work in Section 2.4, and the theory developed
in the present section, show that these random fluctuations do not cause practical difficulties.

We cannot necessarily expect even the expected value of $X(-i)$ to be negative if $i$ is large. To appreciate why, let $q_{\alpha}$ denote the $\alpha$-level quantile of the distribution of $X$, and observe that for any $\epsilon > 0$ we may choose $F$ such that $q_{2\alpha} - 3q_{\alpha} > 0$ for $\alpha$'s in an open interval $J \subseteq (0, \epsilon)$. Therefore, if $X(-i) = X(2i) - 3X(i)$ is the two-point rule suggested in Section 2.2 then we expect $X(-i)$ to be positive if $i/n \in J$. Similar examples may be described for more general rules.

For these reasons, the integer $m = m(n)$ employed in the definition of $\hat{f}$ should be chosen to be of smaller order than $n$. We also ask that $m$ be of larger order than $nh$; this ensures that $\hat{f}$ uses all of the first $O(nh)$ values of pseudodata, these being the ones that are most important to edge correction. When $K$ is an $r$th order kernel it is optimal to take $h$ to be of size $n^{-1/(2r+1)}$, and so we ask that $m$ be of larger order than $n^{2r/(2r+1)}$ but of smaller order than $n^{(2r/(2r+1))\eta}$, for some $\eta \in (0, 1/(2r + 1))$.

Our main result describing the performance of $\hat{f}$ is given in the following theorem:

**Theorem 2.1** Let $K$ be a compactly supported $r$th order kernel with two bounded derivatives. Define $\kappa = \int u^r K(u)du$. Suppose $f$ is supported on $[0, 1]$, has $r$ bounded, Hölder continuous derivatives on $[0, 1]$, and (for simplicity of exposition) $f(0) > 0$. Let $h \sim \text{const.} n^{-1/(2r+1)}$, and let $m$ be an order of magnitude larger than $n^{2r/(2r+1)}$, and an order smaller than $n^{(2r/(2r+1))\eta}$, where $0 \leq \eta \leq 1/(2r + 1)$. Then if $\eta$ is sufficiently small (depending on the index of Hölder continuity), we find that

$$
\hat{f}(x) = f(x) + h^r \left\{ (-1)^r \frac{1}{r!} \kappa f^{(r)}(x) + \gamma(x) \right\} + (nh)^{-1/2} \sigma(x) N_n(x) \\
+ o_p\left\{ h^r + (nh)^{-1/2} \right\}
$$

(2.5)

uniformly in $0 \leq x \leq (1 - \varepsilon')$ for arbitrary fixed $\varepsilon' > 0$, where $E(N_n) = 0, E(N_n^2) = 1, N_n(x)$ is asymptotically Normal $N(0, 1)$ for any sequence $x = x(n) \in [0, 1 - \varepsilon']$,

$$
\gamma(x) = C \int_{x/h}^{\infty} (z - xh^{-1})^{r+1} K'(z)dz,
$$

$$
C = \{(r + 1)!\}^{-1} f(0)^{r+2} \{(d/dy)^{r+1} F^{-1}(y)\}_{y=0} \left\{ \sum_{j=1}^{\infty} a_j A_j^{r+1} - (-1)^{r+1} \right\},
$$

$$
\sigma(x)^2 = f(x) \left( \int K^2 \right) + \tau(x),
$$
\[ \tau(x) = f(0) \left[ \sum_{j_1=1}^{s} \sum_{j_2=1}^{s} a_{j_1} a_{j_2} \int_{x/h}^{\infty} \int_{x/h}^{\infty} \min \{ A_{j_1}(u_1 - xh^{-1}), A_{j_2}(u_2 - xh^{-1}) \} \right. \\
\times K'(u_1) K'(u_2) \, du_1 \, du_2 + 2 \sum_{j=1}^{s} a_j \int_{x/h}^{\infty} \, du_1 \int_{-\infty}^{x/h} \min \{ A_j(u_1 - xh^{-1}), (xh^{-1} - u_2) \} K'(u_1) K'(u_2) \, du_2 \].

To interpret these results, first note that because of the compact support of \( K \), \( \gamma \) and \( \tau \) both vanish on \([ch, \infty)\) for some \( c > 0 \). Therefore, if \( x > ch \) then (2.5) reduces to the classical asymptotic expression for bias and error-about-the-mean of a kernel density estimator

\[ \hat{f}(x) = f(x) + (-1)^r \frac{1}{r!} \kappa h^r f^{(r)}(x) + (nh)^{-1/2} \{ f(x) \left( \int K^2 \right) \}^{1/2} N_n(x) + o_f(h^r + (nh)^{-1/2}), \]

where \( N_n(x) \) has zero mean, unit variance and is asymptotically Normal \( N(0,1) \).

Both \( \gamma \) and \( \tau \) are uniformly bounded on \([0, \infty)\), and so (2.5) shows that \( \hat{f} \) has bias and variance of orders \( h^r \) and \( (nh)^{-1} \) respectively, even arbitrarily close to the origin. Thus \( \hat{f} \) has effectively coped with edge effects.

If the constants \( a_j \) and \( A_j \) are chosen so that

\[ \sum_{j=1}^{s} a_j A_j^{r+1} = (-1)^{r+1}, \]

or equivalently if the constraint (2.2) extends to \( j = r + 1 \), then \( C = 0 \) and so \( \gamma \equiv 0 \) on \([0, \infty)\). In this case, the asymptotic bias of \( \hat{f} \) admits the classical form \((-1)^r \frac{1}{r!} \kappa h^r f^{(r)}(x)\), even for \( x = 0 \).
Figure 2.1: Factors influencing sign of $X_{(i)}$. Two $G(u)$ curves: curve through diamonds has $G'' > 0$, curve through triangles has $G'' < 0$. A pseudodatum $X_{(-i)} = \tilde{G}(-i/n)$ is negative if $G'' < 0$ ($f'(0) > 0$) but may be positive for large $i$ if $G'' > 0$ ($f'(0) < 0$) and the axis of $\tilde{G}$ is close to $u = 0$.

2.4 Numerical results

2.4.1 Location of pseudodata

To consider in more detail the circumstances in which a pseudodatum may be positive, we shall focus on the case $r = 2$, which is generally used in practice, and let $s = 2$ or 3. When $r = 2$, our pseudodatum method is equivalent to fitting a parabola through the origin to 2 or 3 points on $\tilde{G}$, the empirical quantile function, and taking $X_{(-i)}$ to be $\tilde{G}(-i/n)$, where $\tilde{G}(u) = \hat{c}_1 u + \hat{c}_2 u^2$.

If $\hat{c}_2 < 0$, the parabola is "head up", and the axis of the parabola will be to the right of the origin. In this case, we will always obtain negative pseudodata. However, if $\hat{c}_2 > 0$ and the axis of the parabola, $u = -\hat{c}_1/(2\hat{c}_2)$, is close to the vertical axis, then although we will obtain negative pseudodata for small $i$, we will obtain positive values for large values of $i$. These points are illustrated in Figure 2.1.

Equating the coefficients $\hat{c}_i$ in the expression for $\tilde{G}(u)$ above with those in the Taylor expansion for $G$ around $u = 0$, we have $\hat{c}_1 = G'(0)$ and $\hat{c}_2 = G''(0)/2$. Thus we may generate positive pseudodata if $u = -G'(0)/G''(0) = f^2(0)/f'(0)$ is small.
Figure 2.2: Empirical quantile function $\tilde{G}$ and extensions at both boundaries for 2000 observations from $f(x) = \left\{ \frac{2}{3} \sin(2\pi x) + 1 \right\} I_{[0, 1]}(x)$. The pseudodata were generated using $X_{(-i)} = -9X_{(i/3)} + 2X_{(i)}$. The horizontal axis is $i/n$, and the vertical axis is $x(i)$. Thus if $f'(0) < 0$, our method may generate positive pseudodata if $f'(0)$ is large and/or $f(0)$ is small.

In Figure 2.2, we illustrate the empirical quantile function $\tilde{G}$ for 2000 observations from $f(x) = \left\{ \frac{2}{3} \sin(2\pi x) + 1 \right\} I_{[0, 1]}(x)$, and its extensions at both boundaries formed by pseudodata generated using $X_{(-i)} = -9X_{(i/3)} + 2X_{(i)}$. We chose this density since $f'(0) > 0$ at the left boundary and $f'(0) < 0$ at the right boundary. For $0 \leq i/n \leq 1$, the graph shows $\tilde{G}$; for $-1 \leq i/n \leq 0$, the graph shows the extension beyond the left boundary; and for $1 \leq i/n \leq 2$, we see the extension beyond the right boundary. The horizontal lines at $X_{(i)} = 0$ and $X_{(i)} = 1$ indicate the bounds of the estimation interval, between which we do not want pseudodata to lie.

As is typically the case when $f'(0) > 0$, the pseudodata generated at the left boundary are negative and decreasing in $i$ for at least a bandwidth beyond the boundary. However, it was observed that when $f'(0) < 0$, as is the case at the right boundary, the pseudodata do not always increase in $i$ for a bandwidth in all samples before the turning point in the pseudodata path. This occurrence is more common in small data sets. In these cases, the best ad hoc but simple solution was found to be to truncate the pseudodata series at the turning point, indicated...
by a dot in the figure, and to repeat the series until the pseudodata extended for one bandwidth beyond the boundary. This solution is sufficiently efficient for the pseudodata estimator to outperform the reflection estimator for large values of \( f'(0) \), but it is not sufficiently efficient for our estimator to consistently outperform the boundary kernel estimator when \( f'(0) < 0 \). Any pseudodata values which fell in \((0, 1)\) were excluded from \( f(x) \).

### 2.4.2 Simulation study

We undertook a simulation study in order to empirically evaluate the factors surrounding pseudodata rule construction given on page 32, using the eight pseudodata generation rules given there. Following the selection of the optimal pseudodata rule, we compared the boundary performance of the pseudodata, boundary kernel, reflection, cut-and-rescale and uncorrected estimators.

In a complete study of boundary performance, we would compare the ability of the various estimators to reproduce the true values of \( f(0) \), \( f'(0) \) and \( f''(0) \), using densities chosen to vary over each of these factors at a number of levels. However for compactness of exposition, in our simulation study we compared the performance of the estimators over a range of values of \( f'(0) \) only. As discussed in Section 2.4.1, the location of the pseudodata is strongly affected by the magnitude and sign of \( f'(0) \), and so the study should therefore expose any weaknesses of the pseudodata estimator.

The four truncated Normal mixture densities listed below and shown in Figure 2.3 on page 39 were used for the study. They were chosen because they have values of \( f'(0) \) of approximately \(-20, -10, -5, -1, 1, 5, 10, \) and \( 20 \). The actual values of \( f'(0) \) for these densities are given in Table 2.1. Slopes at the right boundary are reported as slopes at the left boundary after reflecting the density through \( x = 0.5 \).

\[
f(x) = \left\{ \begin{array}{ll}
\frac{1}{6} c_1 N\left( \frac{1}{5}, \frac{1}{5} \right) + \frac{5}{6} c_2 N\left( \frac{4}{5}, \frac{1}{5} \right) \right\} I_{[0,1]}(x) \\
\frac{1}{2} c_3 N\left( \frac{1}{10}, \frac{1}{15} \right) + \frac{1}{2} c_4 N\left( \frac{7}{10}, \frac{1}{10} \right) \right\} I_{[0,1]}(x) \\
\frac{1}{3} c_5 N\left( -\frac{2}{5}, \frac{1}{4} \right) + \frac{2}{3} c_6 N\left( \frac{3}{2}, \frac{12}{2} \right) \right\} I_{[0,1]}(x) \\
\frac{1}{2} c_7 N\left( -1, \frac{1}{2} \right) + \frac{1}{2} c_8 N\left( \frac{5}{4}, \frac{12}{2} \right) \right\} I_{[0,1]}(x),
\end{array} \right.
\]
where, $c_1 = \{\Phi(4) - \Phi(-1)\}^{-1}$, $c_2 = \{\Phi(1) - \Phi(-4)\}^{-1}$, $c_3 = \{\Phi(\frac{37}{4}) - \Phi(-\frac{3}{2})\}^{-1}$, $c_4 = \{\Phi(\frac{3}{4}) - \Phi(-\frac{33}{4})\}^{-1}$, $c_5 = \{\Phi(\frac{28}{5}) - \Phi(\frac{3}{2})\}^{-1}$, $c_6 = \{\Phi(-1) - \Phi(-3)\}^{-1}$, $c_7 = \{\Phi(4) - \Phi(2)\}^{-1}$ and $c_8 = \{\Phi(-\frac{1}{2}) - \Phi(-\frac{5}{3})\}^{-1}$.

Using $n = 200$ and $n = 2000$ observations per sample, we drew 500 samples from each density. Comparisons of boundary behaviour were made using the root mean squared error, bias, and standard deviation of $\hat{f}(0)$, and the mean size of the angle between the true and estimated densities at $x = 0$. Note that results at $x = 1$ were reported as results at $x = 0$.

The boundary performance of the eight pseudodata generation rules is shown in Tables 2.2 to 2.5 starting on page 51. Based on these tables, we give the following guidelines for constructing pseudodata generation rules:

1. We suggest using $\max(A_j) = 1$ in (2.1). Rules 2PR1, 2PR2, 3PR1 and 3PR2 with larger $\max(A_j)$ had high bias, especially when $f'(0) > 0$. It seems that estimation of $f(0)$ is based on points too far from the boundary. Rules with $\max(A_j) < 1$ have high coefficients $a_j$ and so they produced pseudodata paths with relatively high stochastic variability, resulting in truncation problems. Also, the rules with large $a_j$, 2PR4 and 3PR3, had large negative biases when $f'(0) < 0$.

2. Rules 3PR2 and 3PR3 obey the constraint (2.2) for $j = r + 1$ and $r = 2$, and therefore have the classical asymptotic bias (1.3) throughout the boundary region. However, in our simulations, we observed no reduction in bias when using these rules, even in samples as large as 2000 observations. It appears
3. The best three-point rule, 3FBA, was found to be marginally better than the best two-point rule, 2FBA. It has lower MSE and better reproduced the slope of the power in the boundary region (measured by the angle between \( f'(0) \) and \( \mathbf{f}'(0) \)).

The non-connected percentile rule is 3FBA.

\[
\begin{align*}
(2.6) & \quad f(x) = \left\{ \frac{1}{2} c_1 \mathcal{N}\left( \frac{x}{\lambda}, \frac{1}{\lambda^2} \right) + \frac{1}{2} c_2 \mathcal{N}\left( \frac{x}{\lambda}, \frac{1}{\lambda^2} \right) \right\} f_{0.1}(x) \\
(2.7) & \quad f(x) = \left\{ \frac{1}{2} c_3 \mathcal{N}\left( \frac{x}{\mu}, \frac{1}{\mu^2} \right) + \frac{1}{2} c_4 \mathcal{N}\left( \frac{x}{\mu}, \frac{1}{\mu^2} \right) \right\} f_{0.1}(x)
\end{align*}
\]

\[
\begin{align*}
(2.8) & \quad f(x) = \left\{ \frac{1}{2} c_5 \mathcal{N}\left( -\frac{x}{\beta}, \frac{1}{\beta^2} \right) + \frac{1}{4} c_6 \mathcal{N}\left( \frac{x}{\beta}, \frac{1}{\beta^2} \right) \right\} f_{0.1}(x) \\
(2.9) & \quad f(x) = \left\{ \frac{1}{2} c_7 \mathcal{N}\left( -1, \frac{1}{\lambda^2} \right) + \frac{1}{2} c_8 \mathcal{N}\left( 1, \frac{1}{\lambda^2} \right) \right\} f_{0.1}(x)
\end{align*}
\]

Figure 2.3: Densities used in simulation study
that larger values of $n$ would be needed for this benefit to become evident.

3. The best three-point rule, 3PR4, was found to be marginally better than the best two-point rule, 2PR3: it had lower MSE and better reproduced the shape of the curve in the boundary region (measured by the angle between $f'(0)$ and $\hat{f}'(0)$).

The recommended pseudodata rule is 3PR4:

$$X_{(-i)} = -5X_{(i/3)} - 4X_{(2i/3)} + \frac{10}{3}X_{(i)}, \quad i = 1, \ldots, n.$$ 

The boundary performance of the pseudodata estimator (using 3PR4) and the uncorrected, boundary kernel, cut-and-rescale and reflection estimators is shown in Tables 2.6 and 2.7.

As expected, in general, the pseudodata estimator has a lower MSE than the reflection estimator because it has lower absolute bias for both large and small sample sizes. Note that for small negative slopes, our truncation method does not perform well enough for the pseudodata estimator to outperform reflection. More unexpectedly, we found that when $f'(0) > 0$, the pseudodata estimator had lower MSE than the boundary kernel estimator, especially in small samples. In this case, the boundary kernel estimator had very high variance: it appears to be more sample dependent than the other estimators.

Figures 2.4 and 2.5 show the estimators for 10 samples from the densities (2.7) and (2.8) using $n = 2000$ observations. In Figure 2.4 we see that for densities with $f'(0) > 0$, reflection and cut-and-rescale have high bias but low variance at the boundaries. Both boundary kernel and pseudodata methods are comparatively unbiased, but the boundary kernel method has a higher variance at the boundaries. Figure 2.5 shows that when $f'(0) < 0$, reflection and cut-and-rescale have high bias but low variance at the boundaries. The boundary kernel and pseudodata methods are relatively unbiased, but the pseudodata method has a much higher variance at the boundary than the boundary kernel method.

To estimate a density known to increase or decrease at the boundary, for best performance, the analyst would use boundary kernel edge correction when it decreases and pseudodata correction when it increases.
Figure 2.4: Comparison of boundary correction methods for $f'(0) > 0$. Estimates for 10 samples from density (2.7), $n = 2000$. True density shown by solid line and estimates by dotted lines.
Figure 2.5: Comparison of boundary correction methods for $f'(0) < 0$. Estimates for 10 samples from density (2.8), $n = 2000$. True density shown by solid line and estimates by dotted lines.
2.4.3 Analysis of coal mining data

Our motivation for treating the boundary bias problem was provided by the classic coal mining disaster data set introduced in Section 1.6. Attempts by Diggle (1985) and Diggle and Marron (1988) to estimate the intensity of the underlying point process using kernel methods employed cut-and-rescale and reflection respectively to correct for edge effects at the ends of the observation interval. Particularly at the left-hand end, where the intensity is relatively high, it is of interest to know whether the intensity is increasing or decreasing, but reflection cannot provide an answer. The aim of Chapter 3 is to develop simultaneous confidence bands for the unknown intensity, and for that purpose behaviour at the ends is critical.

Of course, since data beyond the interval ends are not available, the performance of the boundary bias correction methods cannot be evaluated with certainty on the original data set. Therefore we divided the data into exactly half, and applied boundary kernel, reflection and our pseudodata rules edge correction to either end of the two sub-samples obtained in this way, and compared the performance of the estimators at the breakpoint. We generated the pseudodata using the three-point rule

$$X(-i) = -5X(i/3) - 4X(2i/3) + \frac{10}{3}X(i), \quad i = 1, \ldots, n,$$

which was the best of the rules tested in our simulation study in Section 2.4.2, and used cross validation for bandwidth selection.

Figure 2.6 shows the result of applying the different methods of edge correction to both ends of the two half data sets, and to the complete data set. It is clear that at the breakpoint, the pseudodata method of boundary adjustment gives the smallest difference between the two half data set estimates; and also, the half data set estimates are closest to the full data set estimate. Reflection performed poorly at the breakpoint, where it failed to take account of the direction in which the intensity is changing. The boundary kernel method did not perform well using either the bandwidths selected by cross validation, or the same bandwidths employed for the pseudodata method in D. Note that in the case of the boundary kernel methods, \(\hat{f}(0)\) is highly bandwidth-dependent — compare panels A and B of Figure 2.6.
A: Boundary kernel

B: Boundary kernel

C: Reflection

D: Pseudodata

Figure 2.6: Intensity of coal mining disasters: comparison of boundary kernel, reflection and pseudodata estimates. Solid line is estimate using complete data set. Dotted lines are estimates for the two half data sets in the intervals [0,10539] and [10539,40549]. Bandwidths were chosen using cross validation in A, C and D. The same bandwidths were used in B as in D.
2.5 Proof of Theorem 2.1

Let $F$ be the common distribution function of the variables $X_i$, put $U_i = F(X_i)$, and write $U_{(1)}, \ldots, U_{(n)}$ for the corresponding order statistics. Since $f > 0$ on $(0, \epsilon_1)$ for some $0 < \epsilon_1 < 1$, then $G = F^{-1}$ is well-defined and strictly increasing on $(0, \epsilon_1)$.

Define $G$ on $(-\infty, 0)$ to be the polynomial of degree $r + 1$ such that $G^{(k)}(0-) = G^{(k)}(0+)$ for $0 \leq k \leq r + 1$. By Lemma 2.1, for all $\delta > 0$,

$$X(i) = G(i/n) + (U(i) - E U(i)) G'(i/n) + O(n^{-1+\delta})$$

uniformly in $1 \leq i \leq n$, with probability one. Therefore the random variable $X_{(-i)}$ defined by (2.1) satisfies

$$X_{(-i)} = \nu_i + \delta_i + O(n^{-1+\delta})$$

uniformly in $1 \leq i \leq \epsilon_2 n$, where $\epsilon_2 = (\max A_j)^{-1}$,

$$\nu_i = \sum_{j=1}^s a_j G(A_j i/n), \quad \text{and} \quad Y_i = \sum_{j=1}^s a_j(U(A_ji) - E U(A_j i)) G'(A_j i/n).$$

Now, using Taylor expansion and the Hölder continuity of $G^{(r+1)}$,

$$\nu_i = \sum_{j=1}^s a_j \left[ \sum_{k=0}^{r+1} \frac{1}{k!} (A_j i/n)^k G^{(k)}(0) - \{(r + 1)\}^{-1} (A_j i/n)^{r+1} G^{(r+1)}(\theta A_j i/n) \right], \quad 0 \leq \theta \leq 1$$

$$= \sum_{j=1}^s \sum_{k=1}^{r+1} \frac{1}{k!} a_j (A_j i/n)^k G^{(k)}(0) + O\{(i/n)^{r+1+\xi}\},$$

where $\xi$ is the Hölder exponent of $G^{(r+1)}$. In view of (2.2),

$$\nu_i = \sum_{k=1}^{r+1} \frac{1}{k!} (-i/n)^k G^{(k)}(0) - \{(r + 1)\}^{-1} (-i/n)^{r+1} G^{(r+1)}(0)$$

$$+ \{(r + 1)\}^{-1} \sum_{j=1}^s a_j A_j^{r+1}(i/n)^{r+1} G^{(r+1)}(0) + O\{(i/n)^{r+1+\xi}\}$$

$$= G(-i/n) + \delta_i + O\{(i/n)^{r+1+\xi}\}.$$  (2.11)

uniformly in $1 \leq i \leq \epsilon_3 n$, where $\delta_i = \{(r + 1)\}^{-1} C_1 (i/n)^{r+1} G^{(r+1)}(0)$ and $C_1 = \Sigma_{1 \leq j \leq s} a_j A_j^{r+1} - (-1)^{r+1}$. 

45
Combining (2.10) and (2.11) we deduce that

\[ X_{(-i)} = G(-i/n) + \delta_i + Y_i + O\{n^{-1+\delta} + (i/n)^{r+1+\xi}\} \]

uniformly in \(1 \leq i \leq \epsilon_an\), with probability one for all \(\delta > 0\) and some \(\xi > 0\). From this result and the facts that \(|G(-i/n)| = O(i/n)|\), \(|\delta_i| = O\{(i/n)^{r+1}\}\), \(\sup_i|Y_i| = O(n^{-1+\delta/2})\) for all \(\delta > 0\) with probability one, and for some \(\delta > 0\), \(m/n^{1+\delta/2} \to \infty\), we deduce that \(\sup_{1 \leq i \leq m} |X_{(-i)}| = O(mn^{-1})\) with probability one. Hence if \(nx_0/m \to \infty\) then, since \(K\) is compactly supported, \(\hat{f}(x) = \hat{f}(x)\) for all \(x \geq x_0\), for all sufficiently large \(n\).

More generally, by Taylor expansion,

\[
(nh)^{-1} \sum_{i=1}^{m} K\{(x - X_{(-i)})/h\}
\]

\[
= (nh)^{-1} \sum_{i=1}^{m} K\left\{ \left( x - [G(-i/n) + \delta_i + Y_i + O\{n^{-1+\delta} + (i/n)^{r+1+\xi}\}] \right)/h \right\}
\]

\[
= (nh)^{-1} \sum_{i=1}^{m} \left\{ \left( x - G(-i/n) \right)/h \right\} - (nh^2)^{-1} \sum_{i=1}^{m} \left( \delta_i + Y_i \right) K''\left\{ \left( x - G(-i/n) \right)/h \right\}
\]

\[+ O\{mn^{-2+\delta}h^{-2} + (m/n)^{r+2+\xi}h^{-2} + (m/n)^{2r}h^{-3}\} \quad (2.12) \]

uniformly in \(-\infty < x < \infty\), with probability one. Similarly but more simply,

\[
\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_{(i)})/h\}
\]

\[
= (nh)^{-1} \sum_{i=1}^{n} K\left\{ \left( x - G(i/n) \right)/h \right\}
\]

\[- (nh^2)^{-1} \sum_{i=1}^{n} \left( U_{(i)} - E U_{(i)} \right) G'(i/n) K''\left\{ \left( x - G(i/n) \right)/h \right\} + O(n^{-1+\delta}h^{-2}) \quad (2.13) \]

uniformly in \(x\), with probability one. Adding (2.12) and (2.13) we obtain

\[
\hat{f}(x) = \mu_1(x) - \mu_2(x) - \Delta(x) + O\{n^{-1+\delta}h^{-2} + mn^{-2+\delta}h^{-2}
\]

\[+ (m/n)^{r+2+\xi}h^{-2} + (m/n)^{r+3}h^{-3}\} \quad (2.14) \]

uniformly in \(x\), with probability one, where

\[
\mu_1(x) = (nh)^{-1} \sum_{i=-m}^{n} K\left\{ \left( x - G(i/n) \right)/h \right\},
\]

46
\[
\mu_2(x) = (nh^2)^{-1} \sum_{i=1}^{m} \delta_i K'\left[\{x - G(-i/n)\}/h\right],
\]
\[
\Delta(x) = (nh^2)^{-1} \sum_{i=-m}^{n} Z_i K'\left[\{x - G(i/n)\}/h\right],
\]
\[
Z_i = (U_{(i)} - E U_{(i)}) G'(i/n) \text{ if } i \geq 1, \quad Z_i = Y_{-i} \text{ if } i \leq -1, \text{ and } \sum' \text{ denotes summation over nonzero indices.}
\]

Approximating series by integrals, and employing Taylor expansion, we may show that with \(C_2 = \{(r + 1)!\}^{-1} C_1 G^{(r+1)}(0)\) we have

\[
\mu_1(x) = (nh)^{-1} \int_{-\infty}^{\infty} K[\{x - G(y/n)\}/h] \, dy + O\{(nh)^{-1}\}
\]
\[
= \int_{-\infty}^{\infty} K(u) f(x - hu) \, du + O\{(nh)^{-1}\}
\]
\[
= f(x) + (-1)^r \frac{1}{r!} h^r \kappa f^{(r)}(x) + o(h^r) + O\{(nh)^{-1}\}
\]
\[
\mu_2(x) = C_2 (nh^2)^{-1} \int_{1}^{m} (y/n)^{r+1} K'[\{x - G(-y/n)\}/h] \, dy + O\{(nh)^{-1}\}
\]
\[
= C_2 (nh^2)^{-1} \int_{1}^{m} (y/n)^{r+1} K'[\{x + G'(0) y/n\}/h] \, dy + O\{(nh)^{-1}\}
\]
\[
= C_2 h^r G'(0)^{-1} \int_{x/h}^{\infty} (z - xh^{-1})^{r+1} K'(z) \, dz + o(h^r) + O\{(nh)^{-1}\}.
\]

Using Rényi’s representation of order statistics (David (1981, page 21)), we may show that \(\Delta\) is asymptotically Normal \(N(0, \text{E} \Delta^2)\). If \(x \in (0,1)\) is fixed then

\[
E\{\Delta(x)^2\} \sim (nh)^{-1} f(x) \int K^2. \text{ If } x = x(n) \in [0,1) \text{ satisfies } x \to 0 \text{ as } n \to \infty \text{ then } E\{\Delta(x)^2\} \sim (nh)^{-1} f(x) \int K^2 + v(x), \text{ where, with } V_i = U_{(i)} - E U_{(i)},
\]

\[
v(x) = (nh^2)^{-2} \left\{ E \left( \sum_{i=1}^{m} \sum_{j=1}^{s} a_j V_{A_{ij}} G'(A_{ij}/n) K'[\{x - G(-i/n)\}/h] \right)^2 \right. \\
+ 2 E \left( \sum_{i=1}^{m} \sum_{j=1}^{s} a_j V_{A_{ij}} G'(A_{ij}/n) K'[\{x - G(-i/n)\}/h] \right) \\
\left. \times \left( \sum_{i=1}^{m} V_{i} G'(i/n) K'[\{x - G(i/n)\}/h] \right) \right\}
\]
\[
\sim (nh^2)^{-2} G'(0)^2 \left\{ \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} a_{j_1} a_{j_2} \int_{1}^{m} \int_{1}^{m} E (V_{A_{j_1 y_1}} V_{A_{j_2 y_2}}) \\
\times K'[\{x - G(-y_1/n)\}/h] K'[\{x - G(-y_2/n)\}/h] \, dy_1 \, dy_2 \right. \\
+ 2 \sum_{j=1}^{m} a_j \int_{1}^{m} \int_{1}^{m} E (V_{A_{j y_1}} V_{A_{j y_2}}) K'[\{x - G(-y_1/n)\}/h] \\
\left. \times K'[\{x - G(y_2/n)\}/h] \, dy_1 \, dy_2 \right\}
\]

47
\[
(\eta^2 - 2) G'(0)^2 \left\{ \sum_{j_1=1}^s \sum_{j_2=1}^s a_{j_1} a_{j_2} \int_1^m \int_1^m E(V_{A_j y_1} V_{A_j y_2}) \times K'\{x + G'(0) y_1/n\}/h \right. \\
\left. \int_1^m \int_1^m E(V_{A_j y_1} V_{y_2}) K'\{x + G'(0) y_1/n\}/h \right] dy_1 dy_2 \\
+ 2 \sum_{j=1}^s a_j \int_1^m \int_1^m E(V_{A_j y_1} V_{y_2}) K'\{x + G'(0) y_1/n\}/h \right] dy_1 dy_2 \\
\times K'\{x - G'(0) y_2/n\}/h \right] dy_1 dy_2 \right\} \\
\sim (\eta h)^{-1} G^{-1}(0) \left\{ \sum_{j_1=1}^s \sum_{j_2=1}^s a_{j_1} a_{j_2} \int_1^m \int_1^m \min \{A_{j_1}(u_1 - x h^{-1}) , A_{j_2}(u_2 - x h^{-1})\} K'(u_1) K'(u_2) du_1 du_2 \\
+ 2 \sum_{j=1}^s a_j \int_1^m \int_1^m \min \{A_{j}(u_1 - x h^{-1}) , (x h^{-1} - u_2)\} K'(u_1) K'(u_2) du_2 \right\}. \\
\]

If \( h \) is of size \( n^{-1/(2r+1)} \), and \( m \) is of larger order than \( n^{2r/(2r+1)} \) but of smaller order than \( n^{(2r/(2r+1)) + \eta} \), then, choosing \( \delta \) and \( \eta \) sufficiently small we may deduce (2.5) from the results from (2.14) down.
Appendix 2A.1

Lemmas

Lemma 2.1 is the result upon which our pseudodata method is based. Since \( U(i) \sim \text{Beta}(i, n - i + 1) \), it follows that \( \mathbb{E} U(i) = i/(n + 1) = i/n + O(n^{-1}) \), and \( \text{var} U(i) = i(n - i + 1)/\{(n + 2)(n + 1)^2\} = O(n^{-1}) \).

**Lemma 2.1** Using the notation and definitions of Theorem 2.1, for all \( \delta > 0 \),

\[
X(i) = G(i/n) + (U(i) - \mathbb{E} U(i)) G'(i/n) + O(n^{-1+\delta})
\]

uniformly in \( 1 \leq i \leq n \), with probability one.

**Proof** Since \( X(i) = G(U(i)) \), Taylor expansion about \( \mathbb{E} U(i) \) gives

\[
X(i) = G(\mathbb{E} U(i)) + (U(i) - \mathbb{E} U(i)) G'\{\mathbb{E} U(i) + \theta(U(i) - \mathbb{E} U(i))\}, \quad 0 \leq \theta \leq 1.
\]

We first show that \( \sup_{1 \leq i \leq n} |U(i) - \mathbb{E} U(i)| = O(n^{-\delta/2}) \) with probability one. Using Markov's inequality, we have

\[
P\left( \sup_{1 \leq i \leq n} |U(i) - \mathbb{E} U(i)| > t \right) \leq \sum_{i=1}^{n} P(|U(i) - \mathbb{E} U(i)| > t) \leq t^{-2k} \sum_{i=1}^{n} \mathbb{E} \left(|U(i) - \mathbb{E} U(i)|^{2k}\right) \leq t^{-2k} \sum_{i=1}^{n} C n^{-k} = O(t^{-2k} n^{-k+1}).
\]

Put \( t = C n^{-\delta/2} \). Then

\[
P\left( \sup_{1 \leq i \leq n} |U(i) - \mathbb{E} U(i)| > C n^{-\delta/2}\right) = O(n^{-k\delta+1}) = O(n^{-\lambda}),
\]
since for all $\delta > 0$ we can choose $k$ so that $-k\delta + 1 < 0$. Thus for all $\lambda > 0$, since 
$\sum P(\sup_{1 \leq i \leq n} |U(i) - E U(i)| > C n^{(-1+\delta)/2}) < \infty$, the Borel-Cantelli Lemma gives 
$\sup_{1 \leq i \leq n} |U(i) - E U(i)| = O(n^{(-1+\delta)/2})$ a.s. Therefore, since $\sup_{1 \leq i \leq n} (U(i) - E U(i))^2 = O(n^{-1+\delta})$ with probability one, and $G''$ is bounded by assumption, 

\[(U(i) - E U(i))^2 G'' \{ E U(i) + \theta(U(i) - E U(i)) \} = O(n^{-1+\delta})\]

uniformly in $i$ with probability one.

Finally, from (A.2), the above result, and using $E U(i) = i/n + O(n^{-1})$, we have

\[X(i) = G(E U(i)) + (U(i) - E U(i)) G'(E U(i)) + O(n^{-1+\delta}) = \{ G(i/n) + O(n^{-1}) \} + (U(i) - E U(i)) \{ G'(i/n) + O(n^{-1}) \} + O(n^{-1+\delta}) = G(i/n) + (U(i) - E U(i)) G'(i/n) + O(n^{-1+\delta}), \]

proving the lemma.

**Corollary 2.1** For $1 \leq i \leq \epsilon n$ where $0 < \epsilon < 1$, 

\[E X_{(-i)} \sim -(i/n)f(0)^{-1}. \]

**Proof** Taking expectations in equation (A.1), 

\[E X_{(i)} = G(i/n) + O(n^{-1+\delta}) = G(0) + (i/n)G'(0) + O\{(i/n)^2\} \sim (i/n)f(0)^{-1}. \]

Therefore, 

\[E X_{(-i)} = \sum_{j=1}^s a_j E X_{(A_j i)} \sim \sum_{j=1}^s a_j (A_j i/n) f(0)^{-1} = -(i/n)f(0)^{-1}, \]

as required.
**Appendix 2A.2**

**Tables: simulation results**

<table>
<thead>
<tr>
<th>$f'(0)$</th>
<th>(n = 200) &amp;</th>
<th>(n = 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2PR1</td>
<td>2PR2</td>
</tr>
<tr>
<td>M</td>
<td>0.728</td>
<td>0.758</td>
</tr>
<tr>
<td>B</td>
<td>-0.089</td>
<td>-0.287</td>
</tr>
<tr>
<td>S</td>
<td>0.722</td>
<td>0.702</td>
</tr>
<tr>
<td>A</td>
<td>0.168</td>
<td>0.312</td>
</tr>
<tr>
<td>M</td>
<td>0.692</td>
<td>0.548</td>
</tr>
<tr>
<td>B</td>
<td>0.279</td>
<td>-0.080</td>
</tr>
<tr>
<td>S</td>
<td>0.633</td>
<td>0.543</td>
</tr>
<tr>
<td>A</td>
<td>0.135</td>
<td>0.459</td>
</tr>
<tr>
<td>M</td>
<td>0.745</td>
<td>0.616</td>
</tr>
<tr>
<td>B</td>
<td>0.428</td>
<td>0.266</td>
</tr>
<tr>
<td>S</td>
<td>0.609</td>
<td>0.555</td>
</tr>
<tr>
<td>A</td>
<td>0.143</td>
<td>0.461</td>
</tr>
<tr>
<td>M</td>
<td>0.211</td>
<td>0.302</td>
</tr>
<tr>
<td>B</td>
<td>-0.005</td>
<td>0.076</td>
</tr>
<tr>
<td>S</td>
<td>0.211</td>
<td>0.293</td>
</tr>
<tr>
<td>A</td>
<td>0.600</td>
<td>0.382</td>
</tr>
</tbody>
</table>

Table 2.2: Comparison of two-point rules for $f'(0) < 0$: root mean squared error (denoted M), bias (B), standard deviation (S) and angle (A) between $f'(0)$ and $f'(0)$ at $x = 0$. Simulation results were obtained using 500 independent samples of 200 or 2000 observations from truncated Normal mixture densities on [0,1] having given values of $f'(0)$. 51
<table>
<thead>
<tr>
<th>( f'(0) )</th>
<th>( n = 200 )</th>
<th>( n = 2000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2PR1</td>
<td>2PR2</td>
</tr>
<tr>
<td>M</td>
<td>0.138</td>
<td>0.196</td>
</tr>
<tr>
<td>1.20 B</td>
<td>0.071</td>
<td>0.118</td>
</tr>
<tr>
<td>S</td>
<td>0.118</td>
<td>0.157</td>
</tr>
<tr>
<td>A</td>
<td>0.645</td>
<td>1.012</td>
</tr>
<tr>
<td>M</td>
<td>0.694</td>
<td>0.803</td>
</tr>
<tr>
<td>5.99 B</td>
<td>0.599</td>
<td>0.686</td>
</tr>
<tr>
<td>S</td>
<td>0.350</td>
<td>0.418</td>
</tr>
<tr>
<td>A</td>
<td>1.538</td>
<td>1.933</td>
</tr>
<tr>
<td>M</td>
<td>0.821</td>
<td>0.789</td>
</tr>
<tr>
<td>12.76 B</td>
<td>0.616</td>
<td>0.567</td>
</tr>
<tr>
<td>S</td>
<td>0.544</td>
<td>0.548</td>
</tr>
<tr>
<td>A</td>
<td>1.612</td>
<td>1.544</td>
</tr>
<tr>
<td>M</td>
<td>0.877</td>
<td>0.933</td>
</tr>
<tr>
<td>23.42 B</td>
<td>0.755</td>
<td>0.768</td>
</tr>
<tr>
<td>S</td>
<td>0.446</td>
<td>0.530</td>
</tr>
<tr>
<td>A</td>
<td>0.665</td>
<td>0.878</td>
</tr>
</tbody>
</table>

Table 2.3: Comparison of two-point rules for \( f'(0) > 0 \): root mean squared error (denoted M), bias (B), standard deviation (S) and angle (A) between \( f'(0) \) and \( f'(0) \) at \( x = 0 \). Simulation results were obtained using 500 independent samples of 200 or 2000 observations from truncated Normal mixture densities on \([0,1]\) having given values of \( f'(0) \).
Table 2.4: Comparison of three-point rules for $f'(0) < 0$: root mean squared error (denoted M), bias (B), standard deviation (S) and angle (A) between $f'(0)$ and $	ilde{f}'(0)$ at $x = 0$. Simulation results were obtained using 500 independent samples of 200 or 2000 observations from truncated Normal mixture densities on [0,1] having given values of $f'(0)$.

<table>
<thead>
<tr>
<th>$f'(0)$</th>
<th>$n = 200$</th>
<th>$n = 2000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3PR1</td>
<td>3PR2</td>
</tr>
<tr>
<td>M</td>
<td>1.027</td>
<td>1.096</td>
</tr>
<tr>
<td></td>
<td>-17.27</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-0.979</td>
<td>-1.076</td>
</tr>
<tr>
<td>S</td>
<td>0.310</td>
<td>0.207</td>
</tr>
<tr>
<td>A</td>
<td>2.504</td>
<td>2.244</td>
</tr>
<tr>
<td>M</td>
<td>0.423</td>
<td>0.709</td>
</tr>
<tr>
<td></td>
<td>-9.51</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-0.227</td>
<td>-0.635</td>
</tr>
<tr>
<td>S</td>
<td>0.356</td>
<td>0.316</td>
</tr>
<tr>
<td>A</td>
<td>0.624</td>
<td>1.802</td>
</tr>
<tr>
<td>M</td>
<td>0.344</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>-4.10</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.080</td>
<td>-0.176</td>
</tr>
<tr>
<td>S</td>
<td>0.334</td>
<td>0.311</td>
</tr>
<tr>
<td>A</td>
<td>1.090</td>
<td>1.751</td>
</tr>
<tr>
<td>M</td>
<td>0.172</td>
<td>0.257</td>
</tr>
<tr>
<td></td>
<td>-1.16</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.034</td>
<td>-0.051</td>
</tr>
<tr>
<td>S</td>
<td>0.168</td>
<td>0.252</td>
</tr>
<tr>
<td>A</td>
<td>0.899</td>
<td>1.020</td>
</tr>
</tbody>
</table>
Table 2.5: Comparison of three-point rules for $f'(0) > 0$: root mean squared error (denoted M), bias (B), standard deviation (S) and angle (A) between $f'(0)$ and $f(0)$ at $x = 0$. Simulation results were obtained using 500 independent samples of 200 or 2000 observations from truncated Normal mixture densities on [0,1] having given values of $f'(0)$.
Table 2.6: Comparison of boundary correction methods for $f'(0) < 0$: root mean squared error (denoted M), bias (B), standard deviation (S) and angle (A) between $f'(0)$ and $\hat{f}'(0)$ at $x = 0$ for uncorrected ($\hat{f}$), boundary kernel ($\hat{f}_{BK}$), cut-and-rescale ($\hat{f}_{CR}$), reflection ($\hat{f}_R$) and pseudodata ($\hat{f}_{PS}$) estimators. Simulation results were obtained using 500 independent samples of 200 or 2000 observations from truncated Normal mixture densities on [0,1] having given values of $f'(0)$.

<table>
<thead>
<tr>
<th>$f'(0)$</th>
<th>$n = 200$</th>
<th>$n = 2000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{f}$</td>
<td>$\hat{f}_{BK}$</td>
</tr>
<tr>
<td>M</td>
<td>1.810</td>
<td>0.455</td>
</tr>
<tr>
<td>-17.27 B</td>
<td>-1.809</td>
<td>-0.106</td>
</tr>
<tr>
<td>S</td>
<td>0.075</td>
<td>0.442</td>
</tr>
<tr>
<td>A</td>
<td>2.907</td>
<td>0.126</td>
</tr>
<tr>
<td>M</td>
<td>1.519</td>
<td>0.413</td>
</tr>
<tr>
<td>-9.51 B</td>
<td>-1.517</td>
<td>-0.088</td>
</tr>
<tr>
<td>S</td>
<td>0.079</td>
<td>0.404</td>
</tr>
<tr>
<td>A</td>
<td>2.840</td>
<td>0.279</td>
</tr>
<tr>
<td>M</td>
<td>1.153</td>
<td>0.445</td>
</tr>
<tr>
<td>-4.10 B</td>
<td>-1.148</td>
<td>-0.003</td>
</tr>
<tr>
<td>S</td>
<td>0.101</td>
<td>0.445</td>
</tr>
<tr>
<td>A</td>
<td>2.763</td>
<td>0.693</td>
</tr>
<tr>
<td>M</td>
<td>0.637</td>
<td>0.328</td>
</tr>
<tr>
<td>-1.16 B</td>
<td>-0.633</td>
<td>0.017</td>
</tr>
<tr>
<td>S</td>
<td>0.072</td>
<td>0.328</td>
</tr>
<tr>
<td>A</td>
<td>2.157</td>
<td>0.474</td>
</tr>
<tr>
<td>$f'(0)$</td>
<td>$n = 200$</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>$\hat{f}$</td>
<td>$\hat{f}_{BK}$</td>
</tr>
<tr>
<td>M</td>
<td>0.091</td>
<td>0.182</td>
</tr>
<tr>
<td>B</td>
<td>-0.078</td>
<td>0.006</td>
</tr>
<tr>
<td>S</td>
<td>0.046</td>
<td>0.182</td>
</tr>
<tr>
<td>A</td>
<td>0.153</td>
<td>0.413</td>
</tr>
<tr>
<td>M</td>
<td>0.439</td>
<td>0.398</td>
</tr>
<tr>
<td>B</td>
<td>-0.428</td>
<td>0.041</td>
</tr>
<tr>
<td>S</td>
<td>0.096</td>
<td>0.396</td>
</tr>
<tr>
<td>A</td>
<td>0.045</td>
<td>0.519</td>
</tr>
<tr>
<td>M</td>
<td>0.374</td>
<td>0.588</td>
</tr>
<tr>
<td>B</td>
<td>-0.343</td>
<td>-0.003</td>
</tr>
<tr>
<td>S</td>
<td>0.148</td>
<td>0.588</td>
</tr>
<tr>
<td>A</td>
<td>0.022</td>
<td>0.622</td>
</tr>
<tr>
<td>M</td>
<td>0.256</td>
<td>0.635</td>
</tr>
<tr>
<td>B</td>
<td>-0.200</td>
<td>-0.059</td>
</tr>
<tr>
<td>S</td>
<td>0.158</td>
<td>0.632</td>
</tr>
<tr>
<td>A</td>
<td>0.002</td>
<td>0.270</td>
</tr>
</tbody>
</table>

Table 2.7: Comparison of boundary correction methods for $f'(0) > 0$: root mean squared error (denoted M), bias (B), standard deviation (S) and angle (A) between $f'(0)$ and $\hat{f}'(0)$ at $x = 0$ for uncorrected ($\hat{f}$), boundary kernel ($\hat{f}_{BK}$), cut-and-rescale ($\hat{f}_{CR}$), reflection ($\hat{f}_{R}$) and pseudodata ($\hat{f}_{PS}$) estimators. Simulation results were obtained using 500 independent samples of 200 or 2000 observations from truncated Normal mixture densities on $[0,1]$ having given values of $f'(0)$.
Chapter 3

Bootstrap confidence regions for the intensity of a Poisson point process

3.1 Introduction

Nonparametric methods for estimating a varying point process intensity have been developed by a number of authors, including Leadbetter and Wold (1983), Ramlau-Hansen (1983), Diggle (1985), Ellis (1986), and Diggle and Marron (1988). Point estimates may be difficult to interpret without some idea of their accuracy. That typically requires confidence intervals or bands, and the problem of constructing such regions can be rather awkward in the context of dependent data, such as that derived from a Poisson process. In the present chapter we develop several simple and, we argue, attractive bootstrap methods for computing confidence regions for Poisson intensity functions. They use a variety of different resampling algorithms, and employ different variants of the percentile-t bootstrap.

Thus, the context of this chapter is that of the bootstrap for dependent processes. Although we treat point processes rather than time series, the setting is in principle not unlike that of bootstrapping a time series when a structural model is available for the type of dependence (e.g. an autoregression), yet the distribution of perturbations
is virtually arbitrary. This type of problem has been treated by, for example, Bose (1988), but there are fundamental differences between that work and our own. In particular, in the point process context, with no structural assumptions made about the intensity function, the quantity of interest can only be accessed via statistical smoothing, and so smoothing is an essential feature of all our algorithms. This is not the case with more traditional inferential problems involving dependent data. There are of course other bootstrap approaches to statistical inference under dependence, such as the block bootstrap (for example Hall (1985), Carlstein (1986) and Künsch (1989)), but they are not closely related to the methods that we develop.

Literature in the curve estimation context distantly related to our work includes Härdle and Bowman (1988) and Härdle and Marron (1991) on bootstrap confidence bands for regression curves, and Hall (1993) on bootstrap confidence bands for density estimates. The Poisson process context is distinguished by, among other things, the need to develop a different approach to bootstrapping, with consequently different technology.

A major way in which our work differs from that of earlier authors is that we do not take a "Cox process" view of the intensity estimation problem. Earlier contributions (Diggle (1985) and Diggle and Marron (1988)) considered the intensity of the observed Poisson process to be a realization of a stationary stochastic process, and assessed performance of their estimation procedures by taking averages over that process as well as over the observed data conditional on intensity. This considerably simplified their theoretical development, since the estimation problem was transformed to one for stationary rather than "time-varying" processes. However, we feel that the Cox process view is much less appropriate in the confidence region problem than it was for point estimation, since it makes little sense to construct a confidence region for the average intensity. Therefore we treat the Poisson process as genuinely non-stationary. This requires us to develop new theory for point estimation in this more complex case, complementing work of Diggle and Marron on point estimation, as well as contributing new results on confidence regions.

All our work has a straightforward generalization to the case of multivariate Poisson processes, where intensity is a function from a multivariate Euclidean space
to the set of positive numbers. The manner in which our univariate algorithms should be modified is so straightforward that it seems unnecessary to comment further. Generalizations to processes other than Poisson are possible, but unless they are stationary, or replicated data are available, it would seem to be essential to model the structural nature of dependence. Confidence regions for the (scalar) intensity of a stationary non-Poisson process, without structural assumptions, may be constructed using the block bootstrap; see Hall (1985) for methods in the multivariate case. However, since this approach does not require smoothing techniques then it is really beyond the scope of this thesis, and so will not be treated here.

Section 3.2.1 presents the basic kernel estimator of a Poisson intensity function unadjusted for edge effects, and gives its asymptotic bias and variance. Boundary correction of this estimator is outlined in Section 3.2.5. Our three basic methods for resampling are discussed in Section 3.2.2. These are applied to the development of confidence bands in Section 3.2.3, where one percentile and four different percentile-\(t\) constructions are detailed. Each is suitable for use with each of the resampling methods. The confidence regions introduced in Section 3.2.3 are perhaps more accurately described as regions for \(E(\hat{\lambda})\) rather than for \(\lambda\). However, they may be bias-corrected to produce regions for \(\lambda\), as shown in Section 3.2.4.

Section 3.3 develops the theory upon which these intervals and bands are based. These results are proved in Section 3.5. The different confidence bands are applied to both real and simulated data in Section 3.4, which addresses the problem of finite-sample coverage accuracy.

The six appendices at the end of Chapter 3 contain the lemmas used in the proofs of the theorems in this chapter.
3.2 Methodology

3.2.1 Kernel estimation of a Poisson intensity function

Suppose a Poisson point process \( \{X_i\} \), conceptually on the infinite interval \((-\infty, \infty)\), is observed on \( I = (0,1) \), with points \( 0 < X_1 \leq X_2 \leq \ldots \leq X_N < 1 \). In Section 1.4.2, we showed that a natural kernel estimator of the varying intensity \( \lambda(x) \) is given by

\[
\hat{\lambda}(x) = h^{-1} \sum_{i=1}^{N} K\left\{ (x - X_i)/h \right\},
\]

where \( K \) is a kernel function and \( h \) denotes bandwidth. We shall address the matter of edge effects in Section 3.2.5. For now, take \( \hat{\lambda} \) to be the estimator defined by (3.1), perhaps incorporating a modification towards the ends of \( I \).

Statistical consistency of \( \hat{\lambda} \) for \( \lambda \), under smoothness conditions alone, can only occur in a relative sense, and requires that \( \lambda \) increase. In particular, it may be shown that if \( \lambda = l\mu \) where \( \mu \) is a fixed function and \( l \) is a scalar, then \( \hat{\lambda}/l \) converges to \( \mu \) in mean square if and only if \( h \to 0 \) and \( hl \to \infty \) as \( l \to \infty \).

Under these consistency assumptions and using a second-order kernel, we shall prove in Section 3.5 that the bias and variance of \( \hat{\lambda} \) are given approximately by

\[
\frac{1}{2} \kappa h^2 \lambda'' \quad \text{and} \quad h^{-1} \lambda \int K^2 \quad \text{respectively, where} \quad \kappa = \int y^2 K(y) \, dy.
\]

Thus, both the mean and variance of \( \hat{\lambda} \) are proportional to \( \lambda \), a fact which will be used to motivate various percentile-\( t \) bootstrap approaches to constructing confidence regions.

The bias can be made asymptotic to \((-1)^r \frac{1}{r!} \kappa h^r \lambda^{(r)} \) for some \( r > 2 \) by using an \( r \)th order kernel. However, using a high-order kernel can be problematical in practice, since it does not guarantee a nonnegative estimate of \( \lambda \). This difficulty can make the first bootstrap resampling method in Section 3.2.2 unworkable for high-order kernels unless it is appropriately modified. Therefore, we shall assume for the rest of this chapter that the kernel \( K \) is a compactly supported, bounded, symmetric probability density function such as the quartic kernel,

\[
K(u) = \frac{15}{16} (1 - u^2)^2 I_{[-1,1]}(u).
\]

This guarantees that \( \hat{\lambda} \) is nonnegative, an essential property for an estimate of the intensity of a point process.
Let us continue to use the model \( \lambda = l\mu \) for a fixed function \( \mu \) considered earlier, so that \( \lambda \) is of size \( l \). In view of the bias and variance formulae noted just above, the bias is of order \( h^2l \), and the error of \( \hat{\lambda} \) about its mean is of order \( (l/h)^{1/2} \). These two quantities are of equal size when \( h \sim l^{-1/5} \), which is the optimal order of bandwidth in the sense of minimizing error in any \( L_p \) metric for finite \( p \). More concisely, mean squared error at a point is given by

\[
\text{MSE} = \mathbb{E}(\hat{\lambda} - \lambda)^2 
\approx \frac{1}{4} \kappa^2 h^4 l''^2 + h^{-1} l \int K^2,
\]

and the right-hand side is minimized by taking

\[
h = \left( \kappa^{-2} l''^2 \int K^2 \right)^{1/5} = l^{-1/9} \left( \kappa^{-2} l''^2 \mu \int K^2 \right)^{1/5}.
\]

This is of course the analogue of the more familiar bandwidth choice formula for kernel density estimators. Such analogies are discussed in greater detail by Diggle and Marron (1988), albeit for Cox processes rather than the present context.

The statements above about bias, variance, and optimal bandwidth choice apply of course to estimation of \( \lambda \) at a point, and to simultaneous estimation at any finite number of points. Mean integrated squared error over an interval \( I \) is readily derived by integrating formula (3.2):

\[
\text{MISE} = \int \mathbb{E}(\hat{\lambda} - \lambda)^2 
\approx \frac{1}{4} \kappa^2 h^4 \int l''^2 + h^{-1} \left( \int \lambda \right) \left( \int K^2 \right),
\]

and the minimizer of the right-hand side is given by

\[
h_{opt} = l^{-1/9} \left\{ \kappa^{-2} \left( \int \mu''^2 \right)^{-1} \left( \int \mu \right) \left( \int K^2 \right) \right\}^{1/5};
\]

compare (3.3). In particular, the optimal bandwidth is of size \( l^{-1/5} \).

Asymptotic theory for a simultaneous confidence band is best described in the \( L_\infty \) metric, rather than an \( L_p \) metric for finite \( p \). We shall show in Section 3.3 that if \( I_\epsilon \) denotes the interval \((\epsilon, 1 - \epsilon)\) then the supremum of error-about-the-mean over \( I_\epsilon \) is of size \( (l/h)^{1/2} (\log l)^{1/2} \) (compared with \( (l/h)^{1/2} \) in the case of an \( L_p \) metric), and the supremum of bias is of size \( h^2l \) (the same as in an \( L_p \) metric).
3.2.2 Bootstrap resampling methods

There are several ways of bootstrapping the process \{X_i\}, to obtain a resample \{X_i^*\}. We give three of them below. The second and third are motivated by the fact that conditional on \( N = n \), \( X_1, \ldots, X_N \) denote the order statistics of a sample of size \( n \) drawn from the distribution with density \( f(x) = \lambda(x)/\Lambda(1) \), where

\[
\Lambda(x) = \int_0^x \lambda(y) \, dy.
\]

Method (1). Resample the \( X_i^* \)'s as points of a Poisson process of intensity \( \Lambda \). Thus, the resampled Poisson process on the interval \( I \) may be written as \( X_1^*, \ldots, X_N^* \), where \( N^* \) has a Poisson distribution with intensity \( \Lambda(1) \), and

\[
\Lambda(x) = \int_0^x \Lambda(y) \, dy.
\]

Method (2). Conditional on \( \mathcal{X} = \{X_1, \ldots, X_N\} \), let \( N^* \) have a Poisson distribution with parameter \( \Lambda(1) \). Draw \( X_1^*, \ldots, X_N^* \) by sampling randomly with replacement, \( N^* \) times, from \( \mathcal{X} \).

Method (3). Here we do the same as in method (2), except that we replace \( N^* \) by \( N \). Thus, conditional on \( \mathcal{X} \) we draw \( X_1^*, \ldots, X_N^* \) by sampling uniformly, with replacement, from \( \mathcal{X} \), and take this collection to be our resample on \( I \).

The bootstrap version of \( \Lambda \) is

\[
\hat{\Lambda}^*(x) = h^{-1} \sum_i K\{(x - X_i^*)/h\},
\]

where the sum is over \( 1 \leq i \leq N^* \) in the case of methods (1) and (2), and over \( 1 \leq i \leq N \) for method (3).

Note particularly that in the case of methods (2) and (3), \( E(\hat{\Lambda}^*|\mathcal{X}) = \hat{\Lambda} \). This equality is false for method (1), where to a good approximation,

\[
E(\hat{\Lambda}^*|\mathcal{X}) \sim \int \hat{\Lambda}(x - hy) K(y) \, dy.
\]

(A rigorous proof of this result will be given in Theorem 3.3.) The relative simplicity of \( E(\hat{\Lambda}^*|\mathcal{X}) \) under methods (2) and (3) is perhaps one reason for preferring these over method (1).
Methods (2) and (3) produce resamples that include ties, whereas method (1) does not. Methods (1) and (2) give point processes on \( I \) for which the number of points varies from one resample to the next, whereas that number is fixed in the case of method (3).

Resampling under method (1) may be accomplished by first resampling \( N^* \) as in method (2), and then drawing \( X_1^*, \ldots, X_{N^*}' \) by sampling randomly with replacement, \( N^* \) times, from the distribution with density \( \tilde{\lambda}/\tilde{\Lambda}(1) \). This in turn is equivalent to defining \( X_i^* = Y_i^* + hZ_i^* \), where the \( Y_i^* \)'s are drawn randomly with replacement from \( \chi \), and conditional on \( \chi \) the \( Z_i^* \)'s are independent and identically distributed with density \( K \). Comparing this view of method (1) with method (2) we see that (1) might be termed a smoothed version of (2), with the role of the empirical distribution in (2) replaced in (1) by that of a kernel-smoothed empirical distribution. The smoothing step does not have to be conducted using the same bandwidth as that employed to estimate \( \lambda \) in the final analysis. If the bandwidth for the smoothing step were \( h' \) rather than \( h \) then the corresponding version of method (1) would be:

Method (1'). Define

\[
\tilde{\lambda}(x) = h'^{-1} \sum_{i=1}^{N} K\{(x - X_i)/h'\}.
\]

(Indeed, a different kernel function could be employed as well.) Resample the \( X_i^* \)'s as points of a Poisson process of intensity \( \tilde{\lambda} \). Thus, the resampled Poisson process on the interval \( I \) may be written as \( X_1^*, \ldots, X_{N^*}' \), where \( N^* \) has a Poisson distribution with intensity \( \tilde{\Lambda}(1) \), and

\[
\tilde{\Lambda}(x) = \int_0^x \tilde{\lambda}(y) \, dy.
\]

Method (1') is identical to method (1) if \( h = h' \), and converges to method (2) as \( h' \) tends to zero. For further discussion of the smoothed bootstrap, see Hall (1992a) and the references therein.

Next we discuss techniques for constructing confidence regions for \( \lambda \), using any of the three methods of resampling. There are basically two approaches, percentile or percentile-\( t \), either of which might need to be bias-corrected in some way, and both of which are applicable to various types of region (e.g. symmetric or equal-tailed).

We prefer the percentile-\( t \) approach, and focus most of our attention there.
3.2.3 Confidence bands before bias correction

Let $\mathcal{B}$ denote a connected, voidless random subset of the rectangle $(0, 1) \times (0, \infty)$, such that $\mathcal{B} \cap \{(x, y) : 0 < y < \infty\}$ is nonempty for each $x \in (0, 1)$. We call $\mathcal{B}$ a confidence region for $\lambda$ over the set $S \subset (0, 1)$ with coverage $\alpha$ if

$$P\{(x, \lambda(x)) \in \mathcal{B} \text{ for all } x \in S\} = \alpha.$$ 

The coverage level $\alpha$ is said to be nominal if this identity holds true asymptotically, as $\lambda$ increases. Typically $S$ will be either a point or a finite set of points in the interval $(0,1)$, in which cases confidence intervals result; or an interval such as $I_\epsilon = (\epsilon, 1 - \epsilon)$, where $0 < \epsilon < \frac{1}{2}$, when our construction produces a confidence band. It is often informative to display pointwise confidence intervals in the form of a band (for example Wabha (1983) and Silverman (1985)), even though a simultaneous confidence band will result only by taking $S$ to be an interval.

Define $\hat{\lambda}_1 = E(\hat{\lambda}^*|\mathcal{X})$. Percentile-$t$ bootstrap confidence regions, such as those given by $B_1$ to $B_4$ below, for $\lambda$ are based on the assertion that the bootstrap distribution of the stochastic process

$$T^*(x) = \frac{\hat{\lambda}^*(x) - \hat{\lambda}_1(x)}{\hat{\lambda}^*(x)^{1/2}}, \quad x \in I,$$

is an approximation to the unconditional distribution of

$$T(x) = \frac{\hat{\lambda}(x) - E\hat{\lambda}(x)}{\hat{\lambda}(x)^{1/2}}, \quad x \in I.$$

The validity of this claim is established in an asymptotic sense in Theorem 3.2 in Section 3.3. (Standardization by the square root of intensity is motivated by the fact that $\text{var}(\hat{\lambda})$ is approximately proportional to $\lambda$, so that $T$ and $T^*$ are approximately pivotal.) For example, one may define $t_1$ by

$$P\{|T^*(x)| \leq t_1, \text{ all } x \in S|\mathcal{X}\} = \alpha,$$

and take the band to be

$$B_1 = \{(x, y) : x \in S, \max[0, \hat{\lambda}_1(x) - t_1\hat{\lambda}(x)^{1/2}] < y < \hat{\lambda}_1(x) + t_1\hat{\lambda}(x)^{1/2}\}. \quad (3.5)$$
This produces a confidence region whose width at a point \( x \), being proportional to \( \hat{\lambda}(x)^{1/2} \), reflects the variability of the point estimate \( \hat{\lambda}(x) \). A region of relatively constant width may be constructed by first defining \( t_2 \) by

\[
P\{ |T^*(x)| \leq t_2 / \hat{\lambda}(x)^{1/2}, \ \text{all} \ x \in S | \hat{\lambda} \} = \alpha,
\]

and solving

\[
|\hat{\lambda}_1 - A| / \hat{\lambda}^{1/2} = t_2 / A^{1/2}
\]

for the random function \( A \). When there are three real solutions we take the two largest roots \( A_1(x) \) and \( A_2(x) \) and let

\[
B_2 = \{(x, y) : x \in S, A_1(x) < y < A_2(x) \}.
\]

We ignore the interval \((0, A_0)\), where \( A_0 \) is the smallest root since \( P[(x, \lambda(x)) \in \{(x, y) : x \in S, 0 < y < A_0(x)\}] \) is extremely small. When there is one real root the band is

\[
B_2 = \{(x, y) : x \in S, 0 < y < A(x) \}.
\]

A confidence region similar to \( B_1 \) may be derived by noting that, since the variance of \( \hat{\lambda} \) is very nearly proportional to its mean, the square-root transformation is approximately variance-stabilizing. Thus, we might base percentile-\( t \) confidence regions on the assertion that the bootstrap distribution of \( U^*(x) = \hat{\lambda}_1(x)^{1/2} - \hat{\lambda}_1(x)^{1/2} \) is an approximation to the distribution of \( U(x) = \hat{\lambda}(x)^{1/2} - E \{ \hat{\lambda}(x)^{1/2} \} \). This argument suggests defining \( t_3 \) by

\[
P\{ |U^*(x)| \leq t_3, \ \text{all} \ x \in S | \hat{\lambda} \} = \alpha,
\]

finding the solution for the random function \( B \) of

\[
|\hat{\lambda}_1^{1/2} - B^{1/2}| = t_3,
\]

and taking

\[
B_3 = \{(x, y) : x \in S, (\hat{\lambda}_1^{1/2} - t_3)^2 < y < (\hat{\lambda}_1^{1/2} + t_3)^2 \}.
\]

A fourth band, of the equal-tailed type, may be defined by selecting \( t_4 \) and \( t_5 \) such that

\[
P\{ t_4 \leq T^*(x) \leq t_5, \ \text{all} \ x \in S | \hat{\lambda} \} = \alpha,
\]

65
and
\[ P \{ T^*(x) \leq t_4, \ all \ x \in S|\mathcal{X} \} = P \{ T^*(x) \geq t_5, \ all \ x \in S|\mathcal{X} \}, \]
and taking
\[ B_4 = \{(x, y) : x \in S, \ max[0, \hat{\lambda}_1(x) - t_5 \hat{\lambda}(x)^{1/2}] < y < \hat{\lambda}_1(x) - t_4 \hat{\lambda}(x)^{1/2}\}. \]

All these techniques are of the percentile-t type, in that the Studentized ratio \( \{ \hat{\lambda}^*(x) - \hat{\lambda}_1(x) \}/\hat{\lambda}^*(x)^{1/2}, \) rather than just its numerator, \( V^*(x) = \hat{\lambda}^*(x) - \hat{\lambda}_1(x), \) is bootstrapped. The advantages of percentile-t in other contexts involving curve estimation have been described by Hall (1992a, Chapter 4). While the results there do not apply directly to the present case, the results may be modified so that they do. In particular, percentile-t produces better approximations to second-order features of confidence intervals than does the percentile method. However, for the sake of completeness we mention here the percentile analogue \( B_5 \) of the confidence region \( B_1. \) Define \( t_6 \) by
\[ P \{|V^*(x)| \leq t_6, \ all \ x \in S|\mathcal{X} \} = \alpha, \]
and take the band to be
\[ B_6 = \{(x, y) : x \in S, \ max[0, \hat{\lambda}_1(x) - t_6] < y < \hat{\lambda}_1(x) + t_6\}. \]

Non-bootstrap bands, based on extreme-value approximations, may also be considered. A band of that type, \( B_6, \) will be introduced in Section 3.3 when we discuss the asymptotic distribution of stochastic processes such as \( T. \)

### 3.2.4 Bias correction

The methods described above for constructing confidence regions produce regions for the expected value \( E(\hat{\lambda}) \) rather than for \( \lambda \) itself. Depending on the size of bandwidth, the regions may require correction for bias. This problem is simplest in the case of estimation at a point, or simultaneously at a finite number of points; there, the bias correction is unnecessary if and only if bias is negligible, or equivalently if and only if \( h = o(l^{-1/5}) \). If \( S = \mathcal{I}_\varepsilon, \) so that the confidence region is a band, and if \( B_1 \) is defined as in (3.5), then its coverage probability equals \( \alpha + o(1) \) only if
$h = o\{(l^{-1} \log l)^{1/5}\}$. This follows from the discussion in the last paragraph of Section 3.2.1. The condition demands that $h$ be significantly larger (albeit only by a logarithmic factor) than would be required for optimal pointwise performance. An alternative approach is to note that bias is given approximately by $\frac{1}{2} \kappa h^2 \lambda''$ and to explicitly correct a confidence region $B$ for bias, for example replacing $B$ by

$$B' = \{(x, y) : (x, y + \frac{1}{2} \kappa h^2 \lambda''(x)) \in B\},$$

where $\lambda''$ denotes an estimator of $\lambda''$, perhaps equal to the second derivative of a version of $\hat{\lambda}$ but employing a different, larger value of $h$ in its construction. Theoretical aspects of bias correction will be addressed in Section 3.3.

### 3.2.5 Corrections for edge effects

Although the bias of the estimator in (3.1) is of size $O(h^r)$ in the interior of the estimation interval, it is of size $O(1)$ near the boundaries. This increase in bias towards the ends of the interval is a particular problem in intensity estimation, as the process continues but is not observed outside the observation interval. Ideally, the bias would be the same size in the boundary regions as in the interior.

Diggle (1985) and Diggle and Marron (1988) explicitly correct the estimator for edge effects, using the cut-and-rescale and reflection methods respectively, both of which reduce the bias to $O(h)$ in the boundary regions. However, especially for constructing confidence bands, we found this level of bias correction inadequate. In Chapter 2 we introduced a method of bias correction based on generating pseudodata outside the estimation interval using the order statistics of the sample, then applying the usual estimator to the pseudodata as well as the actual data. This method has bias of $O(h^r)$ in the boundary regions, the same size as in the interior.

In the simulation study for the density estimation context detailed in the previous chapter, we show that the pseudodata method outperforms the reflection method in terms of both reducing bias and reproducing the slope of the true curve at the boundary. Our estimator even outperforms the traditional boundary kernel method of bias correction (see for example Gasser et al (1985)) for moderate and small sample sizes when $f'(0) > 0$. 67
In the simulation study in Section 3.4.1, we use our pseudodata method of edge correction, generating pseudodata using
\[ X_{(-i)} = -5X_{(i/3)} - 4X_{(2i/3)} + \frac{10}{3}X_{(i)}. \]

### 3.3 Asymptotic theory

We cannot expect to estimate \( \lambda \) well, using only local information, unless the points \( X_i \) are closely spaced. Thus, asymptotic theory describing consistency must allow \( \lambda \) to increase without bound. We adopt the asymptotic model

\[ \lambda(x) = l\mu(x), \]

(3.6)

where the smooth function \( \mu \) is held fixed and the scalar parameter \( l \geq 1 \) diverges. In the present section we develop asymptotic theory describing the distribution of the stochastic process \( \hat{\lambda} \), and its bootstrap counterpart \( \hat{\lambda}^* \), under the model in (3.6).

Our first result describes a Gaussian approximation to the distribution of \( \hat{\lambda} \), sufficiently accurate for the development of first-order limit theory. Let \( \hat{\lambda}^* \) be derived by any one of the three different bootstrap algorithms, and write \( \mathbb{E}' \) for expectation conditional on \( \mathcal{X} \).

**Theorem 3.1** Assume that \( K \) is compactly supported with \( \text{supp}(K) \subseteq (-c, c) \), that \( K'' \) exists and is Lipschitz continuous, that \( 0 < \inf_{\mathcal{X}} \mu \leq \sup_{\mathcal{X}} \mu < \infty \), \( \sup_{\mathcal{X}} |\mu'| < \infty \), \( \mu' \) is Hölder continuous, and that for some \( \xi > 0, l^{-\xi} \geq h \geq l^{-(1/4)+\xi} \). Then we may write

\[ \hat{\lambda} - \mathbb{E}(\hat{\lambda}) = (h^{-1} l)^{1/2}(\mu^{1/2} U_n + R_{1n}), \]

\[ \hat{\lambda}^* - \mathbb{E}'(\hat{\lambda}^*) = (h^{-1} l)^{1/2}(\mu^{1/2} U_n^* + R_{1n}^*), \]

where \( U_n \), and \( U_n^* \) conditional on \( \mathcal{X} \), are stationary Gaussian processes on \((0, 1)\) with identical distributions, zero means, and covariances

\[ \text{cov}\{U_n(x_1), U_n(x_2)\} = \gamma(x_1 - x_2) \equiv \int K(y) K\{y + (x_1 - x_2) h^{-1}\} dy; \]

and for some \( \eta > 0 \), and all \( k \geq 1 \) and \( \epsilon \in [ch, \frac{1}{2}) \)

\[ \mathbb{E}\left\{ \sup_{\epsilon \leq \xi \leq 1 - \epsilon} |R_{1n}(x)|^k + \sup_{\epsilon \leq \xi \leq 1 - \epsilon} |R_{1n}^*(x)|^k \right\} = O(l^{-kn}) \]

as \( l \to \infty \).
An immediate consequence of Theorem 3.1 is that $\tilde{\lambda} - E(\tilde{\lambda})$ is asymptotically Normally distributed with zero mean and variance $h^{-1} \int K^2$; and that this asymptotic distribution is estimated consistently by the bootstrap. The theorem may also be used to develop an extreme-value approximation to the distribution of $\sup |(\tilde{\lambda} - E(\tilde{\lambda}))/\tilde{\lambda}^{1/2}|$. Indeed, letting $\epsilon \in [ch, \frac{1}{2}]$ and defining $\mathcal{I}_\epsilon = [\epsilon, 1-\epsilon]$ and

$$a_t = \left(2 \log \left[\left\{ \int (K')^2 / \int K^2 \right\}^{1/2} (1 - 2\epsilon)(2\pi h)^{-1} \right]\right)^{1/2},$$

we claim that

$$S \equiv a_t \left\{ (h/\int K^2)^{1/2} \sup_{\mathcal{I}_\epsilon} |(\tilde{\lambda} - E(\tilde{\lambda}))/\tilde{\lambda}^{1/2}| - a_t \right\}$$

has limiting distribution function $G(z) = \exp(-2e^{-z})$. In principle this result could be used to develop simultaneous confidence bands for $E(\tilde{\lambda})$ over $\mathcal{I}_\epsilon$, circumventing the need for bootstrap methods. However, we shall show shortly that this proposal is impractical, not least because the rate of convergence in the limit theorem $P(S \leq z) \to G(z)$ is exceedingly slow, being only $O((\log l)^{-1})$ as $l \to \infty$.

To describe the asymptotic properties of $S$ in more detail, let $S$ denote any subset of $\mathbb{R}$ and define

$$\Delta_1(z, a, b) = P\left[b \left\{ (h/\int K^2)^{1/2} \sup_{\mathcal{I}_\epsilon} |(\tilde{\lambda} - E(\tilde{\lambda}))/\tilde{\lambda}^{1/2}| - a \right\} \leq z \right] - \exp(-e^{-z}),$$

$$\Delta_2(z, a, b) = P\left[b \left\{ (h/\int K^2)^{1/2} \sup_{\mathcal{I}_\epsilon} |(\tilde{\lambda} - E(\tilde{\lambda}))/\tilde{\lambda}^{1/2}| - a \right\} \leq z \right] - \exp(-2e^{-z}),$$

$$\delta_j(S, a, b) = \sup_{z \in S} |\Delta_j(z, a, b)|.$$

If $S$ contains three or more distinct points then, no matter how $a$ and $b$ are chosen as functions of $h$ or $l$, $(\log l)^{-1} = O(\delta_j(S, a, b))$; and, for $a = b = a_t$ and $S = \mathbb{R}$, $\delta_j(\mathbb{R}, a_t, b_t) = O((\log l)^{-1})$. The case $j = 1$ follows from Theorem 2.1 of Hall (1991), Theorem 3.1 above, and the fact that $E(\tilde{\lambda}) = \lambda + O(h^2l)$. The case $j = 2$ may be treated similarly. See also Bickel and Rosenblatt (1973) and Konakov and Piterbarg (1982, 1983, 1984).

Next we introduce confidence bands based on these extreme-value approximations. The theory recounted just above may be used to describe the coverage accuracy of those bands. Assume the conditions of Theorem 3.1, let $z_\alpha$ denote the
solution of the equation \( \exp(-2e^{-z^2}) = \alpha \), put \( t = t_\alpha = (\int K^2/h)^{1/2}(a_t + o_t^{-1}z_\alpha) \), and define

\[
B_6 = \{(x, y) : x \in S, \quad \hat{\lambda}(x) - t\hat{\lambda}(x)^{1/2} \leq y \leq \hat{\lambda}(x) + t\hat{\lambda}(x)^{1/2}\}.
\]

We may deduce straight away from the results in the previous paragraph that \( B_6 \) is a confidence band for \( E(\hat{\lambda}) \) on \( I_x \), with asymptotic coverage \( \alpha \), but that the coverage accuracy uniformly over \( \alpha \) cannot be better than order \((\log l)^{-1}\). In fact,

\[
P\{(x, E\hat{\lambda}(x)) \in B_6, \quad \text{all } x \in I_x\} = \alpha + O\{\log l\}^{-1}\]

uniformly in \( \alpha \).

\[
(\log l)^{-1} = O\left[\sup_\alpha |P\{(x, E\hat{\lambda}(x)) \in B_6, \quad \text{all } x \in I_x\} - \alpha|\right]. \quad (3.9)
\]

However, like the bootstrap bands introduced in Section 3.2.3, \( B_6 \) is a confidence band for \( E\hat{\lambda} \) rather than for \( \lambda \) itself. Therefore, it will not necessarily cover \( \lambda \) with the desired asymptotic probability \( \alpha \), let alone with accuracy \( O\{\log l\}^{-1}\), unless \( h \) is small enough to make the bias negligible.

Under more stringent conditions, including the requirement that \( \mu'' \) exist and be continuous, it may be shown that

\[
E(\hat{\lambda}) = \lambda + \frac{1}{2} \kappa h^2 \lambda'' + o(h^2l) \quad (3.10)
\]

uniformly in \( x \in I_x \), where \( \kappa = \int z^2 K(z)dz \); see Theorem 3.3 below. Therefore, if we ask that \( h = o\{(l^{-1}\log l)^{1/5}\} \) then bias, which is of order \( h^2l \), will be negligibly small in the context of the confidence band, so that

\[
P\{(x, \lambda(x)) \in B_6, \quad \text{all } x \in I_x\} = \alpha + o(1); \quad (3.11)
\]

compare (3.8). An alternative approach is to explicitly bias-correct the band \( B_6 \), as noted in Section 3.2.4.

Our next result describes the Gaussian approximation of Theorem 3.1 in more detail. Theorem 3.2 below is sufficiently detailed to imply that the percentile-\( t \) bootstrap approximation to the distribution of \( S \) is substantially more accurate than the extreme-value approximation described earlier, as we shall shortly relate. A proof of Theorem 3.2 is given in Section 3.5.1.

Define \( M(x) = \int_0^x \mu(y)dy \).
Theorem 3.2 Assume the conditions of Theorem 3.1. Then we may write

$$\tilde{\lambda} - E(\tilde{\lambda}) = h^{-1/2} \int W\{M(x - hy)\} K(y) dy + (h^{-1/2})^{1/2} R_{2n}, \quad (3.12)$$

$$\tilde{\lambda}^* - E'(\tilde{\lambda}^*) = h^{-1/2} \int W^*\{M(x - hy)\} K'(y) dy + (h^{-1/2})^{1/2} R^*_{2n}, \quad (3.13)$$

where $W$, and $W^*$ conditional on $\mathcal{X}$, are standard Wiener processes on $(0, \infty)$, and for all $k \geq 1$, $\epsilon \in [c h, 1/2]$ and some $\beta > 0$,

$$E\left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |R_{2n}(x)|^k + (h^2 l)^{-k/2} \sup_{\epsilon \leq x \leq 1 - \epsilon} |R^*_{2n}(x)|^k \right\} = O((h l)^{-1/2} \log l)^{k \beta}. \quad (3.14)$$

Construction of the processes $W$ and $W^*$ depends on $l$, and in the case of $W^*$, also on $\mathcal{X}$, although of course they have the same distribution on all occasions. The connection between Theorems 3.1 and 3.2 may be drawn by noting that, under the conditions of Theorem 3.1 (assumed in Theorem 3.2),

$$h^{-1/2} \mu(x)^{-1/2} \int W\{M(x - hy)\} K'(y) dy = Z(x) + R(x),$$

where $Z$ is a stationary Gaussian process with covariance $\gamma$, and

$$E\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |R(x)|^k \} = O(l^{-k \eta})$$

for some $\eta > 0$ and all $k \geq 1$. As Theorem 3.1 is a less detailed version of Theorem 3.2, its proof is a consequence of that theorem and the result above.

The main practical significance of Theorem 3.2 derives from the fact that the first terms on the right-hand sides of (3.12) and (3.13) have identical distributions, thereby ensuring the accuracy of the bootstrap approximation to the distribution of the stochastic process $\hat{\lambda} - E(\hat{\lambda})$. Furthermore, the remainder terms $R_{2n}$ and $R^*_{2n}$ are of much smaller order than $(\log l)^{-1}$, guaranteeing that the bootstrap approximation is more accurate than the extreme-value one. Indeed, it may be proved from Theorem 3.2 that, with $S$ defined by (3.7) and

$$S^* = a_t \left\{ \left( \frac{h}{\int K^2} \right)^{1/2} \sup_{\mathcal{X}} |(\hat{\lambda}^* - E'(\hat{\lambda}^*))/\hat{\lambda}^{1/2}| - a_t \right\},$$

we have

$$\sup_{-\infty < z < \infty} |P(S \leq z) - P(S^* \leq z|\mathcal{X})| = O((hl)^{-\eta})$$

71
for some $\eta > 0$, with probability one. This level of accuracy is exponentially greater than that provided by the extreme-value approximation. The bootstrap does not appear to have any classical competitors in this particular problem.

Take $S = \mathcal{I} = [\epsilon, 1-\epsilon]$, so that the region $B_1$ is a confidence band. The argument above may be used to show that

$$P\{x, E\tilde{\lambda}(x)\in B_1, \text{ all } x \in \mathcal{I}_x\} = \alpha + O\{(hl)^{-\eta}\} \quad (3.14)$$

for some $\eta > 0$, uniformly in $\alpha$. Compare this level of accuracy with that provided by the extreme-value approximation, indicated by (3.8) and (3.9).

Of course, the level of coverage accuracy evidenced by (3.14) is for the expected value of $\tilde{\lambda}$, not for $\lambda$. Undersmoothing (e.g. choosing $h$ to be of order $l^{-(1/5)-\eta}$ for some $\eta > 0$), or explicit bias correction, are required if the coverage error of the band $B_1$ is to be of order $l^{-\eta'}$ for some $\eta' > 0$, compared with $(\log l)^{-1}$ for the extreme-value band $B_e$.

We conclude this section by describing details of bias and variance formulae, such as (3.10). These results lead directly to the bias corrections discussed in Section 3.2.4. There it was claimed that if $K$ is a second-order kernel then $E(\tilde{\lambda})$ is given approximately by

$$\lambda_1(x) = \int \lambda(x - hy) K(y) dy = \lambda(x) + \frac{1}{2} \kappa h^2 \lambda''(x) + o(h^2 l), \quad (3.15)$$

provided $\mu''$ exists and is continuous. The order of this approximation to $E(\tilde{\lambda})$ is described by the following theorem, which also treats the first of the three different bootstrap methods. A proof is given in Section 3.5.2.

Define $\tilde{\lambda}_2(x) = \int \lambda(x - hy) K(y) dy$, and let $\epsilon \in [ch, \frac{1}{2}]$.

**Theorem 3.3** Assume the conditions of Theorem 3.1. Then

$$E(\tilde{\lambda}) = \lambda_1 + O(h^{-3} l^{-1/2} + h^{-1+\epsilon}) \quad (3.16)$$

uniformly on $\mathcal{I}_x$, and in the case of the first of the three bootstrap methods,

$$E'(\tilde{\lambda}^*) = \tilde{\lambda}_2 + O(h^{-3} l^{-1/2} + h^{-1+\epsilon}) \quad (3.17)$$

uniformly on $\mathcal{I}_x$, with probability one.
To appreciate the implications of the order of error in (3.16), observe that under the conditions of Theorems 3.1, \( h \geq l^{-(1/4)+\xi} \) for some \( \xi > 0 \). It follows that \( h^{-3} l^{-1/2} + h^{-1+\xi} = o(h^2 l) \), and so (3.15) and (3.16) together imply that if \( K \) is a second-order kernel with \( \int y^2 K(y) \, dy = \kappa \), say, then (3.10) holds.

Formula (3.10) is not always extendible to the bootstrap case, in the form

\[
E'(\hat{\lambda}^*) - \hat{\lambda} = \frac{1}{2} \kappa h^2 l \mu'' + o(h^2 l) = \frac{1}{2} \kappa h^2 l \mu'' + o(h^2 l) \tag{3.18}
\]

with probability one. The reason is that, depending on choice of \( h \), \( \mu'' \) may not be consistent for \( \mu'' \). It may be shown that \( \hat{\mu}'' = \tilde{\mu}'' / l \) has variance of order \( (h^5 l)^{-1} \), which does not converge to zero unless \( h \) is of larger order than \( l^{-1/5} \). However, if \( l^{-\xi} \geq h \geq l^{-(1/5)+\xi} \) for some \( \xi > 0 \), and \( \mu'' \) exists and satisfies a Lipschitz condition, then (3.18) may be shown to hold with probability one. This situation is entirely analogous to that which arises in density estimation, where the second derivative of a kernel estimator is consistent for the second derivative of the true density only if the bandwidth is of sufficiently large an order of magnitude.

Finally, to complement the bias formulae discussed above, we note that expressions for the variance of \( \tilde{\lambda}(x) \), and for the conditional variance of \( \tilde{\lambda}^* \), are readily deduced from Theorem 3.1: \( \text{var} \{\tilde{\lambda}(x)\} \sim h^{-1} l \mu(x) \int K^2 \), and \( \text{var} \{\tilde{\lambda}^*(x) | \mathcal{X}\} \sim h^{-1} l \mu(x) \int K^2 \) with probability one.
3.4 Numerical results

3.4.1 Simulation study

We conducted a simulation study to obtain empirical confirmation that the proposed methodology produces confidence bands with coverage probabilities close to the nominal levels. We used the three methods of bootstrap resampling described in Section 3.2.2, and the methods of confidence band construction resulting in $B_1, \ldots, B_5$, in Section 3.2.3, and $B_6$ in Section 3.3. The simulations were based on the asymptotic model $\lambda(x) = l\mu(x)$. We employed $l = 100, 300, \text{ and } 500$, and three functions $\mu(x)$ having the same underlying shape but different degrees of roughness:

\[
\begin{align*}
\mu_1(x) &= \frac{3}{2} - x + \frac{1}{4} \sin(4\pi x) \\
\mu_2(x) &= \frac{3}{2} - x + \frac{1}{4} \sin(3\pi x) \\
\mu_3(x) &= \frac{3}{2} - x + \frac{1}{6} \sin(3\pi x).
\end{align*}
\]

(The roughness ratios, measured in terms of $\int \mu''^2$, are $4 : 2.25 : 1.5$.) To reduce computing time in the simulations, we took as our bandwidth the asymptotically optimal choice, $h_{opt}$, given in (3.4). We constructed confidence intervals on $I_\varepsilon = [0.2, 0.8] \subseteq [h_{opt}, 1 - h_{opt}]$, so as not to complicate issues of confidence band construction with issues related to boundary correction. The results for $I_\varepsilon = (0, 1)$ will be presented elsewhere.

In discussing the appropriate amount of smoothing to use at different stages of confidence band construction, we shall use coverage error as our performance criterion, and recommend smoothing parameters which minimise the coverage error of the particular confidence band under construction. Coverage probabilities were calculated by constructing the particular confidence band 200 times and determining how many times the confidence band contained the true intensity.

When resampling from the smoothed empirical distribution (resampling method (1)), we recommend using $h_{opt}/4$. This is because we found that the coverage probabilities for method (1) were a little low using $h_{opt}/2$ in the resampling stage; see Table 3.1 for the comparison. We mentioned in Section 1.5 when discussing the smoothed bootstrap that Hall (1992a) suggests that we should not smooth too much.
The results in Table 3.1 provide empirical confirmation of Hall’s recommendation and give an empirical suggestion for how much to smooth.

Recall that our confidence bands are based on approximating the distribution of $T = (\hat{\lambda} - E(\hat{\lambda})/\hat{\lambda}^{1/2}$ by that of $T^* = \{\hat{\lambda}^* - E(\hat{\lambda}^*|\mathcal{X})\}/\hat{\lambda}^*^{1/2}$. For resampling method (1), $T^* = \{\hat{\lambda}^* - \int \hat{\lambda}(x - hy)K(y)\,dy\}/\hat{\lambda}^*^{1/2}$, and for methods (2) and (3), $T^* = (\hat{\lambda}^* - \hat{\lambda})/\hat{\lambda}^*^{1/2}$. We found that correcting for bootstrap bias by undersmoothing the bootstrap resamples, using $h_{opt}/2$ to calculate $\hat{\lambda}^*$, and $h_{opt}$ for $\hat{\lambda}$ in both $T$ and $T^*$, gave good coverage probabilities and smooth confidence bands: see Table 3.2, discussion below, and Figure 3.2. We tried using $h_{opt}/2$ in $\hat{\lambda}$ in both $T$ and $T^*$, but this produced such oscillatory confidence bands that they only covered the true curve about half the time: see Table 3.3 and Figure 3.3. We employed $B = 200$ bootstrap resamples.

In Table 3.2 we show the actual coverage probabilities of the bias corrected confidence bands on $I_z = [0.2, 0.8]$ with nominal level 0.95. The table shows that all methods except $B_6$ have actual coverage probabilities close to the nominal level. The extreme-value confidence bands $B_6$ are consistently too wide, having actual coverage probability of 1.00, confirming that of the methods we have considered, the bootstrap is the only viable method of determining confidence bands for kernel intensity estimates. An analysis of variance of the actual coverage probabilities in Table 3.2 with main effects $B$, $M$, $l$ and $\mu$ shows that $B_6$ has significantly different coverage probabilities from the other confidence bands. No other contrasts between levels of this or other factors are significant, and the data support the hypothesis that the true coverage probability is 0.95. As bootstrap resampling method (3) is computationally the fastest, it may be preferred over the other resampling methods.

### 3.4.2 Analysis of coal mining data

As stated in Chapter 1, our motivation for treating the problem of confidence bands for the intensity function was provided by the classic coal mining disaster set. Both Barnard (1953) and Cox and Lewis (1966) fitted exponential models to the intensity function in order to detect a trend in the accident rate. The confidence intervals for
their point estimates for the trend indicate that it is highly unlikely that the rate is constant or increasing.

However, in our view, the kernel intensity estimator shown in Figure 2.6 is a more flexible tool for exploratory analysis as it shows not only the general decreasing trend, but also reveals peaks at about 8,000 days and 32,000 days ($x = 0.2$ and $x = 0.78$). By adding confidence bands to the kernel estimator of intensity we can address the same question of global trend that is answered by parametric analysis. Moreover, we can determine whether the additional information about the data, the apparent peaks, is likely to be real or merely a chimera of the non-parametric methodology used.

Figure 3.1 shows the equal-tailed percentile-$t$ confidence band for the coal mining disaster data generated by resampling method (3). The lower confidence limit at $x = 0$ is above the upper confidence limit for $x > 0.4$, providing strong evidence of the general decreasing trend revealed by the parametric analysis. Moreover, the peaks at $x = 0.2$ and $x = 0.78$ are likely to be real features of the data as smoothing out these peaks by hand requires crossing the confidence band.

In our analysis of the coal mining data we selected the bandwidth by least-squares
cross-validation. This is shown in Brooks and Marron (1991) to be asymptotically optimal for kernel intensity estimation, in the sense that the ISE obtained using the cross-validation bandwidth converges almost surely to the minimum ISE.
Figure 3.2: Comparison of bootstrap confidence bands generated using three resampling methods and six confidence band construction methods. True intensity $\lambda = 500\mu_1$ shown by thin solid line, estimated intensity $\hat{\lambda}$ is thin dotted line, and confidence band for $\hat{\lambda}$ on $[0.2,0.8]$ is thick solid line. Bias correction by undersmoothing with recommended choices of bandwidth $h_{opt}/2$ in $\hat{\lambda}^*$ and $h_{opt}$ in $\hat{\lambda}$. 
Figure 3.3: Comparison of bootstrap confidence bands generated using three resampling methods and six confidence band construction methods. True intensity $\lambda = 500\mu_1$ shown by thin solid line, estimated intensity $\hat{\lambda}$ is thin dotted line, and confidence band for $\hat{\lambda}$ on $[0.2,0.8]$ is thick solid line. Bias correction by undersmoothing using suboptimal choices of bandwidth $h_{opt}/2$ in $\hat{\lambda}^*$ and $h_{opt}/2$ in $\hat{\lambda}$ in both $T$ and $T^*$. 
Table 3.1: Comparison of coverage probabilities of confidence bands for $\lambda$ using resampling method (1) with $h_{opt}/2$ and $h_{opt}/4$ in the kernel-smoothed empirical distribution. The nominal level is 0.95.

<table>
<thead>
<tr>
<th></th>
<th>$\mu_1$</th>
<th></th>
<th>$\mu_2$</th>
<th></th>
<th>$\mu_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h_{opt}/2$</td>
<td>$h_{opt}/4$</td>
<td>$h_{opt}/2$</td>
<td>$h_{opt}/4$</td>
<td>$h_{opt}/2$</td>
</tr>
<tr>
<td>$l = 100$</td>
<td>$B_1$</td>
<td>0.940</td>
<td>0.980</td>
<td>0.890</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td>$B_2$</td>
<td>0.870</td>
<td>0.940</td>
<td>0.870</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td>$B_3$</td>
<td>0.950</td>
<td>0.985</td>
<td>0.880</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>$B_4$</td>
<td>0.960</td>
<td>0.970</td>
<td>0.890</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>$B_5$</td>
<td>0.925</td>
<td>0.970</td>
<td>0.870</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td>$B_6$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$l = 300$</td>
<td>$B_1$</td>
<td>0.950</td>
<td>0.970</td>
<td>0.900</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>$B_2$</td>
<td>0.895</td>
<td>0.950</td>
<td>0.845</td>
<td>0.920</td>
</tr>
<tr>
<td></td>
<td>$B_3$</td>
<td>0.915</td>
<td>0.985</td>
<td>0.895</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td>$B_4$</td>
<td>0.925</td>
<td>0.975</td>
<td>0.895</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td>$B_5$</td>
<td>0.915</td>
<td>0.960</td>
<td>0.865</td>
<td>0.920</td>
</tr>
<tr>
<td></td>
<td>$B_6$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$l = 500$</td>
<td>$B_1$</td>
<td>0.905</td>
<td>0.965</td>
<td>0.885</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td>$B_2$</td>
<td>0.875</td>
<td>0.910</td>
<td>0.850</td>
<td>0.910</td>
</tr>
<tr>
<td></td>
<td>$B_3$</td>
<td>0.905</td>
<td>0.960</td>
<td>0.890</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>$B_4$</td>
<td>0.895</td>
<td>0.970</td>
<td>0.865</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td>$B_5$</td>
<td>0.895</td>
<td>0.960</td>
<td>0.865</td>
<td>0.920</td>
</tr>
<tr>
<td></td>
<td>$B_6$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 3.2: Observed coverage probabilities of confidence bands for $\lambda$ with nominal level=0.95. Results are shown for confidence band construction methods $B_1, \ldots, B_6$ and three methods of bootstrap resampling $M_1, M_2$ and $M_3$. Each band was determined from 200 bootstrap resamples and the coverage probabilities were calculated from 200 such bands. We utilised pseudodata to correct edge effects, and undersmoothing for bias correction using recommended choices of bandwidth $h_{opt}/2$ in $\hat{\lambda}$ and $h_{opt}$ in $\hat{\lambda}$.  

<table>
<thead>
<tr>
<th></th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l = 100$</td>
<td>0.980</td>
<td>0.960</td>
<td>0.955</td>
<td>0.970</td>
<td>0.975</td>
<td>0.960</td>
<td>0.955</td>
<td>0.945</td>
<td>0.910</td>
<td>0.940</td>
<td>0.930</td>
<td>0.920</td>
<td>0.970</td>
<td>0.980</td>
<td>0.960</td>
<td>0.955</td>
<td>0.940</td>
<td>0.930</td>
</tr>
<tr>
<td>$l = 300$</td>
<td>0.980</td>
<td>0.960</td>
<td>0.955</td>
<td>0.970</td>
<td>0.975</td>
<td>0.960</td>
<td>0.955</td>
<td>0.945</td>
<td>0.910</td>
<td>0.940</td>
<td>0.930</td>
<td>0.920</td>
<td>0.970</td>
<td>0.980</td>
<td>0.960</td>
<td>0.955</td>
<td>0.940</td>
<td>0.930</td>
</tr>
<tr>
<td>$l = 500$</td>
<td>0.980</td>
<td>0.960</td>
<td>0.955</td>
<td>0.970</td>
<td>0.975</td>
<td>0.960</td>
<td>0.955</td>
<td>0.945</td>
<td>0.910</td>
<td>0.940</td>
<td>0.930</td>
<td>0.920</td>
<td>0.970</td>
<td>0.980</td>
<td>0.960</td>
<td>0.955</td>
<td>0.940</td>
<td>0.930</td>
</tr>
</tbody>
</table>
### Table 3.3: Observed coverage probabilities of confidence bands for $\lambda$ with nominal level=0.95. Results are shown for confidence band construction methods $B_1, \ldots, B_6$ and three methods of bootstrap resampling $M_1$, $M_2$ and $M_3$. Each band was determined from 200 bootstrap resamples and the coverage probabilities were calculated from 200 such bands. We utilised pseudodata to correct edge effects, and undersmoothing for bias correction using suboptimal choices of bandwidth $h_{opt}/2$ in $\hat{\lambda}^*$ and $h_{opt}/2$ in $\hat{\lambda}$ in $T$ and $T^*$. 

<table>
<thead>
<tr>
<th>l</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\mu_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
</tr>
<tr>
<td>100</td>
<td>0.720</td>
<td>0.545</td>
<td>0.490</td>
<td>0.740</td>
<td>0.525</td>
<td>0.505</td>
<td>0.775</td>
<td>0.595</td>
<td>0.500</td>
<td>0.780</td>
<td>0.575</td>
<td>0.555</td>
</tr>
<tr>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0.700</td>
<td>0.490</td>
<td>0.410</td>
<td>0.740</td>
<td>0.495</td>
<td>0.475</td>
<td>0.715</td>
<td>0.465</td>
<td>0.525</td>
<td>0.735</td>
<td>0.480</td>
<td>0.515</td>
</tr>
<tr>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.700</td>
<td>0.385</td>
<td>0.430</td>
<td>0.945</td>
<td>0.945</td>
<td>0.455</td>
<td>0.960</td>
<td>0.530</td>
<td>0.550</td>
<td>0.960</td>
<td>0.525</td>
<td>0.515</td>
</tr>
<tr>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Observed coverage probabilities of confidence bands for $\lambda$ with nominal level=0.95. Results are shown for confidence band construction methods $B_1, \ldots, B_6$ and three methods of bootstrap resampling $M_1$, $M_2$ and $M_3$. Each band was determined from 200 bootstrap resamples and the coverage probabilities were calculated from 200 such bands. We utilised pseudodata to correct edge effects, and undersmoothing for bias correction using suboptimal choices of bandwidth $h_{opt}/2$ in $\hat{\lambda}^*$ and $h_{opt}/2$ in $\hat{\lambda}$ in $T$ and $T^*$. 

82
3.5 Proofs

3.5.1 Proof of Theorem 3.2

As this proof is somewhat lengthy, it is broken into six steps. In the first five steps, which are described below, we derive the non-bootstrap part of the theorem. In the sixth step, we show that the proof can be easily extended to derive the bootstrap part when the bootstrap algorithm is of the first type — involving resampling from a Poisson process. Proofs in the other two bootstrap cases are easier than that in the first, owing to the “random uniform resampling” nature of the algorithm, and so they will not be addressed here.

The lemmas used in the proof of this theorem are given in Appendices A3.1 to A3.6.

1. We approximate \( \hat{\lambda} \) in terms of exponential random variables. Writing \( S_i \) for the sum of independent exponential random variables defined on page 84, we show that

\[
\hat{\lambda}(x) = h^{-1} \sum_i K\{ (x - \alpha_i)/h \} + (h^2 l)^{-1} \sum_i S_i \mu(\alpha_i)^{-1} K'\{ (x - \alpha_i)/h \} \\
- \frac{1}{2} l^{-2} \sum_i S_i^2 [h^{-2} \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{ (x - \alpha_i)/h \} \\
+ h^{-3} \mu(\alpha_i)^{-2} K''\{ (x - \alpha_i)/h \} ] + Q_1(x),
\]

where \( E \{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |Q_1(x)|^k \} \leq C_1(k) (h^{-1} l^{-\zeta/2} + h^{-3} l^{-1/2})^k \) for all \( k \geq 1 \).

2. We remove the quadratic term from the above expression.

3. We approximate \( \sum_i S_i \mu(\alpha_i)^{-1} K'\{ (x - \alpha_i)/h \} \) by \( \sum_i T_i \mu(\alpha_i)^{-1} K'\{ (x - \alpha_i)/h \} \), where \( T_i \) is a sum of independent standard Normal random variables.

4. We replace \( \sum_i T_i \mu(\alpha_i)^{-1} K'\{ (x - \alpha_i)/h \} \) with its integral approximation \( \int W \{ M(x - hy) \} K'(y) dy \), where \( W \) is a standard Wiener process whose construction depends on \( l \).

5. Completion of proof.
Step 1: Approximation to $\hat{\lambda}$ in terms of exponential random variables

Let $Y_1, Y_2, \ldots$ be independent standard exponential random variables, and put $Z_i = Y_i - 1$ and $S_i = \sum_{j \leq i} Y_j$. Note that $\{\sum_{j=1}^{i} Y_j/l, i \geq 1\} = \{(i + S_i)/l, i \geq 1\}$ is a homogeneous Poisson process with intensity function $l$. The function $M(x) = \int_0^x \mu(t) \, dt$ is continuous and strictly increasing, so $M^{-1}$ is uniquely defined. Writing $X_i = M^{-1}\{(i + S_i)/l\}$, we see that the successive points $0 < X_1 < X_2 < \ldots$ form a Poisson process with intensity function $l\mu(t) = \lambda(t)$.

Put $\alpha_i = M^{-1}(i/l)$. Let $0 < \epsilon < \frac{1}{2}$ and let $\zeta, B_1, B_2, B_3, B_4$ denote positive constants such that $B_1 \leq \inf_I \mu \leq \sup_I \mu \leq B_2$, $\sup_I |\mu'| \leq B_3$, and $|(M^{-1})''(x) - (M^{-1})''(y)| \leq B_4 |x - y|^{\zeta}$ for $x, y \in I$. (The existence of $B_4$ and $\zeta$ follows from the Hölder continuity of $\mu'$.) Since $K''$ satisfies a Lipschitz condition of order 1 then there exists a constant $C > 0$ such that $|K''(x) - K''(y)| \leq C|x - y|$ for all real $x$ and $y$. We shall write $\sup |K'''|$ for $C$; this notation is suggestive of the meaning of $C$, even if $K'''$ is not well-defined. Recall that $c > 0$ is sufficiently large for $(-c, c)$ to contain the support of $K$. Given $\epsilon \in (0, \frac{1}{2})$, write $C_1, C_2, \ldots$ for generic positive constants depending on $B_1, \ldots, B_4, c$, $\sup |K'''|$, $\epsilon$ and $\zeta$; any other variables on which they depend will be indicated as arguments.

We may define the Poisson process $\{X_i\}$ arbitrarily on $(1, \infty)$, since points in that range do not influence $\hat{\lambda}(x)$ if $ch \leq x \leq 1 - ch$. In particular, we may assume that $\mu$ is defined on $(0, \infty)$ in such a way that $B_1 \leq \inf_{(0, \infty)} \mu \leq \sup_{(0, \infty)} \mu \leq B_2$, $\sup_{(0, \infty)} |\mu'| \leq B_3$, $|(M^{-1})''(x) - (M^{-1})''(y)| \leq B_4 |x - y|^{\zeta}$ for all $x, y \in (0, \infty)$.

We claim that $C_1(k)$ may be chosen so that the random function

$$Q_1(x) \equiv \hat{\lambda}(x) - h^{-1} \sum_i K\{(x - \alpha_i)/h\} + (h^2 l)^{-1} \sum_i S_i \mu(\alpha_i)^{-1} K'(x - \alpha_i)/h$$

$$- \frac{1}{2} l^{-2} \sum_i S_i^2 \left[ h^{-2} \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'(x - \alpha_i)/h \right]$$

$$+ h^{-3} \mu(\alpha_i)^{-2} K''(x - \alpha_i)/h \right]$$

satisfies

$$\mathbb{E} \left\{ \sup_{0 \leq t \leq 1 - \epsilon} |Q_1(x)|^k \right\} \leq C_1(k) \left( h^{-1} l^{-\zeta/2} + h^{-3} l^{-1/2} \right)^k$$

for all $k \geq 1$. 

84
To verify this claim, put $\Delta_i = -(X_i - \alpha_i)$ and observe that

$$K\{(x - X_i)/h\} = K\{(x - \alpha_i)/h + \Delta_i/h\} = K\{(x - \alpha_i)/h\} + \sum_{j=1}^{2} j^{-1}(\Delta_i/h)^j K^{(j)}\{(x - \alpha_i)/h\} + R_i(x)$$

where

$$R_i(x) = (\Delta_i/h)^2 \int_0^1 (1 - t)[K''\{(x - \alpha_i)/h + t\Delta_i/h\} - K''\{(x - \alpha_i)/h\}] dt.$$ 

Using the Lipschitz continuity of $K''$, and noting that $K\{(x - X_i)/h\} > 0$ if $|x - \alpha_i| \leq ch$ or if $|x - \alpha_i| > ch$ and $|x - X_i| \leq ch$, we obtain

$$|R_i(x)| \leq |\Delta_i/h|^3 (\sup |K'''|) \left\{ I(|x - \alpha_i| \leq ch) + I(|x - \alpha_i| > ch, |x - X_i| \leq ch) \right\}$$

$$\leq |\Delta_i/h|^3 (\sup |K'''|) \left\{ I(|x - \alpha_i| \leq ch) + I(ch < |x - \alpha_i| \leq 2ch, |x - X_i| \leq ch) + I(\Delta > ch, |x - \alpha_i| > 2ch, |x - X_i| \leq ch) + I(\Delta_i \leq ch, |x - \alpha_i| > 2ch, |x - X_i| \leq ch) \right\}$$

$$\leq |\Delta_i/h|^3 (\sup |K'''|) \left\{ I(|x - \alpha_i| \leq 2ch) + I(|\Delta| > ch, |x - X_i| \leq ch) \right\}.$$ 

Therefore, if $ch \leq x \leq 1 - ch$, then

$$\widetilde{\lambda}(x) = h^{-1} \sum_i K\{(x - X_i)/h\} = h^{-1} \sum_i K\{(x - \alpha_i)/h\}$$

$$+ h^{-1} \sum_{j=1}^{2} j^{-1}(\Delta_i/h)^j K^{(j)}\{(x - \alpha_i)/h\} + R_i(x), \quad (3.21)$$

where

$$|R_i(x)| \leq (\sup |K'''|) h^{-4} \sum_i |\Delta_i|^3 \left\{ I(|x - \alpha_i| \leq 2ch) + I(|\Delta| > ch, |x - X_i| \leq ch) \right\}. \quad (3.22)$$

Now,

$$X_i = M^{-1}\{(i + S_i)/l\}$$

$$= \alpha_i + (S_i/l) (M^{-1})'(i/l) + \frac{1}{2} (S_i/l)^2 (M^{-1})''(i/l)$$

$$+ (S_i/l)^2 \int_0^1 (1 - t) \left\{ (M^{-1})''(i/l + tS_i/l) - (M^{-1})''(i/l) \right\} dt,$$
so that by the Hölder continuity of \( \mu' \), we have

\[
|\Delta_i + (S_i/l)(M^{-1})(i/l) + \frac{1}{2} (S_i/l)^2 (M^{-1})''(i/l)| \leq B_4|S_i/l|^{2+\zeta}.
\] (3.23)

Assuming that \( \epsilon \leq x \leq 1 - \epsilon \) and \( 0 < h \leq C_2 \), with \( C_2 \) chosen sufficiently small, by (3.21), (3.22) and (3.23),

\[
\hat{\lambda}(x) = h^{-1} \sum_i K\{(x - \alpha_i)/h\} - (h^2 l)^{-1} \sum_i S_i (M^{-1})'(i/l) K'\{(x - \alpha_i)/h\}
- \frac{1}{2} h^{-2} \sum_i S_i^2 \left[ h^{-2} (M^{-1})''(i/l) K''\{(x - \alpha_i)/h\} \right.
- \left. h^{-3} (M^{-1})'(i/l)^2 K''\{(x - \alpha_i)/h\}\right] + Q_1(x)
\] (3.24)

where

\[
|Q_1(x)| \leq C_3 \sum \left\{ \sum_{j=2}^3 h^{-j}|S_i/l|^{j+\zeta} + h^{-4}(|S_i/l|^3 + |S_i/l|^6 + \zeta) \right\} \times \left\{ I(|x - \alpha_i| \leq 2ch) + I(|S_i| > C_4 hl, |x - X_i| \leq ch) \right\}.
\]

Note that \( |S_i| \leq C_4 hl \) implies \( |\Delta_i| \leq ch \) using (3.23), so that \( I(|\Delta_i| > ch, |x - X_i| \leq ch) \leq I(|S_i| > C_4 hl, |x - X_i| \leq ch) \).

Let \( n \) denote the largest \( i \) such that \( \alpha_i \in (0, 1) \). Then for any \( \beta > 0 \),

\[
T = \sup_{\epsilon \leq x \leq 1-\epsilon} \sum_i |S_i|^{\beta} \left\{ I(|x - \alpha_i| \leq 2ch) + I(|S_i| > C_4 hl, |x - X_i| \leq ch) \right\}
= \sum_i |S_i|^{\beta} \sup_{\epsilon \leq x \leq 1-\epsilon} I(|x - \alpha_i| \leq 2ch)
+ \sum_i |S_i|^{\beta} \sup_{\epsilon \leq x \leq 1-\epsilon} I(|S_i| > C_4 hl, |x - X_i| \leq ch)
\leq \left( \sup_{1 \leq i \leq n} |S_i|^{\beta} \right) \sup_{\epsilon \leq x \leq 1-\epsilon} \sum_i I(|x - \alpha_i| \leq 2ch)
+ \sum_i |S_i|^{\beta} I(|S_i| > C_4 hl, 0 < X_i \leq 1),
\]

whence for any \( k \geq 1 \),

\[
E(T^k) \leq 2^k \left[ E \left( \sup_{1 \leq i \leq n} |S_i|^{\beta} \right) \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} \sum_i I(|x - \alpha_i| \leq 2ch) \right\}^k \right.
+ E \left\{ \sum_i |S_i|^{\beta} I(|S_i| > C_4 hl, 0 < X_i \leq 1) \right\}^k \right]
\leq C_5(k) l^{3k/2}(hl)^k + o\{l^{3k/2}(hl)^k\}.
\]

86
using Corollary 3.3, and Lemmas 3.2 and 3.1. Arguing thus we may prove that for all \( k \geq 1 \),
\[
E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |Q_1(x)|^k \right\} \leq C_5(k) (hl)^k \left( \sum_{j=2}^{3} h^{-j} l^{-(j+1)/2} + h^{-4} l^{-3/2} \right)^k \leq C_7(k) (h^{-1} l^{-\zeta/2} + h^{-3} l^{-1/2})^k.
\]

The claimed results (3.19) and (3.20) follow from this inequality and (3.24), noting that \((M^{-1})'(i/l) = \{\mu(\alpha_i)\}^{-1}\) and \((M^{-1})''(i/l) = -\{\mu(\alpha_i)\}^{-3}\mu'(\alpha_i)\).

**Step 2: Removal of quadratic term from (3.18)**

We shall show that
\[
Q_2(x) \equiv \tilde{\lambda}(x) - h^{-1} \sum_i K\{(x - \alpha_i)/h\} + (h^2 l)^{-1} \sum_i S_i \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\}
\]
satisfies
\[
E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |Q_2(x)|^k \right\} \leq C_8(k) (h^2 l)^{-k} (1 + hl)^k l^\beta
\]
for all \( k \geq 1 \), where \( \beta > 0 \) does not depend on \( k \).

We shall invoke (3.19) and (3.20) and prove that
\[
E \left[ \sup_{\epsilon \leq x \leq 1-\epsilon} \left| \sum_i S_i^2 \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \right|^k \right] \leq C_9(k) (hl)^k (1 + hl)^k l^\beta
\]
(3.26)

\[
E \left[ \sup_{\epsilon \leq x \leq 1-\epsilon} \left| \sum_i S_i^2 \mu(\alpha_i)^{-3} \mu'(\alpha_i) K''\{(x - \alpha_i)/h\} \right|^k \right] \leq C_{10}(k) (1 + h^{1+\zeta} l^k)^k l^{k+\beta}.
\]
(3.27)

We derive (3.26) in detail but give only an outline of the proof of (3.27), as it is almost identical.

Write
\[
\sum_i S_i^2 \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} = v(x) + V_1(x) + 2V_2(x),
\]
where
\[
v(x) = \sum_i i \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\},
\]
\[
V_1(x) = \sum_i \left\{ \sum_{j=1}^{i} (Z_j - 1) \right\} \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\},
\]
\[
V_2(x) = \sum_i \left( \sum_{1 \leq j_1 < j_2 \leq i} Z_{j_1} Z_{j_2} \right) \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\},
\]

87
and let
\[ a_j(x) = \sum_{i=j}^{\infty} \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\}. \]

Now,
\[ V_1(x) = \sum_{i=1}^{\infty} \left\{ \sum_{j=1}^{i} (Z_j^2 - 1) \right\} \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \]
\[ = \sum_{j=1}^{\infty} (Z_j^2 - 1) \sum_{i=j}^{\infty} \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \]
\[ = \sum_{j=1}^{\infty} a_j(x) (Z_j^2 - 1). \]

Applying Lemma 3.4 on page 102, we have
\[ E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_1(x)|^k \right\} \leq C_{11}(k) (hl)^{k/2} (1 + hl)^k \beta, \] (3.28)
for \( k \geq 1 \) and some \( \beta > 0 \) not depending on \( k \).

Observe next that
\[ V_2(x) = \sum_{i=1}^{\infty} \left( \sum_{1 \leq j_1 < j_2 \leq i} Z_{j_1} Z_{j_2} \right) \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \]
\[ = \sum_{1 \leq j_1 < j_2 \leq j_2} Z_{j_1} Z_{j_2} \sum_{i=j_2}^{\infty} \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \]
\[ = \sum_{j_2=2}^{\infty} \left\{ Z_{j_2} \sum_{j_1=1}^{j_2-1} a_{j_2}(x) \right\} \]
\[ = \sum_{j_2=2}^{\infty} a_{j_2}(x) Y_{j_2} \text{ where } Y_{j} = Z_j \sum_{i=1}^{j-1} Z_i. \]

Application of Lemma 3.5 on page 104 gives
\[ E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_2(x)|^k \right\} \leq C_{12}(k) (hl)^{k} (1 + hl)^k \beta. \] (3.29)
Then from (A.8) in Lemma 3.6, (3.28) and (3.29), we have
\[ E \left[ \sup_{\epsilon \leq x \leq 1 - \epsilon} \left| \sum_{i} S_i^2 \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \right|^k \right] \]
\[ \leq \sup_{\epsilon \leq x \leq 1 - \epsilon} |v(x)|^k + E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_1(x)|^k \right\} + \left\{ E \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_2(x)|^k \right\} \]
\[ \leq C_{13}(k) (hl)^{k} (1 + hl)^k \beta. \]

Similarly write,
\[ \sum_{i} S_i^2 \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\} = w(x) + W_1(x) + 2W_2(x) \]
where
\[ w(x) = \sum_i i \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\}, \]
\[ W_1(x) = \sum_i \left\{ \sum_{j=1}^i (Z_j^2 - 1) \right\} \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\}, \]
\[ W_2(x) = \sum_i \left( \sum_{1 \leq j < i \leq i} Z_j Z_i \right) \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\}, \]
and let
\[ b_j(x) = \sum_{i=j}^{\infty} \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\}. \]

In Lemma 3.6 we show that \(|b_j(x)| \leq C_{14}(1 + hl) I(x < C_{15} hl)\). Using the same arguments as those employed in the first part of this step, we find that
\[ |w(x)|^k \leq C_{16}(k)^k (1 + h^{1+\epsilon} l)^k \text{ uniformly in } x, \]
\[ E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |W_1(x)|^k \right\} \leq C_{17}(k) (hl)^{k/2} (1 + hl)^k l^{\beta}, \]
\[ E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |W_2(x)|^k \right\} \leq C_{18}(k) (hl)^k (1 + hl)^k l^{\beta}, \]
and so
\[ E \left[ \sup_{\epsilon \leq x \leq 1-\epsilon} \left| \sum_i S_i^2 \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\} \right|^k \right] \leq C_{19}(k) (1 + h^{1+\epsilon} l)^k l^{k+\beta}. \]

Finally, (3.20), (3.19), (3.26) and (3.27) lead to
\[ E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |Q_2(x)|^k \right\} \leq E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |Q_1(x)|^k \right\} + E \left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |Q_1(x) - Q_2(x)|^k \right\} \]
\[ \leq C_{20}(k) \left( h^{-1} l^{-1} + h^{-3} l^{-1/2} \right)^k \]
\[ + (h^3 l^2)^{-k} C_{21}(k) (hl)^k (1 + hl)^k l^{\beta} \]
\[ + (h^2 l)^{-k} C_{22}(k) (1 + h^{1+\epsilon} l)^k l^{k+\beta} \]
\[ \leq C_{23}(k) (h^2 l)^{-k} (1 + hl)^k l^{\beta}. \]

Step 3: Approximation by Gaussian process

If \( N_1, N_2, \ldots \) are independent standard Normal random variables, and \( T_i = \sum_{j=1}^i N_j \), by Theorem 1 of Komlós et al (1975), we have
\[ P\left( \max_{1 \leq i \leq j} |S_i - T_i| > K_1 \log j + x \right) \leq K_2 e^{-\lambda x}, \]

89
for all \( x > 0 \) and all \( j \geq 1 \), where \( K_1, K_2 \) and \( \lambda \) are positive constants. For these \( T_i \)'s, define
\[
Q_3(x) \equiv \hat{J}(x) - h^{-1} \sum_i K\{(x - \alpha_i)/h\} + (h^2 l)^{-1} \sum_i T_i \mu(\alpha_i)^{-1} K'\{(x - \alpha_i)/h\}.
\]

(3.30)

In the present step we shall prove that
\[
E \left\{ \sup_{e \leq x \leq 1 - \epsilon} |Q_3(x)|^k \right\} \leq C_{24}(k) (h^{-1} \log l)^k l^\beta
\]
for all \( k \geq 1 \) and some \( \beta > 0 \).

Using Lemma 3.10 on page 123, the fact that \( E(X^+)^\alpha = \alpha \int_0^\infty x^{\alpha - 1} \Pr(X \geq x) \, dx \) (see for example Feller (1966, page 148)), and the result of Komlós et al given above, we have
\[
E \left( \max_{1 \leq i \leq j} |S_i - T_i|^k \right) - (K_1 \log j)^k
\]
\[
\leq C_{25}(k) + C_{26}(k) E \left\{ \left( \max_{1 \leq i \leq j} |S_i - T_i| - K_1 \log j \right)^+ \right\}^k
\]
\[
= C_{25}(k) + C_{27}(k) \int_0^\infty x^{k-1} P \left( \max_{1 \leq i \leq j} |S_i - T_i| > K_1 \log j + x \right) \, dx
\]
\[
\leq C_{25}(k) + C_{27}(k) \int_0^\infty x^{k-1} K_2 e^{-\lambda x} \, dx = C_{28}(k),
\]
say. Therefore, if \( n \leq C_{29} l \) denotes the largest \( i \) such that \( \alpha_i \in (0,1) \), using Corollary 3.1 on page 99 we obtain
\[
E \left[ \sup_{e \leq x \leq 1 - \epsilon} \left| \sum_i (S_i - T_i) \mu(\alpha_i)^{-1} K'\{(x - \alpha_i)/h\} \right|^k \right]
\]
\[
\leq E \left( \sup_{1 \leq i \leq n} |S_i - T_i|^k \right) \left[ \sup_{e \leq x \leq 1 - \epsilon} \sum_i \mu(\alpha_i)^{-1} |K'\{(x - \alpha_i)/h\}| \right]^k
\]
\[
\leq \left\{ (K_1 \log n)^k + C_{30}(k) \right\} C_{31}(k) (hl)^k
\]
\[
\leq C_{32}(k) (hl \log l)^k.
\]

Finally, (3.25) and (3.30) and the result above lead to
\[
E \left\{ \sup_{e \leq x \leq 1 - \epsilon} |Q_3(x)|^k \right\} \leq E \left\{ \sup_{e \leq x \leq 1 - \epsilon} |Q_2(x)|^k \right\} + E \left\{ \sup_{e \leq x \leq 1 - \epsilon} |Q_2(x) - Q_3(x)|^k \right\}
\]
\[
\leq C_{33}(k) (h^2 l)^{-k} (1 + hl)^k l^\beta
\]
\[
+ (h^2 l)^{-k} C_{34}(k) (hl \log l)^k
\]
\[
\leq C_{35}(k) (h^{-1} \log l)^k l^\beta.
\]

90
Step 4: Simplified Gaussian approximation

Here we show that there exists a standard Wiener process \( W \), whose construction depends on \( l \), such that with

\[
J(x) = \int W\{M(x - hy)\} K'(y) \, dy
\]

and

\[
Q_4(x) \equiv \tilde{\lambda}(x) - h^{-1} \sum_i K\{(x - \alpha_i)/h\} - h^{-1} \log l \, I^2 J(x)
\]

we have

\[
E \{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |Q_4(x)|^k \} \leq C_36(k) (h^{-1} \log l)^k \, t^\beta
\]

for all \( k \geq 1 \) and some \( \beta > 0 \).

Let \( W_1 \) denote a standard Wiener process such that \( N_i = W_1(i) - W_1(i - 1) \) for \( i \geq 1 \), and define the integral approximation to the sum \( \sum_i T_i \mu(\alpha_i)^{-1} K'\{(x - \alpha_i)/h\} \) in (3.30) by

\[
I(x) = \int W_1(y) \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h] \, dy.
\]

If we take \( W(t) = l^{-1/2} W_1(lt) \) then \( I(x) = -hl^{1/2} J(x) \). For \( y \in [i - 1, i) \) and \( i \geq 1 \), put \( V(y) = W_1(y) - W_1(i - 1) \). Then \( W_1(y) = \sum_{0 \leq j < i - 1} N_j + V(y) \), (where \( N_0 = 0 \)), and so \( I(x) = I_1(x) + I_2(x) \), where

\[
I_1(x) = \sum_{i = 1}^\infty \int_{i - 1}^i \left( \sum_{j = 0}^{i - 1} N_j \right) \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h] \, dy
\]

\[
= \sum_{j = 1}^\infty \int_{j + 1}^\infty \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h] \, dy
\]

\[
= \sum_{j = 1}^\infty \int_j^\infty \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h] \, dy
\]

and

\[
I_2(x) = \sum_{i = 1}^\infty \int_{i - 1}^i V(y) \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h] \, dy.
\]

Since \( V(y) \) is a Gaussian process with mean zero, \( I_2 \) is Normally distributed with mean zero, and variance equal to

\[
\sum_{i = 1}^\infty \text{var} \left( \int_{i - 1}^i V(y) \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h] \, dy \right)
\]
\[
\leq \mathbb{E}\left\{ \sup_{0 \leq y \leq 1} |V(y)|^2 \right\} \sum_{i=1}^{\infty} \left( \int_{-1}^{i} \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h]\,dy \right)^2
\]

Now, \( \mathbb{E}\left\{ \sup_{0 \leq y \leq 1} |V(y)|^2 \right\} \leq 1 \), and terms in the sum are non-zero if either \( |x - \alpha_i| < ch \) or \( |x - \alpha_i - 1| < ch \). Since \( \mu \) and \( K' \) are bounded we have

\[
\text{var} I_2(x) \leq C_{37} \sum_{i=1}^{\infty} I(|x - \alpha_i| \leq ch) \leq C_{38} h l
\]

uniformly in \( 0 < x \leq 1 \). Corollary 3.3 gives \( \mathbb{E}|I_2(x)|^k \leq C_{39}(k) (hl)^{k/2} \) for \( k \geq 1 \).

Using the Hölder continuity of \( M^{-1} \) and \( K' \) and a technique similar to that used in Lemma 3.4, we obtain

\[
\mathbb{E}\left\{ \sup_{\epsilon \leq x \leq 1-\epsilon} |I_2(x)|^k \right\} \leq C_{40}(k) (hl)^{k/2} l^\beta. \tag{3.33}
\]

Define \( \beta_j(x) = \sum_{i \geq j} \mu(\alpha_i)^{-1} K'\{(x - \alpha_i)/h\} \) and

\[
\beta_j^2(x) = \int_0^\infty \mu\{M^{-1}(y/l)\}^{-1} K'[\{x - M^{-1}(y/l)\}/h]\,dy
= hl K\{(x - \alpha_j)/h\}.
\]

Since \( |\beta_j(x) - \beta_j^2(x)| \leq C_{41} I(j \leq C_{42} h l) \) uniformly in \( 0 < x < 1 \) and \( 0 < h \leq C_{43} \), for \( k \geq 1 \) we have

\[
\mathbb{E}\left| \sum_{j=1}^{\infty} N_j(\beta_j(x) - \beta_j^2(x)) \right|^k \leq C_{44}(k) \mathbb{E}\left( \sum_{j=1}^{C_{45} h l} N_j \right)^k
\leq C_{46}(k) (hl)^{k/2} \quad \text{(applying Corollary 3.3).}
\]

Again, noting the Hölder continuity of \( M^{-1} \) and \( K' \) and using the technique of Lemma 3.4 it is easily proved that

\[
\mathbb{E}\left[ \sup_{\epsilon \leq x \leq 1-\epsilon} \left| \sum_{j=1}^{\infty} N_j(\beta_j(x) - \beta_j^2(x)) \right|^k \right] \leq C_{47}(k) (hl)^{k/2} l^\beta. \tag{3.34}
\]

Now,

\[
\sum_{j=1}^{\infty} N_j \beta_j(x) = \sum_i T_i \mu(\alpha_i) K'\{(x - \alpha_i)/h\} \quad \text{and} \quad \sum_{j=1}^{\infty} N_j \beta_j^2(x) = I_1(x).
\]

Therefore,

\[
\sum_i T_i \mu(\alpha_i) K'\{(x - \alpha_i)/h\} + hl^{3/2} J(x) = \sum_{j=1}^{\infty} N_j(\beta_j(x) - \beta_j^2(x)) - I_2(x).
\]

92
Finally, using (3.31) (3.33) and (3.34), we have

\[
E \left\{ \sup_{\varepsilon \leq x \leq 1-\varepsilon} |Q_4(x)|^k \right\} \leq E \left\{ \sup_{\varepsilon \leq x \leq 1-\varepsilon} |Q_3(x)|^k \right\} + E \left\{ \sup_{\varepsilon \leq x \leq 1-\varepsilon} |Q_3(x) - Q_4(x)|^k \right\} \\
\leq C_{46}(k) (h^{-1} \log l)^k l^\beta + (h^2)^{-k} C_{49}(k) (hl)^{k/2} l^\beta \\
\leq C_{50}(k) (h^{-1} \log l)^k l^\beta.
\]

**Step 5: Completion**

Result (3.32) implies that

\[
E \left\{ \sup_{\varepsilon \leq x \leq 1-\varepsilon} |\hat{\lambda}(x) - E\hat{\lambda}(x) - h^{-1/2} J(x)|^k \right\} \leq C_{51}(k) (h^{-1} \log l)^k l^\beta,
\]

which establishes the non-bootstrap part of Theorem 3.2.

**Step 6: Bootstrap part of Theorem 3.2**

To derive the bootstrap part, first put \( \hat{\mu} = l^{-1} \hat{\lambda} \) and \( \hat{M}(x) = \int_{0<y<x} \hat{\mu}(y) \, dy \). Let \( \varepsilon' \in (0, \varepsilon) \), and observe that in the case of the first sort of resampling (where \( X_1^*, X_2^*, \ldots \) are drawn from a Poisson process with intensity \( \hat{\lambda} \), conditional on \( \mathcal{X} \)), there exist positive constants \( B_1, \ldots, B_4 \) and \( \zeta \) such that the event \( \mathcal{E} \) that \( B_1 \leq \inf_{I_{\varepsilon'}} \hat{\mu} \leq \sup_{I_{\varepsilon'}} \hat{\mu} \leq B_2, \sup_{I_{\varepsilon'}} |\hat{\mu}'| \leq B_3, \) and \( |(\hat{M}^{-1})''(x) - (\hat{M}^{-1})''(y)| \leq B_4 |x-y|^{\zeta} \) for all \( x, y \in I_{\varepsilon'} \), satisfies

\[
P(\mathcal{E}) = 1 - O(l^{-k})
\]

for all \( k \geq 1 \). We prove in Lemma 3.9 that part of this result pertaining to Hölder continuity of \( (\hat{M}^{-1})'' \), namely that \( P\{ |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq B |x-y|^{\zeta} \) for all \( x, y \in I_{\varepsilon'} \} = 1 - O(l^{-k}) \) for some \( B, \zeta > 0 \) and all \( k > 0 \). If the constants \( C_j \) in Steps 1–4 are permitted to depend on \( \varepsilon' \) as well as \( B_1, \ldots, B_4, c, \sup |K'''|, \varepsilon \) and \( \zeta \), then the argument given in those steps may be employed to prove the existence of a standard Brownian motion \( \hat{W}^* \) such that, writing

\[
J^*(x) = \int W^* \{ \hat{M}(x - hy) \} K'(y) \, dy,
\]

we have

\[
E \left\{ I(\mathcal{E}) \sup_{\varepsilon \leq x \leq 1-\varepsilon} |\hat{\lambda}^*(x) - E\hat{\lambda}^*(x) - h^{-1/2} J^*(x)|^k \right\} \leq C_{52}(k) (h^{-1} \log l)^k.
\]
An argument founded on the modulus of continuity of a Wiener process is used in Lemma 3.7 to prove that with

$$J^\dagger(x) = \int W^\ast \{M(x - hy)\} K'(y) \, dy$$

we have

$$E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |J^\ast(x) - J^\dagger(x)|^k \right\} \leq C_{53}(k) (\log t)^k,$$

and so

$$E \left\{ I(\mathcal{E}) \sup_{\epsilon \leq x \leq 1 - \epsilon} |\hat{\lambda}^\ast(x) - E' \hat{\lambda}^\ast(x) - h^{-1} t^{1/2} J^\dagger(x)|^k \right\} \leq C_{54}(k) (\log t)^k.$$

In view of (3.36), by simple application of the Cauchy-Schwartz inequality, we may remove the quantity $I(\mathcal{E})$ from the expectation on the left-hand side, so proving a version of (3.35) in which $\hat{\lambda}$, $E \hat{\lambda}$, $J$ are replaced by $\hat{\lambda}^\ast$, $E' \hat{\lambda}^\ast$ and $J^\ast$, respectively, and the right hand side is slightly larger in magnitude. This establishes the bootstrap part of the theorem, in the case of the first type of resampling.
3.5.2 Proof of Theorem 3.3

By (3.19) and (3.20), with \( k = 1 \) in the latter,

\[
E \hat{\lambda}(x) = h^{-1} \sum_i K\{(x - \alpha_i)/h\} + \frac{1}{2} l^{-2} \{h^{-3} v(x) + h^{-2} w(x)\}
+ O(h^{-1+\zeta} + h^{-3} l^{-1/2}),
\]

where \( v \) and \( w \) are as in Step 2 of the proof of Theorem 3.2, and the remainder term is of the stated order uniformly in \( \epsilon \leq x \leq 1 - \epsilon \) for any \( ch < \epsilon < \frac{1}{2} \). In Lemma 3.6 we prove that \( v(x) = O(l + h^{2+\zeta} l^2) \) and \( w(x) = O(l + h^{1+\zeta} l^2) \) uniformly in such \( x \). Thus,

\[
E \hat{\lambda}(x) = h^{-1} \sum_i K\{(x - \alpha_i)/h\} + O(h^{-1+\zeta} + h^{-3} l^{-1/2})
\]

(3.37)

uniformly in \( x \).

Put

\[
L(x) = h^{-1} \int K\{(x - M^{-1}(y/l))/h\} dy = \int l\mu(x - hy) K(y) dy.
\]

Using the method of Appendix A3.3 to bound \( h^{-1} \sum_i K\{(x - \alpha_i)/h\} \) and noting that the integrand in \( L(x) \) above has one derivative, we obtain

\[
h^{-1} \sum_i K\{(x - \alpha_i)/h\} = L(x) + O\{l(hl)^{-2}\} = L(x) + O\{(h^2 l)^{-1}\},
\]

uniformly in \( \epsilon \leq x \leq 1 - \epsilon \). Theorem 3.3 in the nonbootstrap case follows from this formula and (3.37). The argument in the first bootstrap case is similar.
Appendices

A.1

Notation

Throughout the following appendices, we use the same notation, definitions and conditions as in Theorem 3.2. In particular, we write $C_1, C_2, \ldots$ for generic positive constants depending on $B_1, \ldots, B_4, c, \sup |K'''|$, $\epsilon$ and $\zeta$; any other variables on which they depend will be indicated as arguments.

Recall also that it is a condition of all the theorems that for some $\xi > 0$, $l^{-\xi} \geq h \geq l^{-1/4+\xi}$.
Appendix 3A.1

Lemmas bounding \( \mathbb{E} \left\{ \sup_{|x| \leq 1 - \epsilon} |Q_1(x)|^k \right\} \)

Lemma 3.1 and 3.2 are used in the first step of Theorem 3.2 to bound \( \mathbb{E} T_k \) and hence \( \mathbb{E} \left\{ \sup_{|x| \leq 1 - \epsilon} |Q_1(x)|^k \right\} \).

The first lemma holds because when \( i \) is small, \( P(0 < X_i \leq 1) \) is high but \( P(|S_i| > C_1 hl) \) is low, so that \( P(|S_i| > C_1 hl, 0 < X_i \leq 1) \) is low. On the other hand, when \( i \) is large, \( P(0 < X_i \leq 1) \) is low but \( P(|S_i| > C_1 hl) \) is high so that \( P(|S_i| > C_1 hl, 0 < X_i \leq 1) \) is still low.

**Lemma 3.1** For \( k \geq 1 \),

\[
\mathbb{E} \left\{ \sum_i |S_i|^\beta I(|S_i| > C_1 hl, 0 < X_i \leq 1) \right\}^k = o \{l^{\beta k/2}(hl)^k \}.
\]

**Proof** First, note that applying the Cauchy-Schwartz inequality, we have

\[
P(|S_i| > C_1 hl, 0 < X_i \leq 1) = \mathbb{E} \{I(|S_i| > C_1 hl, 0 < X_i \leq 1)\}
\leq \left[ \mathbb{E} \{I^2(|S_i| > C_1 hl)\} \mathbb{E} \{I^2(0 < X_i \leq 1)\} \right]^{1/2}
= \left\{ P(|S_i| > C_1 hl) P(0 < X_i \leq 1) \right\}^{1/2}.
\]  

(A.1)

For brevity, in the step below we shall write \( I_i = I(|S_i| > C_1 hl, 0 < X_i \leq 1) \).

Applying the Cauchy-Schwartz inequality, Hölder’s inequality and (A.1), for \( k \geq 1 \) we have

\[
\mathbb{E} \left( \sum_i |S_i|^\beta I_i \right)^k = \sum_{i_1} \cdots \sum_{i_k} \mathbb{E} \left( |S_{i_1}|^\beta I_{i_1} \times \cdots \times |S_{i_k}|^\beta I_{i_k} \right)
\leq \sum_{i_1} \cdots \sum_{i_k} \left\{ \mathbb{E} |S_{i_1} \cdots S_{i_k}|^{2\beta} \right\} \left\{ \mathbb{E} (I_{i_1} \cdots I_{i_k})^2 \right\}^{1/2}
\]

97
\[ \sum \prod_{i=1}^{k} \left( \mathbb{E}[S_i | S_i]^{2 \beta k} \right)^{1/(2k)} \left( \mathbb{E}(I_{S_i}) \right)^{1/(2k)} \]

\[ = \left[ \sum \left\{ \mathbb{E}[S_i | S_i]^{2 \beta k} P(|S_i| > C_{1} h l, 0 < X_i \leq 1) \right\}^{1/(2k)} \right]^{k} \]

\[ \leq \left[ \sum \left\{ \mathbb{E}[S_i | S_i]^{4 \beta k} P(|S_i| > C_{1} h l) P(0 < X_i \leq 1) \right\}^{1/(4k)} \right]^{k}. \]

Let \( N \) be the number of points of the Poisson process \( \{X_i\} \) in \( I = (0, 1) \), and note that \( N \) is Poisson with mean \( \int_{0}^{1} l \mu(x) dx = l m_1 \), say. Then, using Markov’s inequality, we have

\[ P(0 < X_i \leq 1) = P(N \geq i) \leq \begin{cases} C_2(k_1) l^{k_1} (i - l m_1)^{-2k_1} & \text{if } i > l m_1 \\ 1 & \text{if } i \leq l m_1. \end{cases} \]

In Corollary 3.3 on page 102 we show that \( \mathbb{E}[S_i | S_i]^{k_2} \leq C_3(k_2) \left( \text{var } S_i \right)^{k_2/2} \) for \( k_2 \geq 1 \), so that

\[ P(|S_i| > C_{1} h l) \leq (C_{1} h l)^{-k_2} \mathbb{E}[S_i | S_i]^{k_2} \leq (C_{1} h l)^{-k_2} C_3(k_2) i^{k_2/2} \leq \begin{cases} C_4(k_2) l^{-k_2/5} & \text{if } i \leq C_5 l \\ 1 & \text{if } i > C_5 l. \end{cases} \]

Thus,

\[ \sum \left\{ \mathbb{E}[S_i | S_i]^{4 \beta k} P(|S_i| > C_{1} h l) P(0 < X_i \leq 1) \right\}^{1/(4k)} \]

\[ \leq \sum_{i=1}^{C_4 l} \left\{ i^{2 \beta k} C_4(k_2) l^{-k_2/5} \right\}^{1/(4k)} + \sum_{i=C_4 l+1}^{\infty} \left\{ i^{2 \beta k} C_2(k_1) l^{k_1} \right\}^{1/(4k)} \]

\[ \leq \sum_{i=1}^{C_4 l} \left\{ i^{2 \beta k} C_4(k_2) l^{-k_2/5} \right\}^{1/(4k)} \]

\[ + \sum_{j=l(C_6 - m_1) + 1}^{\infty} \left[ (j + l m_1)^{2 \beta k} C_2(k_1) \{ j/(C_6 - m_1) \}^{k_1} \right]^{1/(4k)} \]

\[ \leq C_7(k, k_2) l^{-k_2/20k} \sum_{i=1}^{C_4 l} i^{\beta/2} \]

\[ + C_8(k, k_1) \sum_{j=l(C_6 - m_1) + 1}^{\infty} \left\{ (j + l m_1)^{2 \beta k} \right\}^{1/(4k)}. \]

98
It is simple to show that both terms in this expression converge to constants for appropriate choices of $\beta$, $k$, $k_1$ and $k_2$, so that
\[
\left[ \sum_i \left\{ E|S_i|^{4\beta k} P(|S_i| > C_1 hl) P(0 < X_i \leq 1) \right\}^{1/(4k)} \right]^k \leq C_9(\beta, k, k_1, k_2) = o\{l^{3k/2}(hl)^k\},
\]
proving the lemma.

Lemma 3.2 gives an upper bound for the number of $\alpha_i$ within a distance $2ch$ of some point $x$. Its corollary is used in Step 3 of the proof of Theorem 3.2.

**Lemma 3.2** For all $k > 0$,
\[
\sup_{\epsilon \leq \epsilon_1 \leq 1-\epsilon} \left\{ \sum_i I(|x - \alpha_i| \leq 2ch) \right\}^k \leq C(hl)^k.
\]

**Proof** We shall start by finding a lower bound for the distance between successive $\alpha_i$:
\[
|\alpha_{i+1} - \alpha_i| = M^{-1}(i + 1)/l - M^{-1}(i/l) = \int_{i/l}^{(i+1)/l} (M^{-1})'(y) dy = \int_{i/l}^{(i+1)/l} \mu\{M^{-1}(y)\}^{-1} dy \geq (B_2l)^{-1},
\]
since $\sup_x \mu \leq B_2$. Thus,
\[
\sup_{\epsilon \leq \epsilon_1 \leq 1-\epsilon} \sum_i I(|x - \alpha_i| \leq 2ch) \leq (4ch)(B_2l),
\]
and the result follows.

**Corollary 3.1** For all $k > 0$,
\[
\sup_{\epsilon \leq \epsilon_1 \leq 1-\epsilon} \left[ \sum_i \mu(\alpha_i)^{-1} \left| K'\{(x - \alpha_i)/h\} \right| \right]^k \leq C(hl)^k.
\]

**Proof**
\[
\sup_{\epsilon \leq \epsilon_1 \leq 1-\epsilon} \left[ \sum_i \mu(\alpha_i)^{-1} \left| K'\{(x - \alpha_i)/h\} \right| \right] \leq B_1^{-1} \sup_u |K'(u)| \sup_{\epsilon \leq \epsilon_1 \leq 1-\epsilon} \left\{ \sum_i I(|x - \alpha_i| \leq ch) \right\},
\]
and the result follows from Lemma 3.2.
Appendix 3A.2
Martingale results and applications

The main lemmas of this section, Lemmas 3.4 and 3.5, depend on martingale results. We therefore preface this appendix with the definition of a martingale and some preliminary results.

Definition (from Hall and Heyde 1981, page 1)

Let $(\Omega, \mathcal{F}, P)$ be a probability space: $\Omega$ is a set, $\mathcal{F}$ a $\sigma$-field of subsets of $\Omega$, and $P$ a probability measure defined on $\mathcal{F}$. Let $\mathcal{I}$ be any interval of the form $(a, b), [a, b), (a, b]$ or $[a, b]$ of the ordered set $\{-\infty, \ldots, -1, 0, 1, \ldots, \infty\}$. Let $\{\mathcal{F}_n, n \in \mathcal{I}\}$ be an increasing sequence of $\sigma$-fields of $\mathcal{F}$ sets. Suppose that $\{Z_n, n \in \mathcal{I}\}$ is a sequence of random variables on $\Omega$ satisfying

1. $Z_n$ is measurable with respect to $\mathcal{F}_n$,
2. $\mathbb{E}|Z_n| < \infty$,
3. $\mathbb{E}(Z_n|\mathcal{F}_m) = Z_m$ a.s. for all $m < n$, where $m, n \in \mathcal{I}$.

Then the sequence $\{Z_n, n \in \mathcal{I}\}$ is said to be a martingale with respect to $\{\mathcal{F}_n, n \in \mathcal{I}\}$. We write that $\{Z_n, \mathcal{F}_n, n \in \mathcal{I}\}$ is a martingale.

Lemma 3.3 follows simply from the definition of a martingale.
Lemma 3.3 If \( \{Y_i, i \in \mathcal{I}\} \) is a sequence of random variables, \( E|Y_i| < \infty \), \( S_i = \sum_{j=1}^{i} Y_j \), and \( \mathcal{F}_i = \mathcal{F}(Y_1, \ldots, Y_i) \), then \( \{S_i, \mathcal{F}_i, i \in \mathcal{I}\} \) is a martingale if and only if \( E(Y_i|\mathcal{F}_{i-1}) = 0 \).

Proof Suppose that \( E(Y_i|\mathcal{F}_{i-1}) = 0 \). Obviously \( S_i \) is measurable with respect to \( \mathcal{F}_i = \mathcal{F}(Y_1, \ldots, Y_i) \), and

\[
E(S_i|\mathcal{F}_{i-1}) = E\left\{ \left( Y_i + \sum_{j=1}^{i-1} Y_j \right)|\mathcal{F}_{i-1} \right\} = E(Y_i|\mathcal{F}_{i-1}) + E(S_{i-1}|\mathcal{F}_{i-1}) = S_{i-1},
\]

and so by the definition above, \( \{S_i, \mathcal{F}_i, i \in \mathcal{I}\} \) is a martingale.

Now suppose that \( \{S_i, \mathcal{F}_i\} \) is a martingale. Since \( Y_i = S_i - S_{i-1} \),

\[
E(Y_i|\mathcal{F}_{i-1}) = E(S_i|\mathcal{F}_{i-1}) - E(S_{i-1}|\mathcal{F}_{i-1}) = S_{i-1} - E(S_{i-1}|\mathcal{F}_{i-1}) = 0,
\]

as required. \( \blacksquare \)

Corollary 3.2 Suppose \( X_1, \ldots, X_n \) are independent random variables with \( E(X_i) = 0 \). Put \( S_i = \sum_{j=1}^{i} X_j \), \( 1 \leq i \leq n \) and let \( \mathcal{F}_i = \mathcal{F}(X_1, \ldots, X_i) \). Then \( \{S_i, \mathcal{F}_i, 1 \leq i \leq n\} \) is a martingale.

The following Theorem is given in Hall and Heyde (1981, page 23). It is stated here without proof. Its corollary is applied a number of times in the proof of Theorem 3.2 and used in the proof of Lemma 3.4.

Theorem 3.4 If \( \{S_i, \mathcal{F}_i, 1 \leq i \leq n\} \) is a martingale and \( k > 0 \), then there exists a constant \( C \) depending only on \( k \) such that

\[
E\left( \max_{i \leq n} |S_n|^k \right) \leq C(k) \left( E\left[ \sum_{i=1}^{n} E(X_i^2|\mathcal{F}_{i-1}) \right]^{k/2} \right) + E\left( \max_{i \leq n} |X_i|^k \right).
\]
Corollary 3.3 Let \( 1 \leq k < \infty \), and let \( X_1, \ldots, X_n \) be independent and identically distributed random variables with \( \mathbb{E}(X_i) = 0 \) and \( \mathbb{E}|X_i|^k < \infty \). Put \( S_i = \sum_{j=1}^i X_j, 1 \leq i \leq n \). Then

\[
\mathbb{E}(|S_n|^k) \leq C(k) (\text{var } S_n)^{k/2} + o(\text{var } S_n)^{k/2}.
\]

Proof Let \( \mathcal{F}_i = \mathcal{F}(X_1, \ldots, X_i) \). By Corollary 3.2 above, \( \{S_i, \mathcal{F}_i, 1 \leq i \leq n\} \) is a martingale. Then for \( k > 0 \), Theorem 3.4 gives

\[
\mathbb{E}(|S_n|^k) \leq \mathbb{E}\left(\max_{i \leq n} |S_i|^k\right) \leq C(k) \left\{ (\text{var } S_n)^{k/2} + \mathbb{E}\left(\max_{i \leq n} |X_i|^k\right)\right\}
\]

\[
\leq C(k) \left\{ (\text{var } S_n)^{k/2} + \sum_{i=1}^n \mathbb{E}|X_i|^k\right\}
\]

\[
= C(k) \left\{ (\text{var } S_n)^{k/2} + n \mathbb{E}|X_i|^k\right\}.
\]

The result follows for \( k > 2 \) as \( \frac{n \mathbb{E}|X_i|^k}{(\text{var } S_n)^{k/2}} = \frac{n \mathbb{E}|X_i|^k}{(n \mathbb{E}|X_i|^k)^{k/2}} \to 0 \).

For \( k = 2 \), the result is trivial since \( \mathbb{E}S_n^2 = \text{var } S_n \).

For \( 1 \leq k < 2 \), let \( p = 2/k \). Since \( p > 1 \), we may apply Hölder’s inequality obtaining

\[
\mathbb{E}(|S_n|^k) \leq \mathbb{E}\left\{((S_n)^p)^{1/p}\right\} = \mathbb{E}(S_n^2)^{k/2} = (\text{var } S_n)^{k/2},
\]

completing the proof.

Lemma 3.4 Suppose \( Z_1, Z_2, \ldots \) are independent random variables with mean 0 and variance 1. Let \( V_1(x) = \sum_{j=1}^\infty a_j(x) (Z_j^2 - 1) \), where \( a_j(x) \) is Lipschitz continuous and \( |a_j(x)| \leq C_1(1 + hl) I(j < C_2 hl) \) uniformly in \( x \in I_\varepsilon \subset [ch, 1 - ch] \). Then

\[
\mathbb{E}\left\{\sup_{c \leq x \leq 1 - c} |V_1(x)|^k\right\} \leq C_3(k) (hl)^{k/2} (1 + hl)^k \beta \tag{A.2}
\]

for \( k \geq 1 \), some \( \beta > 0 \) not depending on \( k \), and \( x \in I_\varepsilon \).

Proof Now,

\[
|V_1(x)|^k \leq C_4(k) (1 + hl)^k C_2 hl \left| \sum_{j=1}^{C_2 hl} (Z_j^2 - 1) \right|^k.
\]

Put \( S_i = \sum_{j=1}^i (Z_j^2 - 1) \) and \( \mathcal{F}_i = \mathcal{F}\{(Z_1^2 - 1), \ldots, (Z_i^2 - 1)\} \). Then \( \{S_i, \mathcal{F}_i, 1 \leq i \leq C_2 hl\} \) is a martingale by Corollary 3.2. Therefore we may apply Corollary 3.3
obtaining, for \( k \geq 1 \),
\[
E \left[ \sum_{j=1}^{C_{2h}^{\beta}} (Z_j^2 - 1) \right]^k \leq C_5(k) \left[ \text{var} \left( \sum_{j=1}^{C_{2h}^{\beta}} (Z_j^2 - 1) \right) \right]^{k/2} \leq C_6(k)(hl)^{k/2},
\]
and so
\[
E |V_1(x)|^k \leq C_7(k)(hl)^{k/2}(1 + hl)^k.
\]

Now suppose that \( \mathcal{E}(l) \subseteq [\epsilon, 1 - \epsilon] \) has \( O(l^\beta) \) regularly spaced elements for some \( \beta > 0 \). Then
\[
E \left\{ \sup_{x \in \mathcal{E}(l)} |V_1(x)|^k \right\} \leq \sum_{x \in \mathcal{E}(l)} E |V_1(x)|^k \leq C_8(k)(hl)^{k/2}(1 + hl)^k l^\beta. \tag{A.3}
\]

Given \( x \in [\epsilon, 1 - \epsilon] \), let \( x' \) be an element of \( \mathcal{E}(l) \) which minimises \( |x - x'| \). Then using the Lipschitz continuity of \( a_j(x) \), we may choose \( \beta \) so large that
\[
\sup_{\epsilon \leq x \leq 1 - \epsilon} |a_j(x) - a_j(x')|^k \leq (C_9 |x - x'|)^k \leq (C_{10} l^{-\beta})^k \leq C_{11}(k) l^{-k/2}.
\]

Therefore,
\[
E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_1(x) - V_1(x')|^k \right\} \leq \sup_{\epsilon \leq x \leq 1 - \epsilon} |a_j(x) - a_j(x')|^k E \left[ \sum_{j=1}^{C_{2h}^{\beta}} (Z_j^2 - 1) \right]^k \leq C_{12}(k) h^{k/2}. \tag{A.4}
\]

Since
\[
E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_1(x)|^k \right\} \leq E \left\{ \sup_{\epsilon \leq x \leq 1 - \epsilon} |V_1(x) - V_1(x')|^k \right\} + E \left\{ \sup_{x' \in \mathcal{E}(l)} |V_1(x')|^k \right\},
\]
the desired result \( (A.2) \) follows from this expression, \( (A.3) \) and \( (A.4) \).

Rosenthal's inequality, the following theorem, is given in Hall and Heyde (1981, page 23) and is stated here without proof. It is used in the proof of Lemma 3.5.

**Theorem 3.5 (Rosenthal's inequality)** If \( \{ S_i, \mathcal{F}_i, 1 \leq i \leq n \} \) is a martingale and \( 2 \leq k < \infty \), then there exist constants \( C_1 \) and \( C_2 \) depending only on \( k \) such that
\[
C_1(k) \left( E \left[ \left\{ \sum_{i=1}^{n} E(X_i^2 | \mathcal{F}_{i-1}) \right\}^{k/2} \right] + \sum_{i=1}^{n} E|X_i|^k \right) \leq E|S_n|^k \leq C_2(k) \left( E \left[ \left\{ \sum_{i=1}^{n} E(X_i^2 | \mathcal{F}_{i-1}) \right\}^{k/2} \right] + \sum_{i=1}^{n} E|X_i|^k \right).
\]
Lemma 3.5 Suppose that $Z_1, Z_2, \ldots$ are independent random variables with mean 0, variance 1 and $\sup_i E|Z_i|^k < \infty$. Let $Y_j = Z_j \sum_{i=1}^{j-1} Z_i$, $j = 2, 3, \ldots$. For $x \in [\varepsilon, 1-\varepsilon]$, let $V_2(x) = \sum_{j=2}^{\infty} a_j(x)Y_j$, where $a_j(x)$ is Lipschitz continuous, and $|a_j(x)| \leq C_1 (1 + hl) I(j < C_2 hl)$ uniformly in $x$. Then for $k \geq 2$ and some $\beta > 0$ not depending on $k$, 

$$E \left\{ \sup_{\varepsilon \leq x \leq 1-\varepsilon} |V_2(x)|^k \right\} \leq C_5(k) (hl)^k (1 + hl)^k l^\beta.$$  

(A.5)

Proof Now,

$$|V_2(x)| \leq C_1 (1 + hl) \left| \sum_{j=2}^{\infty} Y_j \right|.$$ 

Put $S_i = \sum_{j=2}^{i} Y_j$, $i = 2, 3, \ldots, C_2 hl$, and $\mathcal{F}_i = \mathcal{F}(Z_1, \ldots, Z_i)$. As $E(Y_i|\mathcal{F}_{i-1}) = 0$, by Lemma 3.3 $\{S_i, \mathcal{F}_i, i = 2, \ldots, C_2 hl\}$ is a martingale. We shall therefore find $E \left| \sum_{j=2}^{C_2 hl} Y_j \right|^k$ by applying Rosenthal’s inequality.

For $k \geq 1,$

$$E |Y_j^k| = E |Z_j^k| E \left( \sum_{i=1}^{j-1} |Z_i| \right) \leq C_4(k) (j-1)^k \leq C_4(k) j^{k/2},$$

using the independence of the $Z_i$, and finding $E \left| \sum_{i=1}^{j-1} Z_i \right|^k$ by applying Corollary 3.3.

Also, since

$$E (Y_j^2|\mathcal{F}_{j-1}) = E (Z_j^2|\mathcal{F}_{j-1}) E \left\{ \left( \sum_{i=1}^{j-1} Z_i \right)^2 |\mathcal{F}_{j-1} \right\} = \left( \sum_{i=1}^{j-1} Z_i \right)^2,$$

we have

$$E \left\{ \sum_{j=2}^{C_2 hl} E (Y_j^2|\mathcal{F}_{j-1}) \right\}^{k/2} = E \left\{ \sum_{j=2}^{C_2 hl} \left( \sum_{i=1}^{j-1} Z_i \right)^2 \right\}^{k/2} \leq (C_2 hl)^{k/2-1} \sum_{j=2}^{C_2 hl} E \left| \sum_{i=1}^{j-1} Z_i \right|^k$$

applying Lemma 3.11,

$$\leq (C_2 hl)^{k/2-1} \sum_{j=2}^{C_2 hl} j^{k/2} \leq C_8(k) (hl)^k.$$ 

Thus,

$$E \left| \sum_{j=2}^{C_2 hl} Y_j \right|^k \leq C_6(k) \left( E \left[ \sum_{j=2}^{C_2 hl} E (Y_j^2|\mathcal{F}_{j-1}) \right]^{k/2} \right) + \sum_{j=2}^{C_2 hl} E |Y_j|^k \leq C_7(k) \left\{ (hl)^k + C_8(k) \sum_{j=2}^{C_2 hl} j^{k/2} \right\} \leq C_9(k) (hl)^k.$$ 

104
Now suppose that $E(l) \subseteq [\epsilon, 1-\epsilon]$ has $O(l^\beta)$ regularly spaced elements for some $\beta > 0$. Then

$$E \left\{ \sup_{x \in E(l)} |V_2(x)|^k \right\} \leq \sum_{x \in E(l)} \{E|V_2(x)|^k\} = C_{12}(k) (hl)^k (1 + hl)^k l^\beta. \quad (A.6)$$

Given $x \in [\epsilon, 1-\epsilon]$, let $x'$ be an element of $E(l)$ which minimises $|x - x'|$. Using the Lipschitz continuity of $a_j(x)$, we may choose $\beta$ so large that

$$\sup_{e \leq x \leq 1-\epsilon} |a_j(x) - a_j(x')|^k \leq (C_{13}|x - x'|)^k \leq (C_{14} l^{-\beta})^k \leq C_{15}(k) l^{-k},$$

and so

$$E \left\{ \sup_{e \leq x \leq 1-\epsilon} |V_2(x) - V_2(x')|^k \right\} \leq \sup_{e \leq x \leq 1-\epsilon} |a_j(x) - a_j(x')|^k E \left| \sum_{j=2}^{C_{2J}} Y_j \right|^k \leq C_{16}(k) h^k. \quad (A.7)$$

Since

$$E \left\{ \sup_{e \leq x \leq 1-\epsilon} |V_2(x)|^k \right\} \leq E \left\{ \sup_{e \leq x \leq 1-\epsilon} |V_2(x) - V_2(x')|^k \right\} + E \left\{ \sup_{x' \in E(l)} |V_2(x')|^k \right\},$$

the required result (A.5) follows from this expression, (A.6) and (A.7).
Appendix 3A.3

Bounds for sums using integral approximations

Lemma 3.6 gives upper bounds for the absolute value of sums such as

$$\sum_i i \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\}$$

by finding bounds for the absolute value of the corresponding integral approximations

$$\int_{-\infty}^{\infty} y \mu\{M^{-1}(y/l)\}^{-2} K''\{(x - M^{-1}(y/l))/h\} dy,$$

then finding the errors in approximating the sums by the integrals.

Abramowitz and Stegun (1965, p886) give the Euler-Maclaurin summation formula with explicit bounds on the remainder as

$$\int_{x_0}^{x_n} f(x)dx = \delta(f_0/2 + f_1 + \cdots + f_{n-1} + f_n/2) - C_1 \delta^2 (f_n' - f_0') - \cdots$$

$$- C_{2k} \delta^{2k} (f_n^{(2k-1)} - f_0^{(2k-1)}) + R_{2k},$$

where \(\delta\), the step length, is of size \(O(n^{-1})\), and \(|R_{2k}| < 2C_{2(k+1)} \delta^{2(k+1)} (f_n^{(2k+1)} - f_0^{(2k+1)})\). From this formula, it is clear that if we cannot assume the existence of any derivatives of \(f(x)\), then the error in approximating a sum by an integral is of size \(O\{h(f_0 + f_n)\} = O\{n^{-1}(f_0 + f_n)\}\). If we can assume the existence of \(k\) derivatives of \(f(x)\), then the error in approximating a sum by an integral is of size \(O\{h^{k+1} (f_0^{(k)} + f_n^{(k)})\} = O\{n^{-(k+1)} (f_0^{(k)} + f_n^{(k)})\}\).
Lemma 3.6

Using the definitions of Theorem 3.2, let $v(x) = \sum_i \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\}$, $w(x) = \sum_i \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\}$, $a_j(x) = \sum_{j \leq i < \infty} \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\}$, and $b_j(x) = \sum_{j \leq i < \infty} \mu(\alpha_i)^{-3} \mu'(\alpha_i) K'\{(x - \alpha_i)/h\}$. Then

$$|v(x)| \leq C_1(l + h^{2+\zeta} l^2),$$  \hspace{1cm} (A.8)

$$|w(x)| \leq C_2(l + h^{1+\zeta} l^2),$$  \hspace{1cm} (A.9)

$$|a_j(x)| \leq C_3(1 + hl) I(j < C_4 hl),$$  \hspace{1cm} (A.10)

$$|b_j(x)| \leq C_5(1 + hl) I(j < C_6 hl),$$  \hspace{1cm} (A.11)

uniformly in $x \in \mathcal{I}_\epsilon \subseteq [ch, 1 - ch]$.

The proof of (A.8) is given in detail, but the proofs of (A.9), (A.10), and (A.11) are only outlined in view of their similarity to that of (A.8).

Proof of (A.8) The integral approximation to $v(x)$ is

$$v_1(x) = \int_0^\infty y \mu\{M^{-1}(y/l)\}^{-2} K''\{(x - M^{-1}(y/l))/h\} dy.$$  \hspace{1cm}

Put $u = \{x - M^{-1}(y/l)\}/h$, so that $y = lM(x - uh)$, and $dy = -hl \mu(x - uh) du$. Then for a point $x \in \mathcal{I}_\epsilon$,

$$v_1(x) = hl^2 \int_{-c}^c M(x - uh) \mu(x - uh)^{-1} K''(u) du.$$  \hspace{1cm}

Writing $N(t) = M(t) \mu(t)^{-1}$ and integrating by parts, we obtain

$$v_1(x) = \begin{align*}
h l^2 \{ & [N(x - uh)K'(u)]_{-c}^c + h \int_{-c}^c N'(x - uh)K'(u)du \} \\
= & \begin{cases} \begin{align*}
h l^2 \int_{-c}^c N'(x - uh)K'(u)du & \text{(since $K'(c) = K'(-c) = 0$)} \\
= & \begin{cases} \begin{align*}
h l^2 \int_{-c}^c \{N'(x - uh) - N'(x)\}K'(u)du & \text{(since $\int K'(u)du = 0$).} \\
\end{align*} \end{cases} \\
\end{cases} \end{align*}$$

Thus,

$$|v_1(x)| \leq hl^2 \int_{-c}^c |N'(x - uh) - N'(x)||K'(u)|du.$$  \hspace{1cm}

Now, $N'(t) = \{\mu(t)^2 - M(t)\mu'(t)\}\mu(t)^{-2}$ is Hölder continuous with exponent $\zeta$ because $\mu'$ is Hölder continuous. Since $\int_{-c}^c |uK'(u)| du$ is bounded, $|v_1(x)| = O(h^{2+\zeta} l^2)$ uniformly in $x \in \mathcal{I}_\epsilon$.  \hspace{1cm}

107
The second step of the proof involves finding the error in approximating \( v(x) \) by \( v_1(x) \). First, note that we cannot differentiate \( v_1(x) \). There are \( O(hl) \) terms in the sum \( v(x) = \sum_i i \mu(\alpha_i)^{-2} K''\{(x - \alpha_i)/h\} \) (the number of terms in the sum is the number of \( \alpha_i \) within a distance \( ch \) of \( x \); see the proof of Lemma 3.2). Thus, \( |v_1(x) - v(x)| = O(hl^2(hl)^{-1}) = O(l) \) uniformly in \( x \in \mathcal{I}_e \), and so \( |v(x)| \leq C_1(l + h^{2+\xi} l^2) \) uniformly in \( x \in \mathcal{I}_e \).

**Proof of (A.9)** For points \( x \in \mathcal{I}_e \), the integral approximation to \( w(x) \) is

\[
w_1(x) = \int y \mu\{M^{-1}(y/l)\}^{-3} \mu'\{M^{-1}(y/l)\} K''\{(x - M^{-1}(y/l))/h\} dy
= hl^2 \int M(x - u) \mu(x - u)^{-2} \mu'(x - u) K'(u) du.
\]

Let \( N(t) = M(t) \mu(t)^{-2} \mu'(t) \) and note its Hölder continuity. Then as \( \int K'(u) du = 0 \), we may write

\[
w_1(x) = hl^2 \int \{N(x - u) - N(x)\} K'(u) du,
\]

so that \( |w_1(x)| = O(h^{1+\xi} l^2) \) uniformly in \( x \in \mathcal{I}_e \). Also we have \( |w_1(x) - w(x)| = O(hl^2(hl)^{-1}) = O(l) \) uniformly in \( x \in \mathcal{I}_e \), and therefore, \( |w(x)| \leq C_2(l + h^{1+\xi} l^2) \) uniformly in \( x \in \mathcal{I}_e \).

**Proof of (A.10)** For points \( x \in \mathcal{I}_e \), the integral approximation to \( a_j(x) \) is

\[
a_{1j}(x) = \int_j^\infty \mu\{M^{-1}(y/l)\}^{-2} K''\{(x - M^{-1}(y/l))/h\} dy
= hl \int_{-(x-\alpha_j)/h}^{(x-\alpha_j)/h} \mu(x - u)^{-1} K''(u) du.
\]

Note that \( a_{1j}(x) > 0 \) if \( |x - \alpha_j| < ch \), i.e. if \( j < C_4hl \). Writing \( N(t) = \mu(t)^{-1} \) and integrating by parts, we have

\[
a_{1j}(x) = hl \left\{ \left[ N(x - u) K'(u) \right]_{-c}^{(x-\alpha_j)/h} + hl \int_{-c}^{(x-\alpha_j)/h} N'(x - u) K'(u) du \right\}
= hl \left[ \mu(\alpha_j)^{-1} K'(x - \alpha_j) + hl \int_{-c}^{(x-\alpha_j)/h} \left( N'(x - u) - N'(x) \right) K'(u) du \right]
+ hl \int_{-c}^{(x-\alpha_j)/h} N'(x) K'(u) du
= hl \left[ \mu(\alpha_j)^{-1} K'(x - \alpha_j) + hl N'(x) K\{(x - \alpha_j)/h\} \right]
+ hl \int_{-c}^{(x-\alpha_j)/h} \left( N'(x - u) - N'(x) \right) K'(u) du.
\]

108
Now $N'(t) = -\mu'(t)\mu(t)^{-2}$ is Hölder continuous with exponent $\zeta$. Also, $\mu(\alpha_j)^{-1}$, $K'$, $N'$, $K$ and $\int |uK'(u)|$ are bounded. Thus,

$$|a_{1j}(x)| = O(hl) I(j < C_4 hl)$$

uniformly in $x \in I$. Also, $|a_{1j}(x) - a_j(x)| = O\{hl\} I(j < C_4 hl) = O(1) I(j < C_4 hl)$, and so $|a_j(x)| = O(1 + hl) I(j < C_4 hl)$ uniformly in $x \in I$.

**Proof of (A.11)** For points $x \in I$, the integral approximation to $b_j(x)$ is

$$b_{1j}(x) = \int_j^{\infty} \mu\{M^{-1}(y/l)\}^{-3} \mu'\{M^{-1}(y/l)\} K'[\{x - M^{-1}(y/l)\}/h] dy$$

$$= hl \int_{-c}^{(x - \alpha_j)/h} \mu(x - uh)^{-2} \mu'(x - uh) K'(u) du.$$  

Note that $b_{1j}(x) > 0$ if $|x - \alpha_j| < ch$, i.e. if $j < C_6 hl$. Write $N(t) = \mu(t)^{-2}\mu'(t)$ and note its Hölder continuity. Then

$$b_{1j}(x) = hl \left[ \int_{-c}^{(x - \alpha_j)/h} \{N(x - uh) - N(x)\} K'(u) du + \int_{-c}^{(x - \alpha_j)/h} N(x) K'(u) du \right]$$

$$= hl \left[ \int_{-c}^{(x - \alpha_j)/h} \{N(x - uh) - N(x)\} K'(u) du + N(x)K\{(x - \alpha_j)/h\} \right].$$

Using the Hölder continuity of $N$, and the fact that $N$, $K$ and $\int |uK'(u)|$ are bounded, we have

$$|b_{1j}(x)| = O(hl) I(j < C_6 hl)$$

uniformly in $x \in I$. Also, $|b_{1j}(x) - b_j(x)| = O\{hl\} I(j < C_6 hl) = O(1) I(j < C_6 hl)$ uniformly in $x \in I$, and so $|b_j(x)| = O(1 + hl) I(j < C_6 hl)$ uniformly in $x \in I$. □
Appendix 3A.4

A lemma on the closeness of two random processes

In the present appendix we shall prove Lemma 3.7 which is used in the proof of the bootstrap part of Theorem 3.2.

Lemma 3.7 Assuming the conditions of Theorem 3.2, with \( J^*(x) = \int W^*\{\tilde{M}(x - hy)\} K'(y) \, dy \) and \( J^t(x) = \int W^*\{M(x - hy)\} K'(y) \, dy \), we have

\[
E\left\{ \sup_{0 \leq \varepsilon \leq 1 - \varepsilon} |J^*(x) - J^t(x)|^k \right\} \leq C_1(k)(h \log l)^k.
\]

The proof uses Proposition 3.2 — a result on the continuity of Wiener processes — and Lemma 3.8, both of which we give below before proceeding with the proof of Lemma 3.7.

Silverman (1978, page 179) gives the following result for the continuity of the Brownian bridge on \( 0 \leq t \leq 1 \):

Proposition 3.1 Let \( W_0(t) \) be a continuous version of the Brownian bridge, with modulus of continuity \( w_0 \). Let

\[
p(u) = \begin{cases} 
  u^{1/2}(1 - u)^{1/2} & \text{if } 0 \leq u \leq 1/2 \\
  1/2 & \text{if } u > 1/2,
\end{cases}
\]

and let

\[
q(u) = \int_u^1 \{\log(1/v)\}^{1/2} dp(v).
\]
Then, with probability one, $w_0$ is dominated by

$$16 (\log B)^{1/2} p + 16 (2)^{1/2} q,$$

where $B$ is a random variable with $B \geq 1$ a.s. and $EB < 4(2)^{1/2}$.

In Lemma 3.7, we have a Wiener process rather than a Brownian bridge: the sample paths do not have a common endpoint. However, as the Brownian bridge in the proposition above and a Wiener process on $0 \leq t \leq 1$ differ only by the continuous term $tW(1)$, a constant multiple of the above modulus of continuity can be utilized for the Wiener process.

Note that for $0 < u \leq 1$,

$$p(u) = \{u (1 - u)\}^{1/2} \leq u^{1/2} |\log u|^{1/2},$$

and

$$q(u) = \int_u^\infty \{\log(1/v)\}^{1/2} dp(v) \leq \int_0^u |\log v|^{1/2} dv^{1/2} \leq |\log u|^{1/2} u^{1/2}.$$

We shall therefore simplify Proposition 3.1 as follows:

**Proposition 3.2** Let $W(t)$, $0 \leq t \leq t_0$, be a continuous version of the Wiener process, with modulus of continuity $w_0$. Then, with probability one,

$$w_0(u) \leq b \{1 + (\log B)^{1/2}\} (u|\log u|)^{1/2}.$$

where $B$ is a random variable with $B \geq 1$ a.s., $EB < 4(2)^{1/2}$, and $b$ depends only on $t_0$.

Since $\sup_T \mu < \infty$, $M(1)$ is bounded and we may choose $t_0$ in Proposition 3.2 to be greater than $M(1)$. For such a value of $t_0$, we have the following Lemma:

**Lemma 3.8** For all $c > 0$ and some $t_0 > M(1)$,

$$P\{\bar{M}(1) > t_0\} \leq C_1(k) \{t_0 - M(1)\}^{-k} h^k = O(1^{-c}).$$

111
Proof. The following two results are needed for the main proof of the lemma:

\[ E \lambda(t)^k = O(l^k) \quad \text{for } t \in (0, ch), \quad (A.12) \]
\[ E|\lambda(t) - \lambda(t)\|^k = O((h^2 l)^k) \quad \text{for } t \in [ch, 1 - ch]. \quad (A.13) \]

We obtain (A.12) by noting that for \( 0 < t < ch \),

\[ \lambda(t) \leq h^{-1}(\sup K) \# \{ X_i : X_i \leq 2ch \} = h^{-1}(\sup K) N \]

where \( N \) is Poisson with mean \( m = \int_0^{2ch} \lambda(t) \, dt = O(hl) \) and \( k \)th factorial moment \( m^k \). Thus,

\[ E \lambda(t)^k \leq C_2(k) h^{-k}(hl)^k = C_2(k) l^k. \]

To show (A.13), note that for \( ch \leq x \leq 1 - ch \), the non-bootstrap cases of Theorems 3.1 and 3.3 give

\[ \lambda - \lambda = \frac{1}{2} \kappa h^2 \lambda'' + o(h^2 l) + (h^{-1} l)^{1/2}(\mu^{1/2} U_n + R_{1n}), \]

so that using Corollary 3.3 on page 102, we have

\[ E|\lambda - \lambda|^k \leq C_3(k) \left\{ \left[ \frac{1}{2} \kappa h^2 \lambda'' \right]^k + (h^{-1} l)^{k/2}(\mu^{1/2} E U_n)^k + E R_{1n}^k \right\} \]
\[ \leq C_5(k) \left\{ (h^2 l)^k + (h^{-1} l)^{k/2}(1 + l^{-kn}) \right\} \]
\[ = O((h^2 l)^k). \]

We now proceed to the main proof of the lemma. Markov’s inequality gives, for \( t_0 > M(1) \),

\[ P\{ M(1) > t_0 \} = P\{ M(1) - M(1) > t_0 - M(1) \} \]
\[ \leq \{ t_0 - M(1) \}^{-k} E|\lambda(1) - M(1)|^k \]
\[ \leq \{ l \{ t_0 - M(1) \} \}^{-k} E \left[ \left\{ \int_0^1 (\lambda(t) - \lambda(t))^k \, dt \right\}^k \right] \]
\[ \leq \{ l \{ t_0 - M(1) \} \}^{-k} \left[ \int_0^1 \left\{ E|\lambda(t) - \lambda(t)|^k \right\}^{1/k} \, dt \right]^k \]

(applying Lemma 3.12)
\[ \leq C_6(k) \{ l \{ t_0 - M(1) \} \}^{-k} \left[ \left( \int_{0}^{1} + \int_{1-ch}^{1} \right) \left\{ E|\lambda(t)|^k + \lambda(t)^k \right\}^{1/k} \, dt \right]^k \]
\[ + \int_{ch}^{1-ch} \left\{ E|\lambda(t) - \lambda(t)|^k \right\}^{1/k} \, dt \]
\[ \leq C_7(k) \{ l \{ t_0 - M(1) \} \}^{-k} \{(hl)^k + (h^2 l)^k \} \]
\[ = C_8(k) \{ t_0 - M(1) \}^{-k} h^k = O(l^{-c}), \quad \forall c > 0. \]
We now have the tools to prove Lemma 3.8. Choose $t_0$ in Proposition 3.2 to be greater than $M(1)$.

**Proof of Lemma 3.8**  
Put 
\[ Z = \sup_{\varepsilon \leq x \leq \varepsilon + 1 - \varepsilon} \sup_{y \in \text{supp}(K)} |\overline{M}(x - hy) - M(x - hy)|. \]

Now, 
\[
\sup_{\varepsilon \leq x \leq \varepsilon + 1 - \varepsilon} |J^*(x) - J^*(x)| = \sup_{\varepsilon \leq x \leq \varepsilon + 1 - \varepsilon} \left| \int \left[ W^*\{\overline{M}(x - hy)\} - W^*\{M(x - hy)\} \right] K'(y) dy \right|
\leq \int \sup_{\varepsilon \leq x \leq \varepsilon + 1 - \varepsilon} w_0(|\overline{M}(x - hy) - M(x - hy)|) |K'(y)| dy,
\leq w_0(Z) \int |K'(y)| dy,
\leq w_0(Z) \int |K'(y)| dy, \tag{A.14}
\]

provided that 
\[ Z = \sup_{\varepsilon \leq x \leq \varepsilon + 1 - \varepsilon} \sup_{y \in \text{supp}(K)} \max |\overline{M}(x - hy) - M(x - hy)| \leq t_0. \tag{A.15} \]

To treat the case when $Z > t_0$ and (A.15) fails, note that 
\[ \sup_{\varepsilon \leq x \leq \varepsilon + 1 - \varepsilon} |J^*(x) - J^*(x)| \leq 2 \sup_{0 \leq x \leq \bar{M}(1)} |W^*(x)| \int |K'(y)| dy. \]

Now, 
\[
E \left[ \left\{ \sup_{0 \leq x \leq \bar{M}(1)} |W^*(x)| \right\}^k I\{\overline{M}(1) > t_0\} \right]
= E \left[ \sum_{j=1}^{\infty} \left\{ \sup_{0 \leq x \leq j} |W^*(x)| \right\}^k I\{\overline{M}(1) \in (j - 1, j) \text{ and } \overline{M}(1) > t_0\} \right]
\leq E \left[ \sum_{j=1}^{\infty} \left\{ \sup_{0 \leq x \leq j} |W^*(x)| \right\}^k I\{\overline{M}(1) > \max(j - 1, t_0)\} \right]
\leq \sum_{j=1}^{\lfloor t_0 + 1 \rfloor} \left[ E \left\{ \sup_{0 \leq x \leq j} |W^*(x)|^{2k} \right\} \right]^{1/2} \left[ P\{\overline{M}(1) > \max(j - 1, t_0)\} \right]^{1/2}
\leq j^{k/2} \left[ P\{\overline{M}(1) > \max(j - 1, t_0)\} \right]^{1/2}
+ \sum_{j > \lfloor t_0 + 1 \rfloor} j^{k/2} \left[ P\{\overline{M}(1) > \max(j - 1, t_0)\} \right]^{1/2}
\]
(Note that by Corollary 3.3, for all $k \geq 1$ we have 
\[ E \left\{ \sup_{0 \leq x \leq j} |W^*(x)|^{2k} \right\} \leq \left[ E \left\{ \sup_{0 \leq x \leq j} |W^*(x)|^2 \right\} \right]^k = j^k. \]

113
\[
\sum_{j=1}^{[t_0+1]} j^{k/2} \left[ P\{ \widetilde{M}(1) > t_0 \} \right]^{1/2} + \sum_{j>[t_0+1]} j^{k/2} \left[ P\{ \widetilde{M}(1) > j - 1 \} \right]^{1/2}
\]

\[
= C_1(k) \sum_{j=1}^{[t_0+1]} j^{k/2} l^{-k'/2} + C_2(k) \sum_{j>[t_0+1]} j^{k/2} \left\{ j - 1 - M(1) \right\}^{-k'/2} h^{k_2'} \]

\[
= C_1(k) l^{-k'/2} \sum_{j=1}^{[t_0+1]} j^{k/2} + C_2(k) h^{k_2'} \sum_{j>[t_0+1]} j^{k/2} \left\{ j - 1 - M(1) \right\}^{-k'/2}
\]

(using Lemma 3.7)

\[
= O(l^{-c}).
\]

as the first sum is bounded and the second sum converges for suitable choice of \( k \) and \( k_2' \). Combining this with (A.14) we deduce that

\[
E \left\{ \left( \sup_{x \leq x \leq 1} |J^*(x) - J^*(x)|^k \right)^k \right\} \leq E \{ w_0(Z)^k \} \left\{ \int |K'(y)| dy \right\}^k + O(l^{-c}),
\]

for all \( c > 0 \).

From Proposition 3.2,

\[
w_0(Z)^k \leq b^k \left\{ 1 + (\log B)^{1/2} \right\}^k (Z| \log Z|)^{k/2},
\]

and applying Hölder’s inequality, we obtain

\[
E \{ w_0(Z)^k \} \leq b^k \left[ E \left\{ 1 + (\log B)^{1/2} \right\}^{2k} E Z^{3k/2} E |\log Z|^{3k/2} \right]^{1/3}
\]

\[
\leq C_3(k) \left\{ 1 + E (\log B)^{2k} \right\}^{1/3} \left( E Z^{3k/2} E |\log Z|^{3k/2} \right)^{1/3}.
\]

Since \( e^x = \sum_{k=0}^{\infty} x^k / k! \), we have \( x^k / k! \leq e^x \), and

\[
E (\log B)^k = k! \frac{E (\log B)^k}{k!} \leq k! E B = O(1).
\]

Also,

\[
E Z^k = E \left\{ \sup_{x \in I_r} \sup_{y \in \text{supp}(K)} \left| \widetilde{M}(x - hy) - M(x - hy) \right|^k \right\}
\]

\[
= \sup_{x \in I_r} \sup_{y \in \text{supp}(K)} \left\{ \text{var} \widetilde{M}(x - hy) \right\}^{k/2}
\]

\[
+ \sup_{x \in I_r} \sup_{y \in \text{supp}(K)} \left| E \widetilde{M}(x - hy) - M(x - hy) \right|^k
\]

\[
\leq C_4(k) \left\{ (l^{-1} \log l)^{k/2} + h^{2k} \right\}
\]

\[
= C_4(k) h^{2k} \left\{ 1 + (h^4 l)^{-k/2} (\log l)^{k/2} \right\}. \]
Finally then,

$$E \left\{ \sup_{0 \leq x \leq 1 - \epsilon} |J^*(x) - J^+(x)|^k \right\} \leq C_5(k) \left( h^{3k} \log h^{3k} \right)^{1/3} \leq C_6(k) (h \log l)^k$$

proving the lemma.

Hölder continuity of $\nu'$

In Lemma 3.9 in this appendix, we shall prove that $\nu'$ is Hölder continuous with probability one. However, first we shall define the first and second derivatives of $\nu$ and describe how to obtain versions of inequalities (3.15) and (3.16) in Theorem 3.2 and 3.3 for these derivatives. We write $\nu'$ and $\nu''$ for the Hölder exponents of $\nu'$ and $\nu''$, respectively.

Recall that $\nu(x) = -\langle \nu \rangle^m \sum_i \nu_i(x) = \nu_i(x)/\langle \nu \rangle$ and that $\nu_i$ satisfies a Lipschitz condition of order $1$. Then for $p = 1$ we have

$$\nu_i(x) = \nu_i(y) + (x - y) \sum_j \nu_i'(y)(x - y_j)/\nu_i(y)$$

For those values of $\nu$ define $\nu_i'(y)$ by

$$\nu_i'(y) = \nu_i(y) = \left( \frac{1}{2} \right)^m \sum_i \nu_i'(y)(x - y_i)/\nu_i(y).$$

Because $\nu_i$ has the same form as $\nu$ with $\nu$ replaced by $\nu_i$, there are versions of Theorems 3.2, 3.3 and 3.4 for $\nu_i$ rather than for $\nu$, and these lead to corresponding theorems for $\nu''$. For example, the version of equation (3.16) in Theorem 3.2 for $\nu$ is

$$\partial^2 \nu_i(x) = \int \partial^2 \nu_i(y) \partial^2 \nu_i(x) \, dy + O(h^{-2} \gamma^{-2} + h^{-1} \gamma^{-1})$$

giving for $\nu_i''$

$$\partial^2 \nu_i'(y) = \int \partial^2 \nu_i(y) \partial^2 \nu_i'(y) \, dy + O(h^{-2} \gamma^{-2} + h^{-1} \gamma^{-1}),$$

and for $\nu''$

$$\partial^2 \nu''(y) = h^{-2} \int \partial^2 \nu_i(y) \partial^2 \nu_i(x) \, dy + O(h^{-2} \gamma^{-2} + h^{-1} \gamma^{-1}).$$
Appendix 3A.5

Hölder continuity of $\hat{\mu}'$

In Lemma 3.9 in this appendix, we shall prove that $\hat{\mu}'$ is Hölder continuous with probability one. However, first we shall define the first and second derivatives of $\hat{\mu}$, and describe how to obtain versions of equations (3.12) and (3.16), in Theorems 3.2 and 3.3, for these derivatives. We write $\zeta_1$ and $\zeta_2$ for the Hölder exponents of $\mu'$ and $\hat{\mu}'$ respectively.

Recall that $\hat{\mu}(x) = (hl)^{-1} \sum_i K\{(x - X_i)/h\}$ and that $K''$ satisfies a Lipschitz condition of order 1. Then for $j = 1$ and 2,

$$\hat{\mu}^{(j)}(x) = (h^{j+1}l)^{-1} \sum_i K^{(j)}\{(x - X_i)/h\}.$$

For these values of $j$, define $\hat{\mu}_{(j)}(x)$ by

$$\hat{\mu}_{(j)}(x) = h^j \hat{\mu}^{(j)}(x) = (hl)^{-1} \sum_i K^{(j)}\{(x - X_i)/h\}.$$

Because $\hat{\mu}_{(j)}$ has the same form as $\hat{\mu}$ with $K$ replaced by $K^{(j)}$, there are versions of Theorems 3.1, 3.2 and 3.3 for $\hat{\mu}_{(j)}$ rather than for $\hat{\mu}$, and these lead to corresponding theorems for $\hat{\mu}'$. For example, the version of equation (3.16) in Theorem 3.3 for $\hat{\mu}$ is

$$E \hat{\mu}(t) = \int \mu(t - hu) K(u) du + O(h^{-3} t^{-3/2} + h^{-1+\zeta} t^{-1}),$$

giving for $\hat{\mu}_{(2)}$

$$E \hat{\mu}_{(2)}(t) = \int \mu(t - hu) K''(u) du + O(h^{-3} t^{-3/2} + h^{-1+\zeta} t^{-1}),$$

and for $\hat{\mu}''$

$$E \hat{\mu}''(t) = h^{-2} \int \mu(t - hu) K''(u) du + O(h^{-5} t^{-3/2} + h^{-3+\zeta} t^{-1})$$

116
\[= h^{-1} \int \mu'(t - hu)K'(u) \, du + O(h^{-5}t^{-3/2} + h^{-3+c}t^{-1}) \]  

(A.16)

Integrating by parts. Similarly, we obtain the version of equation (3.12) in Theorem 3.2 for \( \hat{\mu}' \), which is

\[
\hat{\mu}' - E(\hat{\mu}') = (h^4t)^{-1/2} \int W\{M(t - hu)\} K''(u) \, du + (h^3t)^{-1/2} R_2n. 
\]  

(A.17)

Since \( K'' \) is (Lipschitz) continuous and, by Lemma 3.13, of bounded variation, the Stieltjes integral \( \int dK''(y) \) exists even though the Riemann integral \( \int K''(y) \, dy \) does not exist; see for example Kendall and Stuart (1977, p15ff) for definitions of these integrals. We therefore obtain for \( \hat{\mu}'' \),

\[
\hat{\mu}'' - E(\hat{\mu}'') = (h^6t)^{-1/2} \int W\{M(t - hu)\} dK''(u) + (h^5t)^{-1/2} R_2n. 
\]  

(A.18)

Lemma 3.9 For some \( B, \zeta_2 > 0 \) and all \( k > 0 \),

\[ P\{|\mu'(x) - \mu'(y)| \leq B|x - y|^{\zeta_2} \text{ for all } x, y \in \mathcal{I}_e\} = 1 - O(t^{-k}). \]

Proof We shall start by proving that for all \( x, y \in \mathcal{I}_e \), \( P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| \leq B|x - y|^{\zeta_2} \} = 1 - O(t^{-k}) \) using Markov’s inequality. Therefore we shall derive a bound for \( E|\hat{\mu}'(x) - \hat{\mu}'(y)|^k \). We do this in two ways. The first method uses \( \int_0^t E|\hat{\mu}''(t)|^k \) and so gives a bound including \( |x - y| \) in all terms. The second uses only (A.17), and so the bound does not necessarily include \( |x - y| \) in all terms. The overall bound for \( E|\hat{\mu}'(x) - \hat{\mu}'(y)|^k \) is given by the first of these bounds when \( |x - y| \) is small and the other elsewhere. We complete the lemma by proving that the result holds uniformly for all \( x, y \in \mathcal{I}_e \).

Method 1 Assuming without loss of generality that \( x > y \), we have

\[ |\hat{\mu}'(x) - \hat{\mu}'(y)| = \left| \int_y^x E\hat{\mu}''(t) \, dt \right| \leq \left| \int_y^x E\hat{\mu}''(t) \, dt \right| + \left| \int_y^x \left| \hat{\mu}''(t) - E\hat{\mu}''(t) \right| \, dt \right|,
\]

so that

\[
E|\hat{\mu}'(x) - \hat{\mu}'(y)|^k \leq \left| \int_y^x E\hat{\mu}''(t) \, dt \right|^k + E\left\{ \int_y^x |\hat{\mu}''(t) - E\hat{\mu}''(t)| \, dt \right\}^k \leq \left| \int_y^x E\hat{\mu}''(t) \, dt \right|^k + \left[ \int_y^x \left( E|\hat{\mu}''(t) - E\hat{\mu}''(t)|^k \right) \, dt \right]^{1/k}.
\]  

(A.19)
applying Lemma 3.12.

The integrand of the first integral in (A.19) is obtained using (A.16), which leads to

\[
|E \hat{\mu}''(t)| \leq h^{-1} \int |\mu'(t - hu) - \mu'(t)| |K'(u)| \, du + O(h^{-5} l^{-3/2} + h^{-3+\zeta} l^{-1})
\]

\[
\leq B_4 h^{-1+\zeta} \int |u K'(u)| \, du + O(h^{-5} l^{-3/2} + h^{-3+\zeta} l^{-1}),
\]

so that

\[
\left| \int_{y}^{x} E \hat{\mu}''(t) \, dt \right|^k \leq C_1(k) (x - y)^k \{(h^{-1+\zeta} l)^k + O(h^{-5} l^{-3/2} + h^{-3+\zeta} l^{-1})\}^k.
\]

The integrand of the second integral in (A.19) is obtained utilizing (A.18), which gives

\[
E |\hat{\mu}'' - E(\hat{\mu}'')|^k \leq (h^6 l)^{-k/2} E \left| \int W\{M(t - hu)\} dK''(u) \right|^k + (h^5 l)^{-k/2} E |R_{2n}|^k,
\]

leading to

\[
\left[ \int_{y}^{x} \left\{ E |\hat{\mu}''(t) - E(\hat{\mu}'')|^k \right\}^{1/k} \, dt \right]^k
\]

\[
\leq C_2(k) (h^6 l)^{-k/2} \left( \int_{y}^{x} \left[ E \left| \int W\{M(t - hu)\} dK''(u) \right|^k \right]^{1/k} \, dt \right)^k
\]

\[
+ C_3(k) (h^5 l)^{-k/2} \left\{ \int_{y}^{x} \left( E |R_{2n}|^k \right)^{1/k} \, dt \right\}^k
\]

\[
= A + B,
\]

say. Now,

\[
A = C_2(k) (h^6 l)^{-k/2} \left\{ \int_{y}^{x} \left( E \left| \int [W\{M(t - hu)\} - W\{M(t)\}] dK''(u) \right|^k \right)^{1/k} \, dt \right\}^k
\]

\[
\leq C_2(k) (h^6 l)^{-k/2} \left\{ \int_{y}^{x} \left( E \int w_0(h |u|) \, dK''(u) \right)^{1/k} \, dt \right\}^k,
\]

where \(w_0\) is the modulus of continuity of a Weiner process, defined in Proposition 3.2 on page 111. Thus

\[
A \leq C_3(k) (h^6 l)^{-k/2} (x - y)^k (h |\log h|)^{k/2} = C_3(k) (h^5 l)^{-k/2} (x - y)^k |\log h|^{k/2}.
\]

Also,

\[
B \leq C_4(k) (h^5 l)^{-k/2} (x - y)^k \{(hl)^{-1/2} \log l\}^{k/2} l^3.
\]

118
Combining these expressions for \(|\int_a^b E \mu''(t) \, dt|^k\), \(A\) and \(B\), we obtain

\[
E |\hat{\mu}'(x) - \hat{\mu}'(y)|^k
\leq C_5(k) (x - y)^k \left[ (h^{-1}+\zeta'_1)^k + (h^5 l)^{-k/2} \{ |\log h|^k + (h l)^{-k/2} (\log l)^k l^3 \} \right]
\leq C_6(k) (x - y)^k (h^{-1}+\zeta'_1)^k,
\]

(A.20)

where \(\zeta'_1 = \min(\zeta_1, 1/2)\).

**Method 2** The version of (3.16) for \(\hat{\mu}'\) is

\[
E \hat{\mu}'(t) = h^{-1} \int \mu(t - hu) K'(u) \, du + O(h^{-4} l^{-3/2} + h^{-2}+\zeta_1 l^{-1})
= \int \mu'(t - hu) K(u) \, du + O(h^{-4} l^{-3/2} + h^{-2}+\zeta_1 l^{-1}),
\]

following integration by parts. Together with (A.17), this gives

\[
\hat{\mu}'(t) = \int \mu'(t - hu) K(u) \, du + o\{(h^4 l)^{-1/2}\} + (h^4 l)^{-1/2} \int W\{\nu(t - hu)\} K''(u) \, du + (h^3 l)^{-1/2} R_{2n}
= d(t) + S(t),
\]

writing \(d(t) = \int \mu'(t - hu) K(u) \, du + o\{(h^4 l)^{-1/2}\}\) for the deterministic component of \(\hat{\mu}'(t)\) and \(S(t) = (h^4 l)^{-1/2} \int W\{\nu(t - hu)\} K''(u) \, du + (h^3 l)^{-1/2} R_{2n}\) for the stochastic terms.

The magnitude of the difference between the deterministic terms of \(\hat{\mu}'(x)\) and \(\hat{\mu}'(y)\) is given by

\[
|d(x) - d(y)| = \left| \int \mu'(x - hu) K(u) \, du - \int \mu'(y - hu) K(u) \, du \right|
\leq \int \left| \mu'(x - hu) - \mu'(y - hu) \right| |K(u)| \, du
\leq C_7(k) (x - y)^{\zeta_1}
\]

since \(\int |u K(u)| \, du\) is bounded. Choosing \(B = 2 C_7(k)\) and \(\zeta_2 \leq \zeta_1\), we have

\[
|d(x) - d(y)| \leq \frac{1}{2} B (x - y)^{\zeta_2}.
\]

As our method 2 bound, we shall use

\[
E |\hat{\mu}'(x) - \hat{\mu}'(y) - \{d(x) - d(y)\}|^k \leq C_8(k) \{E |S(x)|^k + E |S(y)|^k\}
\leq C_9(k) (h^4 l)^{-k/2}.
\]

(A.21)
Overall bound

Using Markov’s inequality and the Method 1 bound in (A.20), we have

\[ P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| > B|x - y|^{\zeta_2}\} \]
\[ \leq \{B(x - y)^{\zeta_2}\}^{-k} E|\hat{\mu}'(x) - \hat{\mu}'(y)|^k \]
\[ = \{B(x - y)^{\zeta_2}\}^{-k} C_{10}(k) (x - y)^k (h^{-1+\zeta_1})^k \]
\[ = C_{11}(k) (x - y)^{(1-\zeta_1)} (h^{-1+\zeta_1})^k \]
\[ \leq C_{11}(k) (h^{1-\zeta_1})^k \text{ if } x - y \leq h^{2(1-\zeta_1)}. \]

Markov’s inequality and the Method 2 bound, (A.21), give

\[ P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| > B|x - y|^{\zeta_2}\} \]
\[ \leq P\left[|\hat{\mu}'(x) - \hat{\mu}'(y)| - \{d(x) - d(y)\} | + |d(x) - d(y)| > \frac{1}{2} B(x - y)^{\zeta_2}\right] \]
\[ \leq \left(\frac{1}{2} B(x - y)^{\zeta_2}\right)^{-k} C_{12}(k) (h^4l)^{-k/2} \]
\[ = C_{13}(k) (x - y)^{-k} (h^4l)^{-k/2} \]
\[ \leq C_{13}(k) (h^4l)^{-k/4} \text{ if } x - y \geq (h^4l)^{-1/(4\zeta_2)}. \]

Choosing \( \zeta_2 < -\frac{\log(h^4l)}{8(1-\zeta_1)\log h} \) so that \((h^4l)^{-1/(4\zeta_2)} < h^{2(1-\zeta_1)}\), and combining these two bounds, we have

\[ P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| > B|x - y|^{\zeta_2}\} \leq C_{14}(k) \max\{(h^{1-\zeta_1})^k, (h^4l)^{-k/4}\} \leq C_{15}(k) l^{-\epsilon k}. \]

Therefore, for all \( \nu > 0 \),

\[ \sup_{x, y \in \mathcal{I}_\nu} P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| > B|x - y|^{\zeta_2}\} = O(l^{-\nu}). \tag{A.22} \]

Completion

We complete the lemma by proving that, for all \( \nu > 0 \),

\[ P\{\forall x, y \in \mathcal{I}_\nu, |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq B|x - y|^{\zeta_2}\} = 1 - O(l^{-\nu}). \]

By routine algebra, making use of the Hölder continuity of \( K' \), we can prove that there exist constants \( C_{16}, C_{17}, \zeta_3 > 0 \) such that

\[ P\{\forall x, y \in \mathcal{I}_\nu, |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq C_{16} l^{\zeta_3} |x - y|^{\zeta_3}\} = 1 - O(l^{-\nu}) \tag{A.23} \]
for all \( \nu > 0 \). Now, if \( |x - y| \leq (C_{16} l^{C_{17}})^{-2/\zeta_3} \) and \( |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq C_{16} l^{C_{17}} |x - y|^{\zeta_3} \), then

\[
|\hat{\mu}'(x) - \hat{\mu}'(y)| \leq C_{16} l^{C_{17}} |x - y|^{\zeta_3/2} |x - y|^\zeta_3/2
\]

\[
\leq C_{16} l^{C_{17}} \{ (C_{16} l^{C_{17}})^{-2/\zeta_3} \}^{\zeta_3/2} |x - y|^{\zeta_3/2}
\]

\[
= |x - y|^{\zeta_3/2}.
\]

Therefore, defining \( \delta = (C_{16} l^{C_{17}})^{-2/\zeta_3} \), we have by (A.23), for all \( \nu > 0 \),

\[
P\{ \forall x, y \in I_{e'} \text{ with } |x - y| \leq \delta, |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq |x - y|^{\zeta_3/2} \} = 1 - O(l^{-\nu}). \quad \text{(A.24)}
\]

Choose a lattice \( I_{e^o} \subseteq I_{e'} \) with edge width \( \delta = (C_{16} l^{C_{17}})^{-2/\zeta_3} \), and given \( x, y \in I_{e'} \), let \( x^o, y^o \) be the nearest elements of \( I_{e^o} \). By (A.22),

\[
P\{ \text{for some } x, y \in I_{e^o}, |\hat{\mu}'(x) - \hat{\mu}'(y)| > B |x - y|^{\zeta_3} \}
\]

\[
\leq \sum_{x, y \in I_{e^o}} P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| > B |x - y|^{\zeta_3} \}
\]

\[
\leq \{(C_{16} l^{C_{17}})^{2/\zeta_3}\}^{2} \sup_{x, y \in I_{e'}} P\{|\hat{\mu}'(x) - \hat{\mu}'(y)| > B |x - y|^{\zeta_3} \}
\]

\[
= O(l^{4C_{17}/\zeta_3 l^{-\nu}}),
\]

for all \( \nu > 0 \). Hence,

\[
P\{ \text{for some } x, y \in I_{e^o}, |\hat{\mu}'(x) - \hat{\mu}'(y)| > B |x - y|^{\zeta_3} \} = O(l^{-\nu}) \quad \text{(A.25)}
\]

for all \( \nu > 0 \). Now for general \( x, y \in I_{e'} \), with \( |x - y| > \delta \)

\[
|\hat{\mu}'(x) - \hat{\mu}'(y)| \leq |\hat{\mu}'(x) - \hat{\mu}'(x^o)| + |\hat{\mu}'(x^o) - \hat{\mu}'(y^o)| + |\hat{\mu}'(y^o) - \hat{\mu}'(y)|
\]

\[
\leq |x - x^o|^{\zeta_3/2} + B |x^o - y^o|^{\zeta_3} + |y^o - y|^{\zeta_3/2} \equiv T
\]

on the event

\[
E = \{ \forall x, y \in I_{e'} \text{ with } |x - y| \leq \delta, |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq |x - y|^{\zeta_3/2} \}
\]

\[
\cap \{ \forall x, y \in I_{e^o}, |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq B |x - y|^{\zeta_3} \},
\]

where, by (A.24) and (A.25),

\[
P(E) = 1 - O(l^{-\nu})
\]

121
for all \( \nu > 0 \). Let \( \zeta_4 = \min(\zeta_2, \zeta_3/2) \). Then

\[
T \leq 2 \delta \zeta_4 \cdot B |(x - y) + (x^\circ - x) + (y - y^\circ)|^{\zeta_4}
\]
\[
\leq 2 \delta \zeta_4 / 2 + 3 B (|x - y|^{\zeta_4} + 2 \delta \zeta_4)
\]
\[
\leq 2 |x - y|^\zeta_4 / 4 + 3 B (|x - y|^\zeta_4 / 4 + 2 |x - y|^\zeta_4)
\]
\[
\leq (2 + 9B) |x - y|^\zeta_4
\]

Hence, on \( \mathcal{E} \), and assuming \( |x - y| > \delta \),

\[
|\hat{\mu}'(x) - \hat{\mu}'(y)| \leq (2 + 9B) |x - y|^\zeta_4.
\]

Therefore, by (A.24),

\[
P\{ \forall x, y \in \mathcal{I}_e, |\hat{\mu}'(x) - \hat{\mu}'(y)| \leq B' |x - y|^\zeta_4 \} = 1 - O(1^{-\nu})
\]

for all \( \nu > 0 \).

Note that since \( \sum P\{ \sup_{x,y \in \mathcal{I}_e} |\hat{\mu}'(x) - \hat{\mu}'(y)| > B' |x - y|^\zeta_4 \} < \infty \), the Borel-Cantelli Lemma gives \( \sup_{x,y \in \mathcal{I}_e} |\hat{\mu}'(x) - \hat{\mu}'(y)| = B' |x - y|^\zeta_4 \) with probability 1.
Appendix 3A.6
Technical lemmas

Lemma 3.10 For all $k \geq 1$ and $a > 0$,

\[ E(X^k - a^k) \leq C_1(k) + C_2(k) E \{(X - a)^+\}^k. \]

Proof
Now,

\[
E(X^k - a^k) \leq E\{(X^k - a^k) I(X > a)\}
= E\{(X^k - a^k) I(a < X \leq 2a)\} + E\{(X^k - a^k) I(X > 2a)\}
\leq (2a)^k + R, \tag{A.26}
\]

where

\[
R = E\{(X - a)(X^{k-1} + X^{k-2}a + \cdots + a^{k-1}) I(X > 2a)\}
\leq E\{(X - a)(kX^{k-1}) I(X > 2a)\}
\leq 2^{k-1} k E\{(X - a)^k I(X > 2a)\},
\]

using the fact that $X \leq 2(X - a)$ if $X > 2a$. Therefore

\[
R \leq 2^{k-1} k E\{(X - a)^+\}^k, \tag{A.27}
\]

and the result follows from (A.26) and (A.27).

Lemma 3.11 For all $k > 1$,

\[
\left| \sum_{i=1}^{n} X_i \right|^k \leq n^{k-1} \sum_{i=1}^{n} |X_i|^k.
\]

123
Proof Since
\[
\left| \sum_{i=1}^{n} X_i \right| \leq \sum_{i=1}^{n} |X_i| \leq \left( \sum_{i=1}^{n} |X_i|^k \right)^{1/k} \left( \sum_{i=1}^{n} 1 \right)^{1-1/k}
\]
by Hölder’s inequality, it follows that
\[
\left| \sum_{i=1}^{n} X_i \right|^{k} \leq \left\{ n^{(k-1)/k} \left( \sum_{i=1}^{n} |X_i|^k \right)^{1/k} \right\}^k = n^{k-1} \sum_{i=1}^{n} |X_i|^k,
\]
as required.

Lemma 3.12 For all \( k > 1 \),
\[
\mathbb{E} \left\{ \int |\hat{g}(x)| \, dx \right\}^k \leq \left[ \int \left\{ \mathbb{E} |\hat{g}(x)|^k \right\}^{1/k} \, dx \right]^k.
\]
Proof Using Hölder’s inequality we obtain,
\[
\mathbb{E} \left\{ \int |\hat{g}(x)| \, dx \right\}^k = \int \ldots \int \mathbb{E} \left\{ |\hat{g}(x_1)| \times \ldots \times |\hat{g}(x_k)| \right\} \, dx_1 \ldots dx_k
\]
\[
\leq \int \ldots \int \prod_{j=1}^{k} \left\{ \mathbb{E} |\hat{g}(x_j)|^k \right\}^{1/k} \, dx_1 \ldots dx_k
\]
\[
= \left[ \int \left\{ \mathbb{E} |\hat{g}(x)|^k \right\}^{1/k} \, dx \right]^k,
\]
the desired result.

Lemma 3.13 A compactly supported function \( f \) which satisfies a Lipschitz condition of order 1 is of bounded variation.

Proof Let \( \text{supp}(f) \subseteq (-c, c) \) and let \( C \) be a constant such that \( |f(x) - f(y)| < C|x - y| \). Suppose that \( -\infty < \ldots < u_i < -c < u_{i+1} < \ldots < u_j < c < u_{j+1} < \ldots < \infty \).

Then
\[
\sum_{k=1}^{m-1} |f(u_{k+1}) - f(u_k)| = \sum_{k=i}^{j} |f(u_{k+1}) - f(u_k)| \leq 2 \sup |f| + 2cC = C_0,
\]
and \( f \) is of bounded variation.
References


Properties of Random Sequences and Processes. Springer-Verlag, Heidel-
berg.


