ON NONPARAMETRIC CURVE ESTIMATION AND DISCRIMINATION

A thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

from

THE AUSTRALIAN NATIONAL UNIVERSITY

by

Matthew Paul Wand, B.Math., *Wollongong*

Department of Statistics

November 1988

DECLARATION

I hereby certify that this thesis is my own original work with the exception of Sections 2.2, 2.4, 2.6 and 5.2 which contain joint original work with Professor Peter Hall.

M.P. Wand

ACKNOWLEDGEMENTS

I wish to extend warm thanks to my supervisor, Professor Peter Hall for his kindness and guidance throughout the course of this research. The assistance of a Commonwealth Postgraduate Research Award was very much appreciated.

Sincere thanks are due to my family, especially my parents Paul and Christine, my brothers Sean, Ben and Tim and my grandparents Freda and Noel Martin, and Amy and Bob Wand, for their continuing love, support and encouragement.

Finally, I wish to extend special thanks to the residents of Ursula College for the warm friendships and good times which I enjoyed there while preparing this thesis.

M. P. Wand

November 1988

TABLE OF CONTENTS

ABSTRACT

This thesis is largely concerned with the study of nonparametric curve estimation. Given a set of data it is often desirable to obtain estimates of various functions which are related to the distribution of the data, such as probability densities and regression curves, without having to impose rigid parametric assumptions.

We begin our research by considering the problem of kernel density estimation. Traditionally, theory for this problem has been centred around the L_2 norm as a measure of loss. However, attention has recently been given to the L_1 norm as a metric for density estimation – due largely to the monograph of Devroye and Györfi (1985). The work of these authors does not, however, provide for the exact minimisation of asymptotic expected L_1 distance. One of our primary concerns is the formulation of the solution to this minimisation problem and the investigation its ramifications, including L_1 based rules for window-size selection.

The classical nonparametric density estimators are the histogram and the frequency polygon, which still both enjoy a great deal of usage. The thesis continues with the development of asymptotic L_1 theory for these two estimators.

We further investigate the L_1 properties of kernel regression function estimators (in both random and fixed design settings), kernel mode estimators and density derivative estimators.

Nonparametric curve estimation has become an important tool in discriminant analysis. When discriminating between two populations with densities f and g, the likelihood ratio classification rule relies on both f and g. If no suitable parametric model is available then the usual approach is to obtain nonparametric estimates of f and g and form the classification rule from these estimates. We observe that, for the purposes of discrimination, it is more natural to estimate the difference between f and g than to estimate the individual densities themselves. A variant of least-squares cross-validation is shown to be an effective means of choosing smoothing parameters for the density difference estimator. We demonstrate that our methodology applies to both discrete and continuous data.

Chapter One

INTRODUCTION

1.1 Introduction

Nonparametric curve estimation has become an important tool in many aspects of statistics. The summary of data in the form of a histogram, the detection of the number of modes in a distribution, the fitting of a surface through a cloud of data points and the construction of a classification rule in statistical discrimination are all examples of problems that have solutions within the realm of nonparametric curve estimation.

The first published work in this field was that of Rosenblatt (1956). This author introduced a kernel method for the estimation of probability densities. Parzen (1962) expanded this idea, resulting in a flood literature on the topic. Several alternative methods for estimating a density were proposed, based on concepts such as orthogonal series expansions, maximisation of likelihood products, nearest neighbour distances and spline smoothing of histograms. This was paralleled by the extension of density estimation ideas to the estimation of other "curves" such as regression functions, hazard functions and discrimination boundaries. Along with new methodology there came an abundance of theoretical analyses, mainly concerned with establishing consistency results and obtaining rates of convergence. One very important realisation brought out by the theory is the crucial dependence of the performance of nonparametric curve estimators on the choice of their inherent smoothing parameters. Consequently, criteria for the automatic selection of smoothing have recently received considerable attention in the literature. For an extensive account of the theoretical development of nonparametric curve estimation see Prakasa Rao (1983). Relevant survey material can also be found in Fryer (1977), Tapia and Thompson (1978), Bean and Tsokos (1980), Hand (1981, 1982), Collomb (1985), Silverman (1986) and Marron (1988).

1.2 Outline of Thesis

The main contributions of this thesis can be classified into two broad areas: minimisation of L_1 distance in nonparametric curve estimation (Chapters 2, 3 and 4) and nonparametric discrimination using density differences (Chapters 5 and 6). We shall briefly discuss each of these topics in turn.

In nonparametric density estimation there has recently been significant interest in the L_1 metric as a measure of loss, given by $J_n = \int |f_n - f|$ for a density estimate f_n of a density f. Much of this interest has been fuelled by a monograph of Devroye and Györfi (1985) which provides a detailed treatment of L_1 density estimation. These authors favour the L_1 metric since it is always well-defined if f_n is a density and it is invariant under monotone transformations. On the other hand, in the context of kernel density estimation, Devroye and Györfi do not derive exact asymptotic formulae for the optimal window-size or the corresponding rate of convergence of $E(J_n)$. Instead they compute approximations to these quantities. In traditional L_2 theory, results for exact rates of convergence are readily available under suitable conditions (see e.g. Parzen (1962)). For example, if the Epanechnikov kernel $K(x) = (3/4)(1-x^2)$, $|x| < 1$, is in use and f has a continuous second derivative then the L_2-optimal rate of convergence of the window-size h is $c_2 n^{-1/5}$ where

$$c_2 = \left\{ 15 \Big/ \int (f'')^2 \right\}^{1/5}. \tag{1.1}$$

One of the major contributions of this thesis is the development of a procedure which permits the calculation of exact rates of convergence of $E(J_n)$ and the L_1-optimal window-size. Specifically, we show that minimisation of $E(J_n)$ is essentially equivalent to solving an equation of the form $\Lambda(v) = 0$, where Λ is a strictly increasing differentiable function with $\Lambda(0) < 0$ and $\lim_{v \to \infty} \Lambda(v) = \infty$. This allows one to obtain formulae similar to (1.1) when L_1 distance is being minimised. As a consequence, a window-size selection rule which asymptotically minimises L_1 distance can be formulated. The procedure is shown to be applicable to other curve estimators such as histograms, frequency polygons and regression function estimators.

When discriminating between two populations Π_X and Π_Y with densities f and g and prior probabilities p and $1-p$, the error rate is minimised by using the likelihood ratio discrimination rule. This involves classifying z as Π_X if and only if

$$f(z)/g(z) \geq (1-p)/p. \tag{1.2}$$

When f and g are unknown the usual approach is to construct estimates $\hat{f}(\cdot|h_X)$ and $\hat{g}(\cdot|h_Y)$, where h_X and h_Y are smoothing parameters chosen to minimise the distances between $\hat{f}(\cdot|h_X)$ and f, and $\hat{g}(\cdot|h_Y)$ and g respectively. The discrimination rule based on these estimates is that which classifies z as Π_X if and only if

$$\hat{f}(z|h_X)/\hat{g}(z|h_Y) \geq (1-p)/p. \tag{1.3}$$

Notice that (1.2) and (1.3) are equivalent to $e(z) \geq 0$ and $\hat{e}(z|h_X, h_Y) \geq 0$ respectively, where $e = pf - (1-p)g$ and

$$\hat{e}(z|h_X, h_Y) = p\hat{f}(z|h_X) - (1-p)\hat{g}(z|h_Y).$$

The second major theme of this thesis is based on selection of the smoothing parameter pair (h_X, h_Y) to minimise the distance between $\hat{e}(\cdot|h_X, h_Y)$ and e. If the L_2 metric is used as the measure of distance then a version of least-squares cross-validation, related to the selection rules of Rudemo (1982) and Bowman (1984), arises as a selection rule for (h_X, h_Y). An appealing feature of this strategy is that data from both training samples are used in the smoothing parameter choice.

Chapter 2 is devoted to analysis of the L_1 metric in kernel density estimation. Results of Devroye and Györfi (1985) are extended to allow exact minimisation of asymptotic L_1 loss. This theory is subsequently used to construct a data-based rule for selecting a window-size. The rule, which is related to the L_2-based "plug-in" rule proposed by Woodroofe (1970), is shown to be asymptotically optimal in terms of minimising L_1 distance. Numerical results are also obtained which indicate that there is usually very little difference between L_1-optimal and L_2-optimal window-sizes. For example, if f is the standard normal density then

the L_2-optimal window-size of the Epanechnikov kernel estimator is asymptotic to $c_2 n^{-1/5}$ where $c_2 = 2.345\cdots$. However, we demonstrate that the L_1-optimal window-size is asymptotic to $c_1 n^{-1/5}$ where $c_1 = 2.279\cdots$, so there is less than a 3% difference in the rate of shrinkage of L_1- and L_2-optimal window-sizes. This small difference is typical of comparisons made between L_1- and L_2-optimality in this thesis.

In Chapter 3 exact asymptotic L_1 theory is developed for the histogram and the frequency polygon. In addition, we establish bounds for the optimal rate of convergence of $E(J_n)$ for the frequency polygon, and obtain a lower bound for $\liminf_{n\to\infty} n^{2/5} E(J_n)$ for all densities. These bounds extend the theory developed by Devroye and Györfi (1985) to the frequency polygon.

Chapter 4 deals firstly with L_1 theory of kernel regression estimators, in both random design and fixed design settings. This is followed by the development of exact asymptotic theory for mean absolute error (MAE) of the kernel mode estimator. Here it is demonstrated that MAE-optimal window-sizes are quite close to the corresponding optimal window-sizes for minimisation of mean squared error (MSE). For example, if we restrict ourselves to second order kernels then the ratio of MAE-optimal and MSE-optimal shrinkage rates is shown to be $0.8453\cdots$ uniformly over all thrice-differentiable densities. The L_1 theory of density derivative estimation is also briefly covered in this chapter.

In Chapter 5 we investigate the use of density differences for the discrimination of categorical data. The kernel estimator of Aitchison and Aitken (1976) is used to construct density difference estimators for the classification of multivariate binary data. A variant of least-squares cross-validation is employed as a means of automatically selecting the smoothing parameter pair (h_X, h_Y). This is subsequently shown to be asymptotically optimal by arguing as in Bowman, Hall and Titterington (1984). A similar treatment of unstructured multinomial data is also given.

In Chapter 6 we adapt the ideas of Chapter 5 to cater for the discrimination of continuous data. The asymptotic optimality of the window-size selection rule

in this case is proved using techniques similar to those of Stone (1984).

There are two appendices to the thesis, both referred to in Chapter 2. An L_1 asymptotic optimality result is proved via Kolmós-Major-Tusnády techniques in Appendix A. In Appendix B we explain the generalisation of least-squares cross-validation to density derivative estimators.

Throughout this thesis the set of real numbers is denoted by $\mathbf{R}$, while the set of integers is denoted by $\mathbf{Z}$. The univariate normal density function and distribution function are represented by ϕ and Φ respectively. Unqualified integrals are over $\mathbf{R}$ or $\mathbf{R}^d$, depending upon the context. The minimum of two real numbers s and t is denoted by $s \wedge t$.

Chapter Two

MINIMISATION OF L_1 DISTANCE IN KERNEL DENSITY ESTIMATION

2.1 Introduction

One of the most widely-studied and best understood estimators of probability density is the kernel estimator. Let $X_1, X_2, \ldots$ be a sequence of independent random variables with common univariate density f. The kernel estimator based on the sample $X_1, \ldots, X_n$ which we consider is given by

$$f_n(x|h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where K is the kernel function and h is the window-size. Conditions on K and h are given in the next section.

The choice of window-size is of fundamental importance when constructing a kernel estimator. An overly small window-size means that the estimator places too much emphasis on the given sample which induces a high degree of variance; too large a choice of window-size has the effect of "smoothing out" much of the detail of the true density, which corresponds to a large amount of bias. The classical setting for the variance-bias trade-off is that in which loss is measured in terms of L_2 distance, $M_n(h) = \int \{f_n(\cdot|h) - f\}^2$. Expected L_2 loss can be expressed simply in terms of variance and bias as

$$E\{M_n(h)\} = V_n(h) + B_n(h),$$

where $V_n(h) \equiv \int \mathrm{Var}\{f_n(\cdot|h)\}$ and $B_n(h) \equiv \int \{Ef_n(\cdot|h) - f\}^2$. Under suitable assumptions, $V_n(h)$ and $B_n(h)$ each have straightforward asymptotic expansions. The optimal rate of convergence of $E\{M_n(h)\}$ is achieved when the orders of magnitude of $V_n(h)$ and $B_n(h)$ are matched. From this, closed form formulae for the exact optimal shrinkage rate of the window-size are easily obtainable. The main purpose of this chapter is to describe solutions to the same problem for L_1 distance, $J_n(h) = \int |f_n(\cdot|h) - f|$, and to discuss their implications.

When minimising expected L_1 distance the principle of balancing orders of magnitude of variance and squared bias still applies, although it is more appropriate to work with bias and standard deviation instead. Consequently, the optimal order of magnitude of an L_1-optimal window-size is identical to the corresponding L_2-optimal window-size. However, since $E\{J_n(h)\}$ is a complicated function of bias and standard deviation, exact closed form formulae for the L_1-optimal shrinkage rate are not obtainable in general. Nevertheless, we show that numerical solutions can be found by firstly observing that the problem is essentially equivalent to solving an equation of the form $\Lambda(v) = 0$, where the solution is unique, and then appealing to Newton's method.

Another important problem which we address is the difference between minimising L_1 distance and L_2 distance. This can be quantified in terms of how close the L_1-optimal window-sizes are to their L_2 counterparts. We show by example that the difference is only a few percent in the majority of cases.

In the context of L_2 loss, a number of window-size selection procedures which utilise asymptotic formulae for optimal window-sizes have been proposed, such as those suggested by Woodroofe (1970), Scott, Tapia and Thompson (1977) and Silverman (1986 p.45). Our algorithm for minimising L_1 distance makes it possible to obtain L_1 versions of each of these procedures.

Section 2 describes our approach to asymptotic minimisation of $E\{J_n(h)\}$. Section 3 extends this approach to the asymptotic minimisation of related measures of loss, including that of general L_q loss for $q \geq 1$. The problem of window-size selection is investigated in Section4, leading to an L_1-asymptotically optimal selection rule. Numerical results are presented and discussed in Section 5. All proofs are deferred to Section 6.

2.2 L_1 Theory of the Kernel Estimator

Throughout this section it is assumed that K is a real-valued, measurable function such that

$$\int z^j K(z)\,dz = \begin{cases} 0, & j = 1, \ldots, p-1; \\ (-1)^p \kappa_1 \neq 0, & j = p. \end{cases}$$

We also assume that K is bounded, has compact support and integrates to unity. The window-size $h = h(n)$ is a sequence of positive real numbers such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

We commence with a brief summary of L_2 theory for the kernel estimate. Assuming $f^{(p)}$ is continuous and non-zero at x, and $f(x) > 0$, straightforward calculations can be used to show that the asymptotic bias and variance of $f_n(\cdot|h)$ are given by

$$Ef_n(x|h) - f(x) \sim (\kappa_1/p!)f^{(p)}(x)h^p$$

and

$$\mathrm{Var}\{f_n(x|h)\} \sim \kappa_2^2 f(x)(nh)^{-1},$$

where $\kappa_2 = (\int K^2)^{\frac{1}{2}}$ (see e.g. Parzen (1962)). Letting b and σ denote the functions $(\kappa_1/p!)f^{(p)}$ and $\kappa_2 f^{\frac{1}{2}}$ we obtain from these that

$$E\{M_n(h)\} = h^{2p}\int b^2 + (nh)^{-1}\int \sigma^2 + o\{h^{2p} + (nh)^{-1}\}. \tag{2.1}$$

To asymptotically minimise M_n, the window-size is chosen so that bias and standard deviation are of the same order of magnitude. This is achieved by taking h equal to $h_u = u^2 n^{-p/(2p+1)}$ to give

$$E\{M_n(h_u)\} = \left(u^{4p}\int b^2 + u^{-2}\int \sigma^2\right) n^{-2p/(2p+1)} + o\{n^{-2p/(2p+1)}\},$$

and then choosing u to minimise the leading coefficient

$$\lambda_2(u) = \int b^2\, u^{4p} + \int \sigma^2\, u^{-2}.$$

The value at which λ_2 attains its minimum, u_2^* say, is easily obtained from calculus to be

$$(u_2^*)^2 = \left[\frac{\int \sigma^2}{2p\int b^2}\right]^{1/(2p+1)} = \left[\frac{(p!)^2\kappa_2^2}{2p\kappa_1^2\int\{f^{(p)}\}^2}\right]^{1/(2p+1)},$$

which provides us with the well-known formula for the L_2 asymptotically optimal window-size

$$h_2^* = h_{u_2^*} = (u_2^*)^2 n^{-1/(2p+1)} \tag{2.2}$$

with corresponding L_2 error

$$E\{M_n(h_2^*)\} = (2p+1)(2p)^{-2p/(2p+1)} \left(\int \sigma^2\right)^{2p/(2p+1)} \left(\int b^2\right)^{1/(2p+1)} n^{-2p/(2p+1)} + o\{n^{-2p/(2p+1)}\}.$$

Window-size selection rules based on (2.2) have been discussed in the literature. The L_2-optimal coefficient may be written

$$(u_2^*)^2 = \alpha(K)\beta(f)$$

where

$$\alpha(K) = \{(p!)^2\kappa_2^2/(2p\kappa_1^2)\}^{1/(2p+1)}$$

depends on K and is known and $\beta(f) = [\int\{f^{(p)}\}^2]^{-1/(2p+1)}$ depends on the unknown density. Woodroofe (1970), in the local density estimation context, suggested using an initial window-size to estimate $f^{(p)}$ and "plugging" this estimate into the formula for the L_2-optimal window-size to obtain the final choice. Scott, Tapia and Thompson (1977), in the case where $p = 2$, proposed choosing h to be the largest solution to the equation

$$h = \alpha(K)\beta\{f_n(\cdot|h)\}n^{-1/5},$$

and showed that a solution could be arrived at by iteration. Silverman (1986, p.45) has suggested using the normal distribution as a standard reference. This involves observing that if f is normal with variance δ^2 then

$$\int (f'')^2 = (3/8)\pi^{-\frac{1}{2}}\delta^{-5},$$

so that if a Gaussian kernel is in use then

$$h_2^* = (4/3)^{1/5}\delta n^{-1/5} = (1.06\cdots)\delta n^{-1/5}.$$

The selection rule involves replacing δ by a data-based estimate $\hat{\delta}$ to select a window-size.

To derive the L_1 equivalent of the above L_2 results we begin with the L_1 analogue of (2.1) which states that

$$E\{J_n(h)\} = \int (nh)^{-\frac{1}{2}}\sigma\psi\left(\frac{(nh^{2p+1})^{\frac{1}{2}}b}{\sigma}\right) + o\{h^p + (nh)^{-\frac{1}{2}}\},$$

where

$$\psi(t) = 2t\Phi(t) + 2\phi(t) - t, \qquad t \in \mathbf{R}.$$

This is the pth order version of the second order kernel result presented in Theorem 5.1 of Devroye and Györfi (1985, p.78). There it was assumed that f has two continuous derivatives and compact support. We shall see later that the assumption of compact support can be replaced by a weak moment condition. The function ψ is symmetric. Note also that

$$\psi'(t) = 2\Phi(t) - 1$$

and

$$\psi''(t) = 2\phi(t),$$

so that ψ is monotonic decreasing for negative t, monotonic increasing for positive t and convex everywhere. The reason for ψ appearing in the formula for asymptotic L_1 loss is essentially the fact that for all $t \in \mathbf{R}$,

$$\psi(t) = E|Z - t| \tag{2.3}$$

where Z is a N(0,1) random variable. One result for ψ which is used extensively in the proofs of this thesis is given in Lemma 5.14 of Devroye and Györfi (1985, p.93). We shall state it here for convenience.

Lemma 2.1. *If t, u, v and w are nonnegative numbers then*

$$|t\psi(u/t) - v\psi(w/v)| \leq |u - w| + (2/\pi)^{\frac{1}{2}}|t - v|.$$

As for L_2 loss the optimum is achieved by taking the bias and standard deviation of $f_n(x|h)$ to be of the same order of magnitude. Thus we again take h equal to h_u and obtain

$$E\{J_n(h_u)\} = u^{-1}\int \sigma\psi\left(\frac{u^{2p+1}b}{\sigma}\right)n^{-p/(2p+1)} + o\{n^{-p/(2p+1)}\}. \tag{2.4}$$

Recall that b and σ are both functions of x, say, and that integration on the right-hand side of (2.4) is with respect to x. The asymptotic L_1-optimal value of u will be the minimiser of

$$\lambda(u) = u^{-1}\int \sigma\psi\left(\frac{u^{2p+1}b}{\sigma}\right).$$

Differentiation with respect to u yields

$$\lambda'(u) = 2u^{-2}\Lambda(u^{2p+1})$$

where

$$\Lambda(v) = \int \sigma[2pvb/\sigma\{\Phi(vb/\sigma) - \tfrac{1}{2}\} - \phi(vb/\sigma)].$$

The value of u which minimises $\lambda(u)$, call it u^*, is given by $u^* = (v^*)^{1/(2p+1)}$ where v^* is the solution of $\Lambda(v) = 0$ $(v > 0)$. Note that $\Lambda(0) = -(2\pi)^{-\frac{1}{2}}\kappa_2 \int f^{\frac{1}{2}} < 0$, $\lim_{v\to\infty} \Lambda(v) = \infty$ and

$$\Lambda'(v) = \int b[2p\{\Phi(vb/\sigma) - \tfrac{1}{2}\} + (2p+1)vb/\sigma\phi(vb/\sigma)],$$

which is positive for all $v > 0$. This entails the existence and uniqueness of v^*. Its value can be found quickly by Newton's method as the limit of the sequence $v_1, v_2, \ldots$ where $v_{i+1} = v_i - \Lambda(v_i)/\Lambda'(v_i)$. The asymptotic L_1-optimal window-size when estimating f is therefore

$$h^* = (u^*)^2 n^{-1/(2p+1)}.$$

If the value of v^*, and hence u^*, is found for a particular pth order kernel K then it is trivial to calculate u_0^*, the minimiser of $\lambda(u)$ if a different kernel K_0 is in use. For K_0 the values of κ_1 and κ_2 will be different, $\kappa_{0,1}$ and $\kappa_{0,2}$ say, in which case

$$u_0^* = u^*\{(\kappa_1\kappa_{2,0})/(\kappa_2\kappa_{1,0})\}^{1/(2p+1)}.$$

The derivation of the exact asymptotic formula for the L_1-optimal window-size makes it possible to create data-based procedures for selecting h along the same lines as those mentioned above in the L_2 case. These possibilities are discussed in Sections 4 and 5.

The following theorem gives sufficient conditions on f for (2.4) to hold in a uniform sense.

Theorem 2.1. *If $E|X_1|^{1+\epsilon} < \infty$ for some $\epsilon > 0$; if f is bounded; and if $f^{(p)}$ is bounded, continuous and integrable; then*

$$\lim_{n\to\infty} \sup_{u\in[C^{-1},C]} |n^{p/(2p+1)} E\{J_n(h_u)\} - \lambda(u)| = 0 \qquad (2.5)$$

for every $C \geq 1$. *Furthermore,*

$$\inf_{h>0} E\{J(h)\} \sim \lambda(u^*) n^{-p/(2p+1)}, \tag{2.6}$$

where u^* *is the unique value of* u *which minimises* $\lambda(u)$.

The condition $E|X_1|^{1+\epsilon} < \infty$ is only slightly stronger than $\int f^{\frac{1}{2}} < \infty$, the latter condition being necessary for the function λ to be well-defined.

Before closing this section we shall briefly describe the extension of the L_1 theory developed here to multivariate nonnegative kernel estimators. For this we assume that the sample points $X_1, \ldots, X_n$ are $\mathbf{R}^d$-valued and K is a symmetric, bounded, compactly supported, d-variate probability density function. For $1 \leq j \leq d$ we assume that $\kappa_1 = \int_{\mathbf{R}^d} z_j^2 K(z)\, dz$ is independent of j and $\kappa_2 = (\int K^2)^{\frac{1}{2}}$. The estimator which we consider is given by

$$f_n(x|h) = (nh^d)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}$$

with corresponding L_1 loss $J_n(h) = \int_{\mathbf{R}^d} |f_n(\cdot|h) - f|$. If f and all second order derivatives are bounded and continuous then, provided $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, we have at each point $x \in \mathbf{R}^d$,

$$E f_n(x|h) - f(x) \sim b_d(x) h^2 \equiv (\kappa_1/2) \nabla^2 f(x) h^2,$$

and

$$\mathrm{Var}\{f_n(x|h)\} \sim \sigma_d^2(x)(nh^d)^{-\frac{1}{2}} \equiv \kappa_2 f(x)(nh^d)^{-1}$$

where $\nabla^2 f(x) = \sum_{j=1}^{d} (\partial/\partial x_j)^2 f(x)$. Taking h equal to $h_u = u^2 n^{-1/(d+4)}$ one obtains

$$E\{J_n(h_u)\} = u^{-d} \int \sigma_d \psi \left(\frac{u^{4+d} b_d}{\sigma_d} \right) n^{-2/(d+4)} + o\{n^{-2/(d+4)}\}. \tag{2.7}$$

Again, σ_d and b_d are functions of x, and integration is with respect to x. The exact asymptotic formula for the L_1-optimal window-size and the corresponding L_1 loss can be found in exactly the same manner as the univariate case by locating the value of u that minimises the coefficient of $n^{-2/(d+4)}$ in (2.7).

2.3 Extensions to Other Measures of Loss

Some simple adaptations of the theory developed in Section 2 to cater for other measures of loss will be made in this section. The estimator with which we deal is the one-dimensional kernel estimator defined at the beginning of the previous section. The smoothness conditions imposed on f in Section 2 will also be assumed throughout this section. The functions b and σ defined there have the same definitions in this section.

2.3.1 *Minimisation of Mean Absolute Error*

When one is interested in estimating f at a single point $x \in \mathbf{R}$ for which $f^{(p)}$ and f are continuous and non-zero then the analogue of expected L_1 loss is mean absolute error (MAE), which is given by

$$\mathrm{MAE}\{f_n(x|h)\} = E|f_n(x|h) - f(x)|$$

and has the asymptotic representation

$$\mathrm{MAE}\{f_n(x|h)\} = (nh)^{-\frac{1}{2}}\sigma(x)\psi\left(\frac{(nh^{2p+1})^{\frac{1}{2}}b(x)}{\sigma(x)}\right) + o\{h^p + (nh)^{-\frac{1}{2}}\}$$

as $h \to 0$ and $nh \to \infty$. Again we balance the bias and standard deviation orders of magnitude by taking h equal to h_u:

$$\mathrm{MAE}\{f_n(x|h_u)\} = u^{-1}\sigma(x)\psi\left(\frac{u^{2p+1}b(x)}{\sigma(x)}\right)n^{-p/(2p+1)} + o\{n^{-p/(2p+1)}\}.$$

It is easily seen that the MAE-optimal window-size for estimating f at x is $(u_x^*)^2 n^{-1/(2p+1)}$ where $u_x^* = (v_x^*)^{1/(2p+1)}$ and v_x^* is the unique solution of

$$2pvb(x)/\sigma(x)\left[\Phi\{vb(x)/\sigma(x)\} - \tfrac{1}{2}\right] - \phi\{vb(x)/\sigma(x)\} = 0.$$

If we let $c_1(x) = (u_x^*)^2$ be the MAE-optimal coefficient and

$$c_2(x) = [\sigma^2(x)/\{2pb^2(x)\}]^{1/(2p+1)}$$

denote the corresponding optimal coefficient for asymptotic minimisation of mean squared error, then graphs of c_1 and c_2 are almost identical (see Hall and Wand

(1988)). In fact, Schucany (1988) has shown that, uniformly in f and second order kernels K, $c_1(x) = (0.985\cdots)c_2(x)$, proving that minimisation of mean absolute error in pointwise kernel density estimation is virtually equivalent to that of mean squared error.

2.3.2 *Minimisation of L_q Loss*

For any value of $q \geq 1$ we shall define L_q loss to be

$$J_{n,q}(h) = \int |f_n(\cdot|h) - f|^q.$$

Note that for $q > 1$ this is not the same as the L_q norm of the difference between $f_n(\cdot|h)$ and f since the latter would be $J_{n,q}(h)^{1/q}$. Extending the theory of L_1 loss to L_q loss, one obtains

$$E\{J_{n,q}(h)\} = \int \{(nh)^{-\frac{1}{2}}\sigma\}^q \psi_q\left(\frac{(nh^{2p+1})^{\frac{1}{2}}b}{\sigma}\right) + o\{h^p + (nh)^{-\frac{1}{2}}\},$$

where $\psi_q(t) = E|Z - t|^q$ and Z is a normal N(0,1) random variable. This implies that

$$E\{J_{n,q}(h_u)\} = u^{-q}\int \sigma^q \psi_q\left(\frac{u^{2p+1}b}{\sigma}\right) n^{-qp/(2p+1)} + o\{n^{-qp/(2p+1)}\} \qquad (3.1)$$

and so optimality is reached by choosing u to minimise

$$\lambda_q(u) = u^{-q}\int \sigma^q \psi_q\left(\frac{u^{2p+1}b}{\sigma}\right).$$

The ease with which this minimisation can be performed for a particular q depends crucially on the form of the function ψ_q. When $q = 2$ we are dealing with expected L_2 loss, or mean integrated squared error, and ψ_q has the simple form $\psi_2(t) = t^2 + 1$ which allows straightforward derivation of closed form expressions for the asymptotically optimal window-size and corresponding asymptotic expected loss. If estimating a density in L_4 then one has to deal with $\psi_4(t) = t^4 + 6t^2 + 3$ which, with the assistance of Newton's method, allows the required L_4-optimal coefficients to be readily computed. Odd integral values of q lead to much more complicated expressions for ψ_q. In the previous section we saw that $\psi_1(t) = \psi(t) = 2t\Phi(t) + 2\phi(t) - t$. When $q = 3$ one obtains

$$\psi_3(t) = 12t\Phi^{(-2)}(t) - 12t^2\Phi^{(-1)}(t) + 8t^3\Phi(t) + 4(2t^2 + 1)\phi(t) - t^3 - 3t$$

where $\Phi^{(-1)}(t) = \int_{-\infty}^{t} \Phi(z)\,dz$ and $\Phi^{(-2)}(t) = \int_{-\infty}^{t} \Phi^{(-1)}(z)\,dz$.

2.3.3 *Minimisation of Weighted L_q Loss*

Suppose that a weight function $w \geq 0$ is included in the L_q loss formula, so that our aim is to minimise

$$E\{J_{n,q,w}(h)\} = E \int |f_n(\cdot|h) - f|^q w.$$

In this case (3.1) is generalised to

$$E\{J_{n,q,w}(h)\} = u^{-q} \int \sigma^q w \psi_q \left(\frac{u^{2p+1} b}{\sigma}\right) n^{-qp/(2p+1)} + o\{n^{-qp/(2p+1)}\},$$

which is minimised in the usual way.

2.4 L_1 Window-size Selection

The most important choice to be made when estimating the density f by the kernel estimator defined in Section 2 is the value of the window-size parameter h, since it controls the trade-off between smoothing of sample noise and estimating the fine detail of the density. It was established in Theorem 2.2 that, as $n \to \infty$,

$$E\{J_n(h_u)\} \sim \lambda(u) n^{-p/(2p+1)} \tag{4.1}$$

uniformly in $u \in [-C, C]$ for each $C \geq 1$, and

$$\inf_{h>0} E\{J_n(h)\} \sim \lambda(u^*), \tag{4.2}$$

so that, in terms of minimising asymptotic expected loss, the optimal choice of h is $(u^*)^2 n^{-1/(2p+1)}$ where u^* is the unique value of u that minimises $\lambda(u)$. Recall that

$$\lambda(u) = u^{-1} \int \sigma \psi \left(\frac{u^{2p+1} b}{\sigma}\right),$$

which depends on the functions b and σ. A window-size selection procedure based on minimisation of λ would clearly require estimation of b and σ since they in turn depend on the unknown functions $f^{(p)}$ and $f^{\frac{1}{2}}$. Suppose that b_n and σ_n are L_1-consistent estimators of b and σ respectively; that is,

$$\lim_{n\to\infty} \int |b_n - b| = 0 \quad \text{and} \quad \lim_{n\to\infty} \int |\sigma_n - \sigma| = 0; \tag{4.3}$$

where convergence is in the almost sure sense in both cases. We shall also assume that $\sigma_n \geq 0$. Examples of b_n and σ_n will be considered shortly. An estimator of λ would then be

$$\lambda_n(u) = u^{-1} \int \sigma_n \psi \left(\frac{u^{2p+1} b_n}{\sigma_n} \right),$$

which leads to the window-size selection rule

$$h_n^* = (u_n^*)^2 n^{-1/(2p+1)} \tag{4.4}$$

where u_n^* is the unique value of u that minimises $\lambda_n(u)$. This minimisation can be performed in exactly the same way as described for minimisation of $\lambda(u)$.

From (4.3) and Lemma 2.1 it follows that, for any $C > 1$,

$$\lim_{n \to \infty} \sup_{u \in [-C,C]} |\lambda_n(u) - \lambda(u)| = 0 \tag{4.5}$$

almost surely. This entails $\lim_{n\to\infty} \lambda_n(u_n^*) = \lambda(u^*)$ almost surely and $\lim_{n\to\infty} u_n^* = u^*$ almost surely. Recall that $h^* = (u^*)n^{-1/(2p+1)}$ is the L_1 asymptotically optimal window-size. Then we also have

$$\lim_{n \to \infty} h_n^*/h^* = 1$$

almost surely and, using (4.1) and (4.5), we obtain

$$\lim_{n \to \infty} [E\{J_n(h)\}]_{h=h_n^*} / E\{J_n(h^*)\} = 1 \tag{4.6}$$

almost surely. Additionally, we have from (4.2) and (4.6),

$$\lim_{n \to \infty} [E\{J_n(h)\}]_{h=h_n^*} / \inf_{h>0} E\{J_n(h)\} = 1 \tag{4.7}$$

almost surely which means that, in the context of minimising expected loss, the selection rule h_n^* is asymptotically optimal. This is a very attractive property since it means that the window-size selected by (4.4) is asymptotically as good as the L_1-optimal window-size. The *stochastic* equivalent of this result is

$$\lim_{n \to \infty} J_n(h_n^*) / \inf_{h>0} J_n(h) = 1 \tag{4.8}$$

almost surely which, if it were true, would mean that h_n^* is also asymptotically optimal in terms of minimising raw L_1 loss $J_n(h)$. The following theorem allows us to prove (4.8) under certain regularity conditions.

Theorem 4.1. *If K is compactly supported and Hölder continuous, and if f is bounded, then*

$$\lim_{n\to\infty} \{\inf_{h>0} J_n(h)\}/[\inf_{h>0} E\{J_n(h)\}] = 1 \tag{4.9}$$

almost surely, and

$$\lim_{n\to\infty} J_n(h_n^*)/[E\{J_n(h)\}]_{h=h_n^*} = 1 \tag{4.10}$$

almost surely.

It follows from (4.7), (4.9) and (4.10) that (4.8) is true under the conditions imposed in Theorem 4.1. A result very similar to (4.8) is proved in an entirely different manner in Appendix A.

Candidates for the estimators b_n and σ_n will now be proposed. Noting that $b = (\kappa_1/p!)f^{(p)}$ and $\sigma = \kappa_2 f^{\frac{1}{2}}$ we shall use $b_n \equiv (\kappa_1/p!)f_n^{(p)}(\cdot|h_1)$ and $\sigma_n \equiv \kappa_2 f_n^{\frac{1}{2}}(\cdot|h_2)$, where

$$f_n^{(p)}(x|h_1) = (nh_1^{p+1})^{-1} \sum_{i=1}^{n} K_1^{(p)}\{(x - X_i)/h_1\}$$

and

$$f_n^{\frac{1}{2}}(x|h_2) = \left[(nh_2)^{-1} \sum_{i=1}^{n} K_2\{(x - X_i)/h_2\}\right]^{\frac{1}{2}},$$

in which K_1 is a p times differentiable kernel and K_2 is a nonnegative kernel. Of course, $f_n^{(p)}(\cdot|h_1)$ and $f_n^{\frac{1}{2}}(\cdot|h_2)$ are obtained respectively by p-fold differentiation and taking the square-root of kernel density estimators with window-sizes h_1 and h_2. Appropriate choices of h_1 and h_2 will be discussed in Section 5.

The nonnegativity of K_2 ensures that $f_n^{\frac{1}{2}}(\cdot|h_2)$ is well-defined and that $\sigma_n \geq 0$. We also require the L_1-consistency of b_n and σ_n, which follows directly from

Theorem 4.2. *Assume K_1 and K_2 are bounded, compactly supported and integrate to unity; $K_1^{(p)}$ exists and is bounded; K_2 is nonnegative; $E|X_1|^{1+\epsilon} < \infty$ for some $\epsilon > 0$; f is bounded; $f^{(p)}$ is bounded, continuous and integrable; $h_1, h_2 \to 0$, $nh_1^{2p+1}/\log n$ and $nh_2 \to \infty$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} \int |f_n^{(p)}(\cdot|h_1) - f^{(p)}| = 0 \tag{4.11}$$

almost surely, and

$$\lim_{n\to\infty} \int |f_n^{\frac{1}{2}}(\cdot|h_2) - f^{\frac{1}{2}}| = 0 \tag{4.12}$$

almost surely.

Our window-size selection rule can now be fully described. Choose K_1, K_2, h_1 and h_2 to satisfy the conditions imposed in Theorem 4.2, to form the estimators b_n and σ_n. These estimators should then be used in the formula for $\lambda_n(u)$ to locate the minimum at $u = u_n^*$. The window-size to be used in the pth order kernel estimator is $h_n^* = (u_n^*)^2 n^{-1/(2p+1)}$. If the pth order kernel K satisfies the conditions imposed at the beginning of Section 2 and is Hölder continuous, and if the density f satisfies the conditions of Theorem 4.2, then the selection rule is asymptotically optimal in terms of minimising L_1 loss and expected L_1 loss.

2.5 Examples and Discussion

Examples of L_1-optimal window-sizes and the corresponding rates of convergence of $E(J_n)$ to zero will be presented in Subsection 1 of this section. We shall then describe an implementation of the L_1 window-size selection rule proposed in the previous section, and apply it to some simulated sets of data (Subsection 2). Comparisons with corresponding L_2 based window-sizes are made in both theoretical and data-based contexts. The section concludes with the application of the L_1 window-size selection rule to the analysis of a real data set (Subsection 3).

2.5.1 *Examples of L_1-optimal Window-sizes and Rates of Convergence*

Throughout this subsection, the only kernel with which we deal is the Bartlett-Epanechnikov kernel

$$K(x) = (3/4)(1 - x^2), \quad -1 \le x \le 1,$$

for which $p = 2$. In this case $\kappa_1 = 1/5$, $\kappa_2^2 = 3/5$, $b = (1/10)f''$ and $\sigma = (3/5)^{\frac{1}{2}} f^{\frac{1}{2}}$. Also, as described in Section 2, the L_1-optimal window-size is asymptotic to $h^* = (u^*)^2 n^{-1/5}$ where u^* is the value of u which minimises

$$\lambda(u) = u^{-1} \int (3/5)^{\frac{1}{2}} f^{\frac{1}{2}} \psi\left(\frac{u^5 f''}{(60f)^{\frac{1}{2}}}\right).$$

To find u in practice we first locate v^* as the limit of the sequence $v_1, v_2, \cdots$, where $v_{i+1} = v_i - H(v_i)$ and the function H is given by

$$H(v) = 60^{\frac{1}{2}} \left[\int f^{\frac{1}{2}} \{4rv\Phi(vr) - \phi(vr)\} \right] \left[\int f'' \{4\Phi(vr) + 5vr\phi(vr)\} \right]^{-1} \quad (5.1)$$

with $r = f''/(60f)^{\frac{1}{2}}$. As shown in Section 2, u^* is then given by $u^* = (v^*)^{1/5}$. We are interested in the coefficient $(u^*)^2$, which we shall call c_1, and in comparing it to the L_2-optimal coefficient c_2 which has the formula

$$c_2 = \left\{ 15 / \int (f'')^2 \right\}^{1/5}.$$

With this notation the L_1- and L_2-optimal window-sizes are asymptotic to $c_1 n^{-1/5}$ and $c_2 n^{-1/5}$ respectively.

We now give example of the values of c_1 and c_2 for several different densities. Three of these will be members of the two-component, equal proportion normal mixture family with means (-1,1) and variances (σ_1^2, σ_2^2). This family of densities has general form

$$f(x; \sigma_1^2, \sigma_2^2) = \tfrac{1}{2}(2\pi\sigma_1^2)^{-\frac{1}{2}} e^{-(x+1)^2/(2\sigma_1^2)} + \tfrac{1}{2}(2\pi\sigma_2^2)^{-\frac{1}{2}} e^{-(x-1)^2/(2\sigma_2^2)}$$

and will be denoted by $\mathrm{NM}(\sigma_1^2, \sigma_2^2)$. The other densities that we consider are the standard normal distribution $(f(x) = (2\pi)^{-\frac{1}{2}} e^{-x^2/2})$, the Beta(4,4) distribution $(f(x) = 140x^3(1-x)^3,\ 0 \le x \le 1)$, the extreme value density $(f(x) = e^x e^{-e^x})$, the logistic density $(f(x) = e^x(e^x+1)^{-2})$ and Student's t_2 distribution $(f(x) = (x^2+2)^{-3/2})$.

In Table 5.1 we list the values of c_1, c_2 and their ratio c_1/c_2. The values of c_1 and c_2 are remarkably close in every case. This closeness means that, from a practical viewpoint, there is virtually no difference between L_1 smoothing and L_2 smoothing in density estimation for the vast majority of cases. However, Devroye and Györfi (1985, p.109) point out that there is a considerable difference between the L_1-optimal coefficients and L_2-optimal coefficients for heavy-tailed distributions. Working with an approximation to c_1 they demonstrate that when the Cauchy density is approached through the family of Student's t densities, the

Table 5.1: Values of c_1, c_2 and c_1/c_2 for the Bartlett-Epanechnikov kernel.

Density	c_1	c_2	c_1/c_2
N(0,1)	2.279	2.345	0.972
NM(1,1)	3.013	3.257	0.925
NM(1/5,1/5)	1.146	1.183	0.969
NM(1,1/10)	0.980	1.033	1.054
Beta (4,4)	0.377	0.422	0.893
Extreme Value	2.236	2.268	0.986
Logistic	3.743	3.630	1.031
Student's t_2	2.792	2.341	1.193

Table 5.2: Values of $CA(K)B(f)$, $D_1(f)$ and $C^*A(K)B(f)$ for the Bartlett-Epanechnikov kernel.

Density	$CA(K)B(f)$	$D_1(f)$	$C^*A(K)B(f)$
$N(0,1)$	1.002	1.022	1.341
Beta (4,4)	0.936	0.993	1.253
Extreme Value	1.090	1.169	1.459
Logistic	1.092	1.119	1.462

L_1-optimal coefficient approaches ∞ whereas c_2 is finite for the Cauchy density. This means that for a small value of $\epsilon > 0$, the Student's $t_{1+\epsilon}$ distribution will exhibit a large value of c_1/c_2, although Table 5.1 verifies that even for ϵ as small as 1 (that is, for Student's t_2 distribution) there is only about a 20% difference between c_1 and c_2.

It is also seen from Table 5.1 that the normal standard reference rule discussed in Section 2 would take the form $\hat{h}_2 = 2.345\hat{\delta}n^{-1/5}$ when the Bartlett-Epanechnikov kernel is in use. The analogue of this rule for L_1 loss takes the form $\hat{h}_1 = 2.279\hat{\delta}n^{-1/5}$, since $c_1 = 2.279\cdots$ for the N(0,1) density.

Observe from (2.6) that the optimal rate of convergence to zero of $E(J_n)$ is $D_1(f)n^{-2/5}$ where $D_1(f) = \lambda(c_1^{\frac{1}{2}})$. Bounds for value of $D_1(f)$ have been derived by Devroye and Györfi (1985, pp.78-79). If $B(f) = \{\frac{1}{2}(\int f^{\frac{1}{2}})^4 \int |f''|\}^{1/5}$ and $A(K) = \kappa_1^{1/5}\kappa_2^{4/5}$ then

$$CA(K)B(f) \leq D_1(f) \leq C^*A(K)B(f) \tag{5.2}$$

where $C = \inf_{t>0} \psi(t)/t^{1/5} = 1.028493\cdots$ and $C^* = 5(8\pi)^{-2/5}$. Since we are working with the Bartlett-Epanechnikov kernel, (5.2) can be written

$$(0.60767\cdots)B(f) \leq D_1(f) \leq (0.81346\cdots)B(f).$$

Table 5.2 compares these three quantities for four particular densities. Notice that in each case the lower bound is remarkably accurate. The upper bound is not quite as good since it relies on a rather crude approximation based on the inequality $\psi(u) \leq u + (2/\pi)^{\frac{1}{2}}$.

2.5.2 *Implementation of the L_1 Window-size Selection Rule and Simulation*

In Section 3 an L_1 asymptotically optimal window-size selection rule was proposed. In this subsection we shall discuss its implementation when $p = 2$, and also its application to some simulated data.

To make the selection rule fully automatic a way had to be found for choosing h_1 and h_2. We decided to choose h_2, for the estimation of $f^{\frac{1}{2}}$, via least-squares cross-validation. The window-size h_1, required for the estimation of f'', was found

by a generalisation of least-squares cross-validation to cater for estimation of density derivatives. Details of this extension are given in Appendix B and Härdle, Marron and Wand (1989). A problem which has been observed when using least-squares cross-validation to select a window-size is the occasional occurrence of a spurious minimum at an unacceptably small window-size. This window-size provides a curve estimate which is far too noisy so we selected h_1 and h_2 to be the largest values at which a minimum occurs. This strategy provided reasonable choices for almost every sample. The chosen window-sizes will be denoted by $\hat{h}_1$ and $\hat{h}_2$. To avoid the numerical computation of integrals in our cross-validatory criterion functions we took K_1 and K_2, as well as K, to be the Gaussian kernel $N(x) = (2\pi)^{-\frac{1}{2}} e^{-x^2/2}$ which, along with its derivatives, has simple convolution properties (see Appendix B). The window-size selected by this procedure is denoted by h_n^* (see (4.4)).

A small simulation was run to test the procedure for the estimation of the logistic density ($f(x) = e^x(1+e^x)^{-2}$) and the extreme value density ($f(x) = e^x e^{-e^x}$). For each of these densities 15 samples of size $n = 200$ and 5 samples of size $n = 400$ were simulated and the resulting estimates compared to the estimate obtained if $\hat{h}_2$ was used instead. The L_1-optimal window-size for the logistic density is asymptotic to $1.691n^{-1/5}$ which assumes the value 0.586 when $n = 200$ and 0.511 when $n = 400$. For the extreme value density the corresponding optimal window-size is $1.010n^{-1/5}$ which is 0.350 when $n = 200$ and 0.305 when $n = 400$. Tables 5.3 and 5.4 list the values of h_n^* and $\hat{h}_2$ obtained from the simulation study. The closeness of the two selected window-sizes for each sample is quite remarkable since h_n^* is obtained from an L_1 "plug-in" rule while $\hat{h}_2$ is based on least-squares (L_2) cross-validation. These rules asymptotically minimise L_1 loss and L_2 loss respectively, so we have further evidence that there is little difference between L_1 and L_2 smoothing.

Figures 5.1 and 5.2 depict "average case" estimates obtained from the study where, for each density and each sample size, we have plotted the graph of the estimates obtained using the median performance value of h_n^*. The performance

Table 5.3 (a): Values of h_n^* and $\hat{h}_2$ for logistic data and $n = 200$ (15 replications).

Rep. no.	h_n^*	$\hat{h}_2$
1	0.6985	0.6917
2	0.4743	0.4550
3	0.6117	0.6131
4	0.4296	0.5074
5	0.6318	0.5998
6	0.6287	0.6423
7	0.7315	0.7136
8	0.5376	0.5509
9	0.7785	0.0870
10	0.7806	0.8243
11	0.3909	0.3955
12	0.6602	0.6547
13	0.8035	0.8390
14	0.5881	0.5677
15	0.6644	0.6361

Table 5.3 (b): Values of h_n^* and $\hat{h}_2$ for logistic data and $n = 400$ (5 replications).

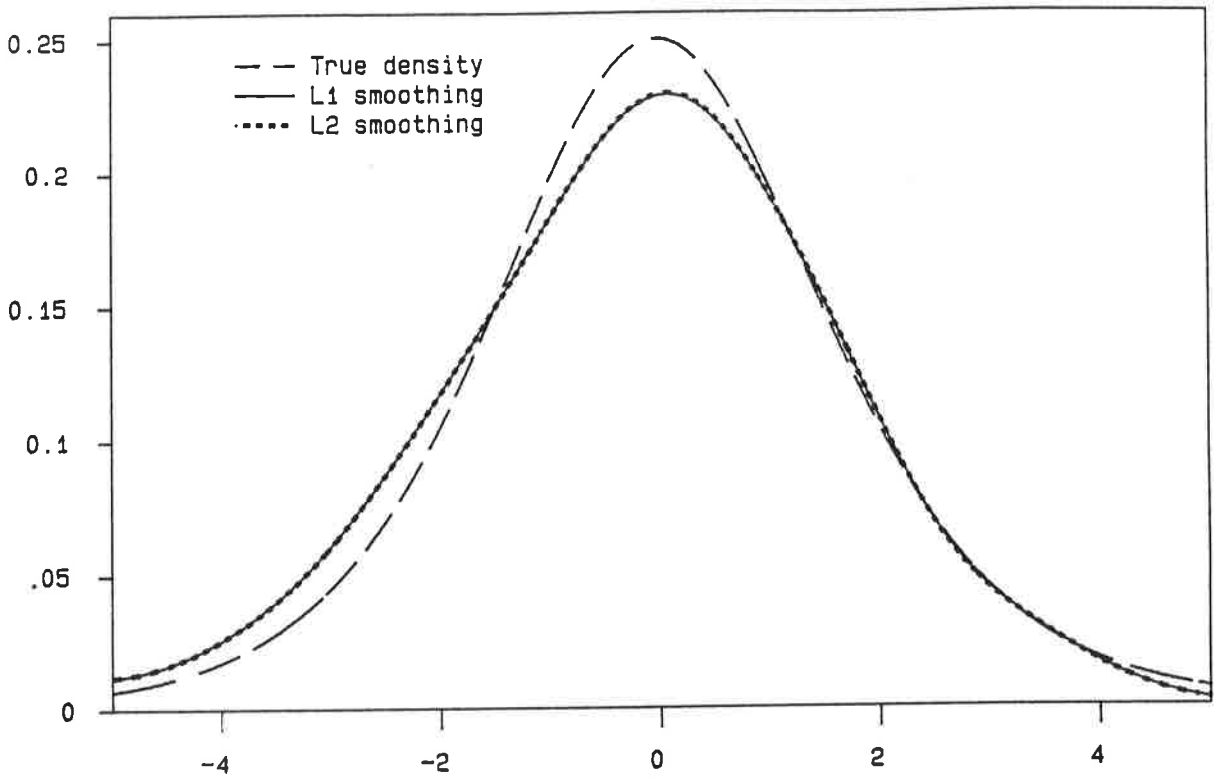
Rep. no.	h_n^*	$\hat{h}_2$
1	0.5879	0.5914
2	0.6193	0.5794
3	0.5692	0.5551
4	0.5785	0.5257
5	0.4081	0.4562

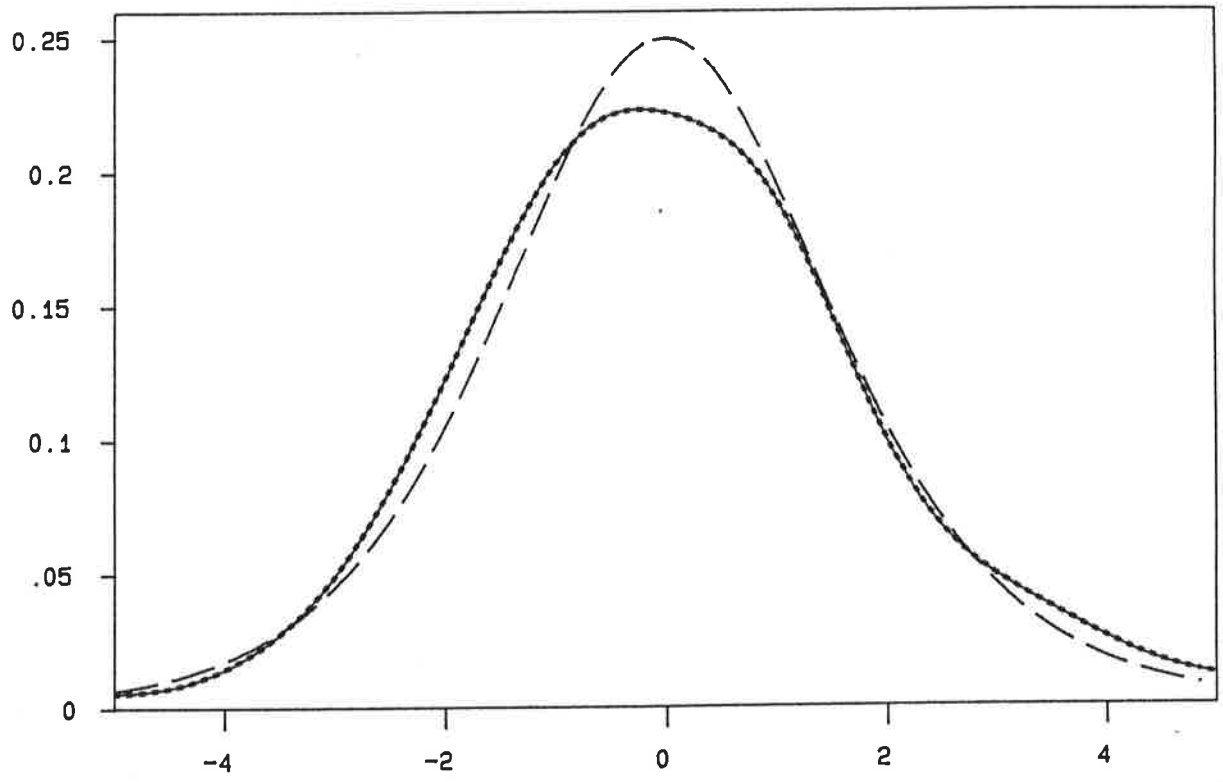
Table 5.4 (a): Values of h_n^* and $\hat{h}_2$ for extreme value data and $n = 200$ (15 replications).

Rep. no.	h_n^*	$\hat{h}_2$
1	0.4296	0.4455
2	0.4086	0.3996
3	0.3629	0.3535
4	0.0751	0.1047
5	0.1933	0.3665
6	0.4480	0.4315
7	0.1055	0.1645
8	0.3684	0.3526
9	0.5065	0.5371
10	0.3993	0.3943
11	0.3693	0.3495
12	0.4778	0.5042
13	0.4529	0.4638
14	0.4200	0.4161
15	0.1564	0.2042

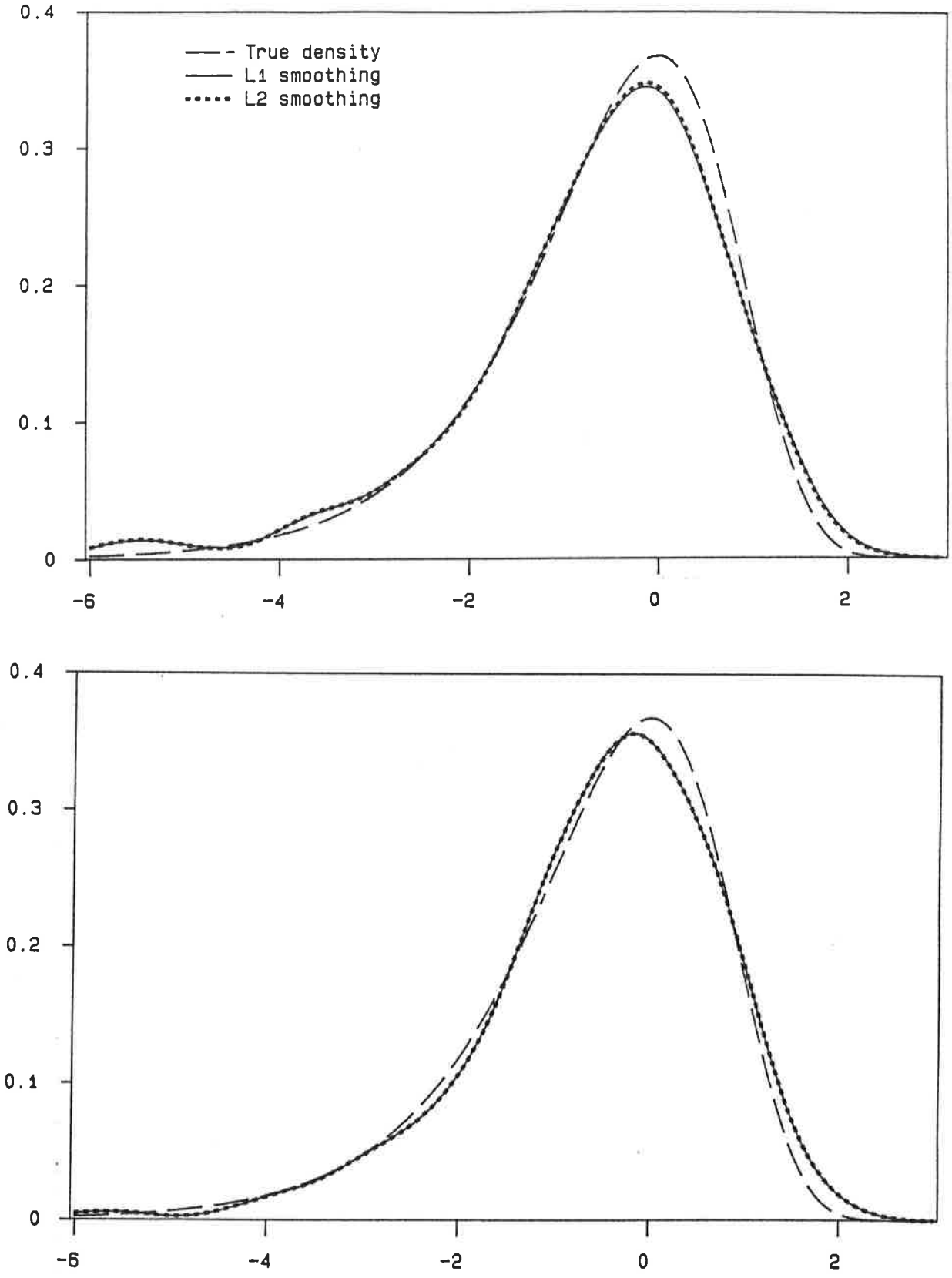
Table 5.4 (b): Values of h_n^* and $\hat{h}_2$ for extreme value data and $n = 400$ (5 replications).

Rep. no.	h_n^*	$\hat{h}_2$
1	0.3747	0.3760
2	0.3679	0.3690
3	0.2355	0.2062
4	0.3316	0.3218
5	0.3793	0.3872

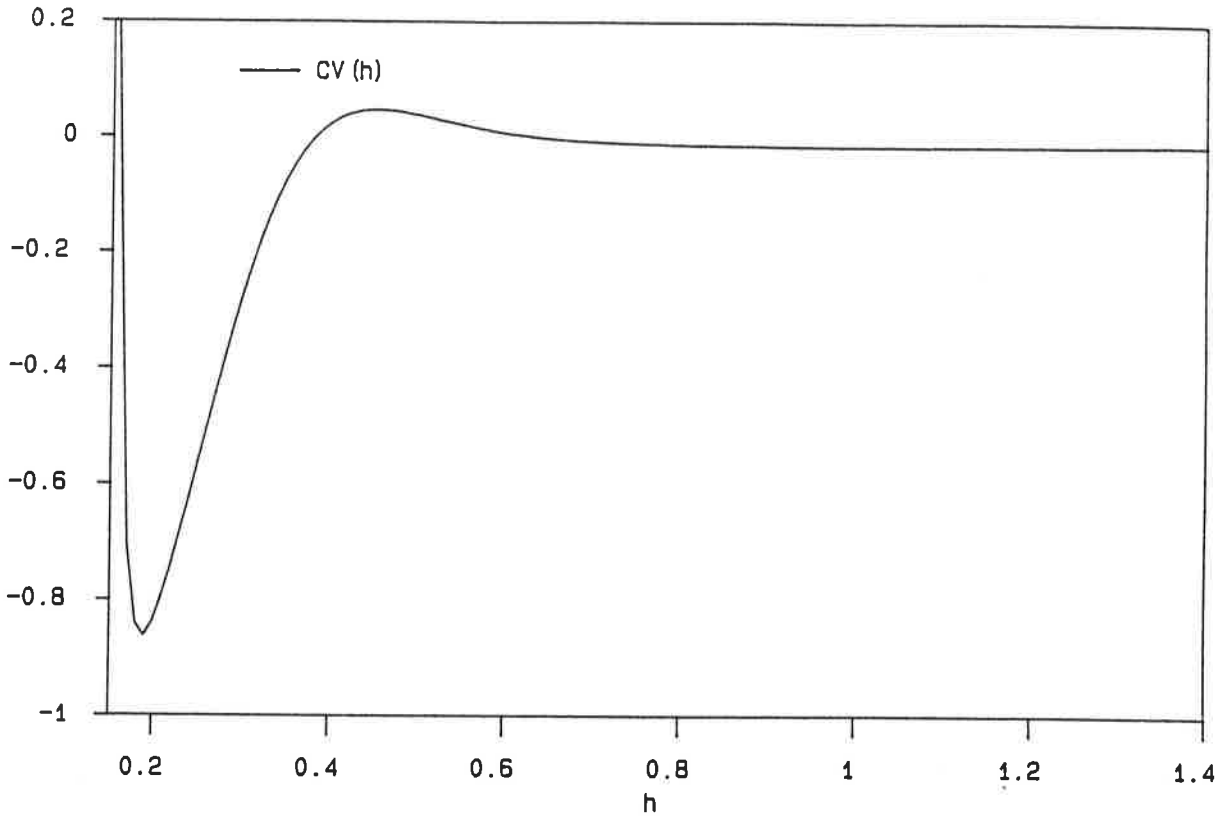


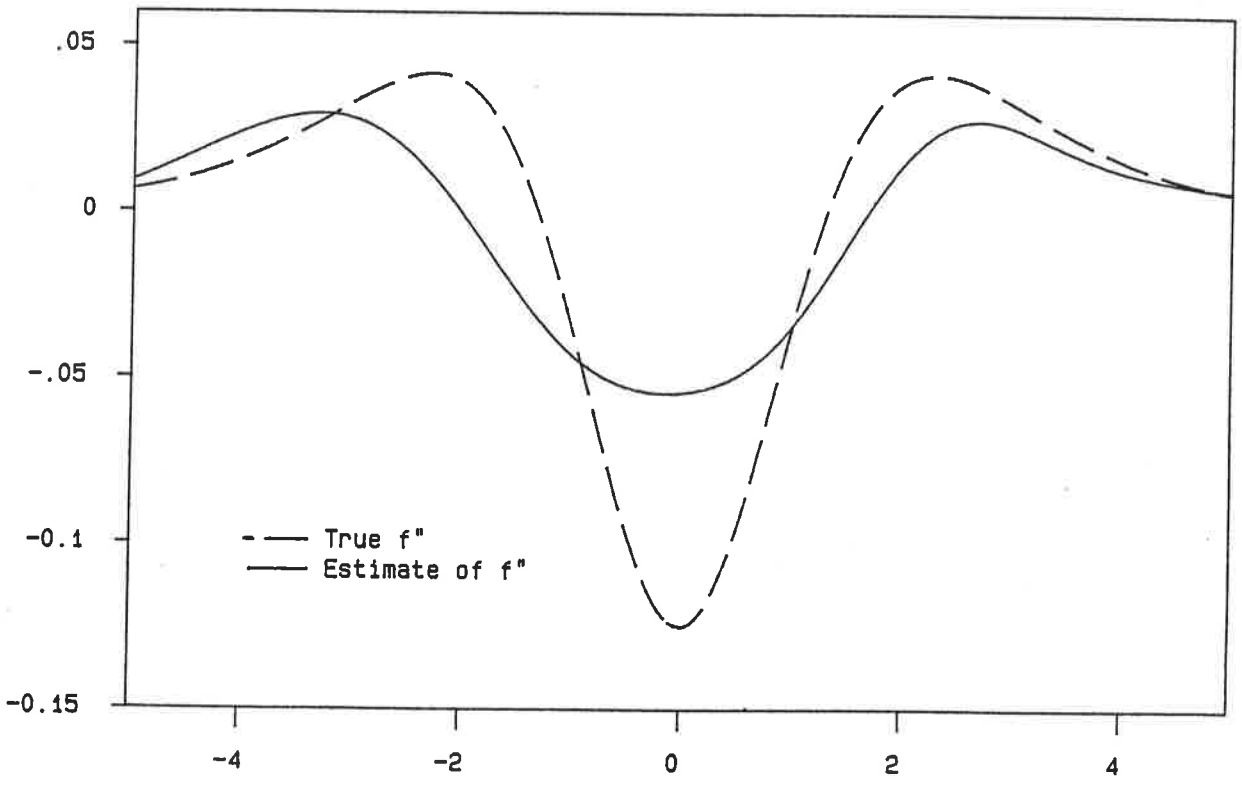


Figures 5.1 (a) and 5.1 (b): Typical estimates of the logistic density when (a) $n = 200$ and (b) $n = 400$. The broken curve is f; the unbroken curve is $f_n(\cdot|h_n^*)$ and the dotted curve is $f_n(\cdot|\hat{h}_2)$.

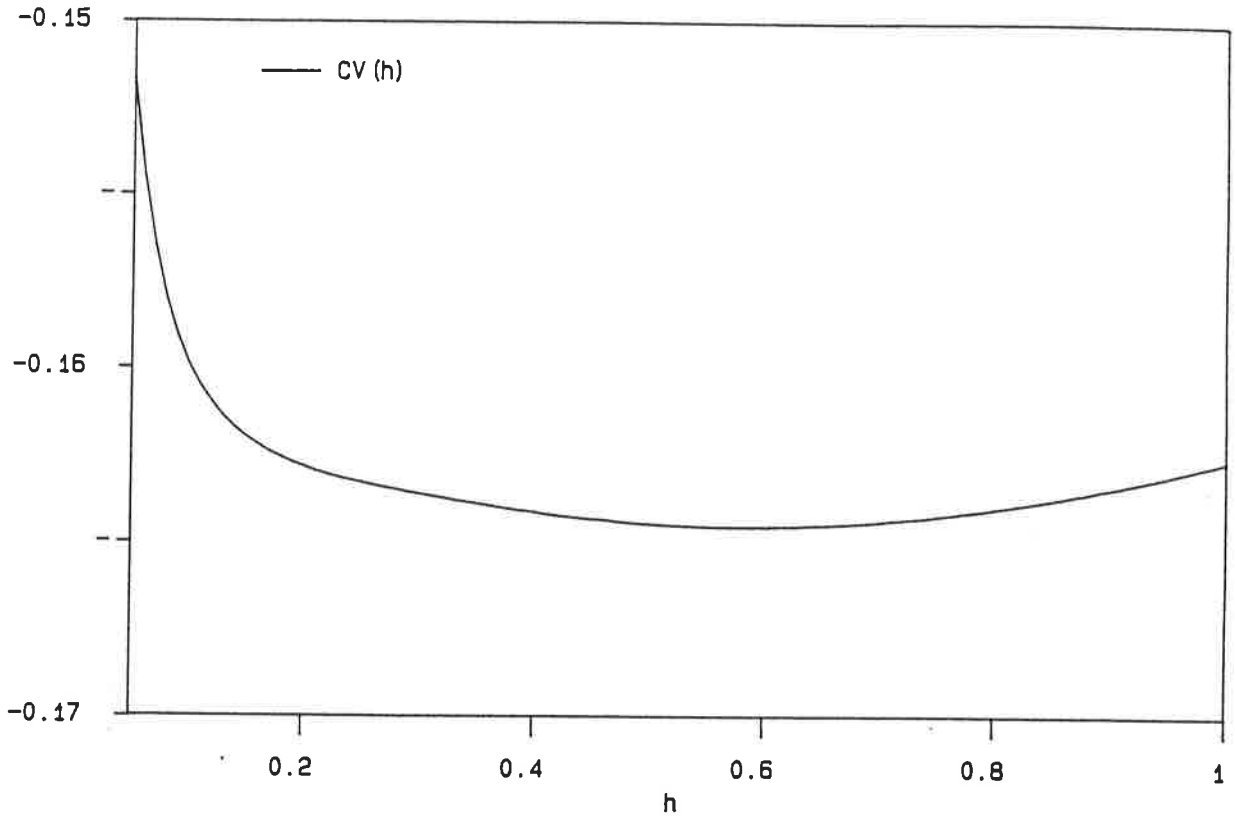


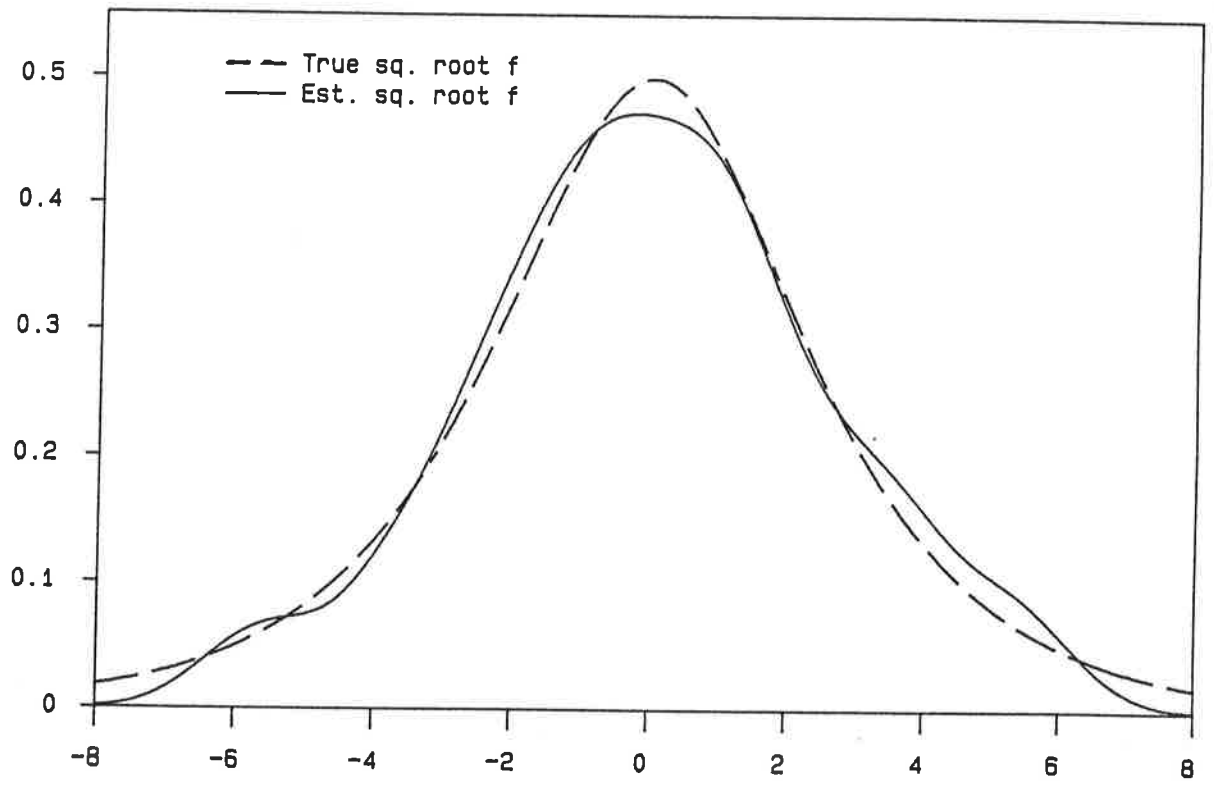
Figures 5.2 (a) and 5.2 (b): Typical estimates of the extreme value density when (a) $n = 200$ and (b) $n = 400$. The broken curve is f; the unbroken curve is $f_n(\cdot|h_n^*)$ and the dotted curve is $f_n(\cdot|\hat{h}_2)$.





Figures 5.3 (a) and 5.3 (b): Typical "pilot" estimator for f'' when f is the logistic density and $n = 400$. Figure 5.3 (a) is the cross-validatory criterion function for the selection of $\hat{h}_1$. In Figure 5.3 (b) the broken curve is f''; the unbroken curve is $f_n''(\cdot|\hat{h}_1)$.





Figures 5.3 (c) and 5.3 (d): Typical "pilot" estimator for $f^{\frac{1}{2}}$ when f is the logistic density and $n = 400$. Figure 5.3 (c) is the cross-validatory criterion function for the selection of $\hat{h}_2$. In Figure 5.3 (d) the broken curve is $f^{\frac{1}{2}}$; the unbroken curve is $f_n{}^{\frac{1}{2}}(\cdot|\hat{h}_1)$.

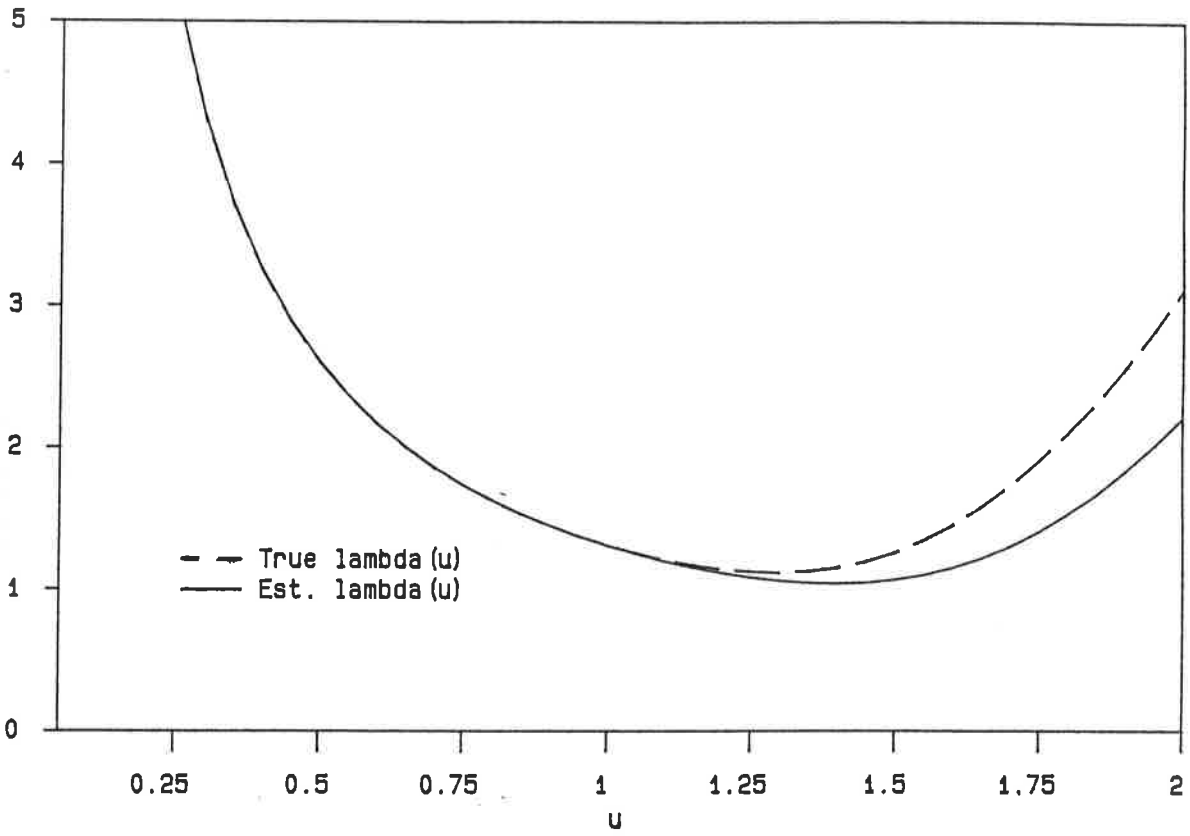


Figure 5.3 (e): Typical estimate of λ when f is the logistic density and $n = 400$. The broken curve is λ; the unbroken curve is λ_n.

Table 5.5: Annual rainfall data for Adelaide, South Australia for years 1839 to 1977. Measurements are in metres.

Year	Rain	Year	Rain	Year	Rain	Year	Rain
1839	1.985	1874	1.723	1909	2.759	1944	1.713
1840	2.423	1875	2.920	1910	2.462	1945	1.785
1841	1.795	1876	1.344	1911	1.599	1946	2.259
1842	2.032	1877	2.492	1912	1.957	1947	2.189
1843	1.709	1878	2.210	1913	1.817	1948	2.140
1844	1.701	1879	2.070	1914	1.148	1949	1.823
1845	1.882	1880	2.247	1915	1.938	1950	1.606
1846	2.621	1881	1.805	1916	2.817	1951	2.554
1847	2.909	1882	1.578	1917	2.890	1952	2.000
1848	1.976	1883	2.696	1918	2.664	1953	2.001
1849	2.555	1884	1.878	1919	1.721	1954	1.674
1850	1.984	1885	1.589	1920	2.670	1955	2.458
1851	3.187	1886	1.442	1921	2.264	1956	2.726
1852	2.745	1887	2.569	1922	2.320	1957	1.675
1853	2.713	1888	1.457	1923	2.979	1958	1.757
1854	1.535	1889	3.087	1924	2.344	1959	1.132
1855	2.315	1890	2.578	1925	2.191	1960	2.307
1856	2.494	1891	1.401	1926	2.220	1961	1.497
1857	2.221	1892	2.153	1927	1.692	1962	1.796
1858	2.155	1893	2.152	1928	1.943	1963	2.455
1859	1.488	1894	2.078	1929	1.751	1964	2.192
1860	1.972	1895	2.128	1930	1.865	1965	1.336
1861	2.360	1896	1.517	1931	2.232	1966	1.951
1862	2.186	1897	1.542	1932	2.504	1967	1.011
1863	2.378	1898	2.075	1933	2.222	1968	2.572
1864	1.983	1899	1.884	1934	2.033	1969	2.068
1865	1.551	1900	2.175	1935	2.345	1970	1.901
1866	2.014	1901	1.801	1936	1.934	1971	2.626
1867	1.910	1902	1.643	1937	2.305	1972	1.756
1868	1.999	1903	2.574	1938	1.926	1973	2.662
1869	1.482	1904	2.031	1939	2.329	1974	2.519
1870	2.386	1905	2.228	1940	1.616	1975	2.060
1871	2.355	1906	2.651	1941	2.256	1976	1.454
1872	2.269	1907	1.778	1942	2.544	1977	1.571
1873	2.106	1908	2.456	1943	1.774		

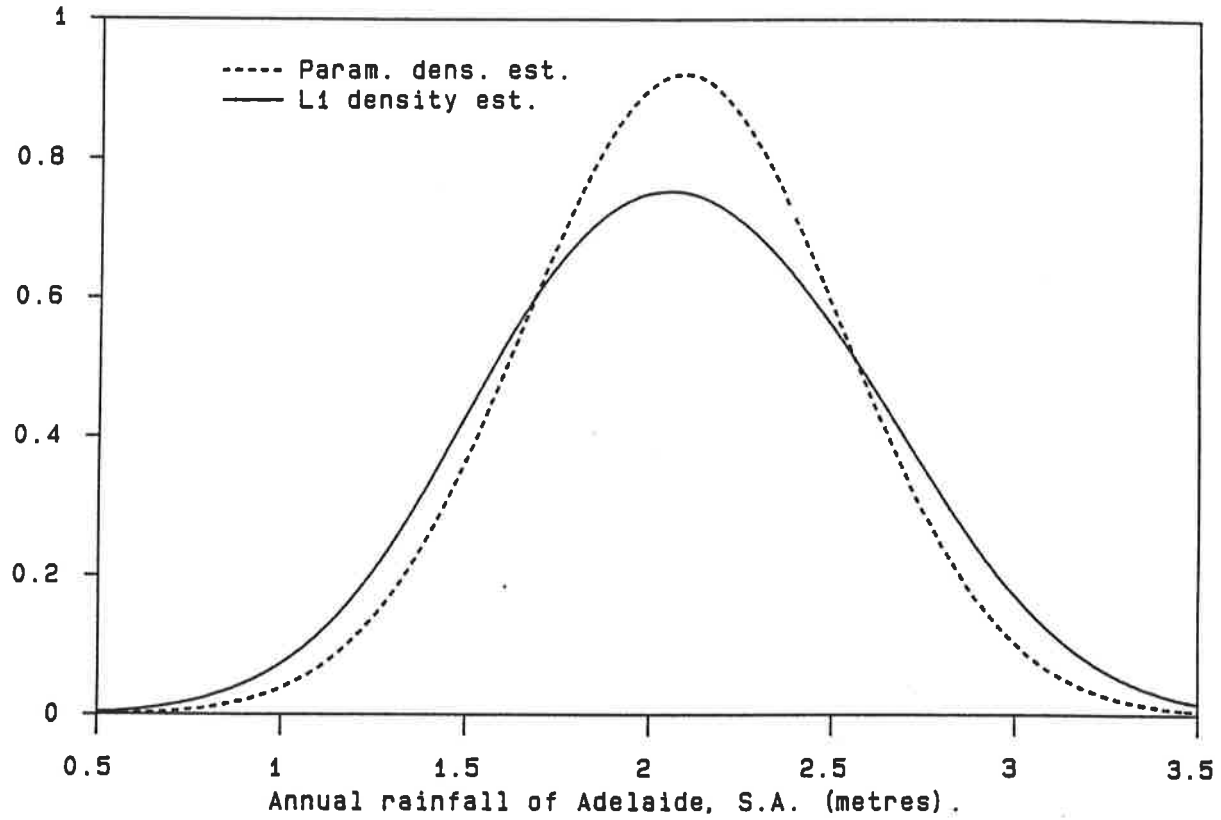


Figure 5.4: Parametric and nonparametric density estimates of annual rainfall of Adelaide, South Australia. The broken curve is the density of the $N(\overline{x}, s^2)$ distribution where $\overline{x}$ and s^2 are the sample mean and variance respectively; the unbroken curve is $f_{139}(\cdot|h^*_{139})$.

of h_n^* is measured in terms of minimising $J_n(h)$ and the median is obtained over replications of the same density and sample size. The graphs of $f_n(\cdot|\hat{h}_2)$ and f are also plotted for comparison. Finally, for the median performance estimate out of the extreme value samples of size $n = 400$ (replication number 1 in Table 5.3 (b)) we also have plotted the cross-validatory criteria and curve estimates involved in the various stages of the selection of h_n^*. The graph in Figure 5.3 (a) is the cross-validatory criterion function needed for selection of $\hat{h}_1$ ($\hat{h}_1 = 1.0035$). Note the occurrence of a spurious minimum near $h = 0.2$. Figure 5.3 (b) shows the "pilot" estimate ${f''}_n(\cdot|\hat{h}_1)$ compared to $f''(x) = e^x(e^{2x} - 4e^x + 1)(e^x + 1)^{-2}$. The cross-validatory criterion function for the selection of $\hat{h}_2$ ($\hat{h}_2 = 0.5914$) is plotted in Figure 5.3 (c) and the "pilot" estimate of $f^{\frac{1}{2}}$ is plotted along with $f^{\frac{1}{2}}$ in Figure 5.3 (d). Figure 5.3 (e) depicts the function $\lambda(u)$ and its estimate $\lambda_n(u)$. The minimum of $\lambda_n(u)$ is attained at $u_n^* = 1.396$, which is reasonably close to $u^* = 1.300$.

2.5.3 *The Adelaide Rainfall Data*

We applied the L_1 window-size selection rule to annual rainfall data for Adelaide, South Australia. The data were obtained from Table 15.1 of Andrews and Herzberg (1985) where rainfall readings, measured in millimetres, are listed in six-day totals for the period 1839 to 1977. The annual totals for these 139 consecutive years constituted our sample. For numerical convenience we converted the totals to metres. See Table 5.5 for these data. In the discussion adjoining the raw data (Andrews and Herzberg (1985, p.105)) it was noted that a 23 year cycle is evident in the data, as well as a gradual trend in winter rainfall. However, since we are using the data for purely illustrative purposes, it was assumed that the annual totals are independent and identically distributed.

With a Gaussian kernel in use the L_1 window-size selection rule chose $h_{139}^* = 0.2697$. The density estimate based on this window-size is plotted in Figure 5.4. The parametric estimate if the assumption of normality is imposed; that is, the normal density function with mean and variance set to their respective sample versions, is also plotted for comparison. It seems very reasonable to conclude that

the annual rainfall for Adelaide has a symmetric distribution. The density estimate appears to be heavier in the tails than the normal distribution. However, this may be a symptom of the fact that, for relatively small sample sizes, window-size selection rules often tend to oversmooth (compare Figure 5.1). Further analysis should include a test for normality of the data.

2.6 Proofs

Throughout this section the symbols $C, C_0, C_1, C_2, \ldots$ will be used to denote positive generic constants, possibly having different values at different appearances.

Proof of Theorem 2.1.

We begin by defining for $C > 0$,

$$I(n,h,C) = \int_{|x|>C} E|f_n(\cdot|h) - f|$$

and

$$\lambda(u,C) = u^{-1} \int_{|x|\le C} \sigma\psi\left(\frac{u^{2p+1}b}{\sigma}\right),$$

where ψ, b and σ are the functions introduced in Section 2. We wish to first establish that for any $C > 0$ and $0 < h \le 1$,

$$I(n,h,C) \le g(C)\{h^p + (nh)^{-\frac{1}{2}}\}, \tag{6.1}$$

where g does not depend on n or h and $\lim_{C\to\infty} g(C) = 0$. Let

$$I_1(n,h,C) = \int_{|x|>C} |Ef_n(\cdot|h) - f|$$

and

$$I_2(n,h,C) = \int_{|x|>C} [\mathrm{Var}\{f_n(\cdot|h)\}]^{\frac{1}{2}}.$$

It follows easily from Liapounov's inequality that $I \le I_1 + I_2$, so (6.1) will hold if it is shown that

$$I_1(n,h,C) \le g_1(C)h^p \tag{6.2}$$

and

$$I_2(n,h,C) \le g_2(C)(nh)^{-\frac{1}{2}}, \tag{6.3}$$

where $g_1(C)$ and $g_2(C)$ each converge to zero as C diverges to infinity.

By Taylor's theorem with remainder and the assumption that K is of order p,

$$\begin{aligned} Ef_n(x|h) &= \int_{-\infty}^{\infty} K(z) f(x - hz)\, dz \\ &= f(x) + \int_{-\infty}^{\infty} K(z) \int_0^1 h^p \frac{(1-t)^{p-1}}{(p-1)!} f^{(p)}(x - tzh)\, dt\, dz, \end{aligned}$$

which leads to

$$(p-1)!|Ef_n(x|h) - f(x)| = \left| h^p \int_{-\infty}^{\infty} K(z) \int_0^1 f^{(p)}(x - thz)(1-t)^{p-1}\, dt\, dz \right|. \quad (6.4)$$

Therefore we have

$$\begin{aligned} I_1(n, h, 2C) \leq & h^p \int_{|x|>2C} \int_{|hz|\leq C} |K(z)| \int_0^1 |f^{(p)}(x - thz)|\, dt\, dz\, dx \\ & + h^p \int_{|x|>2C} \int_{|hz|>C} |K(z)| \int_0^1 |f^{(p)}(x - thz)|\, dt\, dz\, dx. \end{aligned} \quad (6.5)$$

Observe that if $|hz| \leq C$ and $0 < t < 1$ then

$$\{x : |x| > 2C|\} \subseteq \{x : |x - thz| > C\}.$$

Thus the first term on the right-hand side of (6.5) is bounded above by

$$\begin{aligned} & h^p \int_{|hz|\leq C} |K(z)| \int_0^1 \int_{|x-thz|>C} |f^{(p)}(x - thz)|\, dx\, dt\, dz \\ & \qquad \leq h^p \left(\int |K| \right) \int_{|y|>C} |f^{(p)}(y)|\, dy. \end{aligned}$$

If $0 < h \leq 1$ then the second term is no more than

$$h^p \int_{|hz|>C} |K(z)|\, dz \int |f^{(p)}| \leq C^{-p} h^p \int |z^p K(z)|\, dz \int |f^{(p)}|.$$

Therefore $I_1(n, h, 2C) \leq g_1(2C) h^p$ where

$$g_1(2C) = \left(\int |K| \right) \int_{|y|>C} |f^{(p)}(y)|\, dy + C^{-p} \int |z^p K(z)|\, dz \int |f^{(p)}|.$$

Clearly, $\lim_{C\to\infty} g_1(C) = 0$, which concludes the proof of (6.2).

To prove (6.3) we let $\alpha > 1$ and note that by the Cauchy-Schwarz inequality,

$$I_2(n, h, C) \leq g_2^*(C) \left[\int \mathrm{Var}\{f_n(x|h)\}(1 + |x|^\alpha)\, dx \right]^{\frac{1}{2}}$$

where

$$g_2^*(C) = \left\{ \int_{|x|>C} (1+|x|^\alpha)^{-1}\, dx \right\}^{\frac{1}{2}}. \tag{6.6}$$

Trivially,

$$1 + |x|^\alpha \le 2^\alpha (1 + |x - hz|^\alpha + |hz|^\alpha), \tag{6.7}$$

and

$$\mathrm{Var}\{f_n(x|h)\} \le (nh)^{-1} \int K^2(z) f(x - hz)\, dz. \tag{6.8}$$

Combining (6.7) and (6.8) we obtain for $0 < h \le 1$

$$\begin{aligned}
&\int \mathrm{Var}\{f_n(x|h)\}(1+|x|^\alpha)\, dx \\
&\qquad \le (nh)^{-1} 2^\alpha \int\int K^2(z) f(x-hz)(1 + |x-hz|^\alpha + |hz|^\alpha)\, dx\, dz \\
&\qquad \le (nh)^{-1} 2^\alpha \left\{ \left(\int K^2 \right) \int (1+|y|^\alpha) f(y)\, dy \right. \\
&\qquad\qquad \left. + \int |z|^\alpha K^2(z) \int f(x-hz)\, dx\, dz \right\} \\
&\qquad = (nh)^{-1} 2^\alpha \left\{ \left(\int K^2 \right)(1 + E|X_1|^\alpha) + \int |z|^\alpha K^2 \right\},
\end{aligned}$$

which is finite for $\alpha = 1 + \epsilon$. Therefore we have $I_2(n, h, C) \le g_2(C)(nh)^{-\frac{1}{2}}$, where

$$g_2(C) = g_2^*(C) 2^{\alpha/2} \left\{ \left(\int K^2 \right)(1 + E|X_1|^\alpha) + \int |z|^\alpha K^2 \right\}^{\frac{1}{2}}.$$

Clearly $g_2(C)$ approaches zero as C approaches infinity, as required.

To establish (2.5) it is sufficient to show that for every $C_0 \ge 1$ and $C_1 > 0$,

$$\sup_{u \in [C_0^{-1}, C_0]} n^{p/(2p+1)} I(n, h_u, C_1) \le g_3(C_1) \tag{6.9}$$

for all sufficiently large n, where $\lim_{C_1 \to \infty} g_3(C_1) = 0$;

$$\lim_{n\to\infty} \sup_{u \in [C_0^{-1}, C_0]} \left| n^{p/(2p+1)} \int_{|x| \le C_1} E|f_n(\cdot|h_u) - f| - \lambda(u, C_1) \right| = 0; \tag{6.10}$$

and

$$\sup_{u \in [C_0^{-1}, C_0]} |\lambda(u, C_1) - \lambda(u)| \le g_4(C_1), \tag{6.11}$$

where $\lim_{C_1 \to \infty} g_4(C_1) = 0$. The functions g_3 and g_4 are each independent of n. Result (6.9) is easily derived from (6.1) by putting $h = h_u$ and taking n large enough to give

$$\sup_{u \in [C_0^{-1}, C_0]} n^{p/(2p+1)} I(n, h_u, C_1) \leq (C_0 + C_0^{2p}) g(C_1).$$

Condition (6.10) is a simple extension of Theorem 5.1 of Devroye and Györfi (1985, p.78) with $h = h_u$. Devroye and Györfi deal with second order kernels and compactly supported densities. However, the extension to pth order kernels and uniform convergence is straightforward. Observe that the left-hand side of (6.11) is dominated by

$$g_4(C_1) = C_0^{2p}(\kappa_1/p!) \int_{|x|>C_1} |f^{(p)}(x)|\, dx + (2/\pi)^{\frac{1}{2}} C_0 \kappa_2 \int_{|x|>C_1} f^{\frac{1}{2}}(x)\, dx.$$

The function g_4 tends to zero by the assumptions that $f^{(p)}$ is integrable and $E|X_1|^{1+\epsilon}$ is finite for some $\epsilon > 0$.

For the proof of (2.6) note that

$$\begin{aligned} 3E|f_n(\cdot|h) - f| \geq & E\{|f_n(\cdot|h) - Ef_n(\cdot|h)| - |Ef_n(\cdot|h) - f|\} + 2E|f_n(\cdot|h) - f| \\ = & E|f_n(\cdot|h) - Ef_n(\cdot|h)| + |Ef_n(\cdot|h) - f|, \end{aligned}$$

so clearly

$$3E\{J_n(h)\} \geq G_1(h) + G_2(h) \tag{6.12}$$

where $G_1(h) = \int E|f_n(\cdot|h) - Ef_n(\cdot|h)|$ and $G_2(h) = \int |Ef_n(\cdot|h) - f|$. Using Fatou's lemma and the continuity of $f^{(p)}$, it follows from (6.4) that

$$\begin{aligned} & \liminf_{h \to 0} G_2(h)/h^p \\ & \geq \{(p-1)!\}^{-1} \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} K(z) \int_0^1 \liminf_{h \to 0} f^{(p)}(x - thz)(1-t)^{p-1} dt dz \right| dx \\ & = (p!)^{-1} \int |f^{(p)}| \\ & > 0, \end{aligned}$$

so there exist constants C_1, $C_2 > 0$ such that $G_2(h) \geq C_1 h^p$ whenever $0 < h \leq C_2$. Using the boundedness of K it is easily established that $G_2(h) \geq C_3$ for $h \geq C_2$. Therefore we have

$$G_2(h) \geq C_4(h^p \wedge 1). \tag{6.13}$$

Observe that for $C_0 > 1$,

$$\inf_{h>0} E\{J_n(h)\} = \min\left[\inf_{u\in[C_0^{-1},C_0]} E\{J_n(h_u)\}, \inf_{u>C_0} E\{J_n(h_u)\}, \inf_{u<C_0^{-1}} E\{J_n(h_u)\}\right].$$

Choose C_0 so large that $u^* \in [C_0^{-1}, C_0]$ and apply (2.5) to obtain

$$\inf_{u\in[C_0^{-1},C_0]} E\{J_n(h_u)\} \sim n^{-p/(2p+1)}\lambda(u^*). \tag{6.14}$$

We are finished if we show that for C_0 sufficiently large,

$$\inf_{h>0} E\{J_n(h)\} = \inf_{u\in[C_0^{-1},C_0]} E\{J_n(h_u)\}. \tag{6.15}$$

It follows from (6.13) that

$$\begin{aligned}\inf_{u>C_0} E\{J_n(h_u)\} \geq & C_4 \inf_{u>C_0}\{u^{2p}n^{-p/(2p+1)} \wedge 1\} \\ = & C_4\{C_0^{2p}n^{-p/(2p+1)} \wedge 1\}\end{aligned}$$

which, in view of (6.14), is larger than $\inf_{u\in[C_0^{-1},C_0]} E\{J_n(h_u)\}$ for sufficiently large C_0. Also by Lemma 6.1, stated below, we obtain for large n,

$$\begin{aligned}\inf_{u<C_0^{-1}} E\{J_n(h_u)\} \geq & C \inf_{u<C_0^{-1}}\{u^{-1}n^{-p/(2p+1)} \wedge 1\} \\ = & C\{C_0 n^{-p/(2p+1)} \wedge 1\}\end{aligned}$$

which can also be made larger than $\inf_{u\in[C_0^{-1},C_0]} E\{J_n(h_u)\}$ by taking C_0 sufficiently large. This completes the proof of (6.15). ∎

Lemma 6.1. *If K is bounded, has compact support and integrates to unity, and if f is bounded, then there exists a constant $C > 0$ not depending on n or h, such that for $n \geq 1$ and $0 < h \leq 1$,*

$$\int E|f_n(\cdot|h) - Ef_n(\cdot|h)| \geq C\{(nh)^{-\frac{1}{2}} \wedge 1\}.$$

Proof. Applying Lemma 5.8 of Devroye and Györfi (1985, p.90) to the random variables

$$Z_i = h^{-1}K\{(x - X_i)/h\} - E[h^{-1}K\{(x - X_i)/h\}], \qquad i \geq 1$$

we obtain the bound

$$\left|E|f_n(x|h) - Ef_n(x|h)| - (2/\pi)^{\frac{1}{2}}\{\mathrm{Var} f_n(x|h)\}^{\frac{1}{2}}\right| \leq C_1(nh)^{-1},$$

where C_1 does not depend on x, n or h. For $h \geq n^{-1}$ it may be shown using standard arguments based on the asymptotic formula for $\mathrm{Var}\{f_n(\cdot|h)\}$, that for some bounded interval I,

$$\int_I [\mathrm{Var}\{f_n(\cdot|h)\}]^{\frac{1}{2}} \geq C_2(nh)^{-\frac{1}{2}}.$$

Therefore, for $h \geq C_3 n^{-1}$ and C_3 sufficiently large,

$$\int_I E|f_n(\cdot|h) - Ef_n(\cdot|h)| \geq \tfrac{1}{2} C_2(nh)^{-\frac{1}{2}},$$

so the result is true for $h \geq C_3 n^{-1}$. Now let $h \leq C_3 n^{-1}$ and suppose that the support of K is contained in the interval $[-s, s]$. Then

$$P\{f_n(x|h) = 0\} \geq p(x;n)$$

where

$$\begin{aligned} p(x;n) &= P\{|X_j - x| > sh\,,\ 1 \leq j \leq n\} \\ &= \{1 - P(|X_1 - x| \leq sh)\}^n, \end{aligned}$$

Now

$$P(|X_1 - x| \leq sh) = \int_{|y-x|<sh} f(y)\,dy \leq 2Bsh$$

where $B \equiv \sup f$. Noting that $1 - t \geq e^{-t-t^2}$ for $0 \leq t \leq \frac{1}{2}$ we get for large n,

$$\begin{aligned} p(x;n) &\geq \exp\{-2Bsnh - (2Bsnh)^2\} \\ &\geq C_4 \exp(-2Bsnh) \\ &\geq C_4 \exp(-2C_3 Bs) \\ &= C_5 > 0 \end{aligned}$$

where C_5 depends on none of x, n, h. Therefore

$$\begin{aligned} \int E|f_n(\cdot|h) - Ef_n(\cdot|h)| &\geq \int |Ef_n(x|h)| P\{f_n(x|h) = 0\}\,dx \\ &\geq \int |Ef_n(x|h)| p(x;n)\,dx \\ &\geq C_5 \left| \int Ef_n(x|h)\,dx \right| \\ &= C_5, \end{aligned}$$

which completes the proof. ∎

Proof of Theorem 4.1.

Our proof relies upon the following theorem of Devroye (1988). We state it here as a lemma.

Lemma 6.2. *Suppose that* $\int |K| < \infty$. *For all* $\epsilon > 0$,

$$\sup_{h>0, f\in\mathcal{D}} P(|J_n(h) - E\{J_n(h)\}| > \epsilon) \le 2e^{-n\epsilon^2/\{32(\int |K|)^2\}}$$

where $\mathcal{D}$ *is the class of all densities.*

Define h_{opt} and $\hat{h}_{\text{opt}}$ to be the window-sizes which minimise $E\{J_n(h)\}$ and $J_n(h)$ respectively. It may be readily established that for $a > 0$ sufficiently large we have $h_{\text{opt}} \in \mathcal{I}_a \equiv [n^{-a}, n^a]$ and

$$\lim_{n'\to\infty} P(\hat{h}_{\text{opt}}, h_n^* \in \mathcal{I}_a, \text{ for all } n \ge n') = 1.$$

For $c > 0$, let $\mathcal{H} = \mathcal{H}(a, c) = \{h_1, h_2, \ldots\}$ be the strictly increasing sequence given by

$$\begin{cases} h_1 = n^{-a}, \\ h_{i+1} = h_i + n^{-c}, \qquad i \ge 1, \end{cases}$$

and let m be the integer for which $h_{m-1} \le n^a < h_m$. For each $h \in \mathcal{I}_a$ define $h_{\mathcal{H}}$ to be the h_i which is nearest to h. Thus we have $|h - h_{\mathcal{H}}| = \inf_{i\ge 1} |h - h_i|$. Our first aim is to prove that we may choose $c = c(a)$ so large that for some $C > 0$ independent of n and the sample,

$$\sup_{h\in\mathcal{I}_a} |J_n(h) - J_n(h_{\mathcal{H}})| \le Cn^{-1}. \tag{6.16}$$

Suppose that the support of K is confined to $[-s, s]$ for some $s > 0$. Then it can be shown that

$$|J_n(h) - J_n(h_{\mathcal{H}})| \le 2s(\sup|K|)h_{\mathcal{H}}|h^{-1} - h_{\mathcal{H}}^{-1}| + \int |K(x) - K(hx/h_{\mathcal{H}})|\,dx. \tag{6.17}$$

For $h \in \mathcal{I}_a$, the first term is bounded above by $2s(\sup|K|)n^{a-c}$. Using the Hölder continuity of K we obtain for constants $\beta, \gamma > 0$,

$$\begin{aligned} \int |K(x) - K(xh/h_{\mathcal{H}})|\,dx \le& \beta \int |x - xh/h_{\mathcal{H}}|^\gamma I(|x| \le s \text{ or } |hx/h_{\mathcal{H}}| \le s)\,dx \\ \le& \beta n^{(a-c)\gamma} \int |x|^\gamma I(|x| \le s \text{ or } |x| \le sh_{\mathcal{H}}/h)\,dx. \end{aligned}$$

Next choose c so large that for positive constants C_1 and C_2 we have $n^{a-c} \le C_1 n^{-1}$, $n^{(a-c)\gamma} \le C_2 n^{-1}$ and $h_{\mathcal{H}}/h \le 2$. Then an upper bound to the right-hand side of (6.17) is

$$2s(\sup|K|)C_1 n^{-1} + \beta C_2 \int_{|x|\le 2s} |x|^\gamma\,dx n^{-1},$$

from which (6.16) follows immediately.

Letting $\Delta(h)$ denote the difference $J_n(h) - E\{J_n(h)\}$, (6.16) implies that

$$\sup_{h \in \mathcal{I}_a} |\Delta(h) - \Delta(h_{\mathcal{H}})| \leq 2Cn^{-1}. \tag{6.18}$$

Set $\nu, \eta > 0$ and let $\epsilon = \nu n^{-\frac{1}{2}+\eta}$ in Lemma 6.2 so that

$$\begin{aligned} P\{\sup_{1 \leq j \leq m} |\Delta(h_j)| > \nu n^{-\frac{1}{2}+\eta}\} &\leq \sum_{j=1}^{m} P\{|\Delta(h_j)| > \nu n^{-\frac{1}{2}+\eta}\} \\ &\leq 2m \exp\left[-\nu^2 n^{2\eta} \Big/ \left\{32(\int |K|)^2\right\}\right]. \end{aligned}$$

Therefore, since $m = O(n^{a+c})$ as $n \to \infty$,

$$\sum_{n=1}^{\infty} P\{n^{\frac{1}{2}-\eta} \sup_{1 \leq j \leq m} |\Delta(h_j)| > \nu\} < \infty$$

for arbitrary $\nu > 0$. It follows from this and the Borel-Cantelli lemma that

$$\lim_{n \to \infty} n^{\frac{1}{2}-\eta} \sup_{1 \leq j \leq m} |\Delta(h_j)| = 0$$

almost surely. This and (6.18) together imply that

$$\lim_{n \to \infty} n^{\frac{1}{2}-\eta} \sup_{h \in \mathcal{I}_a} |\Delta(h)| = 0$$

almost surely. Suppose it is established that for some $\eta > 0$, $C_0 > 0$ and all sufficiently large n,

$$\inf_{h \in \mathcal{I}_a} E\{J_n(h)\} \geq C_0 n^{-\frac{1}{2}+\eta}. \tag{6.19}$$

Then simple arguments lead to

$$\lim_{n \to \infty} \{\inf_{h \in \mathcal{I}_a} J_n(h)\} / [\inf_{h \in \mathcal{I}_a} E\{J_n(h)\}] = 1$$

almost surely. Taking a sufficiently large we obtain (4.9). The same arguments can be employed to prove (4.10).

It remains to prove (6.19). Recall from (6.12) that

$$3E\{J_n(h)\} \geq G_1(h) + G_2(h) \tag{6.20}$$

where $G_1(h) = \int E|f_n(\cdot|h) - Ef_n(\cdot|h)|$ and $G_2(h) = \int |Ef_n(\cdot|h) - f|$. Lemma 6.1 asserts that for $0 < h \leq 1$, $G_1(h) \geq C_3\{(nh)^{-\frac{1}{2}} \wedge 1\}$. A lower bound for $G_2(h)$ is obtained by first observing that for all $x \in \mathbf{R}$,

$$|Ef_n(x|h) - f(x)| = \left| \int K(z)\{f(x - hz) - f(x)\}\, dz \right| \leq 2(\sup f) \int |K|.$$

Hence

$$G_2(h) \geq \left\{ 2(\sup f) \int |K| \right\}^{-1} \int \{Ef_n(\cdot|h) - f\}^2 \geq C_4(h^b \wedge 1)$$

for some $b > 0$. The second inequality follows from Lemma 1 of Stone (1984). Result (6.19) can be derived easily from (6.20) by considering separately (i) $n^{-a} < h \leq n^{-1}$, (ii) $n^{-1} < h \leq 1$ and (iii) $1 < h \leq n^a$.

Proof of Theorem 4.2.

(i) Proof of (4.11)

To prove (4.11) it is sufficient to show that

$$\lim_{n\to\infty} \int |Ef_n^{(p)}(\cdot|h_1) - f^{(p)}| = 0 \tag{6.21}$$

and

$$\lim_{n\to\infty} \int |f_n^{(p)}(\cdot|h_1) - Ef_n^{(p)}(\cdot|h_1)| = 0 \tag{6.22}$$

almost surely.

To prove (6.21), consider

$$\begin{aligned} Ef_n^{(p)}(x|h_1) &= h_1^{-p} \int_{-\infty}^{\infty} K_1^{(p)}(z) f(x - h_1 z)\, dz \\ &= \int_{-\infty}^{\infty} K_1(z) f^{(p)}(x - h_1 z)\, dz\,, \end{aligned}$$

the last equality coming from p steps of integration by parts. Therefore

$$Ef_n^{(p)}(x|h_1) - f^{(p)}(x) = \int_{-\infty}^{\infty} K_1(z)\{f^{(p)}(x - h_1 z) - f^{(p)}(x)\}\, dz.$$

Suppose that the support of K_1 is contained in the interval $[-s, s]$. Then for each $C > 0$,

$$\sup_{|x|\leq C} |Ef_n^{(p)}(x|h_1) - f^{(p)}(x)| \leq \int_{|z|\leq s} |K_1(z)| \sup_{|x|\leq C} |f^{(p)}(x - h_1 z) - f^{(p)}(x)|\, dz.$$

Now $f^{(p)}$ is continuous and therefore uniformly continuous on $[-C, C]$, so for each $z \in [-s, s]$,

$$\lim_{n\to\infty} \sup_{|x|\le C} |f^{(p)}(x - h_1 z) - f^{(p)}(x)| = 0,$$

since $h_1 \to 0$ as $n \to \infty$. Since $f^{(p)}$ and K_1 are bounded it follows by dominated convergence that

$$\lim_{n\to\infty} \sup_{|x|\le C} |Ef_n^{(p)}(x|h_1) - f^{(p)}(x)| = 0. \tag{6.23}$$

Take h_1 so small that $h_1 s \le \frac{1}{2}C$ and observe that for $|x| > C$ and $|z| \le s$ we must have $|x - h_1 z| > \frac{1}{2}C$. Then

$$\int_{|x|>C} |Ef_n^{(p)}(\cdot|h_1) - f^{(p)}| \le \int_{|x|>C} \int_{|z|\le s} |K_1(z)||f^{(p)}(x - h_1 z)|\, dz\, dx + \int_{|z|\le s} |K_1(z)|\, dz \int_{|x|>C} |f^{(p)}(x)|\, dx,$$

and since both terms of this last expression are dominated by

$$\left(\int |K_1|\right) \int_{|x|\ge \frac{1}{2}C} |f^{(p)}(x)|\, dx,$$

we arrive at the bound

$$\int_{|x|>C} |Ef_n^{(p)}(\cdot|h_1) - f^{(p)}| \le 2\left(\int |K_1|\right) \int_{|x|>\frac{1}{2}C} |f^{(p)}(x)|\, dx. \tag{6.24}$$

Combining (6.23) and (6.24) we get for arbitrary $C > 0$,

$$\limsup_{n\to\infty} \int |Ef_n^{(p)}(\cdot|h_1) - f^{(p)}| \le C_1 \int_{|x|>\frac{1}{2}C} |f^{(p)}(x)|\, dx,$$

where $C_1 \equiv 2(\int |K_1|)$, and so (6.21) follows from this and the integrability of $f^{(p)}$.

The proof of (6.22) uses a version of Bernstein's inequality (see Hoeffding (1963)) which we state as a lemma.

Lemma 6.3. *If $Y_1, \ldots, Y_n$ are independent and identically distributed with zero mean and variance σ^2, and if each $|Y_i| \le c$, then*

$$P\left(\left|\sum_{i=1}^n Y_i\right| > t\right) \le 2\exp\{-\tfrac{1}{2}t^2(n\sigma^2 + ct)^{-1}\}$$

for all $t > 0$.

Note that for any $\xi > 0$ the integral on the left-hand side of (6.22) is bounded above by

$$\int_{|x|\le\xi} |f_n^{(p)}(\cdot|h_1) - Ef_n^{(p)}(\cdot|h_1)| + \int_{|x|>\xi} |f_n^{(p)}(\cdot|h_1)| + \int_{|x|>\xi} |f^{(p)}(x)|\,dx + \int |Ef_n^{(p)}(\cdot|h_1) - f^{(p)}|,$$

so it suffices to prove that for some sequence $\xi = \xi(n)$ tending to infinity,

$$\lim_{n\to\infty} \int_{|x|>\xi} |f_n^{(p)}(\cdot|h_1)| = 0 \tag{6.25}$$

almost surely and

$$\lim_{n\to\infty} \int_{|x|\le\xi} |f_n^{(p)}(\cdot|h_1) - Ef_n^{(p)}(\cdot|h_1)| = 0 \tag{6.26}$$

almost surely. Take h_1 so small and ξ so large that $h_1 s \le \frac{1}{2}\xi$. Then

$$\begin{aligned}
\int_{|x|>\xi} |f_n^{(p)}(\cdot|h_1)| &\le (nh_1^{p+1})^{-1} \sum_{i=1}^n \int_{|x|>\xi} |K_1^{(p)}\{(x - X_i)/h_1\}|\,dx \\
&= (nh_1^p)^{-1} \sum_{i=1}^n \int_{|h_1 x + X_i|>\xi} |K_1^{(p)}(x)|\,dx \\
&\le \sup |K_1^{(p)}| (nh_1^p)^{-1} \sum_{i=1}^n \int_{|h_1 x + X_i|>\xi; |x|\le s} dx \\
&\le 2s \sup |K_1^{(p)}| (nh_1^p)^{-1} \sum_{i=1}^n I(|X_i| > \tfrac{1}{2}\xi),
\end{aligned} \tag{6.27}$$

since $|X_i| \le \frac{1}{2}\xi$ implies that the set $\{x : |h_1 x + X_i| > \xi;\ |x| \le s\}$ is empty. Let $\epsilon > 0$ be such that $E|X_1|^{1+\epsilon} < \infty$ and put $\alpha = 1 + \epsilon$. We take $\xi = h_1^{-p/\beta}$ where $2\alpha/(\alpha+1) < \beta < \alpha$. By Markov's inequality,

$$\rho \equiv P(|X_1| > \tfrac{1}{2}\xi) \le C_2 \xi^{-\alpha}$$

where $C_2 = 2^{-\alpha} E|X_1|^\alpha$. Now,

$$h_1^{-p}\rho \le C_2 h_1^{p(\alpha-\beta)/\beta} \to 0. \tag{6.28}$$

For each $\delta > 0$ we have by Lemma 6.3 (with $Y_i = I(|X_i| > \frac{1}{2}\xi) - \rho$, $c = 2$, $t = \delta n h_1^p$, $\sigma^2 = \rho(1-\rho)$)

$$\begin{aligned} q &\equiv P\left[\left|\sum_{i=1}^{n}\{I(|X_i| > \tfrac{1}{2}\xi) - \rho\}\right| > \delta n h_1^p\right] \\ &\le 2\exp[-\tfrac{1}{2}(\delta n h_1^p)^2\{n\rho(1-\rho) + 2\delta n h_1^p\}^{-1}]. \end{aligned}$$

Note that for $h_1 \le 1$, $\rho(1-\rho) \le \rho \le C_2 h_1^{p\alpha/\beta} \le C_2 h_1^p$ and so $q \le 2\exp\{-C_3(\delta) n h_1^p\}$. Clearly $n h_1^p / \log n \to \infty$, implying that $q = O(n^{-k})$ for all $k > 0$. Therefore

$$\sum_{n=1}^{\infty} P\left[(n h_1^p)^{-1}\left|\sum_{i=1}^{n}\{I(|X_i| > \tfrac{1}{2}\xi) - \rho\}\right| > \delta\right] < \infty$$

for all $\delta > 0$, so by the Borel-Cantelli lemma,

$$\lim_{n\to\infty}(n h_1^p)^{-1}\left|\sum_{i=1}^{n}\{I(|X_i| > \tfrac{1}{2}\xi) - \rho\}\right| = 0$$

almost surely. Combining this with (6.28) we obtain

$$\lim_{n\to\infty}(n h_1^p)^{-1}\sum_{i=1}^{n} I(|X_i| > \tfrac{1}{2}\xi) = 0$$

almost surely. The result at (6.25) is a consequence of this and (6.27).

For the proof of (6.22), define

$$\tau_1^2(x) \equiv \int_{-\infty}^{\infty} K_1^{(p)}(z)^2 f(x - h_1 z)\, dz$$

and

$$\tau^2(x) \equiv \max\{\tau_1^2(x), (1 + |x|^{2\alpha})^{-1}\}.$$

Also let $\mathcal{S}_c \equiv \{x \in (-c, c) : (1 + |x|^\alpha)\tau^2(x) > 2\}$. We shall prove separately that

$$\lim_{n\to\infty}\int_{|x|\le\xi; x\in\mathcal{S}_\infty} |f_n^{(p)}(\cdot|h_1) - E f_n^{(p)}(\cdot|h_1)| = 0 \tag{6.29}$$

almost surely, and

$$\lim_{n\to\infty}\int_{|x|\le\xi; x\notin\mathcal{S}_\infty} |f_n^{(p)}(\cdot|h_1) - E f_n^{(p)}(\cdot|h_1)| = 0 \tag{6.30}$$

almost surely. To derive (6.29) we first prove that the Lebesgue measure of $\mathcal{S}_\infty$, $\mathcal{L}(\mathcal{S}_\infty)$, is bounded. Observe that $(1+|x|^\alpha)\tau^2(x) > 2$ if and only if $(1+|x|^\alpha)\tau_1^2(x) > 2$ and let Y_c be a uniform random variable on $(-c, c)$. Then by Markov's inequality,

$$\begin{aligned}\tfrac{1}{2}E\{(1+|Y_c|^\alpha)\tau_1^2(Y_c)\} &\geq P\{(1+|Y_c|^\alpha)\tau_1^2(Y_c) > 2\} \\ &= P\{(1+|Y_c|^\alpha)\tau^2(Y_c) > 2\} \\ &= \tfrac{1}{2}c^{-1}\mathcal{L}(\mathcal{S}_c).\end{aligned}$$

Therefore

$$\mathcal{L}(\mathcal{S}_c) \leq \tfrac{1}{2}\int_{-c}^{c}(1+|x|^\alpha)\tau_1^2(x)\,dx$$

for all $c > 0$ and hence

$$\begin{aligned}\mathcal{L}(\mathcal{S}_\infty) &\leq \int_{-\infty}^{\infty}(1+|x|^\alpha)\tau_1^2(x)\,dx \\ &= \int_{-\infty}^{\infty}K_1^{(p)}(z)^2\int_{-\infty}^{\infty}(1+|y+h_1z|^\alpha)f(y)\,dy\,dz \\ &\leq 2^\alpha\int_{-\infty}^{\infty}K_1^{(p)}(z)^2\int_{-\infty}^{\infty}(1+|y|^\alpha+h_1^\alpha|z|^\alpha)f(y)\,dy\,dz \\ &= 2^\alpha\left[\int\{K_1^{(p)}\}^2\{E(|X_1|^\alpha)+1\}+h_1^\alpha\int|z|^\alpha K_1^{(p)}\right] \\ &< \infty\end{aligned}$$

uniformly in $h_1 \leq 1$, since $E|X_1|^\alpha < \infty$ and $K_1^{(p)}$ is compactly supported. For each $\delta > 0$ the integral in the left-hand side of (6.29) is no more than

$$\begin{aligned}&\int_{\mathcal{S}_\infty}|f_n^{(p)}(\cdot|h_1)-Ef_n^{(p)}(\cdot|h_1)|I(|f_n^{(p)}(\cdot|h_1)-Ef_n^{(p)}(\cdot|h_1)|\leq\delta) \\ &\qquad+\int_{\mathcal{S}_\infty}|f_n^{(p)}(\cdot|h_1)-Ef_n^{(p)}(\cdot|h_1)|I(|f_n^{(p)}(\cdot|h_1)-Ef_n^{(p)}(\cdot|h_1)|>\delta) \\ &\leq \delta\mathcal{L}(\mathcal{S}_\infty)+2\int_{\mathcal{S}_\infty}|f_n^{(p)}(\cdot|h_1)|I(|f_n^{(p)}(\cdot|h_1)-Ef_n^{(p)}(\cdot|h_1)|>\delta) \\ &\leq \delta\mathcal{L}(\mathcal{S}_\infty)+h_1^{-(p+1)}(2\sup|K_1^{(p)}|)M_1 \qquad (6.31)\end{aligned}$$

where

$$M_1 \equiv \int_{\mathcal{S}_\infty} I\left(\left|\sum_{i=1}^{n}[K_1^{(p)}\{(x-X_i)/h_1\}-EK_1^{(p)}\{(x-X_i)/h_1\}]\right|>\delta nh_1^{p+1}\right)dx.$$

Let $Y_i = K_1^{(p)}\{(x-X_i)/h_1\}-EK_1^{(p)}\{(x-X_i)/h_1\}$, $c = 2\sup|K_1^{(p)}|$ and $t = \delta nh_1^{p+1}$, and note that

$$\sigma^2 \equiv \mathrm{Var}(Y_1) \leq \tau_1^2(x)h_1 \leq C_1h_1$$

for some constant $C_1 > 0$ independent of n and h_1. This leads to

$$\begin{aligned}\tfrac{1}{2}t^2(\sigma^2 n + ct)^{-1} &\geq C_2(\delta)(nh_1^{p+1})^2(C_1 h_1 n + nh_1^{p+1})^{-1} \\ &\geq C_3(\delta)nh_1^{2p+1}.\end{aligned}$$

Therefore by Lemma 6.3,

$$E(M_1) \leq 2\int_{\mathcal{S}_\infty} e^{-C_3(\delta)nh_1^{2p+1}}\,dx = 2\mathcal{L}(\mathcal{S}_\infty)e^{-C_3(\delta)nh_1^{2p+1}} = O(n^{-k})$$

for all $k > 0$, since $nh_1^{2p+1}/\log n \to \infty$. Thus we may conclude by Markov's inequality and the Borel-Cantelli lemma that $\lim_{n\to\infty} h_1^{-(p+1)}M_1 = 0$ almost surely. This implies that

$$\limsup_{n\to\infty} \int_{|x|\leq\xi; x\in\mathcal{S}_\infty} |f_n^{(p)}(\cdot|h_1) - Ef_n^{(p)}(\cdot|h_1)| \leq \mathcal{L}(\mathcal{S}_\infty)\delta$$

almost surely for all $\delta > 0$, from which (6.29) follows immediately.

The integral on the left-hand side of (6.30) may be written as

$$\begin{aligned}&\int_{|x|\leq\xi; x\notin\mathcal{S}_\infty} \left|(nh_1^{p+1})^{-1}\sum_{i=1}^n Y_i\right| I\left\{\left|(nh_1^{p+1})^{-1}\sum_{i=1}^n Y_i\right| \leq \delta\tau(x)^{\beta/\alpha}\right\} dx \\ &\quad + \int_{|x|\leq\xi; x\notin\mathcal{S}_\infty} \left|(nh_1^{p+1})^{-1}\sum_{i=1}^n Y_i\right| I\left\{\left|(nh_1^{p+1})^{-1}\sum_{i=1}^n Y_i\right| > \delta\tau(x)^{\beta/\alpha}\right\} dx \\ &\quad \leq \delta\int \tau^{\beta/\alpha} + h_1^{-(p+1)}(2\sup|K_1^{(p)}|)M_2\end{aligned}$$

where

$$M_2 \equiv \int_{|x|\leq\xi; x\notin\mathcal{S}_\infty} I\left\{\left|\sum_{i=1}^n Y_i\right| > \delta nh_1^{p+1}\tau(x)^{\beta/\alpha}\right\} dx.$$

From Hölder's inequality we have

$$\begin{aligned}\int_{-\infty}^{\infty} \tau(x)^{\beta/\alpha}\,dx \leq &\left\{\int_{-\infty}^{\infty} \tau(x)^2(1+|x|)^\alpha\,dx\right\}^{\beta/(2\alpha)} \\ &\times\left\{\int_{-\infty}^{\infty}(1+|x|)^{-\alpha\beta/(2\alpha-\beta)}\,dx\right\}^{(2\alpha-\beta)/(2\alpha)}. \quad (6.32)\end{aligned}$$

The second factor on the right-hand side of (6.32) is finite by choice of β. To deal with the first factor, observe that

$$\begin{aligned}\int_{-\infty}^{\infty} \tau(x)^2(1+|x|)^\alpha\,dx \leq &\int_{-\infty}^{\infty} \tau_1^2(x)(1+|x|)^\alpha\,dx \\ &+ \int_{-\infty}^{\infty}(1+|x|)^\alpha(1+|x|^{2\alpha})^{-1}\,dx.\end{aligned}$$

Clearly the second term on the right-hand side of this expression is finite. The first term equals

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K_1^{(p)}(z)^2 f(x - h_1 z)(1 + |x|)^{\alpha}\, dx\, dz,$$

which is finite uniformly in $h_1 \leq 1$ using the fact that $E|X_1|^{\alpha} < \infty$ as before. This proves that $\int \tau^{\beta/\alpha} < \infty$. Therefore it suffices to prove that $E(M_2) = O(n^{-k})$ for all $\delta > 0$ and $k > 0$. We apply Lemma 6.3, again with the same Y_i and c but this time with $t = \delta n h_1^{p+1} \tau(x)^{\beta/\alpha}$. From before we have $\sigma^2 \leq \tau_1(x)^2 h_1 \leq \tau(x)^2 h_1$, so that

$$\begin{aligned} T(x) &\equiv \tfrac{1}{2} t^2 (n\sigma^2 + ct)^{-1} \\ &\geq \tfrac{1}{2}\delta \tau(x)^{2\beta/\alpha} (n h_1^{p+1})^2 \{n h_1 \tau(x)^2 + c\delta n h_1^{p+1} \tau(x)^{\beta/\alpha}\}^{-1} \\ &\geq C_4(\delta) \tau(x)^{2\beta/\alpha} (n h_1^{p+1})^2 \{n h_1 \tau(x)^2 + n h_1^{p+1} \tau(x)^{\beta/\alpha}\}^{-1}. \end{aligned}$$

Suppose firstly that $h_1^p \leq \tau(x)^{2-\beta/\alpha}$. Then

$$\begin{aligned} T &\geq \tfrac{1}{2} C_4(\delta) \tau(x)^{-2(\alpha-\beta)/\alpha} n h_1^{2p+1} \\ &\geq 2^{\beta/\alpha - 2} C_4(\delta) n h_1^{2p+1} (1 + |x|^{\alpha})^{(\alpha-\beta)/\alpha} \\ &\geq 2^{\beta/\alpha - 2} C_4(\delta) n h_1^{2p+1}, \end{aligned}$$

since $x \notin \mathcal{S}_{\infty}$ entails $\tau(x)^{-2} \geq \frac{1}{2}(1 + |x|^{\alpha})$. Next suppose that $h_1^p > \tau(x)^{2-\beta/\alpha}$. Then

$$\begin{aligned} T &\geq \tfrac{1}{2} C_4(\delta) n h_1^{p+1} \tau(x)^{\beta/\alpha} \\ &\geq \tfrac{1}{2} C_4(\delta) n h_1^{p+1} (1 + |x|^{2\alpha})^{-\beta/(2\alpha)}, \end{aligned}$$

since $\tau(x)^2 \geq (1 + |x|^{2\alpha})^{-1}$. Since $|x| \leq \xi = h_1^{-p/\beta}$ we have

$$(1 + |x|^{2\alpha})^{-\beta/(2\alpha)} \geq C_4 h_1^p,$$

giving $T \geq C_5(\delta) n h_1^{2p+1}$. Combining both these bounds gives $T(x) \geq C_6(\delta) n h_1^{2p+1}$ for all $x \notin \mathcal{S}_{\infty}$ and $|x| \leq \xi$. Therefore by Lemma 6.3,

$$E(M_2) \leq 2 \int_{|x| \leq \xi; x \notin \mathcal{S}_{\infty}} e^{-T(x)}\, dx \leq 4\xi \exp\{-C_6(\delta) n h_1^{2p+1}\} = O(n^{-k})$$

for all $k > 0$ and $\delta > 0$, as required.

(ii) Proof of (4.12).

Let $\alpha = 1 + \epsilon$, so that $E|X_1|^\alpha < \infty$. Then

$$\int |f_n^{\frac{1}{2}}(\cdot|h_2) - f^{\frac{1}{2}}| \leq \left\{ \int_{-\infty}^{\infty} |f_n(x|h_2) - f(x)|(1+|x|^\alpha)\, dx \right\}^{\frac{1}{2}} \times \left\{ \int_{-\infty}^{\infty} (1+|x|^\alpha)^{-1}\, dx \right\}^{\frac{1}{2}}$$

by the Cauchy-Schwarz inequality. Since $\int_{-\infty}^{\infty}(1+|x|^\alpha)^{-1}\,dx$ is finite we need to show that $\int_{-\infty}^{\infty} |f_n(x|h_2) - f(x)|(1+|x|^\alpha)\,dx \to 0$ almost surely as $n \to \infty$. For $C > 0$ this integral is dominated by

$$(1+C^\alpha)\int |f_n(\cdot|h_2) - f| + \int_{|x|>C} f(x)(1+|x|^\alpha)\,dx + \int_{|x|>C} f_n(x|h_2)(1+|x|^\alpha)\,dx.$$

Under the conditions imposed on f, K_2 and h_2 in the theorem, $\lim_{n\to\infty} \int |f_n(\cdot|h_2) - f| = 0$ almost surely (see e.g. Devroye and Györfi (1985) Theorem 3.1, p.12). Also

$$\lim_{C\to\infty} \int_{|x|>C} f(x)(1+|x|^\alpha)\,dx = 0,$$

since $E|X_1|^\alpha < \infty$, so it remains to show that

$$\lim_{C\to\infty} \limsup_{n\to\infty} \int_{|x|>C} f_n(x|h_2)(1+|x|^\alpha)\,dx = 0 \tag{6.33}$$

almost surely. Suppose that the support of K_2 is contained in $[-s, s]$ and let h_2 be so small that $h_2 s \leq \frac{1}{2}C$. Then the integral in (6.33) can be written as

$$n^{-1}\sum_{i=1}^{n} \int_{|X_i+h_2y|>C;|y|\leq s} K_2(y)(1+|X_i+h_2y|^\alpha)\,dy$$
$$\leq 2^{\alpha+1} s \sup|K_2| n^{-1} \sum_{i=1}^{n} (1+|X_i|^\alpha) I(|X_i| > \tfrac{1}{2}C) + 2^\alpha h_2^\alpha \int |y|^\alpha K_2(y)\,dy.$$

By the strong law of large numbers,

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}(1+|X_i|^\alpha) I(|X_i| > \tfrac{1}{2}C) = \int_{|x|>\frac{1}{2}C} f(x)(1+|x|^\alpha)\,dx$$

almost surely as $n \to \infty$. Therefore, noting that K_2 has compact support,

$$\limsup_{n\to\infty} \int_{|x|>C} f_n(x|h_2)(1+|x|^\alpha) \leq C_1 \int_{|x|>\frac{1}{2}C} f(x)(1+|x|^\alpha)\,dx$$

almost surely for some constant $C_1 > 0$. Result (6.33) follows immediately from this and the existence of $E|X_1|^\alpha$. ∎

Chapter Three

MINIMISATION OF L_1 DISTANCE IN HISTOGRAM AND FREQUENCY POLYGON DENSITY ESTIMATION

3.1 Introduction

The histogram is the exordial nonparametric density estimator. However, it still enjoys a great deal of usage in data analysis and presentation. A closely related density estimator is the frequency polygon which is constructed by straight-line interpolation of the histogram. In the univariate case both estimators are based on partitioning the real line into equal-sized intervals, or "bins", and counting the number of sample points in each bin. The smoothing parameter associated with each of these estimators is the length of the partition intervals, often referred to as the bin-width. This chapter is concerned with using techniques employed in the previous chapter for kernel density estimation to derive similar optimality results for the histogram and frequency polygon. Our theory leads to an L_1 version of the rules of Scott (1979, 1985) for bin-width choice. We also extend the work of Devroye and Györfi (1985) to derive closed form bounds for $\liminf_{n\to\infty} \inf_{h>0} n^{2/5} E\{J_n(h)\}$ and $\limsup_{n\to\infty} \inf_{h>0} n^{2/5} E\{J_n(h)\}$ for the frequency polygon, where $J_n(h)$ is the L_1 distance between the frequency polygon $f_n(\cdot|h)$, with bin-width h, and true density.

Section 2 treats the histogram, followed by an analysis of the frequency polygon in Section 3. Numerical examples are presented in Section 4. Section 5 contains proofs.

3.2 L_1 Theory of the Histogram

We shall consider histogram density estimators defined on the real number line with respect to the partition $\{B_{nr},\ r \in \mathbf{Z}\}$ where $B_{nr} = [rh, (r+1)h)$. The partition elements B_{nr} are the bins of the histogram and h is the bin-width. Since our results are asymptotic in nature we will assume that $h = h(n)$ is a sequence

of positive numbers satisfying

$$\lim_{n\to\infty} h = 0 \qquad \text{and} \qquad \lim_{n\to\infty} nh = \infty. \tag{2.1}$$

The histogram estimator of $f(x)$ is given by

$$f_n(x|h) = h^{-1}\mu_n(B_{nr}), \qquad x \in B_{nr},$$

where μ_n is the empirical measure based on the sample $X_1, \ldots, X_n$ and is given by

$$\mu_n(S) = n^{-1}\text{card}\{i : 1 \le i \le n \text{ and } X_i \in S\}, \qquad S \subseteq \mathbf{R}.$$

Under certain smoothness assumptions on the density f, Scott (1979) and Freedman and Diaconis (1981) have shown that the L_2-optimal bin-width for the histogram is asymptotic to

$$h_2^* = \left\{6 \Big/ \int (f')^2\right\}^{1/3} n^{-1/3} \tag{2.2}$$

and that the optimal rate of convergence to zero of expected L_2 loss $M_n(h) = \int\{f_n(\cdot|h) - f\}^2$ is given by

$$E\{M_n(h_2^*)\} = (3/2)\left\{\int (f')^2 \Big/ 6\right\}^{1/3} n^{-2/3} + o(n^{-2/3}). \tag{2.3}$$

Scott (1979) proposed a data-based rule for choosing h from the formula at (2.1) by using the normal density family as a reference standard. The rule is based on the observation that if f is the normal density with variance σ^2 then

$$h_2^* = (24\pi^{\frac{1}{2}})^{1/3}\sigma n^{-1/3} = (3.49\cdots)\sigma n^{-1/3}.$$

This leads to the data-based choice

$$h_{n,2}^* = 3.49 s n^{-1/3} \tag{2.4}$$

where s is an estimate of the standard deviation.

Let $J_n(h) = \int |f_n(\cdot|h) - f|$ be the L_1 distance between $f_n(\cdot|h)$ and f. Our goal is to derive the L_1 analogues of the formulae given at (2.2), (2.3) and (2.4).

Devroye and Györfi (1985, pp.98,99) have shown that for all densities f having compact support and a bounded, continuous derivative,

$$E\{J_n(h)\} = \int (nh)^{-\frac{1}{2}} f^{\frac{1}{2}} \psi\left(\frac{(nh^3)^{\frac{1}{2}}|z_n|}{2f^{\frac{1}{2}}}\right) + o\{h + (nh)^{-\frac{1}{2}}\}$$

where ψ is the function introduced in Section 2.2 and z_n is given by

$$z_n(x) = 2[\{(r+\tfrac{1}{2})h - x\}/h]f'(x), \qquad x \in B_{nr}.$$

Also, they have established the bounds

$$\liminf_{n\to\infty} n^{1/3} \inf_{h>0} E\{J_n(h)\} \geq (0.880261\cdots)B_H(f) \tag{2.5}$$

and

$$\limsup_{n\to\infty} n^{1/3} \inf_{h>0} E\{J_n(h)\} \leq (1.290381\cdots)B_H(f) \tag{2.6}$$

where

$$B_H(f) = \{\tfrac{1}{2}(\textstyle\int f^{\frac{1}{2}})^2 \int |f'|\}^{1/3}.$$

Therefore the optimal rate of convergence of $E\{J_n(h)\}$ lies between

$$(0.880\cdots)B_H(f)n^{-1/3} \qquad \text{and} \qquad (1.290\cdots)B_H(f)n^{-1/3},$$

and this rate is achieved by choosing the bin-width to be asymptotic to $h_u = u^2 n^{-1/3}$ for some $u > 0$ (thus balancing the orders of magnitude of the bias and standard deviation of $f_n(\cdot|h)$ as for the kernel estimator treated in Section 2.2). The optimal choice of u and the corresponding rate of convergence of $E\{J_n(h)\}$ can be determined from

Theorem 2.1. *Suppose that f has a bounded and Lipschitz continuous first derivative and vanishes outside a compact set. Then for all $u > 0$,*

$$E\{J_n(h_u)\} = \zeta(u)n^{-1/3} + o(n^{-1/3})$$

where

$$\zeta(u) = u^{-1} \int_{-\infty}^{\infty} \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x)^{\frac{1}{2}} \psi\left(\frac{u^3 f'(x) y}{f(x)^{\frac{1}{2}}}\right) dy\, dx. \tag{2.7}$$

The L_1-optimal bin-width is therefore asymptotic to $h^* = (u^*)^2 n^{-1/3}$ where u^* is the value of u that minimises $\zeta(u)$. To locate u^* we observe that $\zeta'(u) = 2u^{-2}\Lambda(u^3)$ where

$$\Lambda(v) = \int_{-\infty}^{\infty} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left\{ 2vf'(x)y\Phi\left(\frac{vf'(x)y}{f(x)^{\frac{1}{2}}}\right) - f(x)^{\frac{1}{2}}\phi\left(\frac{vf'(x)y}{f(x)^{\frac{1}{2}}}\right)\right\} dy\, dx. \quad (2.8)$$

Note that

$$\zeta''(u) = u^{-3} \int_{-\infty}^{\infty} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left\{ \frac{9f'(x)^2y^2u^6}{f(x)^{\frac{1}{2}}}\psi''\left(\frac{u^3f'(x)y}{f(x)^{\frac{1}{2}}}\right) + 2\psi\left(\frac{u^3f'(x)y}{f(x)^{\frac{1}{2}}}\right)\right\} dy\, dx,$$

which is clearly positive for all $u > 0$ since $\psi,\ \psi'' \geq 0$; $\Lambda(0) = -(2\pi)^{-\frac{1}{2}} \int f^{\frac{1}{2}}$ and $\lim_{v\to\infty} \Lambda(v) = \infty$. Together these results imply that u^* exists, is unique and is given by $u^* = (v^*)^{1/3}$ where v^* is the solution of $\Lambda(v) = 0$. In view of Theorem 2.1 the best possible rate of convergence of $E\{J_n(h)\}$ to zero is $\zeta(u^*)n^{-1/3}$.

The assumption of f having compact support can be weakened to the existence of $E|X_1|^{1+\epsilon}$ for some $\epsilon > 0$ by arguing as in the proof of Theorem 2.2.1.

In Section 4, examples of the L_1-optimal bin-widths are presented. In particular it is seen that for normal data with variance σ^2 the L_1-optimal bin-width is asymptotic to

$$h^* = (3.37\cdots)\sigma n^{-1/3}.$$

Therefore the L_1 analogue of the bin-width selection rule of Scott (1979) is

$$h_n^* = 3.37sn^{-1/3},$$

which smooths about 96.5% as much as $h_{n,2}^*$. Consequently, the L_1 based rule and the L_2 based rule are virtually equivalent.

3.3 L_1 Theory of the Frequency Polygon

The frequency polygon is the density estimator obtained by straight-line interpolation of the histogram heights at the middle of each bin. This estimator is defined on the partition $\{G_{nr},\ r \in \mathbf{Z}\}$ where $G_{nr} = [(r - \frac{1}{2})h, (r + \frac{1}{2})h)$ and is given by

$$f_n(x|h) = h^{-1}(r + \tfrac{1}{2} - x/h)\mu_n(B_{n(r-1)}) - h^{-1}(r - \tfrac{1}{2} - x/h)\mu_n(B_{nr}), \quad x \in G_{nr}.$$

Here the B_{nr} have the definition ascribed to them in the previous section. Again it is assumed throughout that the bin-width h satisfies (2.1). The L_2 loss of the frequency polygon is investigated by Scott (1985) where it is established that for $M_n(h) = \int \{f_n(\cdot|h) - f\}^2$,

$$E\{M_n(h)\} = (5/12)\left\{49\int (f'')^2 \Big/ 15\right\}^{1/5} n^{-4/5} + o(n^{-4/5})$$

and the bin-width required to achieve this minimum is asymptotic to

$$h_2^* = 2\left[\frac{15}{49\int (f'')^2}\right]^{1/5} n^{-1/5}. \tag{3.1}$$

Also in Scott (1985) the frequency polygon bin-width selection rule

$$h_{n,2}^* = 2.15sn^{-1/5} \tag{3.2}$$

was proposed (s being an estimate of the standard deviation) and is based on the result

$$h_2^* = 2(40\pi^{\frac{1}{2}}/49)^{1/5}\sigma n^{-1/5} = (2.15\cdots)\sigma n^{-1/5}$$

for Gaussian data having variance σ^2.

The asymptotic theory of L_1 loss, $J_n(h) = \int |f_n(\cdot|h) - f|$, of the frequency polygon developed here is for densities belonging to the function class $\mathcal{F}$ where

$\mathcal{F} = \{f : f$ is a density with two bounded, Lipschitz continuous derivatives$\}$.

For each $f \in \mathcal{F}$ we put $B(f) = \{\frac{1}{2}(\int f^{\frac{1}{2}})^4 \int |f''|\}^{1/5}$ as was done in Section 2.5 in the context of kernel density estimation. We also define γ to be a universal constant given by

$$\gamma = \tfrac{1}{2} + \log_e(1 + 2^{\frac{1}{2}})/2^{3/2}.$$

Theorem 3.1. *For all $f \in \mathcal{F}$,*

$$\begin{aligned}\liminf_{n\to\infty} n^{2/5} \inf_{h>0} E\{J_n(h)\} &\geq A_1 A_2 B(f) \\ &\geq A_1 A_2 A_3\end{aligned} \tag{3.3}$$

where

$$A_1 = \inf_{u>0} u^{-1/5}\psi(u) = 1.028493\cdots,$$

$$A_2 = (\gamma^4/4)^{1/5}$$

and

$$A_3 = \inf_{f \in \mathcal{F}} B(f) = (2^9/3^4)^{1/5}.$$

Our next result involves the functions s_n and t_n given by

$$s_n(x) = \{1 - 3(x - rh)^2/h^2\} f''(x)/6, \quad x \in G_{nr}$$

and

$$t_n(x) = \{\tfrac{1}{2} + 2(x - rh)^2/h^2\}^{\frac{1}{2}} f(x)^{\frac{1}{2}}, \quad x \in G_{nr}.$$

Theorem 3.2. *Suppose $f \in \mathcal{F}$ and has compact support. Then*

$$E\{J_n(h)\} = \int (nh)^{-\frac{1}{2}} t_n \psi\left(\frac{(nh^5)^{\frac{1}{2}} s_n}{t_n}\right) + o\{h^2 + (nh)^{-\frac{1}{2}}\}. \tag{3.4}$$

In addition,

$$\limsup_{n \to \infty} n^{2/5} \inf_{h > 0} E\{J_n(h)\} \leq A_4 B(f) \tag{3.5}$$

where

$$A_4 = 5\{\gamma^2/(16\pi)\}^{2/5}.$$

The bounds given at (3.3) and (3.5) are the frequency polygon analogues of the bounds for the histogram stated at (2.5) and (2.6). It follows from these that optimally, $E\{J_n(h)\}$ converges to zero at a rate between $C\,B(f)n^{-2/5}$ and $C^*\,B(f)n^{-2/5}$ where $C = 0.659\cdots$ and $C^* = 0.882\cdots$ (so that $C^*/C = 1.339\cdots$). This optimal rate of convergence of order $n^{-2/5}$ is achieved when $h \sim h_u = u^2 n^{-1/5}$ for some $u > 0$. To extract the exact optimal rate of convergence of $E\{J_n(h)\}$ and the corresponding optimal bin-width we appeal to

Theorem 3.3. *Suppose $f \in \mathcal{F}$ and has compact support. Then for all $u > 0$,*

$$E\{J_n(h_u)\} = \zeta(u) n^{-2/5} + o(n^{-2/5})$$

where

$$\zeta(u) = u^{-1} \int_{-\infty}^{\infty} \int_0^1 \sigma(x,y) \psi\left(\frac{u^5 b(x,y)}{\sigma(x,y)}\right) dy\, dx, \tag{3.6}$$

$b(x,y) = f''(x)(12y^2 - 12y - 1)/24$ and $\sigma^2(x,y) = f(x)(2y^2 - 2y + 1)$.

It follows from this result that the minimising value of u is $u^* = (v^*)^{2/5}$ where v^* is the unique solution of

$$\int_{-\infty}^{\infty}\int_0^1 \left\{4vb(x,y)\Phi\left(\frac{vb(x,y)}{\sigma(x,y)}\right) - \sigma(x,y)\phi\left(\frac{vb(x,y)}{\sigma(x,y)}\right)\right\} dy\,dx = 0. \qquad (3.7)$$

This leads to $h^* = (u^*)^2 n^{-1/5}$ as the asymptotically optimal bin-width.

Once again we note that the assumption of compact support made in Theorems 3.2 and 3.3 can be weakened to the existence of $E|X_1|^{1+\epsilon}$ for some $\epsilon > 0$ by arguing as in the proof of Theorem 2.2.1.

It is seen from results in Section 4 that the L_1 analogue of Scott's (1985) bin-width selection rule given at (3.2) is

$$h_n^* = 2.07sn^{-1/5},$$

which provides about 96% as much smoothing as $h_{n,2}^*$.

3.4 Numerical Results

In this section we shall present some examples of L_1-optimal bin-widths and rates of convergence for the histogram and frequency polygon. Subsection 1 contains results for the histogram; results for the frequency polygon are given in Subsection 2.

3.4.1 *L_1-optimal Bin-widths and Rates of Convergence for the Histogram*

The following discussion applies to the histogram density estimator defined in Section 2. There it was established that the optimal bin-width for minimising expected L_1 loss is asymptotic to $c_1 n^{-1/3}$ where the constant c_1 is obtained by solving $\Lambda(v) = 0$, with Λ having the definition ascribed to it at (2.8), and then setting $c_1 = (v^*)^{2/3}$. The value of c_1 can be compared to its L_2 counterpart

$$c_2 = \left\{6\Big/\int (f')^2\right\}^{1/3},$$

the coefficient of $n^{-1/3}$ in the formula for the L_2-optimal bin-width given at (2.2).

Table 4.1: Values of c_1, c_2 and c_1/c_2 for the histogram.

Density	c_1	c_2	c_1/c_2
N(0,1)	3.37	3.49	0.97
Beta (4,4)	0.57	0.62	0.92
Extreme Value	3.77	3.63	1.04
Logistic	5.75	5.65	1.01

Table 4.2: Values of $CB_H(f)$, $D_1(f)$ and $C^*B_H(f)$ for the histogram.

Density	$CB_H(f)$	$D_1(f)$	$C^*B_H(f)$
N(0,1)	1.11	1.19	1.63
Beta (4,4)	1.04	1.13	1.53
Extreme Value	1.16	1.28	1.71
Logistic	1.19	1.28	1.74

Table 4.3 (a): Values of $B_H(f)$.

Density	$B_H(f)$
N(0,1)	$2^{1/3}$
Beta (4,4)	$(11025\pi^2/65536)^{1/3}$
Extreme Value	$(2\pi/e)^{1/3}$
Logistic	$(\pi/2)^{2/3}$

Table 4.3 (b): Values of $B(f)$.

Density	$B(f)$
N(0,1)	$(128\pi/e)^{1/10}$
Beta (4,4)	$(83349\pi^4 5^{\frac{1}{2}}/8)^{1/5}$
Extreme Value	$[8\pi^2 e^{-3/2}\{5^{\frac{1}{2}}\cosh(5^{\frac{1}{2}}/2) - 2\sinh(5^{\frac{1}{2}}/2)\}]^{1/5}$
Logistic	$(\pi^8/27)^{1/10}$

The values of c_1 and c_2 were obtained numerically for the following densities: N(0,1), Beta(4,4), extreme value and logistic (see Section 2.5 for their respective definitions) and are presented in Table 4.1. Notice that the percentage difference c_1 and c_2 is quite small in every case. This provides further evidence that there is little difference between estimating a density with respect to the L_1 norm and with respect to the L_2 norm.

Also for these densities we computed the corresponding optimal rate of convergence of $E(J_n)$ to zero. This quantity has leading term $D_1(f)n^{-1/3}$ where $D_1(f) = \zeta(c_1^{\frac{1}{2}})$ and ζ is given by the expression at (2.7). According to (2.5) and (2.6) the bounds

$$CB_H(f) \leq D_1(f) \leq C^*B_H(f)$$

exist, where $C \doteq 0.880$ and $C^* \doteq 1.290$. The value of $B_H(f)$ for each example density is given in Table 4.3 (a). In Table 4.2 we tabulate $D_1(f)$, $CB_H(f)$ and $C^*B_H(f)$. As for the kernel density estimator, the lower bound approximates $D_1(f)$ to a high degree of accuracy.

3.4.2 L_1-*optimal Bin-widths and Rates of Convergence for the Frequency Polygon*

Results for the frequency polygon corresponding to those given for the histogram in the previous subsection are presented here. For the frequency polygon density estimator defined in Section 3, the L_1-optimal bin-width is asymptotic to $c_1n^{-1/5}$ where $c_1 = (v^*)^{2/5}$ and v^* is the solution to the equation at (3.7). The L_2-optimal bin-width is asymptotic to $c_2n^{-1/5}$ where

$$c_2 = 2\left[\frac{15}{49\int(f'')^2}\right]^{1/5}.$$

Examples of values of c_1 and c_2 are listed in Table 4.4 and once again we see that there is virtually no difference between them.

For the frequency polygon the optimal rate of convergence of $E(J_n)$ is $D_1(f)n^{-2/5}$ where $D_1(f) = \zeta(c_1^{\frac{1}{2}})n^{-1/5}$ and ζ is as defined at (3.6). In Section 3 $D_1(f)$ was shown to be bounded below by $CB(f)$ where $C \doteq 0.659$, and above by $C^*B(f)$ where $C^* \doteq 0.883$. Table 4.5 provides a comparison of $D_1(f)$ and its

Table 4.4: Values of c_1, c_2 and c_1/c_2 for the frequency polygon.

Density	c_1	c_2	c_1/c_2
N(0,1)	2.07	2.15	0.96
Beta (4,4)	0.35	0.39	0.89
Extreme Value	2.06	2.08	0.99
Logistic	3.41	3.33	1.02

Table 4.5: Values of $C\,B(f)$, $D_1(f)$ and $C^*\,B(f)$ for the frequency polygon.

Density	$C\,B(f)$	$D_1(f)$	$C^*\,B(f)$
N(0,1)	1.09	1.12	1.46
Beta (4,4)	1.02	1.09	1.36
Extreme Value	1.18	1.28	1.58
Logistic	1.19	1.23	1.59

bounds. The values of $B(f)$ for each of these densities are given in Table 4.3 (b).

3.5 Proofs

The symbols $C, C_1, C_2, \ldots$ will be used to denote positive generic constants throughout this section. For the first proof, $f_n(\cdot|h)$ is the histogram density estimator defined in Section 2.

Proof of Theorem 2.1.

According to Theorem 5.6 of Devroye and Györfi (1985, p.99) we have

$$E\{J_n(h_u)\} = \lambda(u)n^{-1/3} + o(n^{-1/3})$$

so it suffices to prove that $\lambda(u) = \zeta(u) + o(1)$ as $n \to \infty$. Let $\xi_{nr} = (r + \frac{1}{2})h_u$, $r \in \mathbf{Z}$, denote the bin mid-points and put $\mathbf{Z}_C = \mathbf{Z} \cap [-Ch_u^{-1}, Ch_u^{-1} - 1]$. Applying the inequality

$$\psi(v) \le (2/\pi)^{\frac{1}{2}} + v, \qquad v > 0, \tag{5.1}$$

we obtain

$$\sum_{r \notin \mathbf{Z}_C} \int_{[-C,C) \cap B_{nr}} f^{\frac{1}{2}} \psi\left(\frac{u^3 z_n}{2f^{\frac{1}{2}}}\right) \le 2h_u\{(2/\pi)^{\frac{1}{2}} \sup f^{\frac{1}{2}} + (u^3/2) \sup |f'|\} = o(1)$$

which gives

$$\begin{aligned}\lambda(u) =& u^{-1} \sum_{r \in \mathbf{Z}_C} \int_{B_{nr}} f(x)^{\frac{1}{2}} \psi\left(\frac{u^3 f'(x)(x - \xi_{nr})}{h_u f(x)^{\frac{1}{2}}}\right) dx + o(1) \\ =& u^{-1} \sum_{r \in \mathbf{Z}_C} h_u \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\xi_{nr} + h_u y)^{\frac{1}{2}} \psi\left(\frac{u^3 f'(\xi_{nr} + h_u y)y}{f(\xi_{nr} + h_u y)^{\frac{1}{2}}}\right) dy + o(1). \end{aligned} \tag{5.2}$$

Also, by Lemma 2.2.1 and Lipschitz continuity of $|f'|$ and $f^{\frac{1}{2}}$,

$$\begin{aligned}&\left| \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\xi_{nr} + h_u y)^{\frac{1}{2}} \psi\left(\frac{u^3 f'(\xi_{nr} + h_u y)y}{f(\xi_{nr} + h_u y)^{\frac{1}{2}}}\right) dy - \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\xi_{nr})^{\frac{1}{2}} \psi\left(\frac{u^3 f'(\xi_{nr})y}{f(\xi_{nr})^{\frac{1}{2}}}\right) dy \right| \\ &\le u^3 \int_{-\frac{1}{2}}^{\frac{1}{2}} ||f'(\xi_{nr} + h_u y)| - |f'(\xi_{nr})|||y| \, dy + (2/\pi)^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} |f(\xi_{nr} + h_u y)^{\frac{1}{2}} - f(\xi_{nr})^{\frac{1}{2}}| \, dy \\ &\le u^3 \int_{-\frac{1}{2}}^{\frac{1}{2}} C_1 y^2 h_u \, dy + (2/\pi)^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} C_2 |y| h_u \, dy \\ &= C_3 h_u \end{aligned} \tag{5.3}$$

where $C_3 > 0$ does not depend on n. Combining (5.2) and (5.3) leads to

$$\lambda(u) = u^{-1} \sum_{r \in \mathbf{Z}_C} h_u \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\xi_{nr})^{\frac{1}{2}} \psi \left(\frac{u^3 f'(\xi_{nr}) y}{f(\xi_{nr})^{\frac{1}{2}}} \right) dy + o(h_u). \tag{5.4}$$

On account of the absolute continuity of f' and Corollary 2.24 of Freedman and Diaconis (1981) the right hand side of (5.4) can be approximated by an integral so that

$$\begin{aligned} \lambda(u) = & \int \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x)^{\frac{1}{2}} \psi \left(\frac{u^3 f'(x) y}{f(x)^{\frac{1}{2}}} \right) dy\, dx + O(h_u) \\ = & \zeta(u) + o(1), \end{aligned}$$

as had to be shown. ∎

For the rest of this section we take $f_n(\cdot|h)$ to be the frequency polygon density estimator defined in Section 3. The proofs of Theorems 3.1, 3.2 and 3.3 are preceded by six lemmas.

Lemma 5.1. *Suppose that* $\lim_{n\to\infty} \int |Ef_n(\cdot|h) - f| = 0$ *for a given* $f \in \mathcal{F}$. *Then* $\lim_{n\to\infty} h = 0$.

Proof. For $x \in \mathbf{R}$ let $g_h(x) = Ef_n(x|h)$. Assume first that $\lim_{n\to\infty} h = \infty$. Then $g_h(x) \to 0$ for all x, and

$$\liminf_{n\to\infty} \int |g_h - f| \geq \int \liminf_{n\to\infty} |g_h - f| = 1$$

which is a contradiction. Let b be an arbitrary positive constant. Observe that $\int |g_b - f| = 0$ is equivalent to

$$\int_{G_{nr}} |g_b - f| = 0 \quad \text{for all } r \in \mathbf{Z}.$$

Since g_b is linear on each G_{nr} this implies that f is piecewise linear. Consequently f'' vanishes almost everywhere implying that $f = 0$ almost everywhere since $f \in \mathcal{F}$. Therefore we must have $\int |g_b - f| > 0$ for $b > 0$. Now assume that $h \to b$ and observe that

$$\int |g_b - f| \leq \int |g_h - f| + \int |g_h - g_b|.$$

Since g_h is continuous in h, $\int |g_h - g_b| \to 0$ which leads to another contradiction. Thus, invoking the subsequence argument, we conclude that $\lim_{n\to\infty} h = 0$. ∎

Lemma 5.2. *Let $S \subseteq \mathbf{R}$ be a finite interval with non-empty interior and p and q be positive measurable functions on S. For any convex function Ψ on S we have*

$$\int_S p\Psi(q/p) \geq \left(\int_S p\right) \Psi\left(\int_S q \Big/ \int_S p\right).$$

Proof. Let U be a random variable having density $pI_S/\int_S p$ and put $V = q(U)/p(U)$. Then, appealing to Jensen's inequality,

$$\begin{aligned}\int_S p\Psi(q/p) \Big/ \int_S p =& E\{\Psi(V)\} \\ \geq& \Psi\{E(V)\} \\ =& \Psi\left(\int_S q \Big/ \int_S p\right)\end{aligned}$$

which immediately gives the desired result. ∎

Lemma 5.3. *Let (Y_1, Y_2) be a random pair having a trinomial $(n; p_1, p_2)$ distribution. For arbitrary real numbers α and $\beta \neq 0$ we have*

$$\sup_{y \in \mathbf{R}} P(\alpha Y_1 + \beta Y_2 = y) \leq c^*(1-p_1)/\{np_2(1-p_1-p_2)\}^{\frac{1}{2}} + P(Y_1 > n/2)$$

where c^* is a universal constant.

Proof. Observe that

$$\begin{aligned}P(\alpha Y_1 + \beta Y_2 = y) =& \sum_{i=0}^{n} P(\alpha Y_1 + \beta Y_2 = y | Y_1 = i) P(Y_1 = i) \\ =& \sum_{i=0}^{n} P\{Z_i = (y - \alpha i)/\beta\} P(Y_1 = i)\end{aligned}$$

where Z_i is a binomial $\{n-i, p_2/(1-p_1)\}$ random variable. Applying Lemma 5.14 of Devroye and Györfi (1985, p.101) we obtain for some universal constant $c > 0$,

$$\begin{aligned}&\sup_{y \in \mathbf{R}} P(\alpha Y_1 + \beta Y_2 = y) \\ &\leq \sum_{i=0}^{n} \left(c[(n-i)p_2(1-p_1)^{-1}\{1 - p_2/(1-p_1)\}]^{-\frac{1}{2}} \wedge 1\right) P(Y_1 = i) \\ &\leq \frac{c(1-p_1)}{\{(n/2)p_2(1-p_1-p_2)\}^{\frac{1}{2}}} \sum_{i \leq n/2} P(Y_1 = i) + \sum_{i > n/2} P(Y_1 = i) \\ &\leq c^*(1-p_1)/\{np_2(1-p_1-p_2)\}^{\frac{1}{2}} + P(Y_1 > n/2)\end{aligned}$$

where $c^* = 2^{\frac{1}{2}}c$. This completes the proof. ∎

In the following we let

$$b_n(x) = Ef_n(x|h) - f(x) \quad \text{and} \quad \sigma_n^2(x) = \mathrm{Var}\{f_n(x|h)\}.$$

Lemma 5.4. *For each* $r \in \mathbf{Z}$,

$$\left| E \int_{G_{nr}} |f_n(\cdot|h) - f| - \int_{G_{nr}} \sigma_n \psi(b_n/\sigma_n) \right| \leq c^\dagger n^{-1}$$

where $c^\dagger > 0$ *is a universal constant.*

Proof. Suppose that $x \in G_{nr}$. Put

$$\begin{aligned} W_i =& h^{-1}(r + \tfrac{1}{2} - x/h) I_{B_{n(r-1)}}(X_i) - h^{-1}(r - \tfrac{1}{2} - x/h) I_{B_{nr}}(X_i) \\ & - E\{h^{-1}(r + \tfrac{1}{2} - x/h) I_{B_{n(r-1)}}(X_1) - h^{-1}(r - \tfrac{1}{2} - x/h) I_{B_{nr}}(X_1)\} \end{aligned}$$

and observe that

$$n^{-1} \sum_{i=1}^{n} W_i = f_n(x|h) - E\{f_n(x|h)\}, \qquad x \in G_{nr}.$$

Applying Lemma 5.8 of Devroye and Györfi (1985, p.90), with $a = -b_n(x)/\sigma_n(x)$, leads to

$$\left| \int_{G_{nr}} E|f_n(\cdot|h) - f| - \int_{G_{nr}} \sigma_n \psi(b_n/\sigma_n)| \leq \right| cn^{-1} \int_{G_{nr}} \frac{E|W_1^3|}{E(W_1^2)}. \qquad (5.5)$$

Since $x \in G_{nr}$ we have

$$\frac{E|W_1^3|}{E(W_1^2)} \leq 2h^{-1} E\{(|r + \tfrac{1}{2} - x/h| + |r - \tfrac{1}{2} - x/h|) W_1^2\}/E(W_1^2) \leq 4h^{-1},$$

so the right-hand side of (5.5) is no more than $4cn^{-1}$ which completes the proof. ∎

Lemma 5.5. *Under the conditions of Theorem 3.2 and assuming that* $\lim_{n\to\infty} h = 0$ *we have*

$$\int |b_n - h^2 s_n| = o(h^2)$$

and

$$\int |b_n| \sim h^2 \int |s_n| \sim (1/8) h^2 \int |f''|.$$

Proof. Our proof uses arguments similar to those given by Devroye and Györfi (1985) for the proof of their Lemma 5.17. Since f'' is everywhere continuous we have by Taylor expansion with remainder,

$$f(y) = f(x) + (y - x) f'(x) + \tfrac{1}{2}(y - x)^2 f''(x) + \tfrac{1}{2}(y - x)^2 \{f''(\xi) - f''(x)\}$$

for some $\xi \in (x,y)$. Therefore

$$\begin{aligned}
Ef_n(x|h) =&(r+\tfrac{1}{2}-x/h)h^{-1}\int_{B_{n(r-1)}} f(y)\,dy-(r-\tfrac{1}{2}-x/h)h^{-1}\int_{B_{nr}} f(y)\,dy\\
=&f(x)+\tfrac{1}{2}(r+\tfrac{1}{2}-x/h)h^{-1}f''(x)\int_{B_{n(r-1)}}(y-x)^2\,dy\\
&-\tfrac{1}{2}(r-\tfrac{1}{2}-x/h)h^{-1}f''(x)\int_{B_{nr}}(y-x)^2\,dy\\
&+\tfrac{1}{2}(r+\tfrac{1}{2}-x/h)h^{-1}\{f''(\xi_0)-f''(x)\}\int_{B_{n(r-1)}}(y-x)^2\,dy\\
&-\tfrac{1}{2}(r-\tfrac{1}{2}-x/h)h^{-1}\{f''(\xi_1)-f''(x)\}\int_{B_{nr}}(y-x)^2\,dy
\end{aligned}$$

for some $\xi_0 \in B_{n(r-1)}$ and $\xi_1 \in B_{nr}$. From this we obtain

$$b_n(x) = h^2 s_n(x) + R_n(x,h)$$

where

$$\begin{aligned}
R_n(x,h) =& \tfrac{1}{2}(r+\tfrac{1}{2}-x/h)\{(h^2/3)(3r^2-3r+1)-xh(2r-1)+x^2\}\{f''(\xi_o)-f''(x)\}\\
&-\tfrac{1}{2}(r-\tfrac{1}{2}-x/h)\{(h^2/3)(3r^2+3r+1)-xh(2r+1)+x^2\}\{f''(\xi_1)-f''(x)\}.
\end{aligned}$$

Notice that

$$\sup_{x\in G_{nr}}|r+\tfrac{1}{2}-x/h| = \sup_{x\in G_{nr}}|r-\tfrac{1}{2}-x/h| = 1$$

and

$$\begin{aligned}
\sup_{x\in G_{nr}}&|(h^2/3)(3r^2-3r+1)-xh(2r-1)+x^2|\\
&= \sup_{x\in G_{nr}}|(h^2/3)(3r^2-3r+1)-xh(2r-1)+x^2|\\
&= 5h^2/6,
\end{aligned}$$

so for $x \in G_{nr}$,

$$|R_n(x,h)| \le h^2 \sup_{|x-y|\le 2h}|f''(x)-f''(y)|.$$

Since f'' is continuous with compact support, f'' is uniformly continuous so for each $\epsilon > 0$ there exists $\delta > 0$ such that $|x-y| < \delta$ implies $|f''(x)-f''(y)| < \epsilon$ for all $x, y \in \mathbf{R}$. Taking $0 < h < \delta/2$ we obtain for x in G_{nr},

$$\sup_{|x-y|\le 2h}|f''(x)-f''(y)| < \epsilon,$$

implying that $h^{-2}|R_n(x,h)| \to 0$ for all x. Thus, by the assumptions on f and bounded convergence,

$$\int |b_n - h^2 s_n| = o(h^2).$$

Also,

$$\int |b_n| \sim h^2 \int |s_n|$$

so it remains to prove $\int |s_n| \sim (1/8)\int |f''|$. Let $[-C, C)$, $C > 0$, contain the support of f and define

$$\mathbf{Z}_C = \mathbf{Z} \cap [-Ch^{-1} + \tfrac{1}{2}, Ch^{-1} - \tfrac{1}{2}] = \{r \in \mathbf{Z} : G_{nr} \subseteq [-C, C)\}. \tag{5.6}$$

First, note that

$$\sum_{r \notin \mathbf{Z}_C} \int_{[-C,C) \cap G_{nr}} |s_n| \leq 2h \sup |f''| = o(1).$$

Secondly, $|f''|$ is uniformly continuous on $[-C, C)$ so for arbitrary $\epsilon > 0$ and sufficiently small h,

$$\sup_{|x-y|<h} ||f''(x)| - |f''(y)|| < \epsilon$$

for all $x, y \in [-C, C)$. It follows that for $r \in \mathbf{Z}_C$,

$$\sup_{x \in G_{nr}} ||f''(x)| - \sup_{G_{nr}} |f''|| < \epsilon \tag{5.7}$$

for all sufficiently small h. Therefore

$$\begin{aligned}
\sum_{r \in \mathbf{Z}_C} \int_{G_{nr}} |s_n| \leq & (1/6) \sum_{r \in \mathbf{Z}_C} \sup_{G_{nr}} |f''| \int_{(r-\frac{1}{2})h}^{(r+\frac{1}{2})h} |1 - 3(x - rh)^2/h^2| \, dx \\
= & (1/3) \sum_{r \in \mathbf{Z}_C} \sup_{G_{nr}} |f''| \int_0^{h/2} (1 - 3x^2/h^2) \, dx \\
= & (1/8) \sum_{r \in \mathbf{Z}_C} \sup_{G_{nr}} |f''| h \\
\leq & (1/8) \sum_{r \in \mathbf{Z}_C} \int_{G_{nr}} (|f''| + \epsilon) + O(h) \quad \text{(by (5.7))} \\
\leq & (1/8) \int |f''| + C\epsilon/4 + O(h).
\end{aligned}$$

Similarly, we can derive the lower bound $(1/8)\int |f''| - C\epsilon/4 - O(h)$ to give the desired result. ∎

Lemma 5.6. *Under the conditions of Theorem 3.2 and the assumptions that* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$,

$$\int |\sigma_n - (nh)^{-\frac{1}{2}} t_n| = o\{(nh)^{-\frac{1}{2}}\}$$

and

$$\int \sigma_n \sim (nh)^{-\frac{1}{2}} \int t_n \sim \gamma (nh)^{-\frac{1}{2}} \int f^{\frac{1}{2}}.$$

Proof. For $x \in G_{nr}$,

$$\begin{aligned}\sigma_n^2(x) =& (r + \tfrac{1}{2} - x/h)^2 n^{-1} h^{-2} \int_{B_{n(r-1)}} f(y)\,dy \left\{1 - \int_{B_{n(r-1)}} f(y)\,dy\right\} \\ &+ (r - \tfrac{1}{2} - x/h)^2 n^{-1} h^{-2} \int_{B_{nr}} f(y)\,dy \left\{1 - \int_{B_{nr}} f(y)\,dy\right\} \\ &+ 2\{(r - x/h)^2 - 1/4\} n^{-1} h^{-2} \int_{B_{n(r-1)}} f(y)\,dy \int_{B_{nr}} f(y)\,dy.\end{aligned}$$

By Taylor expansion and the smoothness assumptions on f we have

$$\int_{B_{n(r-1)}} f(y)\,dy = hf(x) + o(h)$$

and

$$\int_{B_{nr}} f(y)\,dy = hf(x) + o(h)$$

giving, after some algebra,

$$\sigma_n(x) = (nh)^{-\frac{1}{2}} t_n(x,h) + o\{(nh)^{-\frac{1}{2}}\}$$

for all $x \in \mathbf{R}$. Therefore, using the assumption that f has compact support, and by bounded convergence, we obtain

$$\int |\sigma_n - (nh)^{-\frac{1}{2}} t_n| = o\{(nh)^{-\frac{1}{2}}\}$$

and

$$\int \sigma_n \sim (nh)^{-\frac{1}{2}} \int t_n.$$

Let $\mathbf{Z}_C$ have the same definition ascribed to it at (5.6). Then clearly

$$\sum_{r \notin \mathbf{Z}_C} \int_{[-C,C) \cap G_{nr}} t_n \le 2h \sup f^{\frac{1}{2}} = o(1).$$

Since $f^{\frac{1}{2}}$ is uniformly continuous on $[-C, C)$ then, arguing as in the proof of Lemma 5.5, we obtain

$$\begin{aligned}
\sum_{r\in\mathbf{Z}_C}\int_{G_{nr}} t_n &\leq \sum_{r\in\mathbf{Z}_C}\sup_{G_{nr}} f^{\frac{1}{2}}\int_{(r-\frac{1}{2})h}^{(r+\frac{1}{2})h}\{\tfrac{1}{2}+2(x-rh)^2/h^2\}^{\frac{1}{2}}\,dx\\
&= \sum_{r\in\mathbf{Z}_C}\sup_{G_{nr}} f^{\frac{1}{2}}h^{-1}2^{3/2}\int_0^{h/2}\{(h/2)^2+x^2\}^{\frac{1}{2}}\,dx\\
&= \sum_{r\in\mathbf{Z}_C}\sup_{G_{nr}} f^{\frac{1}{2}}\{\tfrac{1}{2}+2^{-3/2}\sinh^{-1}1\}h\\
&= \sum_{r\in\mathbf{Z}_C}\sup_{G_{nr}} f^{\frac{1}{2}}\gamma h\\
&\leq \gamma\sum_{r\in\mathbf{Z}_C}\int_{G_{nr}}(f^{\frac{1}{2}}+\epsilon)+O(h)\\
&\leq \gamma\int f^{\frac{1}{2}}+2C\gamma\epsilon+O(h).
\end{aligned}$$

In the same manner we may derive the lower bound $\gamma\int f^{\frac{1}{2}}-2C\gamma\epsilon-O(h)$ to provide us with

$$\int t_n \sim \gamma\int f^{\frac{1}{2}}$$

as required. ∎

Proof of Theorem 3.1.

We commence with the observation that

$$\inf_{h>0} E\{J_n(h)\} \geq \min[\inf_{h\geq n^{-\frac{1}{2}}} E\{J_n(h)\},\ \inf_{h<n^{-\frac{1}{2}}} E\{J_n(h)\}].$$

Let $h^* = h^*(n)$ be a sequence of bin-widths satisfying $h^* \geq n^{-\frac{1}{2}}$ and

$$E\{J_n(h^*)\} \sim \inf_{h\geq n^{-\frac{1}{2}}} E\{J_n(h)\}.$$

The condition $h^* \geq n^{-\frac{1}{2}}$ clearly implies that $(nh^*)^{-1} = o(n^{-2/5})$. The latter condition on h^* implies that $\lim_{n\to\infty} h^* = 0$. To see this, let $h^\dagger = h^\dagger(n)$ be another sequence satisfying $h^\dagger \to 0$ and $h^\dagger \geq n^{-\frac{1}{2}}$. Then

$$\begin{aligned}
E\{J_n(h^\dagger)\} &\geq E\{J_n(h^*)\}\{1+o(1)\}\\
&\geq \int |Ef_n(\cdot|h^*) - f|\{1+o(1)\}.
\end{aligned}$$

By the L_1 consistency of the frequency polygon, $E\{J_n(h^\dagger)\} \to 0$, which implies that $\int |Ef_n(\cdot|h^*) - f| \to 0$. Lemma 5.1 asserts that $h^* \to 0$. For all intervals $S = (-C, C)$, $C > 0$,

$$\begin{aligned}
\inf_{h \geq n^{-\frac{1}{2}}} E\{J_n(h)\} \sim & E\{J_n(h^*)\} \\
\geq & E \int_S |f_n(\cdot|h^*) - f| \\
\geq & \int_S \sigma_n \psi(b_n/\sigma_n) - cn^{-1}(2C/h^* + 2) \\
& \text{(by Lemma 5.4)} \\
\geq & \left(\int_S \sigma_n\right) \psi\left(\int_S |b_n| \Big/ \int_S \sigma_n\right) - O\{(nh^*)^{-1}\} \\
& \text{(by Lemma 5.2)} \\
\geq & A_1 \left(\int_S \sigma_n\right)^{4/5} \left(\int_S |b_n|\right)^{1/5} - O\{(nh^*)^{-1}\} \\
& \text{(by definition of } A_1) \\
= & A_1 \left\{\gamma(nh^*)^{-\frac{1}{2}} \int_S f^{\frac{1}{2}}\right\}^{4/5} \left\{(1/8)(h^*)^2 \int_S |f''|\right\}^{1/5} \{1 + o(1)\} \\
& \text{(by Lemmas 5.5 and 5.6)} \\
= & A_1 A_2 B(f) n^{-2/5} + o(n^{-2/5}).
\end{aligned}$$

Letting $C \to \infty$ we obtain

$$\liminf_{n \to \infty} n^{2/5} \inf_{h \geq n^{-\frac{1}{2}}} E\{J_n(h)\} \geq A_1 A_2 B(f). \tag{5.8}$$

Next, let h^* be a sequence for which $h^* < n^{-\frac{1}{2}}$ and

$$E\{J_n(h^*)\} \sim \inf_{h < n^{-\frac{1}{2}}} E\{J_n(h)\}.$$

We have

$$\begin{aligned}
\liminf_{n \to \infty} n^{2/5} E\{J_n(h^*)\} \geq & \int \liminf_{n \to \infty} n^{2/5} E|f_n(\cdot|h^*) - f| \\
& \text{(by Fatou's Lemma)} \\
\geq & \tfrac{1}{2} \int \liminf_{n \to \infty} n^{2/5} E|f_n(\cdot|h^*) - Ef_n(\cdot|h^*)| \qquad (5.9)
\end{aligned}$$

by Jensen's inequality. Let $x \in G_{nr}$, for some $r \in \mathbf{Z}$, and (T_{r-1}, T_r) be a random pair having a trinomial $(n; p_{r-1}, p_r)$ distribution, where

$$p_{r-1} \equiv \int_{B_{n(r-1)}} f, \qquad p_r \equiv \int_{B_{nr}} f.$$

Putting

$$T \equiv (r + \tfrac{1}{2} - x/h)T_{r-1} - (r - \tfrac{1}{2} - x/h)T_r$$

we obtain

$$\begin{aligned}
&n^{2/5}E|f_n(x|h^*) - Ef_n(x|h^*)| \\
&= n^{-3/5}(h^*)^{-1}E|T - E(T)| \\
&\geq M\,P\{|T - E(T)| \geq Mh^*n^{3/5}\} \\
&\quad (M > 0 \text{ arbitrary, by Markov's inequality}) \\
&\geq M\left\{1 - 2Mh^*n^{3/5}\sup_{t\in\mathbf{R}} P(T = t)\right\} \\
&\geq M\left(1 - 2Mh^*n^{3/5}\left[\frac{c^*(1-p_{r-1})}{\{np_{r-1}(1-p_{r-1}-p_r)\}^{\frac{1}{2}}} + P\left(T_{r-1} > \frac{n}{2}\right)\right]\right). \qquad (5.10)
\end{aligned}$$

By the consistency of the histogram density estimator (see, e.g., Devroye and Györfi (1985), Theorem 2.2, pp.7,8) $p_{r-1}/h^* \to f(x)$ and $p_r/h^* \to f(x)$ almost everywhere. Therefore

$$\begin{aligned}
h^*n^{3/5}(1-p_{r-1})/\{np_{r-1}(1-p_{r-1}-p_r)\}^{\frac{1}{2}} \sim& f(x)^{-\frac{1}{2}}(h^*)^{\frac{1}{2}}n^{1/10} \\
\geq& f(x)^{-\frac{1}{2}}n^{-3/20} \\
=& o(1) \qquad (5.11)
\end{aligned}$$

for almost all x such that $f(x) > 0$. Additionally,

$$\begin{aligned}
h^*n^{3/5}P(T_{r-1} > n/2) \leq& h^*n^{3/5}P\{|T_{r-1} - E(T_{r-1})| > n(\tfrac{1}{2} - p_{r-1})\} \\
\leq& \frac{h^*p_{r-1}(1-p_{r-1})}{n^{2/5}(p_{r-1}^2 - p_{r-1} + 1/4)} \\
\sim& 4(h^*)^2 f(x)n^{-2/5} \\
=& o(1). \qquad (5.12)
\end{aligned}$$

From (5.11) and (5.12), the expression at (5.10) is asymptotic to M for almost all x satisfying $f(x) > 0$. Since M is arbitrary, it follows from (5.9) that

$$\liminf_{n\to\infty} \inf_{h<n^{-\frac{1}{2}}} n^{2/5}E\{J_n(h)\} = \infty.$$

This, combined with (5.8), implies (3.3).

To verify the second inequality of the theorem we appeal to Theorem 5.3 of Devroye and Györfi (1985, p.88) which declares that for all densities $f \in \mathcal{F}$, $B(f) \geq (2^9/3^4)^{1/5} = A_3$ with the lower bound being attained by the isosceles triangular density (with a generalised definition of $B(f)$). ∎

Proof of Theorem 3.2.

Let $(-C, C)$ contain the support of f for some constant $C > 0$. Using the bound derived in Lemma 5.4 we obtain

$$\left| E\{J_n(h)\} - \int \sigma_n \psi(b_n/\sigma_n) \right| \leq c^\dagger n^{-1}(2C/h + 2) = O\{(nh)^{-1}\}. \tag{5.13}$$

By Lemmas 2.2.1, 5.5 and 5.6,

$$\begin{aligned} &\left| \int \sigma_n \psi(b_n/\sigma_n) - \int (nh)^{-\frac{1}{2}} t_n \psi\left(\frac{h^2 s_n}{(nh)^{-\frac{1}{2}} t_n} \right) \right| \\ &\qquad \leq \int |b_n - h^2 s_n| + (2/\pi)^{\frac{1}{2}} \int |\sigma_n - (nh)^{-\frac{1}{2}} t_n| \\ &\qquad = o\{h^2 + (nh)^{-\frac{1}{2}}\}. \end{aligned} \tag{5.14}$$

The asymptotic behaviour of $E\{J_n(h)\}$, given by (3.4), is a direct consequence of (5.13) and (5.14).

Using the inequality at (5.1) we obtain

$$\begin{aligned} \int (nh)^{-\frac{1}{2}} t_n \psi\left(\frac{(nh^5)^{\frac{1}{2}} s_n}{t_n} \right) \leq & (2/\pi)^{\frac{1}{2}} (nh)^{-\frac{1}{2}} \int t_n + h^2 \int |s_n| \\ = & (2/\pi)^{\frac{1}{2}} \gamma (nh)^{-\frac{1}{2}} \int f^{\frac{1}{2}} + (1/8) h^2 \int |f''| \\ & + o\{h^2 + (nh)^{-\frac{1}{2}}\}, \end{aligned}$$

leading to the bound

$$E\{J_n(h)\} \leq \left\{ (2/\pi)^{\frac{1}{2}} \gamma (nh)^{-\frac{1}{2}} \int f^{\frac{1}{2}} + (1/8) h^2 \int |f''| \right\} \{1 + o(1)\}. \tag{5.15}$$

The value of h which asymptotically minimises the right-hand side of (5.15) is

$$h = \left[\frac{8\gamma^2 (\int f^{\frac{1}{2}})^2}{\pi (\int |f''|)^2} \right]^{1/5} n^{-1/5}.$$

Therefore

$$\inf_{h>0} E\{J_n(h)\} \leq 5\{\gamma^2/(16\pi)\}^{2/5} \left\{ \tfrac{1}{2} \left(\int f^{\frac{1}{2}} \right)^4 \int |f''| \right\}^{1/5} n^{-2/5} + o(n^{-2/5}),$$

which immediately leads to (3.5). ∎

Proof of Theorem 3.3.

Taking $h = h_u = u^2 n^{-1/5}$ in (3.4) we obtain

$$E\{J_n(h_u)\} = u^{-1} \int_{|x|<C} t_n(x)\psi\left(\frac{u^5 s_n(x)}{t_n(x)}\right) dx\, n^{-2/5} + o(n^{-2/5})$$

where the interval $[-C, C)$, $C > 0$, contains the support of f. It therefore suffices to show that

$$\int_{|x|<C} t_n(x)\psi\left(\frac{u^5 s_n(x)}{t_n(x)}\right) dx = \int_{|x|<C}\int_0^1 \sigma(x,y)\psi\left(\frac{u^5 b(x,y)}{\sigma(x,y)}\right) dy\, dx + o(1). \tag{5.16}$$

Define $\tau_{nr} = (r - \frac{1}{2})h_u$, $r \in \mathbf{Z}$, the frequency polygon bin edges, and let

$$\mathbf{Z}_C = \mathbf{Z} \cap [-Ch_u^{-1} + \tfrac{1}{2}, Ch_u^{-1} - \tfrac{1}{2}] = \{r \in \mathbf{Z} : G_{nr} \subseteq [-C, C)\}.$$

Proceeding as in the proof of Theorem 2.1, the left-hand side of (5.16) equals

$$\begin{aligned}
&\sum_{r\in\mathbf{Z}_C} \int_{G_{nr}} [2\{(x-\tau_{nr})/h_u\}^2 - 2(x-\tau_{nr})/h_u + 1]^{\frac{1}{2}} f(x)^{\frac{1}{2}} \\
&\qquad \times \psi\left(\frac{u^5 |12\{(x-\tau_{nr})/h_u\}^2 - 12(x-\tau_{nr})/h_u - 1||f''(x)|}{24[2\{(x-\tau_{nr})/h_u\}^2 - 2(x-\tau_{nr})/h_u + 1]^{\frac{1}{2}} f(x)^{\frac{1}{2}}}\right) dx + o(1) \\
&= \sum_{r\in\mathbf{Z}_C} h_u \int_0^1 (2y^2 - 2y + 1)^{\frac{1}{2}} f(\tau_{nr} + h_u y)^{\frac{1}{2}} \\
&\qquad \times \psi\left(\frac{u^5 |12y^2 - 12y - 1||f''(\tau_{nr} + h_u y)|}{24(2y^2 - 2y + 1)^{\frac{1}{2}} f(\tau_{nr} + h_u y)^{\frac{1}{2}}}\right) dy + o(1) \\
&= \sum_{r\in\mathbf{Z}_C} h_u \int_0^1 (2y^2 - 2y + 1)^{\frac{1}{2}} f(\tau_{nr})^{\frac{1}{2}} \\
&\qquad \times \psi\left(\frac{u^5 |12y^2 - 12y - 1||f''(\tau_{nr})|}{24(2y^2 - 2y + 1)^{\frac{1}{2}} f(\tau_{nr})^{\frac{1}{2}}}\right) dy + o(h_u) \\
&\qquad \text{(by Lemma 2.2.1 and since } |f''| \text{ and } f^{\frac{1}{2}} \text{ are Lipschitz continuous)} \\
&= \int_{|x|<C}\int_0^1 (2y^2 - 2y + 1)^{\frac{1}{2}} f(x)^{\frac{1}{2}} \\
&\qquad \times \psi\left(\frac{u^5 |12y^2 - 12y - 1||f''(x)|}{(2y^2 - 2y + 1)^{\frac{1}{2}} f(x)^{\frac{1}{2}}}\right) dy\, dx + o(1),
\end{aligned}$$

with the last inequality holding since $f^{\frac{1}{2}}$ and f'' are absolutely continuous. The last written expression equals the right-hand side of (5.16) as required. ∎

Chapter Four

MINIMISATION OF L_1 DISTANCE IN KERNEL ESTIMATION OF DENSITY FUNCTIONALS

4.1 Introduction

The kernel method of density estimation has been adapted to allow nonparametric estimation of a selection of related functions. In this chapter we analyse the L_1 convergence properties of kernel-based estimation for regression functions, density modes and density derivatives.

Section 3 discusses the random design kernel regression function estimator; the fixed design case is treated in Section 3. In Section 4 we investigate the convergence properties of mean absolute error of kernel mode estimators. Section 5 briefly discusses the L_1 convergence of kernel density derivative estimation. All proofs are given in Section 6.

4.2 L_1 Theory of the Kernel Regression Function Estimator (Random Design)

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a sequence of independent, identically distributed random pairs. We shall consider the problem of estimating the regression function

$$r(x) \equiv E(Y|X = x).$$

The kernel regression function estimator, first proposed by Nadaraya (1964) and Watson (1964) and based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, may be written

$$r_n(x|h) = a_n(x|h)/f_n(x|h)$$

where $f_n(x|h)$ is the kernel density estimator,

$$f_n(x|h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

and

$$a_n(x|h) = (nh)^{-1} \sum_{i=1}^{n} Y_i K\{(x - X_i)/h\}.$$

The kernel K is assumed to be a symmetric probability density function throughout this section. The extension to higher-order kernels is straightforward. The window-size h is taken to be a sequence constrained to lie in some set $H_n \subseteq \mathbf{R}^+$.

Our theory for the L_1 loss incurred by $r_n(\cdot|h)$ will be confined to the problem of estimating r over compact intervals having non-empty interior and on which the marginal density f is bounded away from zero. Throughout our discussion we shall let S denote an interval in $\mathbf{R}$ having these properties but which is otherwise arbitrary. Notice that r is always well-defined on S since the possibility of r having a zero denominator has, for sufficiently large samples, been eliminated by the above condition on f. The corresponding L_1 loss for this problem is

$$J_n(h) = \int_S |r_n(\cdot|h) - r|.$$

The asymptotic minimisation of this quantity, however, is thwarted by the random denominator of the estimator $r_n(\cdot|h)$ which takes the form of the marginal density estimator $f_n(\cdot|h)$. If K has compact support, as is required for many of the technical results of this section, then for all large n and small h there is a positive chance that $f_n(\cdot|h) \equiv 0$ over some set of positive measure in S. This means that $E\{J_n(h)\}$ will be infinite. To overcome this we shall instead consider the minimisation of $E\{\tilde{J}_n(h)\}$ where

$$\tilde{J}_n(h) = \int_S |r_n(\cdot|h) - r| f_n(\cdot|h) f^{-1}.$$

The modified loss $\tilde{J}_n(h)$ can be thought of as an asymptotic version of $J_n(h)$ since, under certain regularity conditions,

$$\tilde{J}_n(h) = J_n(h) + o(1) \tag{2.1}$$

almost surely as $n \to \infty$, uniformly in $h \in H_n$. An outline of the proof of this statement is given in Section 6. We shall preface the statement of the first main result of this section with some notation. Define the functions a, b, s and σ by

$$a \equiv rf, \qquad b \equiv (\kappa_1/2)(a''f - af'')f^{-2},$$

$$s(x) \equiv E(Y_1^2 | X_1 = x), \qquad \sigma^2 \equiv \kappa_2^2 (s - r^2) f^{-1}.$$

As in Chapter Two, κ_1 and κ_2 are given by

$$\kappa_1 = \int z^2 K \qquad \text{and} \qquad \kappa_2 = \left(\int K^2 \right)^{\frac{1}{2}}.$$

Also, with ψ having the definition ascribed to it in Section 2.2, we put

$$\lambda(u) = u^{-1} \int_S \sigma \psi(u^5 b / \sigma)$$

and $h_u = u^2 n^{-1/5}$ for $u > 0$. Assumptions which will be used in this section are:

(A1) For $n \geq 1$, $H_n = [G^{-1} n^{\delta - 1}, G n^{-\delta}]$ for some constants $G \geq 1$ and $\delta > 0$.

(A2) The function K is Hölder continuous and compactly supported.

(A3) The function f is Hölder continuous and strictly positive on an open interval containing S.

(A4) The functions f and r each have a continuous second derivative on S.

(A5) The functions f and s are each continuous on S.

(A6) The random variables Y_i, $1 \leq i \leq n$, are bounded.

(A7) The function r is bounded on S.

These conditions are obviously not mutually exclusive. However, this arrangement allows convenient statement of this results in this section and in Section 7. Condition (A1) is a common assumption made for the analysis of the regression estimator. Since h must satisfy a similar condition to ensure consistency of $r_n(\cdot | h)$, it is not as restrictive as might at first appear.

Theorem 2.1. *Under conditions (A1) – (A6) we have*

$$\lim_{n \to \infty} \sup_{u \in [C^{-1}, C]} |n^{2/5} E\{\tilde{J}_n(h_u)\} - \lambda(u)| = 0 \tag{2.2}$$

for all $C \geq 1$. Furthermore,

$$\inf_{h \in H_n} E\{\tilde{J}_n(h)\} \sim \lambda(u^*) n^{-2/5} \tag{2.3}$$

where u^* is the minimiser of $\lambda(u)$.

The optimal window-size for minimising asymptotic expected L_1 loss $E\{\tilde{J}_n(h)\}$ is clearly given by $h^* = (u^*)^2 n^{-1/5}$. Arguments given for the density estimators

considered in Chapters Two and Three can easily be adapted to show that u^* exists, is unique and satisfies $u^* = (v^*)^{2/5}$ where v^* is the unique solution of $\Lambda(v) = 0$ and

$$\Lambda(v) = \int_S [4vb\{\Phi(vb/\sigma) - \tfrac{1}{2}\} - \sigma\phi(vb/\sigma)]. \tag{2.4}$$

An L_1-based window-size selection rule for the regression function can be developed from the algorithm described in the previous section by obtaining an estimate of the function Λ. Let $\hat{b}_n$ and $\hat{\sigma}_n$ be L_1-consistent estimates of b and σ respectively and define

$$\Lambda_n(v) = \int_S [4v\hat{b}_n\{\Phi(v\hat{b}_n/\hat{\sigma}_n) - \tfrac{1}{2}\} - \hat{\sigma}_n(v\hat{b}_n/\hat{\sigma}_n)].$$

Our proposed window-size selection rule is

$$h_n^* = (v_n^*)^{2/5} n^{-1/5}$$

where v_n^* satisfies $\Lambda_n(v) = 0$. Arguments identical to those given in Section 2.4 in the context of density estimation imply that h_n^* is asymptotically optimal in the sense that

$$\lim_{n\to\infty} [E\{\tilde{J}(h)\}]_{h=h_n^*} / \inf_{h\in H_n} E\{\tilde{J}_n(h)\} = 1. \tag{2.5}$$

The stochastic analogue of this asymptotic optimality result is

$$\lim_{n\to\infty} \tilde{J}_n(h_n^*) / \inf_{h\in H_n} \tilde{J}_n(h) = 1 \tag{2.6}$$

almost surely. This can be established by appealing to

Theorem 2.2. *Under conditions (A1) - (A3), (A6) and (A7) we have*

$$\lim_{n\to\infty} \{\inf_{h\in H_n} \tilde{J}_n(h)\} / [\inf_{h\in H_n} E\{\tilde{J}_n(h)\}] = 1 \tag{2.7}$$

almost surely, and

$$\lim_{n\to\infty} \tilde{J}_n(h_n^*) / [E\{\tilde{J}_n(h)\}]_{h=h_n^*} = 1 \tag{2.8}$$

almost surely.

The asymptotic optimality result at (2.6) holds under assumptions (A1) - (A3), (A6) and (A7) by virtue of (2.5), (2.7) and (2.8).

The window-size selection rule h_n^* and its asymptotic optimality properties are dependent on L_1-consistent estimators $\hat{b}_n$ and $\hat{\sigma}_n$ for b and σ. The choice of these estimators is now discussed. Let K_0 be a twice-differentiable probability density function having compact support, and set

$$\begin{aligned}
f_n(x|h_1) =& (nh_1)^{-1} \sum_{i=1}^{n} K_0\{(x - X_i)/h_1\}, \\
a_n(x|h_1) =& (nh_1)^{-1} \sum_{i=1}^{n} Y_i K_0\{(x - X_i)/h_1\}, \\
f_n''(x|h_2) =& n^{-1}h_2^{-3} \sum_{i=1}^{n} K_0''\{(x - X_i)/h_2\}, \\
a_n''(x|h_2) =& n^{-1}h_2^{-3} \sum_{i=1}^{n} Y_i K_0''\{(x - X_i)/h_2\}, \\
s_n(x|h_1) =& (nh_1)^{-1} \sum_{i=1}^{n} Y_i^2 K_0\{(x - X_i)/h_1\}/f_n(x|h_1), \\
r_n(x|h_1) =& a_n(x|h_1)/f_n(x|h_1),
\end{aligned}$$

where h_1 and h_2 are each positive. Our proposed estimators $\hat{b}_n$ and $\hat{\sigma}_n$ are

$$\hat{b}_n(\cdot|h_1, h_2) \equiv (\kappa_1/2)\{a_n''(\cdot|h_2)f_n(\cdot|h_1) - a_n(\cdot|h_1)f_n''(\cdot|h_2)\}f_n(\cdot|h_1)^{-2}$$

and

$$\hat{\sigma}_n(\cdot|h_1) \equiv \kappa_2\{s_n(\cdot|h_1) - r_n(\cdot|h_1)^2\}^{\frac{1}{2}} f_n(\cdot|h_1)^{-\frac{1}{2}}.$$

The L_1-consistency of $\hat{b}_n(\cdot|h_1, h_2)$ and $\hat{\sigma}_n(\cdot|h_1)$ is guaranteed by

Theorem 2.3. *Assume that (A1), (A3), (A4) and (A6) are true; $h_1 \in H_n$; and $h_2 \to 0$, $nh_2^5/\log n \to \infty$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} \int_S |\hat{b}_n(\cdot|h_1, h_2) - b| = 0 \tag{2.9}$$

almost surely.

Theorem 2.4. *Assume that (A1), (A3), (A5) and (A6) are true; and $h_1 \in H_n$. Then*

$$\lim_{n\to\infty} \int_S |\hat{\sigma}_n(\cdot|h_1) - \sigma| = 0 \tag{2.10}$$

almost surely.

Effective selection of h_1 and h_2 can be accomplished by using least-squares cross-validation as was suggested in Section 2.5 in the context of density estimation.

4.3 L_1 Theory of the Kernel Regression Function Estimator (Fixed Design)

Consider the model

$$Y_i = m(x_i) + \epsilon_i, \qquad 1 \leq i \leq n,$$

where m is an unknown regression function defined on $[0,1]$, $x_i \equiv i/n$, $1 \leq i \leq n$, are fixed design points, the ϵ_i are independent and identically distributed random variables, each having mean zero and variance σ^2, and the Y_i are the observable response data. A class of kernel estimators of m is defined by

$$m_n(x|h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - x_i)/h\} Y_i$$

where K is a symmetric positive kernel integrating to unity and supported on $[-1,1]$. Gasser and Müller (1979) observe that the asymptotic analysis of expected L_2 loss of this estimator , $M_n(h) = \int_0^1 \{m(\cdot|h) - m\}^2$, is hindered the fact that the integrated squared bias near the boundaries of [0,1] dominates the integrated squared bias in the interior. To overcome this problem these authors suggest modifying the kernel estimator near the boundaries of [0,1] by using a second order kernel K_α, with support confined to $[-1,\alpha]$, to estimate $m(\alpha h)$ for each $0 \leq \alpha < 1$ and the kernel K_α^-, given by $K_\alpha^- = K_\alpha(-x)$, to estimate $m(1 - \alpha h)$. The modified kernel estimator is therefore

$$m_n(x|h) = \begin{cases} (nh)^{-1} \sum_{i=1}^{n} K_\alpha\{(x - x_i)/h\} Y_i, & x = \alpha h,\ 0 \leq \alpha < 1; \\ (nh)^{-1} \sum_{i=1}^{n} K\{(x - x_i)/h\} Y_i, & h \leq x \leq 1 - h; \\ (nh)^{-1} \sum_{i=1}^{n} K_\alpha\{(x_i - x)/h\} Y_i, & x = 1 - \alpha h,\ 0 \leq \alpha < 1. \end{cases} \tag{3.1}$$

Assume K and K_α are of second order, and satisfy

(K1) The functions K and K_α, $0 \leq \alpha < 1$, each satisfy a Lipschitz condition of order 1.

(K2) $\int_{-\alpha}^{1} K_\alpha(x)^2\,dx \le C$ where $C > 0$ is a constant not depending on α.

(K3) The kernels K_α depend continuously on α and $K_\alpha \to K$ as $\alpha \to 1$.

Examples of kernels satisfying these properties are given in Gasser and Müller (1979). They also show that under (K1) – (K3) and the assumption that m has two continuous derivatives on [0,1] the L_2 loss of the estimator at (3.1), $M_n(h) = \int_0^1 \{m_n(\cdot|h) - m\}^2$, satisfies

$$E\{M_n(h)\} = (\kappa_1^2/4)h^4 \int_0^1 (m'')^2 + \sigma^2\kappa_2(nh)^{-1} + o\{h^4 + (nh)^{-1}\} + O(n^{-1}) \quad (3.2)$$

where $\kappa_1 = \int z^2 K$ and $\kappa_2 = \left(\int K^2\right)^{\frac{1}{2}}$. The L_2-optimal value of h is therefore asymptotic to

$$h_2^* = \left[\frac{\kappa_2^2\sigma^2}{\kappa_1^2 \int_0^1 (m'')^2}\right]^{1/5} n^{-1/5}. \quad (3.3)$$

The derivation of (3.2) makes use of the following approximations for the mean and variance of $m_n(x|h)$:

$$E\{m_n(x|h)\} = h^{-1} \int_0^1 K^\dagger\{(x-w)/h\} m(w)\,dw + O(n^{-1}), \quad (3.4)$$

where $K^\dagger$ stands for either K or K_α, depending on the value of x, and

$$\mathrm{Var}\{m_n(x|h)\} = \sigma^2 \int (K^\dagger)^2 (nh)^{-1} + O\{(nh)^{-2}\}. \quad (3.5)$$

The analogue of (3.2) for L_1 loss $J_n(h) = \int_0^1 |m_n(\cdot|h) - m|$ is provided by

Theorem 3.1. *Assuming (K1) – (K3) are satisfied; the function m has a continuous second derivative on [0,1]; the random variables Y_i, $1 \le i \le n$, are bounded; and $h \to 0$, $nh \to \infty$ as $n \to \infty$, we have*

$$E\{J_n(h)\} = \sigma\kappa_2(nh)^{-\frac{1}{2}} \int_0^1 \psi\left(\frac{(nh^5)^{\frac{1}{2}}\kappa_1 m''}{2\kappa_2\sigma}\right) + o\{h^2 + (nh)^{-\frac{1}{2}}\} + O(n^{-1}). \quad (3.6)$$

In particular,

$$E\{J_n(h_u)\} = \sigma\kappa_2 u^{-1} \int_0^1 \psi\left(\frac{u^5\kappa_1 m''}{2\kappa_2\sigma}\right) n^{-2/5} + o(n^{-2/5}),$$

where $h_u = u^2 n^{-1/5}$.

The asymptotic expansion of $E\{J_n(h)\}$ at (3.6) can be used to show that the L_1-optimal window-size is asymptotic to $h^* = (v^*)^{2/5}n^{-1/5}$ where v^* is the solution to

$$\int_0^1 \left[2v\kappa_1 m'' \left\{\Phi\left(\frac{v\kappa_1 m''}{2\kappa_2\sigma}\right) - \frac{1}{2}\right\} - \kappa_2\sigma\phi\left(\frac{v\kappa_1 m''}{2\kappa_2\sigma}\right)\right] = 0.$$

Table 3.1 lists values of $c_1\sigma^{-2/5}$ and $c_2\sigma^{-2/5}$ for a selection of regression functions when the Epanechnikov kernel is in use. Here c_1 and c_2 are the coefficients of $n^{-1/5}$ in the formulae for h^* and h_2^* respectively. The functions are (i) $m(t) = \cos(\pi t)$, (ii) $m(t) = e^t$, (iii) $m(t) = e^t/(e^t + e^{\frac{1}{2}})$ and (iv) $m(t) = (t+1)^{-1}$.

4.4 L_1 Theory of the Kernel Mode Estimator

The mode of a univariate density f is defined to be $M(f)$ where

$$M(f) = \inf\left\{t : f(t) = \sup_{s\in\mathbf{R}} f(s)\right\}.$$

Let θ stand for $M(f)$ and consider the estimator for θ,

$$\theta_n(h) = M\{f_n(\cdot|h)\},$$

where $f_n(x|h) = (nh)^{-1}\sum_{i=1}^n K\{(x - X_i)/h\}$ and $X_1, \ldots, X_n$ is a sample of independent random variables, each having density f. The function K is assumed to be a pth order kernel with a continuous first derivative. The analogue of expected L_1 loss for mode estimation is mean absolute error (MAE), given by

$$\mathrm{MAE}\{\theta_n(h)\} = E|\theta_n(h) - \theta|.$$

Asymptotic theory for the kernel mode estimator in the context of minimising the mean squared error of $\theta_n(h)$ has been developed by Eddy (1980) where, under the assumptions of h converging to zero at a rate slower than $n^{-1/5}$, it was established that

$$\lim_{n\to\infty}(nh^{2p+3})^{\frac{1}{2}} = d < \infty,$$

and K and f satisfying certain regularity conditions (see Theorem 2.1, Eddy (1980) for details), that

$$(nh^3)^{\frac{1}{2}}\{\theta_n(h) - \theta\} \to \mathrm{N}\left(\frac{d\kappa_1 f^{(p+1)}(\theta)}{p!f''(\theta)}, \frac{f(\theta)\kappa_{2,1}}{\{f''(\theta)\}^2}\right). \tag{4.1}$$

In this context $\kappa_1 = (-1)^p \int z^p K$ and $\kappa_{2,1} = \left[\int \{K^{(1)}\}^2\right]^{\frac{1}{2}}$. Defining

$$b(\theta) = \frac{\kappa_1 f^{(p+1)}(\theta)}{p! f''(\theta)} \qquad \text{and} \qquad \sigma^2(\theta) = \frac{f(\theta)\kappa_{2,1}^2}{\{f''(\theta)\}^2}$$

we may transform (4.1) to

$$\theta_n(h) - \theta = (nh^3)^{-\frac{1}{2}}\{db(\theta) + \sigma(\theta)Z\} \tag{4.2}$$

where Z is a random variable having an asymptotically N(0,1) distribution. The relation at (4.2) can be used to asymptotically minimise mean absolute error by taking h asymptotic to $h_u = u^2 n^{-1/(2p+3)}$ and observing that

$$\theta_n(h_u) - \theta = n^{-p/(2p+3)}\{u^{2p} b(\theta) + u^{-3}\sigma(\theta)Z\}.$$

Therefore, as $n \to \infty$,

$$E|\theta_n(h_u) - \theta| = \lambda(u) n^{-p/(2p+3)} + o\{n^{-p/(2p+3)}\}$$

where

$$\begin{aligned}\lambda(u) =& E|u^{2p} b(\theta) + u^{-3}\sigma(\theta)Z| \\ =& u^{-3}\sigma(\theta)\psi\{u^{2p+3} b(\theta)/\sigma(\theta)\}.\end{aligned}$$

The MAE-optimal window-size is therefore asymptotic to

$$h^* = (u^*)^2 n^{-p/(2p+3)}$$

where $u^* = (v^*)^{1/(2p+3)}$ and v^* is the unique solution to $\Lambda(v) = 0$ and

$$\Lambda(v) = 2pvb(\theta)[\Phi\{vb(\theta)/\sigma(\theta)\} - \tfrac{1}{2}] - \sigma(\theta)\phi\{vb(\theta)/\sigma(\theta)\}.$$

Consequently $v^* = \alpha_p \sigma(\theta)/b(\theta)$ where α_p is the unique solution to

$$2p\alpha_p\{\Phi(\alpha_p) - \tfrac{1}{2}\} - \phi(\alpha_p) = 0.$$

In the special case where $p = 2$ we obtain $v^* = \alpha_2 \sigma(\theta)/b(\theta)$ where

$$\alpha_2 = 0.480949\cdots.$$

Table 3.1: Values of $c_1\sigma^{-2/5}$, $c_2\sigma^{-2/5}$ and for fixed design regression estimation using the Epanechnikov kernel.

Reg. funct.	$c_1\sigma^{-2/5}$	$c_2\sigma^{-2/5}$	c_1/c_2
(i)	0.781	0.790	0.989
(ii)	1.345	1.363	0.987
(iii)	6.557	6.620	0.991
(iv)	1.801	1.809	0.996

Table 4.1: Values of c_1, c_2 and for mode estimation using the Gaussian kernel.

Density	c_1	c_2
Extreme Value	0.862	1.020
Gamma (1,5)	1.710	2.023
Lognormal (0,1)	0.246	0.291

Therefore the MAE-optimal window-size is asymptotic to $h^* = c_1 n^{-1/7}$ where

$$c_1 = \left[\frac{(0.9252\cdots) f(\theta) \kappa_{2,1}^2}{\kappa_1^2 \{f'''(\theta)\}^2} \right]^{1/7}.$$

Comparing this with the corresponding asymptotic MSE-optimal window-size $h_2^* = c_2 n^{-1/7}$ where

$$c_2 = \left[\frac{3 f(\theta) \kappa_{2,1}^2}{\kappa_1^2 \{f'''(\theta)\}^2} \right]^{1/7}$$

we see that the ratio of h^* to h_2^* is given by

$$c_1/c_2 = (0.95252\cdots/3)^{1/7} = 0.8453\cdots$$

for all sufficiently smooth densities f such that $f''(\theta)$, $f'''(\theta) \neq 0$ and second-order kernels K. Table 4.1 lists values of c_1 and c_2 for a selection of densities when K is the Gaussian kernel. The extreme value density is as defined in Section 2.5, the Gamma (1,5) density is given by

$$f(x) = (1/24) x^4 e^{-x}, \ x > 0,$$

whereas the lognormal (0,1) density is

$$f(x) = \{x(2\pi)^{\frac{1}{2}}\}^{-1} e^{-(\log x)^2/2}, \ x > 0.$$

4.5 L_1 Theory of the Kernel Density Derivative Estimator

Let $X_1, \ldots, X_n$ be a sample of independent real-valued random variables having common density f. A kernel estimator of $f^{(r)}$, the rth derivative of f, is

$$f_n^{(r)}(x|h) = n^{-1} h^{-r-1} \sum_{i=1}^{n} K^{(r)}\{(x - X_i)/h\}$$

where K is a pth order kernel that is r times differentiable and the window-size $h = h(n)$ satisfies $h \to 0$ and $nh^{2r+1} \to \infty$ as $n \to \infty$. Assume that $f^{(r+p)}$ is continuous, and define the functions b and σ by

$$b \equiv (\kappa_1/p!) f^{(r+p)}, \qquad \sigma \equiv \kappa_{2,r} f^{\frac{1}{2}},$$

where $\kappa_1 = (-1)^p \int z^p K$ and $\kappa_{2,r} = \left[\int \{K^{(r)}\}^2\right]^{\frac{1}{2}}$. The L_1 loss of $f_n^{(r)}(\cdot|h)$ is given by $J_n(h) = \int |f_n^{(r)}(\cdot|h) - f^{(r)}|$ and, by a straightforward extension of the theory of Section 2.2, satisfies

$$E\{J_n(h)\} = \int (nh^{2r+1})^{-\frac{1}{2}} \sigma \psi \left(\frac{(nh^{2r+2p+1})^{\frac{1}{2}} b}{\sigma} \right) + o\{h^p + (nh^{2p+1})^{-\frac{1}{2}}\}.$$

The L_1 optimal window-size has leading term

$$h^* = (v^*)^{2/(2r+2p+1)} n^{-1/(2r+2p+1)}$$

where v^* is the unique solution of

$$\int \{2pvb\Phi(vb/\sigma) - (2r+1)\sigma\phi(vb/\sigma)\} = 0.$$

The optimal rate of convergence of $E(J_n)$ is therefore given by

$$E\{J_n(h^*)\} \sim (v^*)^{-(2r+1)/(2r+2p+1)} \int \sigma\psi(v^* b/\sigma) n^{-p/(2r+2p+1)}.$$

4.6 Proofs

In the following, $C, C_0, C_1, \ldots$ will be used to denote positive generic constants. The interval S is assumed to have the properties ascribed to it in Section 2. Its Lebesgue measure will be denoted throughout by $\mathcal{L}(S)$.

Proof of (2.1).

We give an outline only. The required assumptions are (A1)–(A3), (A6), (A7) and the Hölder continuity of $a \equiv rf$. It may be shown that

$$|\tilde{J}_n(h) - J_n(h)| \leq \mathcal{L}(S)(\inf_S f_n)^{-1}(\inf_S f)^{-1} \sup_S |f_n - f|(\sup_S |a_n - a| + \sup_S |r| \sup_S |f_n - f|). \tag{6.1}$$

Lemma 1 of Härdle and Marron (1985) asserts that, under the above conditions,

$$\lim_{n\to\infty} \sup_S \sup_{h \in H_n} |f_n(\cdot|h) - f| = 0 \tag{6.2}$$

almost surely. Also, it may be established in a similar fashion using (A6) that

$$\lim_{n\to\infty} \sup_S \sup_{h \in H_n} |a_n(\cdot|h) - a| = 0 \tag{6.3}$$

almost surely. From (A3), $\inf_S f \geq C_0 > 0$. Therefore, we have from (6.2) that

$$\inf_S f_n(\cdot|h) \geq \tfrac{1}{2}C_0 \tag{6.4}$$

with probability tending to 1 as $n \to \infty$, uniformly in $h \in H_n$. Result (2.1) is a direct consequence of (6.1), (6.2), (6.3) and (6.4). ∎

In the following we shall write

$$\tilde{r}_n(\cdot|h) \equiv r_n(\cdot|h)f_n(\cdot|h)f^{-1} + r\{1 - f_n(\cdot|h)f^{-1}\}$$

so that

$$\tilde{J}_n(h) = \int_s |\tilde{r}_n(\cdot|h) - r|.$$

In addition we put

$$\tilde{b}_n(\cdot|h) \equiv E\tilde{r}_n(\cdot|h) - r, \qquad \tilde{\sigma}_n^2(\cdot|h) \equiv \mathrm{Var}\{\tilde{r}_n(\cdot|h)\}.$$

The first three lemmas are required for the proof of Theorem 2.1.

Lemma 6.1. *Assume that* $\lim_{n\to\infty} h = 0$ *and conditions (A3) and (A4) hold. Then*

$$\int_S |\tilde{b}_n - h^2 b| = o(h^2).$$

Proof. Observe that

$$\begin{aligned} Ea_n(x|h) =& h^{-1}E[E(Y_1|X_1)K\{(x - X_1)/h\}] \\ =& h^{-1}\int_{-\infty}^{\infty} r(u)K\{(x-u)/h\}f(u)\,du \\ =& \int_{-\infty}^{\infty} r(x - hz)f(x - hz)K(z)\,dz \\ =& r(x)f(x) + \tfrac{1}{2}h^2(rf)''(x)\kappa_1 + o(h^2) \end{aligned}$$

uniformly in $x \in S$. Similar calculations give

$$Ef_n(x|h) = f(x) + \tfrac{1}{2}h^2 f''(x)\kappa_1 + o(h^2).$$

Therefore,

$$\begin{aligned} \tilde{b}_n(x) &= f(x)^{-1}\{Ea_n(x|h) - r(x)Ef_n(x|h)\} \\ &= \tfrac{1}{2}h^2\kappa_1 f(x)^{-1}\{a''(x) - r(x)f''(x)\} + o(h^2) \\ &= h^2 b(x) + o(h^2) \end{aligned}$$

uniformly in S. The required result follows by dominated convergence. ∎

Lemma 6.2. *Assume that $nh \to \infty$ as $n \to \infty$ and conditions (A3) and (A5) are true. Then*

$$\int_S |\tilde{\sigma}_n - (nh)^{-\frac{1}{2}}\sigma| = o\{(nh)^{-\frac{1}{2}}\}.$$

Proof. As in the proof of the previous lemma, Taylor expansion can be used to establish that

$$\mathrm{Var}\{a_n(x|h)\} = (nh)^{-1}s(x)f(x)\kappa_2^2 + o\{(nh)^{-1}\},$$

$$\mathrm{Var}\{f_n(x|h)\} = (nh)^{-1}f(x)\kappa_2^2 + o\{(nh)^{-1}\}$$

and

$$\mathrm{Cov}\{a_n(x|h), f_n(x|h)\} = (nh)^{-1}r(x)f(x)\kappa_2^2 + o\{(nh)^{-1}\},$$

where all expansions are uniform in $x \in S$. Thus,

$$\begin{aligned}\tilde{\sigma}_n^2(x) =& f(x)^{-2}[\mathrm{Var}\{a_n(x|h)\} + r(x)^2\mathrm{Var}\{f_n(x|h)\} - 2r(x)\mathrm{Cov}\{a_n(x|h), f_n(x|h)\}] \\ =& (nh)^{-1}\kappa_2^2 f(x)^{-1}\{s(x) - r(x)^2\} + o\{(nh)^{-1}\}\end{aligned}$$

from which the required result follows. ∎

Lemma 6.3. Let (A3), (A6) and (A7) hold and $x \in S$. We have

$$\left| E|\tilde{r}_n(x|h) - r(x)| - \tilde{\sigma}_n(x)\psi\left(\frac{\tilde{b}_n(x)}{\tilde{\sigma}_n(x)}\right)\right| \leq c(nh)^{-1} \tag{6.5}$$

where c is a positive constant independent of x, n and h.

Proof. Put

$$\begin{aligned}W_i =& h^{-1}f(x)^{-1}\{Y_i - r(x)\}K\{(x - X_i)/h\} \\ & - E[h^{-1}f(x)^{-1}\{Y_i - r(x)\}K\{(x - X_i)/h\}]\end{aligned}$$

for $1 \leq i \leq n$. The W_i's are clearly independent and identically distributed random variables each having mean zero and variance $n\tilde{\sigma}_n(x)$. From Lemma 5.8 of Devroye and Györfi (1985, p.90) we obtain

$$\left| E|\tilde{r}_n(x|h) - r(x)| - \tilde{\sigma}_n(x)\psi\left(\frac{\tilde{b}_n(x)}{\tilde{\sigma}_n(x)}\right)\right| \leq \frac{c^* E|W_1|^3}{nE(W_1^2)}$$

for a universal constant $c^* > 0$. Also,

$$\begin{aligned}E|W_1|^3 \leq& 2h^{-1}f(x)^{-1}E[\{|Y_1| + |r(x)|\}|K\{(x - X_i)/h\}|W_1^2] \\ \leq& 2h^{-1}B\,E(W_1^2)\end{aligned}$$

where B is an upper bound to $(|Y_1| + |r|)|K|f^{-1}$ on S. The left-hand side of (6.5) is therefore bounded above by $c = 2Bc^*$. ∎

Proof of Theorem 2.1.

Clearly $T_{n1} + T_{n2}$, where

$$T_{n1} = n^{2/5} \sup_{u \in [C^{-1}, C]} \left| \int_S E|\tilde{r}_n(\cdot|h_u) - r| - \int_S \tilde{\sigma}_n \psi(\tilde{b}_n/\tilde{\sigma}_n) \right|$$

and

$$T_{n2} = \sup_{u \in [C^{-1}, C]} \left| n^{2/5} \int_S \tilde{\sigma}_n \psi(\tilde{b}_n/\tilde{\sigma}_n) - \lambda(u) \right|.$$

A straightforward application of Lemma 6.3 gives

$$T_{n1} \leq cC^2 \mathcal{L}(S) n^{-2/5}. \tag{6.6}$$

By Lemmas 2.2.1, 6.1 and 6.2,

$$\begin{aligned}
T_{n2} \leq & \sup_{u \in [C^{-1}, C]} \int_S |n^{2/5} \tilde{\sigma}_n \psi(\tilde{b}_n/\tilde{\sigma}_n) - u^{-1} \sigma \psi(u^5 b/\sigma)| \\
\leq & \sup_{u \in [C^{-1}, C]} \int_S |n^{2/5} |\tilde{b}_n| - u^4 |b|| \\
& + (2/\pi)^{\frac{1}{2}} \sup_{u \in [C^{-1}, C]} \int_S |n^{2/5} \tilde{\sigma}_n - u^{-1} \sigma| \\
\leq & C^4 \sup_{u \in [C^{-1}, C]} h_u^{-2} \int_S |\tilde{b}_n - h_u^2 b| \\
& + (2/\pi)^{\frac{1}{2}} C (nh_u)^{\frac{1}{2}} \int_S |\tilde{\sigma}_n - (nh_u)^{-\frac{1}{2}} \sigma| \\
= & o(1).
\end{aligned} \tag{6.7}$$

Combining (6.6) and (6.7) we obtain (2.2).

The proof of (2.3) follows the same arguments used to prove the analogous result in the kernel density estimation case (Theorem 2.2.1). A substantial part of these is the establishment of the following inequalities:

$$\int_S |E\tilde{r}_n(\cdot|h) - r| \geq C_1(h^2 \wedge 1) \tag{6.8}$$

and

$$\int_S E|\tilde{r}_n(\cdot|h) - E\tilde{r}_n(\cdot|h)| \geq C_2\{(nh)^{-\frac{1}{2}} \wedge 1\}. \tag{6.9}$$

These inequalities are readily derived using essentially the same ideas as those employed in the proofs of Theorem 2.2.1 and Lemma 2.6.1. ∎

Proof of Theorem 2.2. The proofs of (2.7) and (2.8) can be performed by straightforward adaptation of the proof of Theorem 2.4.1. However, we require the regression analogue of Lemma 2.6 which is stated and proved below. ∎

Lemma 6.4. *Assume that f is strictly positive on S and conditions (A6) and (A7) are satisfied. Then for all $\epsilon > 0$,*

$$\sup_{h>0} P[|\tilde{J}_n(h) - E\{\tilde{J}_n(h)\}| > \epsilon] \leq 2e^{-C_0 n\epsilon^2}$$

where C_0 is a positive constant not depending on n or ϵ.

Proof. For $1 \leq i \leq n$ we let $\mathcal{F}_i$ denote the σ-field generated by the random variables $\{X_1, \ldots, X_i, Y_1, \ldots, Y_i\}$; $\mathcal{F}_0$ is defined to be the trivial σ-field. Also, we shall let

$$V_n \equiv \tilde{J}_n(h) - E\{\tilde{J}_n(h)\}$$

and

$$Z_i \equiv E\{\tilde{J}_n(h)|\mathcal{F}_i\} - E\{\tilde{J}_n(h)|\mathcal{F}_{i-1}\}.$$

Notice that $V_n = \sum_{i=1}^n Z_i$ and the Z_i's form a martingale difference sequence. Lemma 2 of Devroye (1988) declares that for such a sequence,

$$P\left(\left|\sum_{i=1}^n Z_i\right| > \epsilon\right) \leq 2\exp\left[-\epsilon^2\left\{2\sum_{i=1}^n (\text{ess sup}\,|Z_i|)^2\right\}^{-1}\right]. \tag{6.10}$$

Letting

$$W_{i,k} = n^{-1}h^{-1}f(x)^{-1}\sum_{j=i}^{k}\{Y_j - r(x)\}K\{(x - X_j)/h\}$$

one obtains, for $1 \leq i \leq n$,

$$\begin{aligned}
|Z_i| &\leq \int_S |E(|\tilde{r}_n - r||\mathcal{F}_i) - E(|\tilde{r}_n - r||\mathcal{F}_{i-1})| \\
&= \int_S |E(|W_{1,i-1} + W_{i,i} + W_{i+1,n}||\mathcal{F}_i) - E(|W_{1,i-1} + W_{i,i} + W_{i+1,n}||\mathcal{F}_{i-1})| \\
&\leq \int_S \sup_{a\in\mathbf{R}} ||a + W_{i,i}| - E|a + W_{i,i}|| \\
&\leq \int_S |W_{i,i} - E(W_{i,i})| + \int_S E|W_{i,i} - E(W_{i,i})|.
\end{aligned}$$

The last step is a consequence of Lemma 1 of Devroye (1988). Therefore

$$\begin{aligned}|Z_i| \leq & \int_S |W_{i,i}| + 3E \int_S |W_{i,i}| \\ \leq & C_1 n^{-1}\end{aligned}$$

where $C_1 = 4 \int |K| \sup_S f^{-1}(\sup_S |r| + B)$ and B is an upper bound to the Y_i's, $1 \leq i \leq n$. This implies that

$$\sum_{i=1}^{n} (\text{ess sup } |Z_i|)^2 \leq C_1^2 n^{-1},$$

so that the right-hand side of (6.10) is bounded above by $2\exp(-n\epsilon^2 C_0)$ where $C_0 = \frac{1}{2} C_1^{-2}$. The left-hand side of (6.10) is simply $P[|\tilde{J}_n(h) - E\{\tilde{J}_n(h)\}| > \epsilon]$ so the proof is finished. ∎

For the proofs of Theorems 2.3 and 2.4 we require the following lemma.

Lemma 6.5. *Under conditions (A1) – (A3) we have*

$$\lim_{n \to \infty} \sup_{x \in S} \sup_{h \in H_n} \left| \frac{f_n(x|h) - f(x)}{f_n(x|h)} \right| = 0$$

almost surely.

Proof. According to a result of Härdle and Marron (1985, Lemma 1),

$$\lim_{n \to \infty} \sup_{x \in S} \sup_{h \in H_n} |f_n(x|h) - f(x)| = 0 \tag{6.11}$$

almost surely. Therefore, for all sufficiently large n,

$$\sup_{x \in S} \sup_{h \in H_n} |f_n(x|h) - f(x)| \leq \inf_{x \in S} f(x).$$

Hence,

$$\begin{aligned}\sup_{x \in S} \sup_{h \in H_n} \left| \frac{f_n(x|h) - f(x)}{f_n(x|h)} \right| \leq & \frac{\sup_{x \in S} \sup_{h \in H_n} |f_n(x|h) - f(x)|}{\sup_{x \in S} \sup_{h \in H_n} |f(x) - |f_n(x|h) - f(x)||} \\ \leq & \frac{\sup_{x \in S} \sup_{h \in H_n} |f_n(x|h) - f(x)|}{\inf_{x \in S} f(x) - \sup_{x \in S} \sup_{h \in H_n} |f_n(x|h) - f(x)|}\end{aligned}$$

for all large n. The required result follows from (6.11). ∎

Proof of Theorem 2.3.

From the triangle inequality we have

$$2\kappa_1^{-1}|\hat{b}_n(\cdot|h_1,h_2)-b| \leq |a_n''(\cdot|h_2)f_n(\cdot|h_1)^{-1}-a''f^{-1}| + |a_n(\cdot|h_1)f_n''(\cdot|h_2)f_n''(\cdot|h_1)^{-2}-af''f^{-2}|$$

so it suffices to show that

$$\lim_{n\to\infty}\int_S |a_n''(\cdot|h_2)f_n(\cdot|h_1)^{-1}-a''f^{-1}| = 0 \tag{6.12}$$

almost surely, and

$$\lim_{n\to\infty}\int_S |a_n(\cdot|h_1)f_n''(\cdot|h_2)f_n(\cdot|h_1)^{-2}-af''f^{-2}| = 0 \tag{6.13}$$

almost surely. In the following we suppress the dependence of $a_n(\cdot|h_1)$ and $f_n(\cdot|h_1)$ on h_1 and $a_n''(\cdot|h_2)$ and $f_n''(\cdot|h_2)$ on h_2. Observe that

$$a_n''f_n^{-1}-a''f^{-1} = f^{-1}\{a_n''-a''+a''f^{-1}(f-f_n)\}\{1+(f-f_n)f_n^{-1}\}$$

which implies

$$\int_S |a_n''f_n^{-1}-a''f^{-1}| \leq \sup_S(f^{-1})\int_S |a_n''-a''|\{1+\sup_S|(f_n-f)f_n^{-1}|\} + \sup_S(|a''|f^{-2})\int_S |f_n-f|\{1+\sup_S|(f_n-f)f_n^{-1}|\}.$$

Since $h_1 \in H_n$ we have by Lemma 6.5,

$$\lim_{n\to\infty}\sup_S|(f_n-f)f_n^{-1}| = 0 \tag{6.14}$$

almost surely. Also, by Theorem 3.1 of Devroye and Györfi (1985, p.12) we have

$$\lim_{n\to\infty}\int_S |f_n-f| = 0 \tag{6.15}$$

almost surely. Therefore, to prove (6.12) it remains to show that

$$\lim_{n\to\infty}\int |a_n''-a''| = 0$$

almost surely. However, this is a consequence of

$$\lim_{n\to\infty}\int_S |Ea_n''-a''| = 0 \tag{6.16}$$

and

$$\lim_{n\to\infty} \int_S |a_n'' - Ea_n''| = 0 \tag{6.17}$$

almost surely. Using integration by parts we obtain

$$Ea_n''(x|h_2) = \int_{-\infty}^{\infty} a''(x - h_2 z) K_0(z)\, dz.$$

Suppose that the support of K_0 is contained in $[-w, w]$ for some $w > 0$. Then

$$\int_S |Ea_n''(x|h_2) - a''(x)|\, dx \le \mathcal{L}(S) \int_{|z|\le w} |K_0(z)| \sup_{x\in S} |a''(x - h_2 z) - a''(x)|\, dz$$

On account of (A4), a'' is uniformly continuous on S, so for each $z \in [-w, w]$,

$$\lim_{n\to\infty} \sup_{x\in S} |a''(x - h_2 z) - a''(x)| = 0.$$

From the dominated convergence theorem and the boundedness of K_0 and a'' we obtain (6.16). For the proof of (6.17) we apply Bernstein's inequality (Lemma 2.6.3) to

$$T_i = Y_i K_0''\{(x - X_i)/h_2\} - E[Y_i K_0''\{(x - X_i)/h_2\}]$$

for $1 \le i \le n$, with $c = 2 \sup |K_0''| B$, where B is an upper bound to the Y_i's (which exists by assumption (A6)) and $t = \epsilon n h_2^3$ for arbitrary $\epsilon > 0$. Also,

$$\begin{aligned} \mathrm{Var}(T_1) \le & h_2 \int_{|z|\le w} s(x - h_2 z) f(x - h_2 z) K_0''(z)^2\, dz \\ \le & C_1 h_2 \end{aligned}$$

since sf and K_0'' are bounded. Hence, for $h_2 \le 1$,

$$\begin{aligned} \tfrac{1}{2} t^2 \{n\mathrm{Var}(T_1) + ct\}^{-1} \ge & C_2(\epsilon)(nh_2^3)^2 \{C_3(\epsilon) n h_2\}^{-1} \\ = & C_4(\epsilon) n h_2^5. \end{aligned}$$

For each $\epsilon > 0$, observe that

$$\begin{aligned} & \int_S |a_n''(x|h_2) - Ea_n''(x|h_2)|\, dx \\ & \quad \le \epsilon \mathcal{L}(S) + \int_S |a_n''(x|h_2) - Ea_n''(x|h_2)| I\{|a_n''(x|h_2) - Ea_n''(x|h_2)| > \epsilon\}\, dx \\ & \quad \le \epsilon \mathcal{L}(S) + 2h_2^{-3} B \sup |K_0''| U \end{aligned}$$

where $U = \int_S I\left(|\sum_{i=1}^n T_i| > t\right)$. By Bernstein's inequality,

$$\begin{aligned} E(U) = &\int_S P\left(\left|\sum_{i=1}^n T_i\right| > t\right) \\ \leq &2\mathcal{L}(S)\exp\{-C_4(\epsilon)nh_2^5\} \\ = &O(n^{-\lambda}) \end{aligned}$$

for all $\lambda > 0$ since $nh_2^5/\log n \to \infty$. Markov's inequality yields

$$\sum_{i=1}^n P(|h_2^{-3}U| > \xi) < \infty$$

for all $\xi > 0$. Therefore we have, via the Borel-Cantelli Lemma, $h_1^{-3}U \to 0$ almost surely. Consequently,

$$\limsup_{n\to\infty} \int_S |a_n''(\cdot|h_2) - Ea_n''(\cdot|h_2)| \leq \mathcal{L}(S)\epsilon$$

for all $\epsilon > 0$, which implies (6.17) and concludes the proof of (6.12).

We shall prove (6.13) by writing

$$\begin{aligned} a_n f_n'' f_n^{-2} - af''f^{-2} = &f^{-2}(a_n f_n'' - af'')[1 + 2(f - f_n)f_n^{-1} + \{(f - f_n)f_n^{-1}\}^2] \\ &+ rf''f^{-2}(f - f_n)[2 + 3(f - f_n)f_n^{-1} + \{(f - f_n)f_n^{-1}\}^2] \end{aligned}$$

which leads to

$$\begin{aligned} \int_S |a_n f_n'' f_n^{-2} - af''f^{-2}| \leq &(\sup_S f^{-1})^2 \int_S |a_n f_n'' - af''| \\ &\times [1 + 2\sup_S |(f_n - f)f_n^{-1}| + \{\sup_S |(f_n - f)f_n^{-1}|\}^2] \\ &+ \sup_S |rf''f^{-2}| \int_S |f_n - f| \\ &\times [2 + 3\sup_S |(f_n - f)f_n^{-1}| + \{\sup_S |(f_n - f)f_n^{-1}|\}^2]. \end{aligned}$$

In view of (6.14), (6.15) and the boundedness of r, f'' and f^{-1} it is clear that (6.13) will follow from

$$\lim_{n\to\infty} \int_S |a_n f_n'' - af''| = 0 \tag{6.18}$$

almost surely. The integral in this expression is dominated by

$$\int_S |a_n - a||f_n''| + \sup_S |a| \int_S |f_n'' - f|. \tag{6.19}$$

The second term converges to zero almost surely by Theorem 2.4.2. For the first term we use the Cauchy-Schwarz inequality, followed by the inequality $(u+v)^2 \leq 2(u^2+v^2)$, to obtain

$$\begin{aligned}\int_S |a_n - a||f_n''| &\leq \left\{\int_S (a_n-a)^2 \int_S (f_n'')^2\right\}^{\frac{1}{2}} \\ &\leq \left[2\int_S (a_n-a)^2 \left\{\int_S (f_n''-Ef_n'')^2 + \int_S (Ef_n'')^2\right\}\right]^{\frac{1}{2}}. \quad (6.20)\end{aligned}$$

Techniques used to deal with $\int_S |a_n''-a''|$ in the first part of this proof can be readily adapted to show that $\lim_{n\to\infty}\int_S (a_n-a)^2 = 0$ almost surely and $\lim_{n\to\infty}\int_S (f_n''-Ef_n'')^2 = 0$ almost surely. Also, it is easy established that

$$\int_S (Ef_n'')^2 \leq 4w^2\mathcal{L}(S)(\sup|K_0|)^2(\sup|f''|)^2 < \infty.$$

These results, combined with the estimates at (6.19) and (6.20), imply (6.18) as required. ∎

Proof of Theorem 2.4.

We commence with

$$\begin{aligned}\kappa_2^{-1}\int_S |\hat{\sigma}_n - \sigma| &= \int_S |\{(s_n - r_n^2)f_n^{-1}\}^{\frac{1}{2}} - \{(s-r^2)f^{-1}\}^{\frac{1}{2}}| \\ &\leq \int_S |(s_n-r_n^2)f_n^{-1} - (s-r^2)f^{-1}|^{\frac{1}{2}} \\ &\leq \int_S \{|s_nf_n^{-1} - sf^{-1}| + |r_n^2f_n^{-1} - r^2f^{-1}|\}^{\frac{1}{2}} \\ &\leq \left\{\int_S |s_nf_n^{-1} - sf^{-1}| + \int_S |r_n^2f_n^{-1} - r^2f^{-1}|\right\}^{\frac{1}{2}} \mathcal{L}(S)^{\frac{1}{2}},\end{aligned}$$

with the last step coming from the Cauchy-Schwarz inequality. The required result is a direct consequence of

$$\lim_{n\to\infty}\int_S |s_nf_n^{-1} - sf^{-1}| = 0$$

almost surely, and

$$\lim_{n\to\infty}\int_S |r_n^2f_n^{-1} - r^2f^{-1}| = 0$$

almost surely. The proof of these can be accomplished using exactly the same arguments employed in the proof of Theorem 2.3. ∎

Proof of Theorem 3.1.

For each $x \in [0,1]$ let

$$B_n(x) \equiv Em_n(x|h) - m(x) \quad \text{and} \quad S_n(x) \equiv [\text{Var}\{m_n(x|h)\}]^{\frac{1}{2}}$$

denote the bias at x and variance at x respectively. We also let b stand for the function $\frac{1}{2}\kappa_1 m''$. Result (3.6) is therefore equivalent to

$$\left| E\{J_n(h)\} - (nh)^{-\frac{1}{2}}\kappa_2\sigma \int_0^1 \psi\left(\frac{(nh^5)^{\frac{1}{2}}b}{\kappa_2\sigma}\right)\right| = o\{h^2 + (nh)^{-\frac{1}{2}}\} + O(n^{-1}).$$

The left-hand side of this expression is dominated by

$$\left|E\{J_n(h)\} - \int_0^1 S_n\psi(B_n/S_n)\right| + \int_0^1 \left|S_n\psi(B_n/S_n) - (nh)^{-\frac{1}{2}}\kappa_2\sigma\psi\left(\frac{(nh^5)^{\frac{1}{2}}b}{\kappa_2\sigma}\right)\right|. \tag{6.21}$$

We deal with the first term of (6.21) by applying Lemma 5.8 of Devroye and Györfi (1985, p.90) to the random variables

$$Z_i = h^{-1}Y_iK^\dagger\{(x - x_i)/h\} - E[h^{-1}Y_iK\{(x - x_i)/h\}]$$

for $1 \le i \le n$, where $K^\dagger$ is either K or K_α, depending of the value of x. We obtain

$$\left|E|m_n(x|h) - m(x)| - S_n(x)\psi\left(\frac{B_n(x)}{S_n(x)}\right)\right| \le \frac{cE|Z_1|^3}{nE(Z_1^2)}$$

uniformly in $x \in [0,1]$, where c is a universal constant. Since $K^\dagger$ is bounded and the Y_i's are bounded by a constant B, say, we have

$$E|Z_1|^3 \le 2h^{-1}\sup|K^\dagger|B\,E(Z_1^2).$$

Therefore, with integration over $[0,1]$,

$$|E\{J_n(h)\} - \int_0^1 S_n\psi(B_n/S_n)| = O\{(nh)^{-1}\} = o\{(nh)^{-\frac{1}{2}}\}$$

since $nh \to \infty$ as $n \to \infty$.

Lemma 2.2.1 asserts that the second term on the right-hand side of (6.21) is no more than

$$\int_0^1 ||B_n| - h^2|b|| + (2/\pi)^{\frac{1}{2}}\int_0^1 |S_n - (nh)^{-\frac{1}{2}}\kappa_2\sigma|$$

so it is sufficient to show that

$$\int_0^1 |B_n - h^2 b| = o(h^2) + O(n^{-1}) \tag{6.22}$$

and

$$\int_0^1 |S_n - (nh)^{-\frac{1}{2}} \kappa_2 \sigma| = o\{(nh)^{-\frac{1}{2}}\} + O(n^{-1}). \tag{6.23}$$

For the proof of (6.22) we first consider $x = \alpha h$ where $0 \leq \alpha < 1$. Then it follows from (3.4) that for all small h,

$$\begin{aligned} E\{m_n(x|h)\} &= \int_{-1}^{\alpha} K_\alpha(z) m(x - hz)\, dz + O(n^{-1}) \\ &= m(x) + \tfrac{1}{2} h^2 \int_{-1}^{\alpha} z^2 K_\alpha(z)\, dz m''(0) + o(h^2) + O(n^{-1}) \end{aligned}$$

uniformly in α, the second step following from the assumption that K_α is a second order kernel. We then have from (K2),

$$\begin{aligned} \int_0^h |B_n| &\leq \tfrac{1}{2} h^3 \sup_{\alpha \in [0,1)} \int_{-1}^{\alpha} z^2 K_\alpha(z)\, dz |m''(0)| + o(h^3) + O(n^{-1}h) \\ &= o(h^2). \end{aligned}$$

Similarly, we obtain

$$\int_{1-h}^1 |B_n| = o(h^2).$$

For $x \in [h, 1-h]$ and small h, we have from (3.4),

$$\begin{aligned} E\{m_n(x|h)\} &= \int_{-1}^{1} K(z) m(x - hz)\, dz + O(n^{-1}) \\ &= m(x) + h^2 b(x) + o(h^2) + O(n^{-1}) \end{aligned}$$

so that

$$|B_n(x) - h^2 b(x)| = o(h^2) + O(n^{-1})$$

uniformly in $x \in [h, 1-h]$. The left-hand side of (6.22) is dominated by

$$\begin{aligned} &(1 - 2h) \sup_{x \in [h, 1-h]} |B_n(x) - h^2 b(x)| + \int_0^h |B_n| \\ &\quad + \int_{1-h}^h |B_n| + h^3 \sup_{x \in [0,h)} |b(x)| + h^3 \sup_{x \in (1-h,1]} |b(x)| = o(h^2) + O(n^{-1}) \end{aligned}$$

so (6.22) obtains. Result (6.23) can be derived in the same way from the approximation given at (3.5). ■

Chapter Five

NONPARAMETRIC DISCRIMINATION OF CATEGORICAL DATA USING DENSITY DIFFERENCES

5.1 Introduction

Consider the problem of discriminating between two populations Π_X and Π_Y having discrete probability functions f and g respectively. Let p be the prior probability that an unclassified observation z is from Π_X and let $\mathcal{Z}$ denote the set of all possible values of z. A discrimination rule for this problem is a partition $\{\mathcal{Z}_X, \mathcal{Z}_Y\}$ of $\mathcal{Z}$ for which z is assigned to Π_X if $z \in \mathcal{Z}_X$, and is assigned to Π_Y if $z \in \mathcal{Z}_Y$. The probability of misclassification, or error rate, of this discrimination rule is

$$\begin{aligned} \mathrm{ER}(\mathcal{Z}_X, \mathcal{Z}_Y) =& p \sum_{z \in \mathcal{Z}_Y} f(z) + (1-p) \sum_{z \in \mathcal{Z}_X} g(z) \\ =& p - \sum_{z \in \mathcal{Z}_X} e(z) \end{aligned}$$

where $e \equiv pf - (1-p)g$. The function e is referred to as the *density difference*. The ideal discrimination rule is therefore $\{\mathcal{Z}_X^*, \mathcal{Z}_Y^*\}$ where $\mathcal{Z}_X^* = \{z : e(z) \geq 0\}$ and $\mathcal{Z}_Y^* = \{z : e(z) < 0\}$. The function e is of fundamental importance to the discrimination problem. When e is unknown, the usual nonparametric approach is to obtain estimates $\hat{f}(\cdot|h_X)$ and $\hat{g}(\cdot|h_Y)$ of f and g respectively, based on training samples from each population. Here h_X and h_Y are smoothing parameters with h_X chosen to minimise the distance between $\hat{f}(\cdot|h_X)$ and f and h_Y chosen to minimise the distance between $\hat{g}(\cdot|h_Y)$ and g. The resulting discrimination rule is that which classifies z as coming from Π_X if and only if

$$\hat{e}(z|h_X, h_Y) \equiv p\hat{f}(z|h_X) - (1-p)\hat{g}(z|h_Y) \geq 0. \tag{1.1}$$

Condition (1.1) is, of course, equivalent to the more familiar likelihood ratio criterion

$$\hat{f}(z|h_X)/\hat{g}(z|h_Y) \geq (1-p)/p.$$

In this chapter we propose the following alternative approach to choosing h_X and h_Y. Select the smoothing parameter pair (h_X, h_Y) to minimise the distance

between $\hat{e}(\cdot|h_X, h_Y)$ and e. Notice that, under this scheme, attention is focussed on the estimation of a discrimination rule, rather than on probability functions, and data in training samples from both populations are used in the selection of the smoothing parameters. We measure distance between $\hat{e}(\cdot|h_X, h_Y)$ and e with the L_2 metric, giving rise to variants of least-squares cross-validation for the selection of (h_X, h_Y). Justification of these selection rules is provided by asymptotic results similar to those discussed by Bowman, Hall and Titterington (1984).

Section 2 deals with density difference estimation for binary data, and Section 3 discusses the same approach to unstructured multinomial data. Examples of each are given in Section 4. A proof of an asymptotic optimality result in the case of binary data is presented in Section 4.

5.2 Nonparametric Discrimination of Binary Data

We begin our analysis of the estimation of density differences for discrimination by considering the nonparametric classification of binary data. The sample space for such data is $\mathcal{B} = \{0, 1\}^d$ where $d \geq 1$. A data vector from this space would often represent the answers to a set of d questions (0="no", 1="yes"). According to the notation introduced in Section 1, for each $z \in \mathcal{B}$, $f(z)$ is the probability that an observation z from population Π_X assumes the value of z; $g(z)$ is the corresponding probability for population Π_Y. The functions f and g are assumed to be non-identical. For a d-vector $a = (a_1, \ldots, a_d)$ we define $|a| = \sum_{i=1}^{d} a_i$. Based on training samples $\mathcal{X} = \{X_i, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ the kernel density estimates used in this setting are

$$f_m(z|h_X) = m^{-1} \sum_{i=1}^{m} h_X^{|z-X_i|}(1 - h_X)^{d-|z-X_i|}$$

and

$$g_n(z|h_Y) = n^{-1} \sum_{i=1}^{n} h_Y^{|z-Y_i|}(1 - h_Y)^{d-|z-Y_i|}$$

(see Aitchison and Aitken (1976)). The smoothing parameters for these estimators are the window-sizes h_X and h_Y. The classical relative frequency estimators are achieved when the window-sizes are equal to zero. Our estimator for the density

difference $e = pf - (1-p)g$ is

$$e_{mn}(z|h_X, h_Y) = pf_m(z|h_X) - (1-p)g_n(z|h_Y).$$

The quality of the discrimination rule based on $e_{mn}(\cdot|h_X, h_Y)$ can be assessed in terms of its distance from the function e according to some global loss criterion. The loss with which we shall work is summed mean squared error given by

$$Q_{mn}(h_X, h_Y) = \sum_{z \in \mathcal{B}} E\{e_{mn}(z|h_X, h_Y) - e(z)\}^2.$$

Alternative distance measures are those based on Kullback-Leibler distance and absolute error, however they will not be considered here since the asymptotic theory for each of these criteria is not so well understood as it is for squared error.

It is of interest to derive the optimal choice of (h_X, h_Y) when aiming to asymptotically minimise $Q_{mn}(h_X, h_Y)$. Define, for $1 \leq j \leq d$ and $z \in \mathcal{B}$,

$$f_j(z) = \sum_{x:|x-z|=j} f(x)$$

and

$$g_j(z) = \sum_{x:|x-z|=j} g(x).$$

Standard asymptotic theory leads to the following asymptotic expressions for bias and variance:

$$Ef_m(z|h_X) - f(z) = \{f_1(z) - df(z)\}h_X + O(h_X^2),$$

$$Eg_n(z|h_Y) - g(z) = \{g_1(z) - dg(z)\}h_Y + O(h_Y^2),$$

$$\begin{aligned}\mathrm{Var}\{f_m(z|h_X)\} =& m^{-1}f(z)\{1-f(z)\} - 2h_X m^{-1}[df(z)\{1-f(z)\} + f(z)f_1(z)] \\ &+ O(m^{-1}h_X)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\{g_n(z|h_Y)\} =& n^{-1}g(z)\{1-g(z)\} - 2h_Y n^{-1}[dg(z)\{1-g(z)\} + g(z)g_1(z)] \\ &+ O(n^{-1}h_Y).\end{aligned}$$

Observing that

$$\begin{aligned}E\{e_{mn}(z|h_X, h_Y) - e(z)\}^2 =& p^2\mathrm{Var}\{f_m(z|h_X)\} + (1-p)^2\mathrm{Var}\{g_n(z|h_Y)\} \\ &+ [p\{Ef_m(z|h_X) - f(z)\} - (1-p)\{Eg_n(z|h_Y) - g(z)\}]^2,\end{aligned}$$

it is clear that the summed mean squared error of $e_{mn}(\cdot|h_X, h_Y)$ can be expressed as

$$\begin{aligned} Q_{mn}(h_X, h_Y) = {} & p^2 m^{-1} \sum_{z \in \mathcal{B}} ((f(z)\{1 - f(z)\} - 2h_X[df(z)\{1 - f(z)\} + f(z)f_1(z)]) \\ & + (1-p)^2 n^{-1} \sum_{z \in \mathcal{B}} (g(z)\{1 - g(z)\} - 2h_Y[dg(z)\{1 - g(z)\} + g(z)g_1(z)]) \\ & + \sum_{z \in \mathcal{B}} [ph_X\{f_1(z) - df(z)\} - (1-p)h_Y\{g_1(z) - dg(z)\}]^2 \\ & + O(m^{-1}h_X^2 + n^{-1}h_Y^2 + h_X^3 + h_Y^3). \end{aligned}$$

Ignoring the remainder term and then minimising $Q_{mn}(h_X, h_Y)$ with respect to (h_X, h_Y) we find that the optimal window-sizes $h_{X,\mathrm{opt}}$ and $h_{Y,\mathrm{opt}}$ satisfy

$$h_{X,\mathrm{opt}} \sim \tilde{h}_{X,\mathrm{opt}} \equiv (T_{XX}T_{YY} - T_{XY}^2)^{-1}(T_{YY}S_X m^{-1} + \rho T_{XY}S_Y n^{-1}) \qquad (2.1)$$

and

$$h_{Y,\mathrm{opt}} \sim \tilde{h}_{Y,\mathrm{opt}} \equiv (T_{XX}T_{YY} - T_{XY}^2)^{-1}(T_{XX}S_Y n^{-1} + \rho^{-1} T_{XY}S_X m^{-1}) \qquad (2.2)$$

provided $T_{XX}T_{YY} - T_{XY}^2 \neq 0$, where $\rho = (1-p)/p$;

$$T_{XX} = \sum (f_1 - df)^2, \qquad T_{YY} = \sum (g_1 - dg)^2,$$

$$T_{XY} = \sum (f_1 - df)(g_1 - dg),$$

$$S_X = d + \sum (f_1 - df)f, \qquad S_Y = d + \sum (g_1 - dg)g.$$

It is interesting to note that it is possible for either $\tilde{h}_{X,\mathrm{opt}}$ or $\tilde{h}_{Y,\mathrm{opt}}$ to assume a negative value. If, for example, we take $d = 2$, $f(0,0) = g(1,0) = 0.1$, $f(0,1) = g(1,1) = 0.2$, $f(1,0) = g(0,0) = 0.3$ and $f(1,1) = g(0,1) = 0.4$ then

$$\tilde{h}_{X,\mathrm{opt}} = (45/32)(5m^{-1} - 3\rho n^{-1}) \qquad \text{and} \qquad \tilde{h}_{Y,\mathrm{opt}} = (45/32)(5n^{-1} - 3\rho^{-1}m^{-1}).$$

If $5\rho n < 3m$ then $\tilde{h}_{X,\mathrm{opt}} < 0$ while $5\rho m < 3n$ implies that $\tilde{h}_{Y,\mathrm{opt}} < 0$. Such possibilities do not exist when smoothing parameters are chosen separately for each population, however it must be remembered that (h_X, h_Y) is being chosen for the estimation of the *difference* between two densities and not the individual densities themselves.

When choosing (h_X, h_Y) in practice the "optimal" formulae given at (2.1) and (2.2) cannot be applied since they depend on the unknown probabilities f and g. Note however that minimisation of $Q_{mn}(h_X, h_Y)$ is equivalent to minimisation of

$$\begin{aligned} S_{mn}(h_X, h_Y) =& Q_{mn}(h_X, h_Y) - \sum_{z \in \mathcal{B}} e(z)^2 \\ =& \sum_{z \in \mathcal{B}} E\{e_{mn}(z|h_X, h_Y)^2\} - 2 \sum_{z \in \mathcal{B}} E\{e_{mn}(z|h_X, h_Y)\} e(z). \end{aligned}$$

An unbiased estimator of the right-hand side is

$$\begin{aligned} \hat{S}_{mn}(h_X, h_Y) = & \sum_{z \in \mathcal{B}} e_{mn}(z|h_X, h_Y)^2 \\ & - 2 \Bigg[p^2 m^{-1} \sum_{i=1}^{m} f_{m,i}(X_i|h_X) + (1-p)^2 n^{-1} \sum_{i=1}^{n} g_{n,i}(Y_i|h_Y) \\ & - p(1-p) \left\{ m^{-1} \sum_{i=1}^{m} g_n(X_i|h_Y) + n^{-1} \sum_{i=1}^{n} f_n(Y_i|h_X) \right\} \Bigg] \end{aligned}$$

where

$$f_{m,i}(z|h_X) = (m-1)^{-1} \sum_{j \neq i} h_X^{|z - X_j|} (1 - h_X^{d - |z - X_j|})$$

and

$$g_{n,i}(z|h_Y) = (n-1)^{-1} \sum_{j \neq i} h_Y^{|z - Y_j|} (1 - h_Y^{d - |z - Y_j|})$$

are the "leaving-one-out" estimators for f and g respectively. The selection of (h_X, h_Y) can be accomplished by choosing $(\hat{h}_X, \hat{h}_Y)$, the window-size pair at which $\hat{S}_{mn}(h_X, h_Y)$ is minimised. This window-size selection rule is a version of least-squares cross-validation, first proposed in different settings by Rudemo (1982) and Bowman (1984). Brown and Rundell (1985) have also proposed a version of this technique for the classification of categorical data.

A desirable property of any window-size selection rule is that it be asymptotically optimal in some sense. This means that the window-sizes chosen by the rule should be asymptotically equivalent to the optimal window-sizes with respect to a particular measure of loss. The following result ensures that this is indeed the case for the rule proposed in the previous paragraph.

Theorem 2.1. *Let $(\hat{h}_X, \hat{h}_Y)$ and $(h_{X,\mathrm{opt}}, h_{Y,\mathrm{opt}})$ denote the window-size pairs which minimise $\hat{S}_{mn}(h_X, h_Y)$ and $Q_{mn}(h_X, h_Y)$ respectively. If, for some positive constant*

ξ, $m/n \to \xi$ *as* $m, n \to \infty$ *then*

$$\hat{h}_X / h_{X,\text{opt}} \to 1 \qquad \text{and} \qquad \hat{h}_Y / h_{Y,\text{opt}} \to 1 \tag{2.3}$$

in probability as $m, n \to \infty$.

Our proposed prescription for discrimination is now complete. Assign an incoming observation Z to population Π_X if $e_{mn}(Z|\hat{h}_X, \hat{h}_Y) \geq 0$ and to population Π_Y if $e_{mn}(Z|\hat{h}_X, \hat{h}_Y) < 0$.

A simpler alternative to the above proposal is to estimate the density difference using only one window-size h and take $\hat{h}$, the minimiser of $S_{mn}(h, h)$, as our data-based choice. This procedure will also produce asymptotically optimal window-sizes in the sense that $\hat{h}/h_{\text{opt}} \to 1$ in probability as $m, n \to \infty$ and $m/n \to \xi > 0$ where h_{opt} is the window-size that minimises $Q_{mn}(h, h)$ and satisfies

$$h_{\text{opt}} \sim \tilde{h}_{\text{opt}} \equiv [m^{-1}\{d + \sum f(f_1 - df)\} + n^{-1}\rho^2\{d + \sum g(g_1 - dg)\}]$$
$$\times [\sum \{(f_1 - df) - \rho(g_1 - dg)\}^2]^{-1}.$$

In this case we classify the new observation Z as coming from population Π_X or Π_Y according as

$$e_{mn}(Z|\hat{h}, \hat{h}) \geq 0 \text{ or } < 0.$$

5.3 Nonparametric Discrimination of Unstructured Multinomial Data

In this section the data are assumed to come from a set of c unordered cells which we label $1, \ldots, c$. Such data will be called unstructured c-nomial data. For example, a data set which records the eye colour of a group of individuals may be interpreted as a sample from the four cells (1) blue, (2) brown, (3) grey and (4) green. If we do not take account of any natural ordering of these cells then the data can be taken as unstructured 4-nomial. Given that we again have two populations Π_X and Π_Y and two training samples $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ of unstructured c-nomial data we shall define $f(i)$ and $g(i)$ to be the probability attached to cell i for populations Π_X and Π_Y respectively. Also, we let $\mathcal{M}_i$ and $\mathcal{N}_i$ denote the numbers of observations in cell i from $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Appropriate density estimates for this model are

$$f_m(i|h_X) = m^{-1}\{\mathcal{M}_i(1-h_X) + (c-1)^{-1}(m-\mathcal{M}_i)h_X\}$$

and

$$g_n(i|h_Y) = n^{-1}\{\mathcal{N}_i(1-h_Y) + (c-1)^{-1}(n-\mathcal{N}_i)h_Y\}$$

for $1 \le i \le c$ (see Aitchison and Aitken (1976)).

Our goal is to minimise the mean summed squared error of the density difference estimator $e_{mn}(\cdot|h_X, h_Y) = pf_m(\cdot|h_X) - (1-p)g_n(\cdot|h_Y)$ given by

$$Q_{mn}(h_X, h_Y) = \sum_{i=1}^{c} E\{e_{mn}(i|h_X, h_Y) - e(i)\}^2,$$

where $e = pf - (1-p)g$. Straightforward calculations lead to

$$\begin{aligned} Q_{mn}(h_X, h_Y) =& p^2 m^{-1}\{1 - 2c(c-1)^{-1}h_X\}\sum f(1-f) \\ &+ (1-p)^2 n^{-1}\{1 - 2c(c-1)^{-1}h_X\}\sum g(1-g) \\ &+ (c-1)^{-2}\sum\{p(1-cf)h_X - (1-p)(1-cg)h_Y\}^2 \\ &+ O(m^{-1}h_X^2 + n^{-1}h_Y^2). \end{aligned}$$

From this it may be shown that the optimal window-sizes $h_{X,\mathrm{opt}}$ and $h_{Y,\mathrm{opt}}$ are such that

$$h_{X,\mathrm{opt}} \sim \tilde{h}_{X,\mathrm{opt}} \equiv (T_{XX}T_{YY} - T_{XY}^2)^{-1}(T_{YY}S_X m^{-1} + \rho T_{XY}S_Y n^{-1}),$$

$$h_{Y,\mathrm{opt}} \sim \tilde{h}_{Y,\mathrm{opt}} \equiv (T_{XX}T_{YY} - T_{XY}^2)^{-1}(T_{XX}S_Y n^{-1} + \rho^{-1}T_{XY}S_X m^{-1}),$$

where $\rho = (1-p)/p$ and

$$T_{XX} = (c-1)^{-2}\sum(1-cf)^2, \qquad T_{YY} = (c-1)^{-2}\sum(1-cg)^2,$$

$$T_{XY} = (c-1)^{-2}\sum(1-cf)(1-cg),$$

$$S_X = c(c-1)^{-1}\sum f(1-f), \qquad S_Y = c(c-1)^{-1}\sum g(1-g).$$

It is easily verified that an unbiased estimator of $S_{mn}(h_X, h_Y) \equiv Q_{mn}(h_X, h_Y) - \sum e^2$ is

$$\begin{aligned}\hat{S}_{mn}(h_X, h_Y) &= \sum e_{mn}(\cdot|h_X, h_Y)^2 \\ &\quad - 2\Bigg(\frac{p^2}{m(m-1)} \sum_{i=1}^{c} \left[\mathcal{M}_i(\mathcal{M}_i - 1) \left\{ 1 - \frac{ch_X}{c-1} \right\} \right] + p^2(c-1)^{-1}h_X \\ &\quad + \frac{(1-p)^2}{n(n-1)} \sum_{i=1}^{c} \left[\mathcal{N}_i(\mathcal{N}_i - 1) \left\{ 1 - \frac{ch_Y}{c-1} \right\} \right] + (1-p)^2(c-1)^{-1}h_Y \\ &\quad - p(1-p) \left\{ m^{-1} \sum_{i=1}^{c} \mathcal{M}_i g_n(i|h_Y) + n^{-1} \sum_{i=1}^{c} \mathcal{N}_i f_m(i|h_X) \right\} \Bigg)\end{aligned}$$

so a practical procedure for choosing (h_X, h_Y) is to set it equal to $(\hat{h}_X, \hat{h}_Y)$, the window-size pair at which $\hat{S}_{mn}(h_X, h_Y)$ is minimised. This minimisation problem has the exact solution

$$\hat{h}_X = (\hat{T}_{XX}\hat{T}_{YY} - \hat{T}_{XY}^2)^{-1}\{\hat{T}_{YY}\hat{S}_X(m-1)^{-1} + \rho\hat{T}_{XY}\hat{S}_Y(n-1)^{-1}\},$$

$$\hat{h}_Y = (\hat{T}_{XX}\hat{T}_{YY} - \hat{T}_{XY}^2)^{-1}\{\hat{T}_{XX}\hat{S}_Y(n-1)^{-1} + \rho^{-1}\hat{T}_{XY}\hat{S}_X(m-1)^{-1}\}$$

where

$$\hat{T}_{XX} = (c-1)^{-2} \sum_{i=1}^{c} (1 - m^{-1}\mathcal{M}_i c)^2, \qquad \hat{T}_{YY} = (c-1)^{-2} \sum_{i=1}^{c} (1 - n^{-1}\mathcal{N}_i c)^2,$$

$$\hat{T}_{XY} = (c-1)^{-2} \sum_{i=1}^{c} (1 - m^{-1}\mathcal{M}_i c)(1 - n^{-1}\mathcal{N}_i c),$$

$$\hat{S}_X = c(c-1)^{-1}m^{-1} \sum_{i=1}^{c} \mathcal{M}_i(1 - m^{-1}\mathcal{M}_i), \quad \hat{S}_Y = c(c-1)^{-1}n^{-1} \sum_{i=1}^{c} \mathcal{N}_i(1 - n^{-1}\mathcal{N}_i).$$

By the weak law of large numbers, $\mathcal{M}_i/m \to f(i)$ in probability, and $\mathcal{N}_i/n \to g(i)$ in probability, for each $1 \le i \le c$. Therefore, as each sample increases in size we have $\hat{T}_{XX} \to T_{XX}$, $\hat{T}_{YY} \to T_{YY}$, $\hat{T}_{XY} \to T_{XY}$, $\hat{S}_X \to S_X$ and $\hat{S}_Y \to S_Y$, each convergence being in probability. This immediately entails

$$\hat{h}_X / h_{X,\mathrm{opt}} \to 1 \qquad \text{and} \qquad \hat{h}_Y / h_{Y,\mathrm{opt}} \to 1$$

in probability as $m, n \to \infty$, provided $m/n \to \xi$ for some constant $\xi > 0$. Thus, $(\hat{h}_X, \hat{h}_Y)$ is an asymptotically optimal window-size selection rule in the context of minimising mean summed squared error.

The theory for the estimator $e_{mn}(\cdot|h)$ – that is, the estimator for e with $h_X = h_Y = h$ – can be developed in the same way.

5.4 Example

Data from Anderson *et al* (1972) consist of two training samples of ten-dimensional binary data. The first training sample represents the presence or absence of ten ocular symptoms of forty patients known to have the disease *keratoconjunctivitis sicca* (KCS). This constitutes the $\mathcal{X}$ sample with $m = 40$. The $\mathcal{Y}$ sample is the corresponding symptom data for thirty-seven non-KCS ($n = 37$). We assume that $p = \frac{1}{2}$. A rule for classifying an incoming patient as a KCS victim based on his/her symptom vector z is to observe the sign of

$$e_{40,37}(z|\hat{h}_X, \hat{h}_Y) = \tfrac{1}{2} f_{40}(z|\hat{h}_X) - \tfrac{1}{2} g_{37}(z|\hat{h}_Y),$$

where $(\hat{h}_X, \hat{h}_Y)$ is the window-size pair which minimises $\hat{S}_{40,37}(h_X, h_Y)$. The discrimination rule based on "smoothing" the data has an advantage over that based on the classical cell proportion estimators which often lead to zero estimates of the probabilities $f(z)$ and $g(z)$, due to the sparsity of the data (there are 1024 points in the sample space in this example). Zero cell counts in the training samples will often lead to an indeterminant classification rule when smoothing is not applied. Table 4.1 lists the values of $(\hat{h}_X, \hat{h}_Y)$ obtained by minimisation of $\hat{S}_{40,37}(h_X, h_Y)$. We have also tabulated the window-sizes obtained if least-squares cross-validation (see Brown and Rundell (1985)) and likelihood cross-validation (see Aitchison and Aitken (1976)) are applied to each training sample separately. There is a considerable difference between the window-sizes for each method.

To give an indication of the efficacy of the classification rules we omitted each patient in turn from the training samples and used the reduced sample to select (h_X, h_Y) and estimate $e = \frac{1}{2} f - \frac{1}{2} g$. The omitted patient was then reclassified according to the resulting discrimination rule. Table 4.2 lists the number of misclassifications for each of the three methods mentioned above. It is seen that for this data set the density difference approach outperforms the analogous approach based on estimating the densities individually. These two rules are slightly

Table 4.1: Window-sizes selected by the KCS data.

Method	$\hat{h}_X$	$\hat{h}_Y$
(i)	0.2161	0.0124
(ii)	0.1950	0.0083
(iii)	0.1570	0.0400

Table 4.2: Number of misclassifications when patients are sequentially omitted from the KCS data.

Method	$\mathcal{X}$ misclass.	$\mathcal{Y}$ misclass.
(i)	4	2
(ii)	4	3
(iii)	4	1

Methods are (i) minimisation of $\hat{S}_{40,37}(h_X, h_Y)$, (ii) least squares cross validation applied to $\mathcal{X}$ and $\mathcal{Y}$ individually and (iii) likelihood cross-validation applied to $\mathcal{X}$ and $\mathcal{Y}$ individually.

bettered by the rule formed by using likelihood cross-validation to estimate the individual densities.

5.5 Proof of Theorem 2.1

Our initial aim is to find an asymptotic expansion for $\hat{S}_{mn}(h_X, h_Y)$, which may be written as

$$\begin{aligned}\hat{S}_{mn}(h_X, h_Y) =& p^2 \sum_{z \in \mathcal{B}} f_m(z|h_X)^2 + (1-p)^2 \sum_{z \in \mathcal{B}} g_n(z|h_Y)^2 \\ & - 2p(1-p) \sum_{z \in \mathcal{B}} f_m(z|h_X) g_n(z|h_Y) \\ & - 2p^2 m^{-1} \sum_{i=1}^{m} f_{m,i}(X_i|h_X) - 2(1-p)^2 n^{-1} \sum_{i=1}^{n} g_{n,i}(Y_i|h_Y) \\ & + 2p(1-p) n^{-1} \sum_{i=1}^{n} f_m(Y_i|h_X) + 2p(1-p) m^{-1} \sum_{i=1}^{m} g_n(X_i|h_Y). \end{aligned} \tag{5.1}$$

We shall treat each summation in this expression separately. For each $z \in \mathcal{B}$ and $0 \le j \le d$ define

$$\mathcal{M}_j(z) = \sum_{i=1}^{m} I(|z - X_i| = j)$$

and

$$\mathcal{N}_j(z) = \sum_{i=1}^{n} I(|z - Y_i| = j).$$

Note that $\mathcal{M}_j(z)$ represents the number of observations in the sample $\mathcal{X}$ which are exactly j units away from z, where distance is measured in terms of $|\cdot|$, and $\mathcal{N}_j(z)$ is the same quantity for $\mathcal{Y}$. The following asymptotic expansion exists for $f_m(z|h_X)$:

$$\begin{aligned} f_m(z|h_X) = & m^{-1}\mathcal{M}_0(z) + m^{-1}h_X\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\} \\ & + m^{-1}h_X^2 \left\{ \binom{d}{2} \mathcal{M}_0(z) - (d-1)\mathcal{M}_1(z) + \mathcal{M}_2(z) \right\} + o_p(m^{-1}h_X^2). \end{aligned} \tag{5.2}$$

Therefore

$$\begin{aligned} f_m(z|h_X)^2 = & m^{-2}\mathcal{M}_0^2(z) + 2m^{-2}h_X\mathcal{M}_0(z)\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\} \\ & + 2m^{-2}h_X^2\mathcal{M}_0(z) \left\{ \binom{d}{2} \mathcal{M}_0(z) - (d-1)\mathcal{M}_1(z) + \mathcal{M}_2(z) \right\} \\ & + m^{-2}h_X^2\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\}^2 + o_p(m^{-2}h_X^2). \end{aligned}$$

Similar expansions exist for $g_n(z|h_Y)$ and $g_n(z|h_Y)^2$. Multiplying the expansions for $f_m(z|h_X)$ and $g_n(z|h_Y)$ we obtain

$$\begin{aligned}
f_m(z|h_X)g_n(z|h_Y) = {} & (mn)^{-1}\mathcal{M}_0(z)\mathcal{N}_0(z) + (mn)^{-1}h_Y\mathcal{M}_0(z)\{\mathcal{N}_1(z) - d\mathcal{N}_0(z)\} \\
& + (mn)^{-1}h_Y^2\mathcal{M}_0(z)\left\{\binom{d}{2}\mathcal{N}_0(z) - (d-1)\mathcal{N}_1(z) + \mathcal{N}_2(z)\right\} \\
& + (mn)^{-1}h_X\mathcal{N}_0(z)\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\} \\
& + (mn)^{-1}\mathcal{N}_0(z)h_X^2\left\{\binom{d}{2}\mathcal{M}_0(z) - (d-1)\mathcal{M}_1(z) + \mathcal{M}_2(z)\right\} \\
& + (mn)^{-1}h_Xh_Y\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\}\{\mathcal{N}_1(z) - d\mathcal{N}_0(z)\} \\
& + o_p\{(mn)^{-1}h_X^2 + (mn)^{-1}h_Y^2 + (mn)^{-1}h_Xh_Y\}.
\end{aligned}$$

We also have

$$m^{-1}\sum_{i=1}^{m} f_{m,i}(X_i|h_X) = \{m(m-1)\}^{-1}\sum\sum\nolimits_{i\neq j}\Lambda(i,j) + o_p(m^{-1}h_X^2)$$

where

$$\begin{aligned}
\Lambda(i,j) \equiv {} & I(X_i = X_j) + h_X\{I(|X_i - X_j| = 1) - dI(X_i = X_j)\} \\
& + h_X^2\left\{\binom{d}{2}I(X_i = X_j) - (d-1)I(|X_i - X_j| = 1) + I(|X_i - X_j| = 2)\right\}.
\end{aligned}$$

Observing that

$$\begin{aligned}
\sum\sum\nolimits_{i\neq j} I(X_i = X_j) &= \sum_{z\in\mathcal{B}}\mathcal{M}_0^2(z) - m, \\
\sum\sum\nolimits_{i\neq j} I(|X_i - X_j| = 1) &= \sum_{z\in\mathcal{B}}\mathcal{M}_0(z)\mathcal{M}_1(z), \\
\sum\sum\nolimits_{i\neq j} I(|X_i - X_j| = 2) &= \sum_{z\in\mathcal{B}}\mathcal{M}_0(z)\mathcal{M}_2(z),
\end{aligned}$$

we derive

$$\begin{aligned}
m^{-1}\sum_{i=1}^{m} f_{m,i}(X_i|h_X) = {} & \{m(m-1)\}^{-1}\sum_{z\in\mathcal{B}}\mathcal{M}_0^2(z) - (m-1)^{-1} \\
& + \{m(m-1)\}^{-1}h_X\sum_{z\in\mathcal{B}}\mathcal{M}_0(z)\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\} + (m-1)^{-1}h_Xd \\
& + \{m(m-1)\}^{-1}h_X^2\sum_{z\in\mathcal{B}}\mathcal{M}_0(z)\left\{\binom{d}{2}\mathcal{M}_0(z) - (d-1)\mathcal{M}_1(z) + \mathcal{M}_2(z)\right\} \\
& - h_X^2\binom{d}{2}(m-1)^{-1} + o_p(m^{-1}h_X^2).
\end{aligned}$$

The asymptotic expansion for $n^{-1}\sum_{i=1}^{n} g_{n,i}(Y_i|h_Y)$ is found in the same way. On account of (5.2) we have

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n} f_m(Y_i|h_X) &= (mn)^{-1}\sum_{i=1}^{n}\mathcal{M}_0(Y_i) \\
&\quad + (mn)^{-1}h_X\sum_{i=1}^{n}\{\mathcal{M}_1(Y_i) - d\mathcal{M}_0(Y_i)\} \\
&\quad + (mn)^{-1}h_X^2\sum_{i=1}^{n}\left\{\binom{d}{2}\mathcal{M}_0(Y_i) - (d-1)\mathcal{M}_1(Y_i) + \mathcal{M}_2(Y_i)\right\} \\
&\quad + o_p(m^{-1}h_X^2) \\
&= (mn)^{-1}\sum_{z\in\mathcal{B}}\mathcal{M}_0(z)\mathcal{N}_0(z) + (mn)^{-1}\sum_{z\in\mathcal{B}}\mathcal{N}_0(z)\{\mathcal{M}_1(z) - d\mathcal{M}_0(z)\} \\
&\quad + (mn)^{-1}h_X^2\sum_{z\in\mathcal{B}}\mathcal{N}_0(z)\left\{\binom{d}{2}\mathcal{M}_0(z) - (d-1)\mathcal{M}_1(z) + \mathcal{M}_2(z)\right\} \\
&\quad + o_p(m^{-1}h_X^2).
\end{aligned}
$$

Analogous working applies to $m^{-1}\sum_{i=1}^{m} g_n(X_i|h_Y)$. Substitution of these expansions into (5.1) gives, after some algebra,

$$
\begin{aligned}
\hat{S}_{mn}(h_X,h_Y) &= \frac{-p^2(m+1)}{m^2(m-1)}\sum_{z\in\mathcal{B}}\mathcal{M}_0^2(z) - \frac{(1-p)^2(n+1)}{n^2(n-1)}\sum_{z\in\mathcal{B}}\mathcal{N}_0^2(z) \\
&\quad + \frac{2p^2}{m-1} + \frac{2(1-p)^2}{n-1} + 2p(1-p)(mn)^{-1}\sum_{z\in\mathcal{B}}\mathcal{M}_0(z)\mathcal{N}_0(z) \\
&\quad - \frac{2p^2h_X}{m-1}\left[d + \sum_{z\in\mathcal{B}}\frac{\mathcal{M}_0(z)}{m}\left\{\frac{\mathcal{M}_1(z)}{m} - \frac{d\mathcal{M}_0(z)}{m}\right\}\right] \\
&\quad - \frac{2(1-p)^2h_Y}{n-1}\left[d + \sum_{z\in\mathcal{B}}\frac{\mathcal{N}_0(z)}{n}\left\{\frac{\mathcal{N}_1(z)}{n} - \frac{d\mathcal{N}_0(z)}{n}\right\}\right] \\
&\quad + p^2h_X^2\sum_{z\in\mathcal{B}}\left\{\frac{\mathcal{M}_1(z)}{m} - \frac{d\mathcal{M}_0(z)}{m}\right\}^2 + (1-p)^2h_Y^2\sum_{z\in\mathcal{B}}\left\{\frac{\mathcal{N}_1(z)}{n} - \frac{d\mathcal{N}_0(z)}{n}\right\}^2 \\
&\quad - 2p(1-p)h_Xh_Y\sum_{z\in\mathcal{B}}\left\{\frac{\mathcal{M}_1(z)}{m} - \frac{d\mathcal{M}_0(z)}{m}\right\}\left\{\frac{\mathcal{N}_1(z)}{n} - \frac{d\mathcal{N}_0(z)}{n}\right\} \\
&\quad + o_p(m^{-1}h_X + n^{-1}h_Y + h_X^2 + h_Y^2 + h_Xh_Y).
\end{aligned}
$$

Ignoring the remainder term and setting $\partial/\partial h_X\hat{S}_{mn}(h_X,h_Y)$ and $\partial/\partial h_Y\hat{S}_{mn}(h_X,h_Y)$ to zero gives

$$
\hat{h}_X \sim \frac{\hat{T}_{YY}\hat{S}_X(m-1)^{-1} + \rho\hat{T}_{XY}\hat{S}_Y(n-1)^{-1}}{\hat{T}_{XX}\hat{T}_{YY} - \hat{T}_{XY}^2}, \tag{5.3}
$$

and

$$
\hat{h}_Y \sim \frac{\hat{T}_{XX}\hat{S}_Y(n-1)^{-1} + \rho^{-1}\hat{T}_{XY}\hat{S}_X(m-1)^{-1}}{\hat{T}_{XX}\hat{T}_{YY} - \hat{T}_{XY}^2}. \tag{5.4}
$$

where

$$\hat{T}_{XX} = \sum_{z \in \mathcal{B}} \left\{ \frac{\mathcal{M}_1(z)}{m} - \frac{d\mathcal{M}_0(z)}{m} \right\}^2, \qquad \hat{T}_{YY} = \sum_{z \in \mathcal{B}} \left\{ \frac{\mathcal{N}_1(z)}{n} - \frac{d\mathcal{N}_0(z)}{n} \right\}^2,$$

$$\hat{T}_{XY} = \sum_{z \in \mathcal{B}} \left\{ \frac{\mathcal{M}_1(z)}{m} - \frac{d\mathcal{M}_0(z)}{m} \right\} \left\{ \frac{\mathcal{N}_1(z)}{n} - \frac{d\mathcal{N}_0(z)}{n} \right\},$$

$$\hat{S}_X = d + \sum_{z \in \mathcal{B}} \frac{\mathcal{M}_0(z)}{m} \left\{ \frac{\mathcal{M}_1(z)}{m} - \frac{d\mathcal{M}_0(z)}{m} \right\},$$

$$\hat{S}_Y = d + \sum_{z \in \mathcal{B}} \frac{\mathcal{N}_0(z)}{n} \left\{ \frac{\mathcal{N}_1(z)}{n} - \frac{d\mathcal{N}_0(z)}{n} \right\}.$$

By the weak law of large numbers we have as $m, n \to \infty$, $\mathcal{M}_0(z)/m \to f(z)$, $\mathcal{N}_0(z) \to g(z)$, $\mathcal{M}_1(z)/m \to f_1(z)$ and $\mathcal{N}_1(z)/n \to g_1(z)$ where each convergence is in probability. As a consequence, $\hat{T}_{XX} \to T_{XX}$, $\hat{T}_{YY} \to T_{YY}$, $\hat{T}_{XY} \to T_{XY}$ $\hat{S}_X \to S_X$ and $\hat{S}_Y \to S_Y$, in probability as $m, n \to \infty$ and $m/n \to \xi > 0$. These results, when combined with the expressions at (2.1), (2.2), (5.3) and (5.4), imply (2.3) as required. ■

Chapter Six

NONPARAMETRIC DISCRIMINATION OF CONTINUOUS DATA USING DENSITY DIFFERENCES

6.1 Introduction

The density difference approach to discrimination introduced in the previous chapter is applied to the discrimination of continuous data in this chapter. Once again we see that selection of the smoothing parameter pair (h_X, h_Y) can effectively be performed via a version of least-squares cross-validation. Asymptotic optimality of this selection rule is provided by a result similar to that of Stone (1984).

The kernel-based development of density difference estimation for continuous data is the topic of Section 2, culminating in a completely automatic prescription for discrimination between continuous populations Π_X and Π_Y. In Section 3, the efficacy of our method is exemplified by some applications to simulated and real data sets. Section 4 contains the proof of asymptotic optimality of the smoothing parameter selection criterion.

6.2 Nonparametric Discrimination of Continuous Data

Given independent training samples $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ of continuous d-variate data, having distinct densities f and g respectively, the usual kernel density estimators have the form

$$f_m(z|h_X) = m^{-1} \sum_{i=1}^{m} K_{h_X}(x - X_i)$$

and

$$g_n(z|h_Y) = n^{-1} \sum_{i=1}^{n} K_{h_Y}(x - Y_i),$$

where $K_h(x) = v_h^{-1} K(x/h)$ and K is a symmetric d-variate kernel function which is assumed to integrate to unity. In general, we shall take the window-size $h = (h_1, \ldots, h_d)$ to be a vector in $\mathbf{R}_+^d$ and define $v_h = \prod_{i=1}^{d} h_i$ to be its volume. For any vector $x = (x_1, \ldots, x_d)$ we define x/h to be $(x_1/h_1, \ldots, x_d/h_d)$ and $|x| =$

$(x_1^2, \ldots, x_d^2)^{\frac{1}{2}}$. The estimator for the density difference $e = pf - (1-p)g$ is the function

$$e_{mn}(z|h_X, h_Y) = pf_m(z|h_X) - (1-p)g_n(z|h_Y),$$

which again requires the selection of the window-size pair (h_X, h_Y) for effective implementation. The appropriate squared-error loss criterion for continuous data is one based on integrated mean squared error, or L_2 loss, given by

$$M_{mn}(h_X, h_Y) = \int \{e_{mn}(\cdot|h_X, h_Y) - e\}^2.$$

The integrated mean squared error of $e_{mn}(\cdot|h_X, h_Y)$ is therefore $E\{M_{mn}(h_X, h_Y)\}$.

The optimal window-size pair associated with the minimisation of the asymptotic representation of $E\{M_{mn}(h_X, h_Y)\}$ is easily obtainable in the important special case where it is assumed that the components of the vectors h_X and h_Y are identical so that the window-size pair (h_X, h_Y) involves only the two scalar parameters h_X and h_Y. In particular, if K is a bounded, square-integrable probability density which is symmetric in each variable and is such that

$$\kappa_1 \equiv \int z_i^2 K(z)\, dz < \infty$$

is independent of $i \in \{1, \ldots, d\}$; if f and g are such that each of their second-order derivatives are bounded, continuous and square-integrable; and if $m/n \to \xi > 0$ as $m, n \to \infty$, then it may be established by standard asymptotic arguments that the optimal window-size pair $(h_{X,\mathrm{opt}}, h_{Y,\mathrm{opt}})$ satisfies

$$(h_{X,\mathrm{opt}}, h_{Y,\mathrm{opt}}) \sim (c_X m^{-1/(d+4)}, c_Y n^{-1/(d+4)}). \tag{2.1}$$

Here c_X and c_Y are the positive solutions of the equations

$$\kappa_1^2 \left\{ c_X^2 p^2 \int (\nabla^2 f)^2 - c_Y^2 p(1-p) \int (\nabla^2 f)(\nabla^2 g) \right\} = d c_X^{-(d+2)} \kappa_2^2 p^2 \tag{2.2}$$

and

$$\kappa_1^2 \left\{ c_Y^2 (1-p)^2 \int (\nabla^2 g)^2 - c_X^2 p(1-p) \int (\nabla^2 f)(\nabla^2 g) \right\} = d\xi c_Y^{-(d+2)} \kappa_2^2 (1-p)^2, \tag{2.3}$$

where $\kappa_2^2 = \int K^2$ and $\nabla^2 a = \sum_{i=1}^d (\partial^2/\partial x_i^2) a(x)$ for a d-variate function a. The optimal rate of convergence of $E\{M_{mn}(h_X, h_Y)\}$ to zero is $C^* n^{-4/(4+d)}$ where

$$C^* = \kappa_2^2 \{p^2 c_X^{-d} \xi^{-4/(d+4)} + (1-p)^2 c_Y^{-d}\} + (\kappa_1^2/4) \int \{p\xi^{-2/(d+4)} c_X^2 \nabla^2 f - (1-p)^2 c_Y^2 \nabla^2 g\}^2.$$

In practice the formulae at (2.1), (2.2) and (2.3) are difficult to use to select a window-size pair since estimators of certain functionals of the unknown densities, namely $\nabla^2 f$ and $\nabla^2 g$, are required. Fortunately, the ideas used to motivate the window-size selection rules proposed in Sections 5.2 and 5.3 are easily extended to the case of continuous data. The mean of $M_{mn}(h_X, h_Y) - \int e^2$ is estimated without bias by

$$\int e_{mn}(\cdot|h_X, h_Y)^2 - 2\{m(m-1)\}^{-1} p^2 \sum\sum_{i \neq j} K_{h_X}(X_i - X_j) - 2\{n(n-1)\}^{-1}(1-p)^2 \sum\sum_{i \neq j} K_{h_Y}(Y_i - Y_j) + 2p(1-p)n^{-1} \sum_{i=1}^n f_m(Y_i|h_X) + 2p(1-p)m^{-1} \sum_{i=1}^m g_n(X_i|h_Y).$$

A slight modification leads to the estimator

$$\hat{S}_{mn}(h_X, h_Y) \equiv \int e_{mn}(\cdot|h_X, h_Y)^2 - 2m^{-2} p^2 \sum\sum_{i \neq j} K_{h_X}(X_i - X_j) - 2n^{-2}(1-p)^2 \sum\sum_{i \neq j} K_{h_Y}(Y_i - Y_j) + 2p(1-p)n^{-1} \sum_{i=1}^n f_m(Y_i|h_X) + 2p(1-p)m^{-1} \sum_{i=1}^m g_n(X_i|h_Y),$$

so it is proposed that the window-size pair be selected from $\mathbf{R}_+^d \times \mathbf{R}_+^d$ to minimise $\hat{S}_{mn}(h_X, h_Y)$. This method is a variant of least-squares cross-validation. The window-size pair selected by this rule will be denoted by $(\hat{h}_X, \hat{h}_Y)$. We classify an incoming observation Z as coming from Π_X if $e_{mn}(Z|\hat{h}_X, \hat{h}_Y) \geq 0$ and from Π_Y otherwise. Once again we may show that this selection rule is asymptotically optimal if certain mild restrictions are imposed on f, g and K. In this context we say that $(\hat{h}_X, \hat{h}_Y)$ is asymptotically optimal with respect to M_{mn} if

$$\lim_{m,n \to \infty} \left[\frac{M_{mn}(\hat{h}_X, \hat{h}_Y)}{\inf\{M_{mn}(h_X, h_Y) : (h_X, h_Y) \in \mathbf{R}_+^d \times \mathbf{R}_+^d\}} \right] = 1 \qquad (2.4)$$

almost surely. Theorem 2.1 provides conditions under which (2.4) is true.

Theorem 2.1. *If the kernel K is symmetric, compactly supported and Hölder continuous; if the densities f and g and their one-dimensional marginals are bounded; and if the sample sizes m and n satisfy $m/n \to \xi$ for some $\xi > 0$ as m and n diverge to infinity; then $(\hat{h}_X, \hat{h}_Y)$ is asymptotically optimal with respect to M_{mn}.*

This result is analogous to the result of Stone (1984) for density estimation. The remarkable feature of such a result is its absence of smoothness restrictions on f and g. We therefore have a window-size selection rule which performs optimally for large samples regardless of the smoothness of the underlying densities.

The curve estimate for e can also be formed by using only one window-size vector h instead of the pair (h_X, h_Y). In this case our estimator is

$$e_{mn}(z|h,h) = pf_m(z|h) - (1-p)g_n(z|h) \tag{2.3}$$

and h can be chosen to minimise $\hat{S}_{mn}(h,h)$. In this case it would be advantageous to standardise each data set for scale before applying the rule. The optimality result analogous to that presented in Theorem 2.1 holds for this selection rule as well.

6.3 Examples and Discussion

We shall illustrate the efficacy of the density difference discrimination rule described in the previous section by applying it to some example data sets. The first problem we consider is that of discriminating between the standard Cauchy and standard normal normal distributions. In Subsection 1 simulated data are used to assess the performance of the density difference discrimination rule, and comparisons with the discrimination rule based on estimating the densities individually are made. The application of each of these rules to a set of real data is discussed in Subsection 2.

6.3.1 *Discrimination Between Cauchy and Normal Distributions.*

Consider the problem where an incoming observation Z is either from a population Π_X having the standard Cauchy distribution $f(z) = \pi^{-1}(1+z^2)^{-1}$ or a

population Π_Y having the standard normal distribution $g(z) = (2\pi)^{-\frac{1}{2}} e^{-z^2/2}$. The prior probabilities of Z coming from each population are assumed to be equal. We are therefore interested in estimating $e = (f - g)/2$.

For this discrimination problem the ideal discrimination rule is to classify Z as coming from Π_Y if and only if $Z \in R_Y$, where

$$R_Y = \{z : e(z) < 0\} \doteq (-1.85, 1.85).$$

The best obtainable error rate is therefore

$$\begin{aligned}
\mathrm{ER}^* =& \tfrac{1}{2} P(Z \notin (-1.85, 1.85) | Z \in \Pi_X) + \tfrac{1}{2} P(Z \in (-1.85, 1.85) | Z \in \Pi_Y) \\
=& \tfrac{1}{2} \left(\int_{-\infty}^{-1.85} + \int_{1.85}^{\infty} \right) \pi^{-1} (1 + z^2)^{-1} \, dz + \tfrac{1}{2} \int_{-1.85}^{1.85} (2\pi)^{-\frac{1}{2}} e^{-z^2/2} \, dz \\
\doteq& 37.44\%.
\end{aligned}$$

We applied the two discrimination rules to twenty-five training samples of simulated data. The sample sizes were $m = n = 100$. The first rule, which we shall label Rule 1, was based on choosing (h_X, h_Y) jointly to estimate the density difference. The second rule, based on choosing h_X and h_Y to estimate the densities separately, will be called Rule 2. The Gaussian kernel $K(z) = (2\pi)^{-\frac{1}{2}} e^{-z^2/2}$ was used throughout since it admits an explicit formula for the cross-validatory score functions. Rule 1 involved locating the window-size pair $(\hat{h}_{X,1}, \hat{h}_{Y,1})$ as the minimiser of

$$\begin{aligned}
\hat{S}_{100,100}(h_X, h_Y) =& (1/40000) \sum_{i=1}^{100} \sum_{j=1}^{100} \{ N(X_i - X_j, 2h_X^2) - 2N(X_i - Y_j, h_X^2 + h_Y^2) \\
&+ N(Y_i - Y_j, 2h_Y^2) + 2N(X_i - Y_j, h_X^2) + 2N(X_i - Y_j, h_Y^2) \} \\
&- (1/20000) \sum \sum_{i \neq j} \{ N(X_i - X_j, h_X^2) + N(Y_i - Y_j, h_Y^2) \},
\end{aligned}$$

where $N(z, h^2) = h^{-1} K(z/h)$, and classifying Z as coming from Π_Y if $Z \in \hat{R}_Y(\hat{h}_{X,1}, \hat{h}_{Y,1})$, where

$$\hat{R}_Y(\hat{h}_{X,1}, \hat{h}_{Y,1}) = \{z : e_{100,100}(z | \hat{h}_{X,1}, \hat{h}_{Y,1}) < 0\}. \tag{3.1}$$

For each sample the exact error rate of the rule was then computed from the formula

$$\mathrm{ER}(\hat{h}_{X,1}, \hat{h}_{Y,1}) = \tfrac{1}{2} \int_{\hat{R}_Y^c} \pi^{-1} (1 + z^2)^{-1} \, dz + \tfrac{1}{2} \int_{\hat{R}_Y} (2\pi)^{-\frac{1}{2}} e^{-z^2/2} \, dz. \tag{3.2}$$

Implementation of Rule 2 required choosing $\hat{h}_{X,2}$ as the minimiser of

$$\hat{S}_{X,100}(h) = (1/10000)\sum_{i=1}^{100}\sum_{j=1}^{100} N(X_i - X_j, 2h^2) - (1/5000)\sum\sum_{i\neq j} N(X_i - X_j, h^2)$$

and $\hat{h}_{Y,2}$ as the minimiser of

$$\hat{S}_{Y,100}(h) = (1/10000)\sum_{i=1}^{100}\sum_{j=1}^{100} N(Y_i - Y_j, 2h^2) - (1/5000)\sum\sum_{i\neq j} N(Y_i - Y_j, h^2).$$

The discrimination region $\hat{R}_Y(\hat{h}_{X,2}, \hat{h}_{Y,2})$ and corresponding error rate $\mathrm{ER}(\hat{h}_{X,2}, \hat{h}_{Y,2})$ were then obtained using formulae analogous to (3.1) and (3.2).

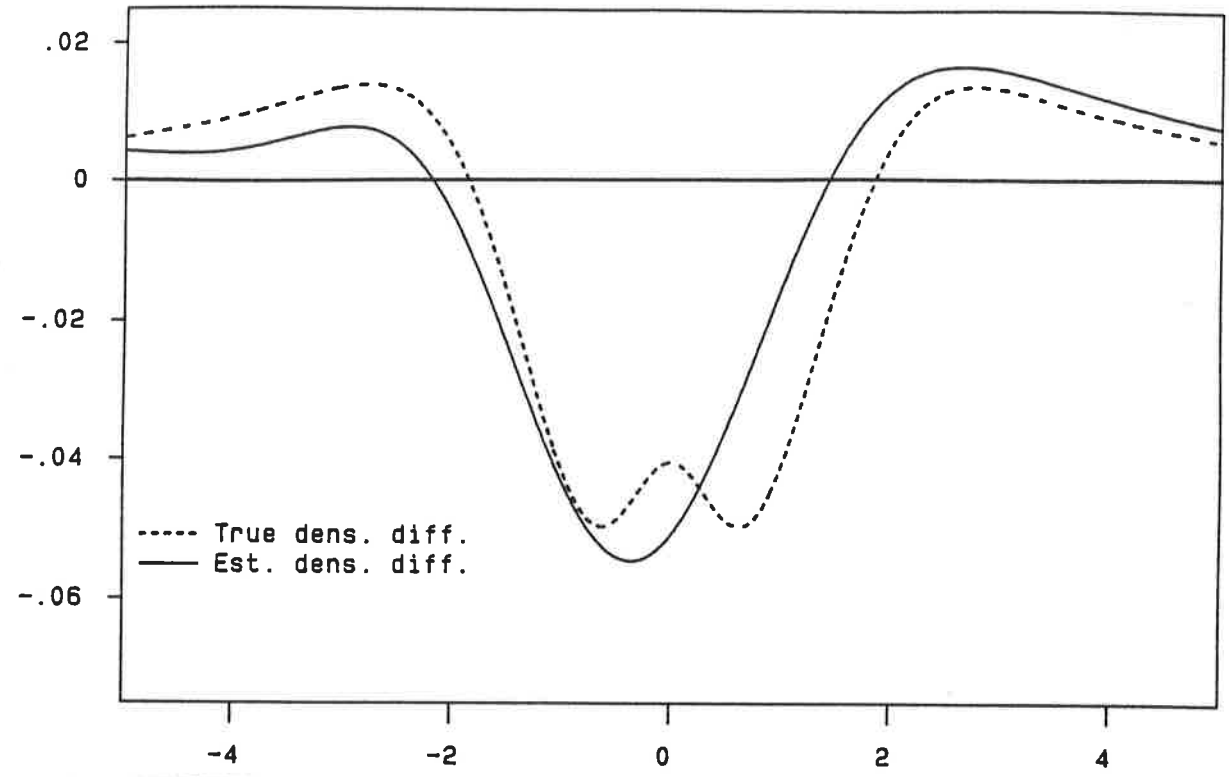
Table 3.1 lists the values of the selected window-sizes and error rates obtained for each of the twenty-five replications. As shown above, the lowest possible error rate is 37.44%. For Rule 1 the average error rate was 38.15% with standard error 0.55% while the average error rate for Rule 2 was 38.90% with standard error 1.05%. For 16 of the 25 samples Rule 1 had a lower error rate than Rule 2. Figure 3.1 (a) shows the density difference estimate corresponding to Rule 1 for replication number 5. This was the thirteenth best estimate, out of the twenty-five replications, in terms of minimising error rate and was chosen in an effort to depict average case performance of the estimator. The estimated classification region $\hat{R}_Y$ in this case is $\hat{R}_Y = (-2.18, 1, 43)$. The estimate of e required for Rule 2 based on the same sample is graphed in Figure 3.1 (b). The estimate of R_Y for this rule is $\hat{R}_Y = (-4.35, -3.41) \cup (-1.87, 1, 13)$. The observation that the window-size selection rule for Rule 2 undersmooths the data in this case was typical of behaviour throughout the study. Rule 1 tended to choose larger window-sizes leading to a somewhat less noisy curve estimate than the one formed using the window-sizes of Rule 2. The asymptotic theory, summarised by the formulae at (2.1), (2.2) and (2.3), lends weight to this observation since the optimal window-size pair satisfies

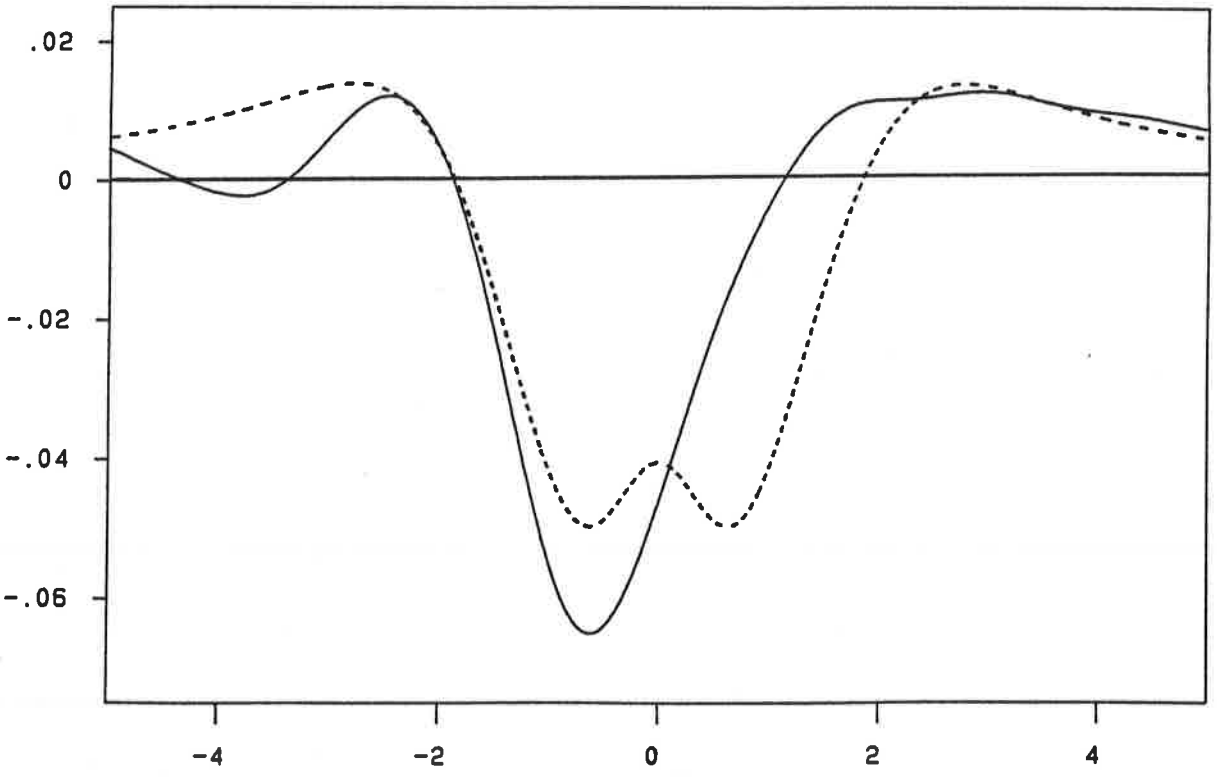
$$(h_{X,\mathrm{opt}}, h_{Y,\mathrm{opt}}) \sim (1.4077m^{-1/5}, 1.4486n^{-1/5})$$

when h_X and h_Y are chosen jointly and $\xi = 1$. If they are chosen separately then $h_{X,\mathrm{opt}} \sim 1.0339m^{-1/5}$ and $h_{Y,\mathrm{opt}} \sim 1.0592n^{-1/5}$. Therefore, for large samples,

Table 3.1: Selected window-sizes and corresponding error rates for discrimination between Cauchy and normal distributions.

Rep	$h_{X,1}$	$h_{Y,1}$	$\mathrm{ER}(h_{X,1}, h_{Y,1})$	$h_{X,2}$	$h_{Y,2}$	$\mathrm{ER}(h_{Y,1}, h_{Y,2})$
1	1.0728	0.7777	37.71	0.4445	0.1305	39.96
2	0.4517	0.5180	37.58	0.2989	0.4476	40.01
3	1.6663	1.0086	37.70	0.7052	0.4420	37.88
4	1.5769	1.1242	37.95	0.3951	0.4912	37.79
5	1.1575	0.8258	37.99	0.5987	0.5025	39.58
6	1.3660	0.8898	37.68	0.5713	0.3048	38.74
7	0.8647	0.1175	38.64	0.8503	0.2135	38.01
8	1.2138	0.8734	38.41	0.3880	0.4420	39.50
9	2.2949	1.6599	39.48	0.4537	0.4929	38.25
10	1.8300	1.2693	38.37	0.7009	0.6145	38.09
11	0.4671	0.1292	39.10	0.4478	0.1808	38.72
12	1.9966	1.4192	38.49	0.6149	0.5803	38.26
13	1.0915	0.7119	38.42	0.2382	0.4186	40.88
14	2.1960	1.6063	38.75	0.3713	0.4165	38.86
15	1.6413	1.0321	37.67	0.7191	0.5610	37.45
16	1.6111	1.0200	37.72	0.4310	1.0000	40.51
17	1.5442	0.8049	37.45	0.6250	0.4502	37.66
18	1.3764	1.0269	37.45	0.4985	0.1777	39.86
19	1.8332	1.4069	38.18	0.4808	0.5045	37.69
20	1.3161	0.8114	37.93	0.2747	0.2435	39.46
21	1.7953	1.4172	37.84	0.4824	0.4217	39.78
22	1.6175	1.0767	37.56	0.5185	0.3733	37.65
23	1.1703	0.8188	38.24	0.4973	0.2446	40.65
24	2.3368	1.7655	39.11	0.4705	0.5763	38.55
25	2.0659	1.4020	38.25	0.6222	0.2357	38.66





Figures 3.1 (a) and 3.1 (b): Typical estimates of the difference between the Cauchy and the normal density with (a) h_X and h_Y are chosen jointly to minimise $\hat{S}_{100,100}(h_X, h_Y)$ and (b) h_X chosen to minimise $\hat{S}_{X,100}$ and h_Y chosen to minimise $\hat{S}_{Y,100}$. The broken curve is $e = (f-g)/2$; the unbroken curves are (a) $e_{100,100}(\cdot|\hat{h}_{X,1}, \hat{h}_{Y,1})$ and (b) $e_{100,100}(\cdot|\hat{h}_{X,2}, \hat{h}_{Y,2})$.

Table 3.2: Calcium concentrations (millimoles/litre) of urines with and without crystals.

With crystals		Without crystals	
2.45	1.32	6.96	4.18
4.49	1.55	4.45	13.00
2.36	1.52	5.54	0.27
2.15	0.77	6.19	7.64
1.16	2.17	7.31	6.63
3.34	0.17	8.53	14.34
1.40	0.83	4.74	9.04
8.48	3.04	2.50	0.58
1.16	1.06	1.27	7.82
2.21	3.93	4.18	12.20
1.93	5.38	3.10	9.39
1.27	3.53	3.01	
1.03	4.54	6.81	
1.47	3.98	8.28	
1.53	1.02	2.33	
5.09	3.46	7.18	
1.05	1.19	5.67	
2.03	5.64	12.68	
7.68	2.66	8.94	
1.45	1.22	3.16	
5.16	2.64	3.30	
0.81	2.31	6.99	
1.32		0.65	

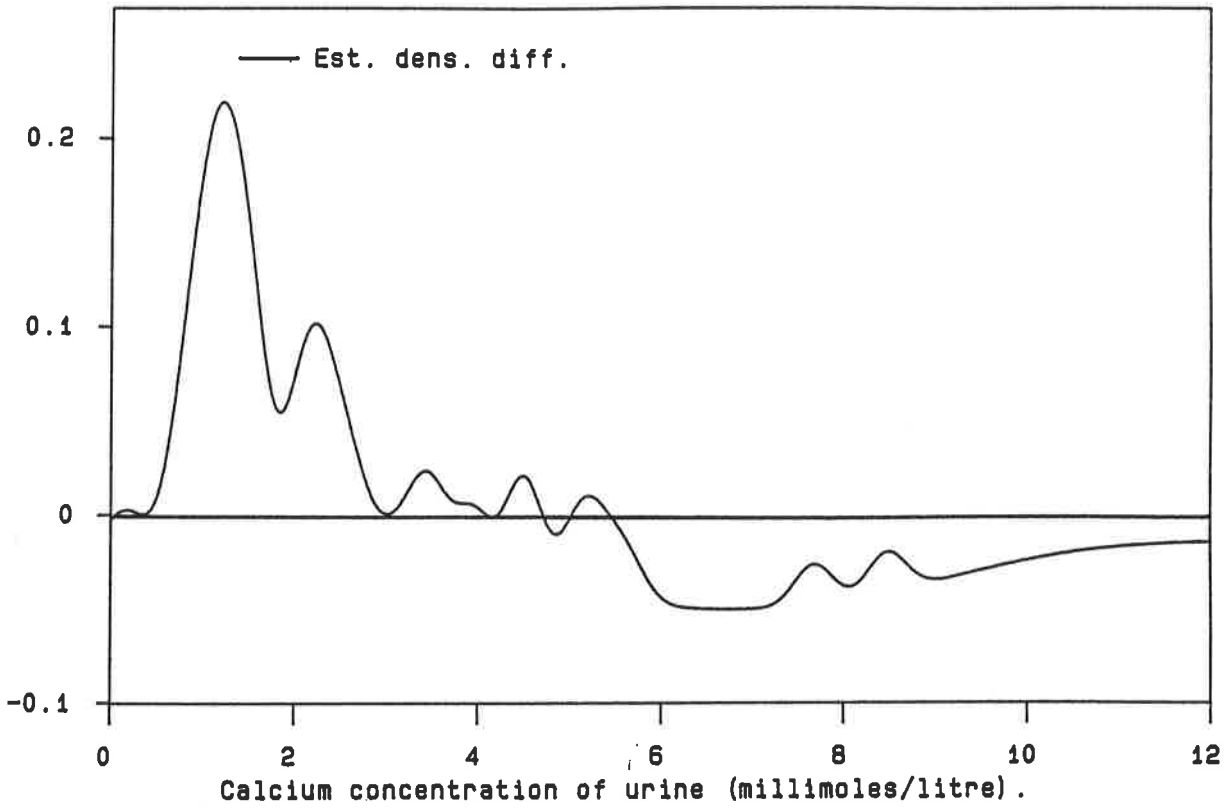


Figure 3.2: Density difference estimate obtained from the urine crystal data (see Table 3.2) where the window-size pair (h_X, h_Y) is chosen to minimise $\hat{S}_{45,34}(h_X, h_Y)$.

the window-sizes chosen by cross-validation to minimise $\hat{S}_{\mathcal{X},m}(h)$ and $\hat{S}_{\mathcal{Y},n}(h)$ will tend to produce window-sizes which are about 25-30% lower than the optimal window-sizes when estimating e for this particular problem.

6.3.2 *Formation of Crystals in Urines*

Table 3.2 displays the calcium concentration in millimoles/litre of 79 urine specimens. In 34 of the specimens the formation of calcium oxalate crystals has been observed. There is no presence of crystals in the remaining 45 specimens. The data were obtained from Table 44.1 of Andrews and Herzberg (1985, p.251). We decided to use the density difference estimator developed in this chapter to form a rule for classifying a urine as either "not to form crystals" or "to form crystals" based on the observed calcium concentration. The calcium concentration readings of those urine specimens without crystals constituted the $\mathcal{X}$ sample with $m = 45$. The $\mathcal{Y}$ sample consisted of the calcium concentrations of the urine specimens containing crystals with $n = 34$. We assumed equal prior probabilities and used a Gaussian kernel for our estimate. The window-size pair selected by minimising $\hat{S}_{34,45}(h_X, h_Y)$ was $(h_X, h_Y) = (0.2076, 1.6675)$. The density difference estimate is plotted in Figure 3.2. The graph shows that a reasonable discrimination rule is to decide that a urine will form crystals if the calcium concentration is above 5 millimetres/litre.

6.4 Proof of Theorem 2.1

Throughout this section we let $C, C_1, C_2, \ldots$ denote positive generic constants. We define the functions f_h and g_h to be

$$f_h(x) = E\{f_m(x|h)\} = \int K_h(x-y)f(y)\,dy$$

and

$$g_h(x) = E\{g_n(x|h)\} = \int K_h(x-y)g(y)\,dy.$$

In addition we put $e_{h_X h_Y} = pf_{h_X} - (1-p)g_{h_Y}$.

The proof of Theorem 2.1 will be preceded by six lemmas. For each lemma it is assumed that f, g, K, m and n satisfy the conditions ascribed to them in the

statement of Theorem 2.1. Most of these results will only be stated and proved for the population Π_X and the sample $\mathcal{X}$. The analogous results for Π_Y and $\mathcal{Y}$ are a trivial consequence.

Lemma 4.1. *There exist positive constants C and ρ such that*

$$\int (e_{h_X h_Y} - e)^2 \geq C(v_{h_X}^{\rho} \wedge v_{h_Y}^{\rho} \wedge 1)$$

for all $(h_X, h_Y) \in \mathbf{R}_+^d \times \mathbf{R}_+^d$.

Proof. Let ψ, χ and ω denote the Fourier transforms of K, f and g respectively. The Fourier transform of K_h is ψ_h, given by $\psi_h(t) = \psi(ht)$ where $ht = (h_1 t_1, \ldots, h_d t_d)$. Also, the Fourier transforms of f_{h_X} and g_{h_Y} are $\psi_{h_X}\chi$ and $\psi_{h_Y}\omega$ respectively. Since χ and ω are not identical, but satisfy $\chi(0) = \omega(0) = 1$ and are continuous, then there exists a non-empty bounded sphere S centred at the origin of $\mathbf{R}^d$ for which

$$\inf_{c>0} \int_{t \in S} |p\chi(t) - c(1-p)\omega(t)|^2 \, dt > 0.$$

Because K is symmetric, compactly supported and integrates to unity, there is a smallest integer $k \geq 1$ such that

$$\int (1 \cdot x)^{2k} K(x) \, dx \neq 0$$

where $1 \cdot x$ is the inner product of x and the d-vector with all entries equal to unity. For this k,

$$\begin{aligned}
\psi_h(t) - 1 &= \int \{cos(ht \cdot x) - 1\} K(x) \, dx \\
&= (-1)^k \{(2k)!\}^{-1} \int (ht \cdot x)^{2k} K(x) \, dx + o(|h|^{2k}) \\
&= (-1)^k \{(2k)!\}^{-1} |h|^{2k} \int (ut \cdot x)^{2k} K(x) \, dx + o(|h|^{2k})
\end{aligned}$$

for some unit vector u, uniformly in $t \in S$, as $|h| \to 0$. Therefore, by Parseval's identity,

$$\begin{aligned}
(2\pi)^d \int (e_{h_X h_Y} - e)^2 &= \int |(\psi_{h_X} - 1)p\chi - (\psi_{h_Y} - 1)(1-p)\omega|^2 \\
&\geq \{(2k!)\}^{-2} |h_X|^{4k} \int_{t \in S} \Big| \int (u_X t \cdot x) K(x) \, dx \, p\chi(t) \\
&\qquad - (|h_Y|/|h_X|)^{2k} \int (u_Y t \cdot x)^{2k} K(x) \, dx \, (1-p)\omega(t) \Big|^2 \, dt \\
&\qquad + o(|h_X|^{4k} + |h_Y|^{4k})
\end{aligned}$$

where u_X and u_Y are unit vectors. Assume, without loss of generality, that $|h_X| \geq |h_Y|$. Then

$$\begin{aligned}(2\pi)^d \int (e_{h_X h_Y} - e)^2 \geq & \{(2k)!\}^{-2} |h_X|^{4k} \inf_{c>0} \int_{t \in S} \left\{ \int (u_X t \cdot x)^{2k} K(x)\, dx \right\}^2 \\ & \times |p\chi(t) - c(1-p)\omega(t)|^2\, dt + o(|h_X|^{4k}) \\ = & C|h_X|^{4k} + o(|h_X|^{4k}) \\ \geq & C|h_Y|^{4k} + o(|h_Y|^{4k})\end{aligned}$$

where $C > 0$. For small values of $|h_X| \wedge |h_Y|$ the required result is a consequence of $|h|^d \geq v_h$. The result for large $|h_X|$ and $|h_Y|$ follows from the observation that $\int (e_{h_X h_Y} - e)^2$ is bounded away from zero whenever both h_X and h_Y lie outside any neighbourhood of the origin. ∎

Let

$$\begin{aligned}\theta_{mn}(h_X, h_Y) &= \int (e_{h_X h_Y} - e)^2 + m^{-1} v_{h_X}^{-1} + n^{-1} v_{h_Y}^{-1}, \\ B_{X,m}(h_X, h_Y) &= m^{-1} \sum_{i=1}^{m} e_{h_X h_Y}(X_i) - E\{e_{h_X h_Y}(X_1)\}, \\ B_{Y,n}(h_X, h_Y) &= n^{-1} \sum_{i=1}^{n} e_{h_X h_Y}(Y_i) - E\{e_{h_X h_Y}(Y_1)\}, \\ D_{X,m} &= m^{-1} \sum_{i=1}^{m} e(X_i) - E\{e(X_1)\}, \\ D_{Y,n} &= n^{-1} \sum_{i=1}^{n} e(Y_i) - E\{e(Y_1)\}.\end{aligned}$$

Also let $H_{X,m}$ and $H_{Y,n}$ be subsets of $\mathbf{R}_+^d$ satisfying $\text{card}(H_{X,m}) \leq Am^a$ and $\text{card}(H_{Y,n}) \leq Bn^b$ for positive constants A, a, B and b. The product set $H_{X,m} \times H_{Y,n}$ will be denoted by H_{mn}.

Lemma 4.2.

$$\lim_{m,n \to \infty} \sup\{\theta_{mn}(h_X, h_Y)^{-1} |B_{X,m}(h_X, h_Y) - D_{X,m}| : (h_X, h_Y) \in H_{mn}\} = 0 \quad (4.1)$$

almost surely, and

$$\lim_{m,n \to \infty} \sup\left\{ \theta_{mn}(h_X, h_Y)^{-1} \left| \int \{e_{mn}(\cdot | h_X, h_Y) - e_{h_X h_Y}\}(e_{h_X h_Y} - e) \right| : (h_X, h_Y) \in H_{mn} \right\} = 0 \quad (4.2)$$

almost surely.

Proof. Define

$$Z_{ih_Xh_Y} = e_{h_Xh_Y}(X_i) - e(X_i) - E\{e_{h_Xh_Y}(X_1) - e(X_1)\}$$

and $u^2_{h_Xh_Y} = \int (e_{h_Xh_Y} - e)^2$. Then clearly $Z_{ih_Xh_Y}$, $i \geq 1$, is a sequence of independent and identically distributed random variables, each having zero mean. It follows from the boundedness of f and g that $|Z_{ih_Xh_Y}| \leq C_1$ and $\mathrm{Var}(Z_{ih_Xh_Y}) \leq C_2 u^2_{h_Xh_Y}$, where C_1 and C_2 are positive constants not depending on (h_X, h_Y). Notice that

$$\overline{Z}_{mh_Xh_Y} \equiv m^{-1} \sum_{i=1}^{m} Z_{ih_Xh_Y} = B_{X,m}(h_X, h_Y) - D_{X,m}.$$

We have from Markov's inequality for all $t > 0$ and $\alpha > 0$,

$$P(|\overline{Z}_{mh_Xh_Y}| > t) \leq t^{-\alpha} m^{-\alpha} E \left| \sum_{i=1}^{m} Z_{ih_Xh_Y} \right|^{\alpha}.$$

Assuming from now on that $\alpha > 1$ we obtain from Rosenthal's inequality (Hall and Heyde (1980), pp.23–24),

$$E \left| \sum_{i=1}^{m} Z_{ih_Xh_Y} \right|^{\alpha} \leq C_3 \left[\left\{ \sum_{i=1}^{m} E(Z^2_{ih_Xh_Y}) \right\}^{\alpha/2} + \sum_{i=1}^{m} E|Z_{ih_Xh_Y}|^{\alpha} \right].$$

This implies that

$$P(|\overline{Z}_{mh_Xh_Y}| > t) \leq C_4 t^{-\alpha} (m^{-\alpha/2} u^{\alpha}_{h_Xh_Y} + m^{1-\alpha})$$

where C_4 depends only on α. Setting $t = \epsilon \theta_{mn}(h_X, h_Y)$, where $\epsilon > 0$, leads to

$$\begin{aligned} P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_Xh_Y}| > \epsilon\} \leq & C_5 (u^2_{h_Xh_Y} + m^{-1} v^{-1}_{h_X} + n^{-1} v^{-1}_{h_Y})^{-\alpha} \\ & \times (m^{-\alpha/2} u^{\alpha}_{h_Xh_Y} + m^{1-\alpha}) \end{aligned} \tag{4.3}$$

By the assumption that $m/n \to \xi > 0$ we may assume, without loss of generality, that $v_{h_X} \leq v_{h_Y}$. Suppose first that $v_{h_X} \geq 1$. Then from Lemma 4.1 $u^2_{h_Xh_Y} \geq C_6$ which, on application of (4.3), gives

$$P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_Xh_Y}| > \epsilon\} \leq C_7 (m^{-\alpha/2} + m^{1-\alpha}). \tag{4.4}$$

Next suppose $0 < v_{h_X} < 1$. Let $\tau = \frac{1}{2} - 1/\{2(1+\rho^{-1})\}$ where ρ is the constant appearing in Lemma 4.1, and consider separately the cases (i) $u_{h_X h_Y} \leq m^{\tau - \frac{1}{2}}$ and (ii) $u_{h_X h_Y} > m^{\tau - \frac{1}{2}}$. If (i) is true then it can be shown that

$$\begin{aligned} P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_X h_Y}| > \epsilon\} \leq & C_5 (m v_{h_X})^{\alpha} (m^{-\alpha/2} u_{h_X h_Y}^{\alpha} + m^{1-\alpha}) \\ \leq & C_8 [m^{-\alpha/\{2(\rho+1)\}} + m^{1-\alpha/(\rho+1)}], \end{aligned} \tag{4.5}$$

while under (ii) we have

$$\begin{aligned} P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_X h_Y}| > \epsilon\} \leq & C_5 u_{h_X h_Y}^{-2\alpha} (m^{-\alpha/2} u_{h_X h_Y}^{\alpha} + m^{1-\alpha}) \\ \leq & C_5 (m^{-\alpha\tau} + m^{1-2\alpha\tau}). \end{aligned} \tag{4.6}$$

It follows from (4.4), (4.5) and (4.6) that for all $v_{h_X} > 0$,

$$P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_X h_Y}| > \epsilon\} = O(m^{-\beta})$$

for every $\beta > 0$. Because of the restriction imposed on the cardinality of the set H_{mn} and the assumption that $m/n \to \xi > 0$ we obtain

$$\sum_{(h_X, h_Y) \in H_{mn}} P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_X h_Y}| > \epsilon\} = O(m^{-\beta})$$

for each $\beta > 0$ implying, by Boole's inequality, that

$$\sum_{m=1}^{\infty} \sup\{P\{\theta_{mn}(h_X, h_Y)^{-1} |\overline{Z}_{mh_X h_Y}| > \epsilon\} : (h_X, h_Y) \in H_{mn}\} < \infty$$

for every $\epsilon > 0$ and $n \geq 1$. The result at (4.1) follows from the Borel-Cantelli lemma.

The proof of (4.2) can be accomplished by applying the same argument to

$$Z_{ih_X h_Y} = \int \{pK_{h_X}(z - X_i) - (1-p)K_{h_Y}(z - Y_i) - e_{h_X h_Y}(z)\}\{e_{h_X h_Y}(z) - e(z)\}\, dz. \blacksquare$$

For $r > 0$ we shall let

$$\theta_{mr}(h) = v_h^r \wedge 1 + m^{-1} v_h^{-1}$$

and

$$\theta_{mnr}(h_X, h_Y) = v_{h_X}^r \wedge v_{h_Y}^r \wedge 1 + m^{-1} v_{h_X}^{-1} + n^{-1} v_{h_Y}^{-1}.$$

Let F be the distribution function of X_1 and F_m be the empirical distribution function of the sample $\mathcal{X}$. The functions G and G_n are defined analogously for Y_1 and $\mathcal{Y}$.

Lemma 4.3. *For all $r > 0$,*

$$\lim_{m\to\infty} \sup_{h\in H_{X,m}} \theta_{mr}(h)^{-1} \left| \int\int_{x\neq y} K_h(x-y)\{dF_m(x) - dF(x)\}\{dF_m(y) - dF(y)\} \right| = 0$$

almost surely.

Proof. Let X^* and X^{**} be independent random variables having the same distribution as X_1. Define the function μ_h by

$$\mu_h(x) = E\{K_h(x - X^*)\}$$

and let

$$\nu_h = E\{\mu_h(X^{**})\} = E\{K_h(X^{**} - X^*)\}.$$

It is easily verified that

$$\int\int_{x\neq y} K_h(x-y)\{dF_m(x) - dF(x)\}\{dF_m(y) - dF(y)\} = 2V_{mh} + W_{mh}$$

where

$$V_{mh} = m^{-2}\sum\sum_{i<j}\{K_h(X_i - X_j) - \mu_h(X_i) - \mu_h(X_j) + \nu_h\}$$

and

$$W_{mh} = m^{-1}\nu_h - 2m^{-2}\sum_{i=1}^{m}\mu_h(X_i).$$

For $1 \le i, j \le m$ let

$$U_{ij} = K_h(X_i - X_j) - \mu_h(X_i) - \mu_h(X_j) + \nu_h.$$

For $2 \le k \le m$ put $T_k = \sum_{j=2}^{k}\sum_{i=1}^{j-1} U_{ij}$ and let $\mathcal{F}_k$ denote the σ-field generated by $\{X_1, \ldots, X_k\}$. Then one may establish that $\{(T_k, \mathcal{F}_k),\ 2 \le k \le m\}$ is a martingale and $\{(Z_k, \mathcal{F}_k),\ 2 \le k \le m\}$ is the corresponding martingale difference sequence with $Z_2 = T_2$ and $Z_k = T_k - T_{k-1}$, $3 \le k \le m$. Since $T_m = \sum_{k=2}^{m} Z_k$, we have for all positive t and all $\alpha > 1$,

$$P(|T_m| > t) \le t^{-2\alpha}E\left|\sum_{k=2}^{m} Z_k\right|^{2\alpha} \le C_1 t^{-2\alpha} m^{\alpha-1}\sum_{k=2}^{m} E|Z_k|^{2\alpha}. \qquad (4.7)$$

The second inequality follows from Hölder's and Burkholder's inequalities (see Hall and Heyde (1980) p.87) with C_1 depending only on α. Next observe that

$$Z_k = \sum_{i=1}^{k-1} U_{ik} = \sum_{i=1}^{k-1} \{K_h(X_i - X_k) - \mu_h(X_i) - \mu_h(X_k) + \nu_h\},$$

so that conditional on X_k, Z_k is a sum of $k-1$ independent and identically distributed random variables. Therefore, by Rosenthal's inequality (see Hall and Heyde (1980) pp.23-24) with conditioning on X_k,

$$E(|Z_k|^{2\alpha}|X_k) \leq C_2[\{(k-1)E(U_{1k}^2|X_k)\}^{\alpha} + (k-1)E(|U_{1k}|^{2\alpha}|X_k)]. \tag{4.8}$$

We shall consider separately the cases (i) $0 < v_h \leq m^{-1/(2r+2)}$, (ii) $m^{-1/(2r+2)} < v_h \leq 1$ and (iii) $v_h > 1$. From the boundedness and compact support of K and the boundedness f we have

$$E(|U_{1k}|^{2\alpha}|X_k) \leq C_3 v_h^{-2\alpha}$$

and

$$E(U_{1k}^2|X_k) \leq C_4 v_h^{-1},$$

where both bounds are in uniform in X_k. Substitution of these estimates into (4.7) and (4.8) gives

$$\begin{aligned} P(|T_m| < t) &\leq C_5 t^{-2\alpha} m^{\alpha-1} \left\{ \left(\sum_{k=1}^{m} k^{\alpha} \right) v_h^{-\alpha} + \left(\sum_{k=1}^{m} k \right) v_h^{-2\alpha} \right\} \\ &\leq C_5 t^{-2\alpha} (m^{2\alpha} v_h^{-\alpha} + m^{\alpha+1} v_h^{-2\alpha}). \end{aligned}$$

Now choose $t = m^2 \theta_{mr}(h)\epsilon$ where $\epsilon > 0$ is arbitrary. Then, noting that $V_{mh} = m^{-2} T_m$, we have

$$P\{\theta_{mr}(h)^{-1}|V_{mh}| > \epsilon\} \leq C_6 (m v_h^r + v_h^{-1})^{-2\alpha} (v_h^{-\alpha} + m^{1-\alpha} v_h^{-2\alpha}).$$

Let (i) be true. Then we have the bound

$$\begin{aligned} P\{\theta_{mr}(h)^{-1}|V_{mh}| > \epsilon\} &\leq C_6 (v_h^{-1})^{-2\alpha} (v_h^{-\alpha} + m^{1-\alpha} v_h^{-2\alpha}) \\ &\leq C_6 \{m^{-\alpha/(2r+2)} + m^{1-\alpha}\}. \end{aligned} \tag{4.9}$$

If (ii) holds then

$$P\{\theta_{mr}(h)^{-1}|V_{mh}| > \epsilon\} \leq C_6(mv_h^r)^{-2\alpha}(v_h^{-\alpha} + m^{1-\alpha}v_h^{-2\alpha})$$
$$\leq C_6(m^{-\alpha} + m^{1-2\alpha}). \tag{4.10}$$

Finally, if (iii) holds then we have

$$P\{\theta_{mr}(h)^{-1}|V_{mh}| > \epsilon\} \leq C_6(m + v_h^{-1})^{-2\alpha}(1 + m^{1-\alpha})$$
$$\leq C_6(m^{-2\alpha} + m^{1-3\alpha}). \tag{4.11}$$

Combining (4.9), (4.10) and (4.11) yields that

$$P\{\theta_{mr}(h)^{-1}|V_{mh}| > \epsilon\} = O(m^{-\beta})$$

for all $\beta > 0$. As in the proof of Lemma 4.2, Boole's inequality and the Borel-Cantelli lemma can be used to establish that

$$\lim_{m\to\infty} \sup_{h\in H_{X,m}} \theta_{mr}(h)^{-1}|V_{mh}| = 0$$

almost surely.

The proof will be complete if it is established that

$$\lim_{m\to\infty} \sup_{h\in H_{X,m}} \theta_{mr}(h)^{-1}|W_{mh}| = 0$$

almost surely. This, however, is quite easy in comparison to the above arguments and follows quickly from the boundedness of K. ∎

Lemma 4.4. *For all $r > 0$,*

$$\lim_{m,n\to\infty} \sup\left\{\theta_{mnr}(h,h)^{-1}\left|\int\int K_h(x-y)\{dF_m(x)-dF(x)\}\{dG_n(y)-dG(y)\}\right| : h \in H_{X,m}\right\} = 0$$

almost surely.

Proof. Let X^* and Y^* be independent random variables such that X^* has the same distribution as X_1 and Y^* has the same distribution as Y_1. Define

$$\mu_{X,h}(x) = E\{K_h(x - X^*)\}, \qquad \mu_{Y,h}(x) = E\{K_h(x - Y^*)\},$$

$$\nu_h = E\{\mu_{X,h}(Y^*)\} = E\{\mu_{Y,h}(X^*)\} = E\{K_h(X^* - Y^*)\}$$

and

$$Z_{nhi} = n^{-1}\sum_{j=1}^{n}\{K_h(X_i - Y_j) - \mu_{Y,h}(X_i) - \mu_{X,h}(Y_j) + \nu_h\}.$$

Then we have

$$\overline{Z}_{mnh} \equiv m^{-1}\sum_{i=1}^{m} Z_{nhi} = \int\int K_h(x-y)\{dF_m(x) - dF(x)\}\{dG_n(y) - dG(y)\}.$$

For $t > 0$ we have by Markov's inequality,

$$P(|\overline{Z}_{mnh}| > t) \leq t^{-2\alpha}m^{-2\alpha}E\left|\sum_{i=1}^{m} Z_{nhi}\right|^{2\alpha},$$

where it is assumed that $\alpha > 1$. Conditional on $\{Y_1, \ldots, Y_n\}$ the random variables Z_{nhi}, $1 \leq i \leq m$, are independent and identically distributed with mean zero. In the following we shall use E' to denote expectation conditional on $\{Y_1, \ldots, Y_n\}$. According to the conditional version of Rosenthal's inequality,

$$\begin{aligned} E'\left|\sum_{i=1}^{m} Z_{nhi}\right|^{2\alpha} &\leq C_1\left[\left\{\sum_{i=1}^{m} E'(Z_{nhi}^2)\right\}^{\alpha} + \sum_{i=1}^{m} E'|Z_{nhi}|^{2\alpha}\right] \\ &= C_1\{m^{\alpha}E'(Z_{nh1}^2)^{\alpha} + mE'|Z_{nh1}|^{2\alpha}\}. \end{aligned}$$

Taking expectations we obtain

$$E\left|\sum_{i=1}^{m} Z_{nhi}\right|^{2\alpha} \leq C_1[m^{\alpha}E\{E'(Z_{nh1}^2)\}^{\alpha} + mE|Z_{nh1}|^{2\alpha}]. \tag{4.12}$$

Set

$$U_{ij} = K_h(X_i - Y_j) - \mu_{Y,h}(X_i) - \mu_{X,h}(Y_j) + \nu_h$$

and note that $Z_{nh1} = n^{-1}\sum_{j=1}^{n} U_{1j}$. Conditional on X_1 the random variables U_{1j}, $1 \leq j \leq n$, are identically distributed with zero mean. Let E'' denote expectation conditional on X_1 and apply Rosenthal's inequality again to yield

$$E''|Z_{nh1}|^{2\alpha} \leq C_2 n^{-2\alpha}\{n^{\alpha}E''(U_{1,1}^2)^{\alpha} + nE''|U_{1,1}|^{2\alpha}\},$$

so that

$$E|Z_{nh1}|^{2\alpha} \leq C_2 n^{-2\alpha}[n^{\alpha}E\{E''(U_{1,1}^2)\}^{\alpha} + nE|U_{1,1}|^{2\alpha}]. \tag{4.13}$$

Also, from the conditional version of Hölder's inequality, $\{E'(Z_{nh1}^2)\}^\alpha \le E'|Z_{nh1}|^{2\alpha}$, implying that

$$E\{E'(Z_{nh1}^2)\}^\alpha \le E|Z_{nh1}|^{2\alpha}. \tag{4.14}$$

Combining the estimates at (4.12), (4.13) and (4.14) provides

$$E\left|\sum_{i=1}^m Z_{nhi}\right|^2 \le C_3[m^\alpha n^{-\alpha} E\{E''(U_{1,1}^2)\}^\alpha + m^\alpha n^{1-2\alpha} E|U_{1,1}|^{2\alpha}]. \tag{4.15}$$

The boundedness of K implies that $E|U_{1,1}|^{2\alpha} \le C_4 v_h^{-2\alpha}$. Also, since the densities f and g are bounded, we can show that $E''(U_{1,1}^2) \le C_6 v_h^{-1}$ uniformly in X_1. Hence $E\{E''(U_{1,1}^2)\}^\alpha \le C_7 v_h^{-\alpha}$. Therefore, in view of (4.15), we obtain

$$E\left|\sum_{i=1}^m Z_{nhi}\right|^2 \le C_8(m^\alpha n^{-\alpha} v_h^{-\alpha} + m^\alpha n v_h^{-2\alpha}),$$

so that

$$P(|\overline{Z}_{mnh}| > t) \le C_8 t^{-2\alpha}(m^{-\alpha} n^{-\alpha} v_h^{-\alpha} + m^{-\alpha} n v_h^{-2\alpha}).$$

Let $t = \theta_{mnr}(h,h)\epsilon$ for $\epsilon > 0$. Then

$$P\{\theta_{mnr}(h,h)^{-1}|\overline{Z}_{mnh}| > \epsilon\}$$
$$\le C_8(v_h^r \wedge 1 + m^{-1}v_h^{-1} + n^{-1}v_h^{-1})^{-2\alpha}(m^{-\alpha}n^{-\alpha}v_h^{-\alpha} + m^{-\alpha}n^{1-2\alpha}v_h^{-2\alpha}).$$

Assume, without loss of generality, that $m \ge n$. Then

$$P\{\theta_{mnr}(h,h)^{-1}|\overline{Z}_{mnh}| > \epsilon\} \le C_8(v_h^r \wedge 1 + n^{-1}v_h^{-1})^{-2\alpha}(n^{-2\alpha}v_h^{-\alpha} + n^{1-3\alpha}v_h^{-2\alpha}).$$

Arguments identical to those used in the proof of Lemma 4.3 lead to

$$P\{\theta_{mnr}(h,h)^{-1}|\overline{Z}_{mnh}| > \epsilon\} = O(n^{-\beta})$$

for all $\beta > 0$. This is sufficient for the required result. ∎

For a function $J : \mathbf{R}^d \to \mathbf{R}$ we shall let $J^\dagger$ denote the convolution of J with itself. As before we let $\kappa_2^2 = K^\dagger(0) = \int K^2$.

Lemma 4.5. *For all $r > 0$,*

$$\lim_{m,n\to\infty} \sup\left\{\theta_{mnr}(h_X,h_Y)^{-1}\left|\int\{e_{mn}(\cdot|h_X,h_Y) - e_{h_X h_Y}\}^2\right.\right.$$
$$\left.\left. - \kappa_2^2\{p^2 m^{-1}v_{h_X}^{-1} + (1-p)^2 n^{-1}v_{h_Y}^{-1}\}\right| : (h_X,h_Y) \in H_{mn}\right\} = 0$$

almost surely.

Proof. It may be shown that

$$\begin{aligned}
&\int \{e_{mn}(\cdot|h_X, h_Y) - e_{h_X h_Y}\}^2 \\
&\quad = p^2 \int\int K_{h_X}^{\dagger}(x-y)\{dF_m(x) - dF(x)\}\{dF_m(y) - dF(y)\} \\
&\qquad + (1-p)^2 \int\int K_{h_Y}^{\dagger}(x-y)\{dG_n(x) - dG(x)\}\{dG_n(y) - dG(y)\} \\
&\qquad - 2p(1-p) \int\int (K_{h_X} * K_{h_Y})(x-y)\{dF_m(x) - dF(x)\}\{dG_n(y) - dG(y)\}.
\end{aligned}$$

Observing that

$$\int\int_{x=y} K_{h_X}^{\dagger}(x-y)\{dF_m(x) - dF(x)\}\{dF_m(y) - dF(y)\} = \kappa_2^2 m^{-1} v_{h_X}^{-1}$$

and

$$\int\int_{x=y} K_{h_Y}^{\dagger}(x-y)\{dG_n(x) - dG(x)\}\{dG_n(y) - dG(y)\} = \kappa_2^2 n^{-1} v_{h_Y}^{-1}$$

we obtain

$$\begin{aligned}
&\int \{e_{mn}(\cdot|h_X, h_Y) - e_{h_X h_Y}\}^2 - \kappa_2^2\{p^2 m^{-1} v_{h_X}^{-1} + (1-p)^2 n^{-1} v_{h_Y}^{-1}\} \\
&\quad = p^2 \int\int_{x\neq y} K_{h_X}^{\dagger}(x-y)\{dF_m(x) - dF(x)\}\{dF_m(y) - dF(y)\} \\
&\qquad + (1-p)^2 \int\int_{x\neq y} K_{h_Y}^{\dagger}(x-y)\{dG_n(x) - dG(x)\}\{dG_n(y) - dG(y)\} \\
&\qquad - 2p(1-p) \int\int (K_{h_X} * K_{h_Y})(x-y)\{dF_m(x) - dF(x)\}\{dG_n(y) - dG(y)\}.
\end{aligned}$$

Application of Lemma 4.3 to the first two terms on the right-hand side, with $K^{\dagger}$ instead of K, and a simple adaptation on Lemma 4.4 to the last term gives the desired result. ∎

We also require the following technical result for the proof of Theorem 2.1.

Lemma 4.6. *Suppose that a_n and b_n are both sequences of real valued functions defined on some set $\mathcal{U}$ and satisfy*

$$\lim_{n\to\infty} \sup_{u\in\mathcal{U}} \left| \frac{a_n(u)}{b_n(u)} - 1 \right| = 0.$$

Then

$$\lim_{n\to\infty} \frac{\inf_{u\in\mathcal{U}} a_n(u)}{\inf_{u\in\mathcal{U}} b_n(u)} = 1.$$

Proof. Assume, without loss of generality, that $a_n, b_n > 0$ for all $n \geq 1$. Let $0 < \epsilon < 1$ and suppose that

$$\sup_{u\in\mathcal{U}} \left| \frac{a_n(u)}{b_n(u)} - 1 \right| < \epsilon. \tag{4.16}$$

It is sufficient to show that this implies

$$\left| \frac{\inf_{u\in\mathcal{U}} a_n(u)}{\inf_{u\in\mathcal{U}} b_n(u)} - 1 \right| < \epsilon. \tag{4.17}$$

From (4.16) we have for all $u \in \mathcal{U}$, $a_n(u) < (1+\epsilon)b_n(u)$. Hence $\inf_{u\in\mathcal{U}} a_n(u) < (1+\epsilon)b_n(u)$ for all $u \in \mathcal{U}$, which implies that

$$\inf_{u\in\mathcal{U}} a_n(u) < (1+\epsilon) \inf_{u\in\mathcal{U}} b_n(u). \tag{4.18}$$

Similarly, we may show that

$$\inf_{u\in\mathcal{U}} a_n(u) > (1-\epsilon) \inf_{u\in\mathcal{U}} b_n(u). \tag{4.19}$$

Combining (4.18) and (4.19) we obtain (4.17) as required. ∎

We shall first prove an altered form of Theorem 2.1 where (h_X, h_Y) is confined to the set H_{mn} rather than the whole of $\mathbf{R}_+^d \times \mathbf{R}_+^d$.

Observe that

$$M_{mn}(h_X, h_Y) = \int \{e_{mn}(\cdot|h_X, h_Y) - e_{h_X h_Y}\}^2 + 2\int \{e_{mn}(\cdot|h_X, h_Y) - e_{h_X h_Y}\}(e_{h_X h_Y} - e) + \int (e_{h_X h_Y} - e)^2,$$

leading to

$$M_{mn}(h_X, h_Y) - \left[\int (e_{h_X h_Y} - e)^2 + \kappa_2^2\{p^2 m^{-1} v_{h_X}^{-1} + (1-p)^2 n^{-1} v_{h_Y}^{-1}\}\right]$$
$$= 2\int \{e_{mn}(\cdot|h_X, h_Y) - e_{h_X h_Y}\}(e_{h_X h_Y} - e) + \int \{e_{mn}(\cdot|h_X, h_Y) - e_{h_X h_Y}\}^2$$
$$- \kappa_2^2\{p^2 m^{-1} v_{h_X}^{-1} + (1-p)^2 n^{-1} v_{h_Y}^{-1}\}.$$

It follows from Lemmas 4.1, 4.2 and 4.5 that

$$\lim_{m,n\to\infty} \sup\{\theta_{mn}(h_X, h_Y)^{-1} |M_{mn}(h_X, h_Y) - \psi_{mn}(h_X, h_Y)| : (h_X, h_Y) \in H_{mn}\} = 0 \tag{4.20}$$

almost surely, where $\psi_{mn}(h_X, h_Y) = \int (e_{h_X h_Y} - e)^2 + \kappa_2^2 \{p^2 m^{-1} v_{h_X}^{-1} + (1-p)^2 n^{-1} v_{h_Y}^{-1}\}$. It is clear on comparison that we can replace $\theta_{mn}(h_X, h_Y)$ by $\psi_{mn}(h_X, h_Y)$ in (4.20) so that

$$\lim_{m,n\to\infty} \sup\{|\psi_{mn}(h_X, h_Y)^{-1} M_{mn}(h_X, h_Y) - 1| : (h_X, h_Y) \in H_{mn}\} = 0 \qquad (4.21)$$

almost surely.

It can be seen from the formula for $\hat{S}_{mn}(h_X, h_Y)$ that

$$\hat{S}_{mn}(h_X, h_Y) + C_{mn} = M_{mn}(h_X, h_Y) + 2R_{mn}(h_X, h_Y)$$

where $C_{mn} = \int e^2 + 2pD_{X,m} - 2(1-p)D_{Y,n}$ and

$$\begin{aligned} R_{mn}(h_X,h_Y) &= p\{D_{X,m} - B_{X,m}(h_X, h_Y)\} - (1-p)\{D_{Y,n} - B_{Y,n}(h_X, h_Y)\} \\ &\quad - p^2 \int\int_{x\neq y} K_{h_X}(x-y)\{dF_m(x) - dF(x)\}\{dF_m(y) - dF(y)\} \\ &\quad - (1-p)^2 \int\int_{x\neq y} K_{h_Y}(x-y)\{dG_n(x) - dG(x)\}\{dG_n(y) - dG(y)\} \\ &\quad + p(1-p) \int\int K_{h_X}(x-y)\{dF_m(x) - dF(x)\}\{dG_n(y) - dG(y)\} \\ &\quad + p(1-p) \int\int K_{h_Y}(x-y)\{dF_m(x) - dF(x)\}\{dG_n(y) - dG(y)\}. \end{aligned}$$

Application of Lemmas 4.1, 4.2, 4.3 and 4.4 to the terms on the right-hand side gives

$$\lim_{m,n\to\infty} \sup\{\theta_{mn}(h_X, h_Y)^{-1} |R_{mn}(h_X, h_Y)| : (h_X, h_Y) \in H_{mn}\} = 0 \qquad \text{almost surely}$$

and so

$$\lim_{m,n\to\infty} \sup\{\psi_{mn}(h_X, h_Y)^{-1} |R_{mn}(h_X, h_Y)| : (h_X, h_Y) \in H_{mn}\} = 0$$

almost surely. Consequently

$$\lim_{m,n\to\infty} \sup \left\{ \frac{M_{mn}(h_X, h_Y)}{\psi_{mn}(h_X, h_Y)} \left| \frac{\hat{S}_{mn}(h_X, h_Y) + C_{mn}}{M_{mn}(h_X, h_Y)} - 1 \right| : (h_X, h_Y) \in H_{mn} \right\} = 0$$

almost surely. Combining this with (4.21) gives us

$$\lim_{m,n\to\infty} \sup \left\{ \left| \frac{\hat{S}_{mn}(h_X, h_Y) + C_{mn}}{M_{mn}(h_X, h_Y)} - 1 \right| : (h_X, h_Y) \in H_{mn} \right\} = 0 \qquad (4.22)$$

almost surely. Recall that $(\hat{h}_X, \hat{h}_Y)$ is the minimiser of $\hat{S}_{mn}(h_X, h_Y)$ over H_{mn}. Then from (4.22)

$$\lim_{m,n\to\infty}\left[\frac{\hat{S}_{mn}(\hat{h}_X,\hat{h}_Y)+C_{mn}}{M_{mn}(\hat{h}_X,\hat{h}_Y)}\right]=1 \qquad \text{almost surely.} \tag{4.23}$$

Also from (4.22) and Lemma 4.6,

$$\lim_{m,n\to\infty}\left[\frac{\inf\{\hat{S}_{mn}(h_X,h_Y)+C_{mn}:(h_X,h_Y)\in H_{mn}\}}{\inf\{M_{mn}(h_X,h_Y):(h_X,h_Y)\in H_{mn}\}}\right]=1 \qquad \text{almost surely}$$

or equivalently,

$$\lim_{m,n\to\infty}\left[\frac{\hat{S}_{mn}(\hat{h}_X,\hat{h}_Y)+C_{mn}}{\inf\{M_{mn}(h_X,h_Y):(h_X,h_Y)\in H_{mn}\}}\right]=1 \qquad \text{almost surely.} \tag{4.24}$$

Combining (4.23) and (4.24) we obtain

$$\lim_{m,n\to\infty}\left[\frac{M_{mn}(\hat{h}_X,\hat{h}_Y)}{\inf\{M_{mn}(h_X,h_Y):(h_X,h_Y)\in H_{mn}\}}\right]=1 \qquad \text{almost surely.}$$

The result stated at (2.4) can be obtained from this one by using the assumption that K is Hölder continuous. The argument used to establish this is along the same lines as that given in the proof of Theorem 4.2.1. ■

Appendix A

PROOF OF AN L_1 OPTIMALITY RESULT USING THE KOMLÓS-MAJOR-TUSNÁDY BROWNIAN BRIDGE APPROXIMATION

This appendix is devoted to the proof of an asymptotic optimality result for the window-size selection rule proposed in Section 2.4. Using technology completely different to that employed in the proof of Theorem 2.4.1 we shall prove a result similar to that stated at (2.4.8).

Throughout we let

$$f_n(x|h) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i),$$

where K is a pth order kernel as defined in Section 2.2 and $K_h(x) = h^{-1}K(x/h)$. Also let $b_n(x|h) = Ef_n(x|h) - f(x)$ denote the bias at x. Recall that the L_1-optimal window-size for this estimator is asymptotic to

$$h^* = c_1 n^{-1/(2p+1)} \tag{A.1}$$

for some constant $c_1 > 0$, and that the data-base window-size h_n^* defined at (2.4.4) is such that

$$\lim_{n\to\infty} h_n^*/h^* = 1 \tag{A.2}$$

almost surely. Our concern is to prove the following asymptotic optimality result for L_1 loss,

$$J_n(h) = \int |f_n(\cdot|h) - f|.$$

In the proofs we shall use $C_1, C_2, \ldots$ to denote positive generic constants.

Theorem A.1. *If the density f has compact support and p continuous derivatives and the kernel K is of bounded variation and has compact support, then*

$$\lim_{n\to\infty} \{J_n(h_n^*)/J_n(h^*)\} = 1$$

in probability.

Our proof or Theorem is based on a Brownian bridge approximation of the empirical distribution function due to Komlós, Major and Tusnády (1975) and further investigated by M. Csörgő, S. Csörgő, Horváth and Mason (1986). The following lemma makes use of such an approximation.

Lemma A.1. *Suppose that the random sample $X_1, \ldots, X_n$ is drawn from a population having density f and distribution function F for which $E|X_1|^{2+p^{-1}+\epsilon} < \infty$ for some $\epsilon > 0$ and that the kernel K is of bounded variation (that is, $\int |dK| < \infty$). Then*

$$J_n(h^*) = A_n(h^*) + o_p\{n^{-p/(2p+1)}\} \tag{A.3}$$

where

$$A_n(h) = \int_{-\infty}^{\infty} \left| n^{-\frac{1}{2}} h^{-1} \int_{-\infty}^{\infty} W_n^o\{F(x-hz)\} dK(z) + b_n(x|h) \right| dx$$

and $\{W_n^o\}_{n=1}^{\infty}$ is a sequence of Brownian bridges.

Proof. Let $X_{n,1} \leq \cdots \leq X_{n,n}$ be the order statistics of the sample and F_n be the corresponding empirical distribution function. Notice that

$$\begin{aligned} f_n(x|h) &= h^{-1} \int_{-\infty}^{\infty} K\{(x-y)/h\} dF_n(y) \\ &= h^{-1} \int_{-\infty}^{\infty} F_n(x-hz) dK(z), \end{aligned}$$

with the last step involving integration by parts and a change of variable. Therefore

$$f_n(x|h) - Ef_n(x|h) = h^{-1} \int_{-\infty}^{\infty} \{F_n(x-hz) - F(x-hz)\} dK(z). \tag{A.4}$$

Consider a random sample of size n from the uniform-(0,1) distribution for which the order statistics are $U_{n,1} \leq \cdots \leq U_{n,n}$ and the empirical distribution function is G_n. Theorem 2.2 of M.Csörgő *et al* (1986 p.43) asserts that for every $0 \leq \nu < \frac{1}{4}$ and as $n \to \infty$,

$$\sup_{U_{n,1} \leq s \leq U_{n,n}} n^{\nu} |n^{\frac{1}{2}}\{G_n(s) - s\} - W_n^o(s)| \{s(1-s)\}^{\nu - \frac{1}{2}} = O_p(1)$$

for some sequence of Brownian bridges W_n^o. Taking $s = F(x)$ in this equation we obtain, as $n \to \infty$,

$$F_n(x) - F(x) = n^{-\frac{1}{2}} W_n^o\{F(x)\} + O_p(1) n^{-(\nu+\frac{1}{2})} [F(x)\{1-F(x)\}]^{\frac{1}{2}-\nu} \qquad (A.5)$$

for all $0 \le \nu < \frac{1}{4}$ and uniformly in $X_{n,1} \le x \le X_{n,n}$. Let

$$\Lambda(x) = \{z : (x - X_{n,n})/h \le z \le (x - X_{n,1})/h\}.$$

Then

$$\begin{aligned} F_n(x-hz) - F(x-hz) =& n^{-\frac{1}{2}} W_n^o\{F(x-hz)\} \\ &+ O_p(1) n^{-(\nu+\frac{1}{2})} [F(x-hz)\{1-F(x-hz)\}^{\frac{1}{2}-\nu} \end{aligned}$$

uniformly in x, $z \in \Lambda(x)$ and h. Therefore

$$\begin{aligned} \int_{\Lambda(x)} \{F_n(x-hz) - F(x-hz)\}\, dK(z) &= n^{-\frac{1}{2}} \int_{\Lambda(x)} W_n^o\{F(x-hz)\}\, dK(z) \\ &+ O_p(1) n^{-(\nu+\frac{1}{2})} \int_{\Lambda(x)} [F(x-hz)\{1-F(x-hz)\}]^{\frac{1}{2}-\nu}\, dK(z), \end{aligned}$$

so from this and (A.3) it follows that for all $x \in \mathbf{R}$,

$$f_n(x|h) - E f_n(x|h) = n^{-\frac{1}{2}} h^{-1} \int_{-\infty}^{\infty} W_n^o\{F(x-hz)\}\, dK(z) + R_n(x,h) \qquad (A.6)$$

where

$$\begin{aligned} \int_{-\infty}^{\infty} |R_n(x,h)|\, dx \le\ & O_p(1) n^{-(\nu+\frac{1}{2})} h^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F(x-hz)\{1-F(x-hz)\}]^{\frac{1}{2}-\nu} |dK(z)| dx \\ &+ n^{-\frac{1}{2}} h^{-1} \int_{-\infty}^{\infty} \int_{\Lambda(x)^c} |W_n^o\{F(x-hz)\}|\, |dK(z)|\, dx \qquad (A.7) \\ &+ h^{-1} \int_{-\infty}^{\infty} \int_{\Lambda(x)^c} |F_n(x-hz) - F(x-hz)|\, |dK(z)|\, dx. \end{aligned}$$

Assume from now on that

$$\frac{1}{2(2p+1)} < \nu < \frac{1+\epsilon}{2(2p+1+\epsilon p)} < \frac{1}{4} \qquad (A.8)$$

for $0 < \epsilon < 1$. Observe that

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F(x-hz)\{1-F(x-hz)\}]^{\frac{1}{2}-\nu} |dK(z)|\, dx \\ &\quad \le h^{-1} \left\{ \int_{-\infty}^{\infty} |dK(u/h)| \right\} \int_{-\infty}^{\infty} [F(y)\{1-F(y)\}]^{\frac{1}{2}-\nu}\, dy \\ &\quad \le \left(\int |dK| \right) \int_0^{\infty} \{P(|X_1| > y)\}^{\frac{1}{2}-\nu}\, dy \\ &\quad \le \left(\int |dK| \right) \left[1 + E|X_1|^{2+p^{-1}+\epsilon} \int_1^{\infty} y^{-(2+p^{-1}+\epsilon)(\frac{1}{2}-\nu)}\, dy \right] \\ &\quad < \infty \end{aligned}$$

from our assumptions on K and f and since $-(2+p^{-1}+\epsilon)(\frac{1}{2}-\nu) < -1$ from the assumption made at (A.8). Therefore the first term on the right-hand side of (A.7), $T_{n1}(h)$ say, is such that

$$T_{n1}(h) = O_p\{h^{-1}n^{-(\nu+\frac{1}{2})}\}. \tag{A.9}$$

Let $T_{n2}(h)$ and $T_{n3}(h)$ denote, respectively, the second and third terms on the right-hand side of (A.7). To bound $T_{n2}(h)$ we first note that for a Brownian bridge $\{W^\circ(t) : 0 \le t \le 1\}$,

$$|W^\circ(t)| = O_p(1)t^{\frac{1}{2}-\delta} \tag{A.10}$$

uniformly in $0 \le t \le 1$, for any $\delta > 0$. This yields

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{hz<x-X_{n,1}} |W_n^\circ\{F(x-hz)\}||dK(z)|\,dx \\
&\quad = O_p(1)h^{-1}\int_{-\infty}^{\infty}\int_{y<X_{n,1}} F(y)^{\frac{1}{2}-\delta}|dK\{(x-y)/h\}|\,dx \\
&\quad = O_p(1)\left(\int|dK|\right)\int_{y<X_{n,1}} F(y)^{\frac{1}{2}-\delta}\,dy \\
&\quad \le O_p(1)\left(\int|dK|\right)F(X_{n,1})^{\nu}\int_{y<X_{n,1}} F(y)^{\frac{1}{2}-\nu-\delta}\,dy.
\end{aligned}$$

It is straightforward to verify that

$$F(X_{n,1}) = U_{n,1} = O_p(n^{-1})$$

as $n \to \infty$ and that $\int_{y<X_{n,1}} F(y)^{\frac{1}{2}-\nu-\delta}\,dy < \infty$ for a sufficiently small choice of δ. Treating the integral over $\{(x,z) : hz < x - X_{n,n},\ -\infty < x < \infty\}$ in a similar way we obtain

$$T_{n2}(h) = O_p\{h^{-1}n^{-(\nu+\frac{1}{2})}\}. \tag{A.11}$$

Finally,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{hz<x-X_{n,1}} |F_n(x-hz)-F(x-hz)||dK(z)|\,dx \\
&\qquad = h^{-1}\int_{-\infty}^{\infty}\int_{y<X_{n,1}} F(y)|dK\{(x-y)/h\}|\,dx \\
&\qquad \le \left(\int|dK|\right)F(X_{n,1})^{\nu+\frac{1}{2}}\int_{y<X_{n,1}} F(y)^{\frac{1}{2}-\nu}\,dy \\
&\qquad = O_p\{n^{-(\nu+\frac{1}{2})}\}.
\end{aligned}$$

Applying a similar treatment to the integral over $\{(x,z) : hz < x - X_{n,n},\ -\infty < x < \infty\}$ it follows that

$$T_{n3}(h) = O_p\{h^{-1}n^{-(\nu+\frac{1}{2})}\}, \tag{A.12}$$

and so after combining (A.7), (A.9), (A.11) and (A.12) one obtains

$$\int_{-\infty}^{\infty} |R_n(x,h)|\,dx = O_p\{h^{-1}n^{-(\nu+\frac{1}{2})}\}.$$

Putting h equal to $h^* = c_1 n^{-1/(2p+1)}$ gives

$$\int_{-\infty}^{\infty} |R_n(x,h^*)|\,dx = O_p\{n^{1/(2p+1)-(\nu+\frac{1}{2})}\} = o_p\{n^{-p/(2p+1)}\} \tag{A.13}$$

by choice of $\nu > 1/\{2(2p+1)\}$. Now

$$\begin{aligned}
&|J_n(h^*) - A_n(h^*)| \\
&\quad \le \int_{-\infty}^{\infty} \left| f_n(x|h^*) - Ef_n(x|h^*) - n^{-\frac{1}{2}}(h^*)^{-1} \int_{-\infty}^{\infty} W_n^{\circ}\{F(x-h^*z)\}\,dK(z) \right| dx \\
&\quad \le \int_{-\infty}^{\infty} |R_n(x,h^*)|\,dx
\end{aligned}$$

and the required result follows from this and (A.13). ∎

For a Brownian bridge W° we will let ξ be the random function given by

$$\xi(x|h) = h^{-\frac{1}{2}} \int_{-\infty}^{\infty} W^{\circ}\{F(x-hz)\}\,dK(z).$$

The distribution of $\xi(x|h)$ is of interest and the next lemma provides us with this.

Lemma A.2. *For each real number x, the random variable $(nh)^{-\frac{1}{2}}\xi(x|h)$ has a normal distribution with mean zero and variance equal to $Var\{f_n(x|h)\}$. Furthermore,*

$$Cov\{(nh)^{-\frac{1}{2}}\xi(x|h), (nh)^{-\frac{1}{2}}\xi(y|h)\} = Cov\{f_n(x|h), f_n(y|h)\}$$

for all $x, y \in \mathbf{R}$.

Proof. By treating the definite integral as a limit of a sum of areas of rectangles it is apparent that $\xi(x|h)$ is the limit of a sum of normal random variables and so itself is normally distributed. Clearly

$$E\xi(x|h) = h^{-\frac{1}{2}} \int_{-\infty}^{\infty} E[W^{\circ}\{F(x-hz)\}]\,dK(z) = 0$$

which implies that

$$\begin{aligned}\mathrm{Cov}\{\xi(x|h),\xi(y|h)\} =& h^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[W^{\circ}\{F(x-hz)\}W^{\circ}\{F(y-hw)\}]\,dK(z)\,dK(w)\\ =& h^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F\{(x-hz)\wedge(y-hw)\}\,dK(z)\,dK(w)\\ &- h^{-1}\int_{-\infty}^{\infty} F(x-hz)\,dK(z)\int_{-\infty}^{\infty} F(y-hw)\,dK(w) \qquad \text{(A.15)}\end{aligned}$$

where we have used the property

$$E\{W^{\circ}(s)W^{\circ}(t)\} = s(1-t), \qquad s \le t$$

of a Brownian bridge W°. It is well-known that $E\{F_n(x)\} = F(x)$ and

$$\mathrm{Cov}\{F_n(s),F_n(t)\} = n^{-1}F(s)\{1-F(t)\}; \qquad s \le t.$$

From these results it follows that the expression at (A.15) is equal to

$$\begin{aligned}nh\Big[E\Big\{&\int_{-\infty}^{\infty} h^{-1}F_n(x-hz)\,dK(z)\int_{-\infty}^{\infty} h^{-1}F_n(y-hw)\,dK(w)\Big\}\\ &-\int_{-\infty}^{\infty} h^{-1}F(x-hz)\,dK(z)\int_{-\infty}^{\infty} h^{-1}F(y-hw)\,dK(w)\Big]\\ &= nh[E\{f_n(x|h)f_n(y|h)\} - \{Ef_n(x|h)\}\{Ef_n(y|h)\}]\\ &= nh\mathrm{Cov}\{f_n(x|h),f_n(y|h)\},\end{aligned}$$

giving rise to

$$\mathrm{Cov}\{(nh)^{-\frac{1}{2}}\xi(x|h),(nh)^{-\frac{1}{2}}\xi(y|h)\} = \mathrm{Cov}\{f_n(x|h),f_n(y|h)\}.$$

In particular

$$\mathrm{Var}\{(nh)^{-\frac{1}{2}}\xi(x|h)\} = \mathrm{Var}\{f_n(x|h)\},$$

completing the proof. ∎

The last lemma we need concerns the random variable $G(h)$ given by

$$G(h) = \int_{-\infty}^{\infty} |(nh)^{-\frac{1}{2}}\xi(x|h) + h^p b_x|\,dx$$

where $b_x \equiv (\kappa_1/p!)f^{(p)}(x)$. Note that $G(h)$ is simply the expression for $A(h)$ with the bias $b_n(x|h)$ replaced by the leading term in its asymptotic expansion.

Lemma A.3. *Suppose that both f and K are bounded with compact support and that $f^{(p)}$ is continuous and bounded. Then*

$$Var\{G(h^*)\} = o\{n^{-2p/(2p+1)}\}.$$

Proof. Since f and K each have compact support we may assume that for some $r > 0$,

$$G(h) = \int_{-r}^{r} |(nh)^{-\frac{1}{2}}\xi(x|h) + h^p b_x|\, dx.$$

Let $T \equiv [-r, r]^2$ and define for all $\delta > 0$,

$$U_\delta \equiv \{(x, y) \in T : |x - y| \leq \delta\}.$$

Also we will write $v_x^2(h) \equiv \text{Var}\{f_n(x|h)\}$ and $N_x \equiv v_x(h)^{-1}(nh)^{-\frac{1}{2}}\xi(x|h)$, so that from Lemma A.2 N_x is an N(0,1) random variable. Consider the bivariate normal random pair (N_x, N_y) where $(x, y) \in T - U_\delta$. From Lemma A.2,

$$\rho_{xy}(h) \equiv \text{Cov}(N_x, N_y) = \{v_x(h)v_y(h)\}^{-1}\text{Cov}\{f_n(x|h), f_n(y|h)\}.$$

Simple algebra leads to

$$\begin{aligned}\text{Cov}\{f_n(x|h), f_n(y|h)\} =& -n^{-1}\{Ef_n(x|h)\}\{Ef_n(y|h)\} \\ &+ n^{-1}h^{-2}E[K\{(x - X_1)/h\}K\{(y - X_1)/h\}].\end{aligned}$$

Suppose that the support of K is contained in the interval $[-s, s]$ and let $h \leq \delta(2s)^{-1}$. Then $K\{(x-X_1)/h\}K\{(y-X_1)/h\} \equiv 0$ implying that, for $x, y \in T - U_\delta$ and small h,

$$\rho_{xy}(h) = \frac{-Ef_n(x|h)Ef_n(y|h)}{nv_x(h)v_y(h)}.$$

Let N_y^o be another N(0,1) random variable which is independent of N_y. It is trivial to show that the distribution of (N_x, N_y) is identical to that of $((1 - \rho_{xy}^2)^{\frac{1}{2}}N_y^o + \rho_{xy}N_y, N_y)$. The variance of G can then be expanded as follows:

$$\begin{aligned}\text{Var}\{G(h)\} =& E\{G(h)^2\} - \{EG(h)\}^2 \\ =& \int\int_T E\{|v_x(h)N_x + h^p b_x||v_y(h)N_y + h^p b_y|\}\, dx\, dy \\ &- \left\{\int_{-r}^{r} E|v_x(h)N_x + h^p b_x|\, dx\right\}^2 \\ =& I_1(\delta, h) + I_2(\delta, h) + I_3(h, \delta) + I_4(\delta, h)\end{aligned}$$

where

$$I_1(\delta,h) = \int\int_{U_\delta} E\{|v_x(h)N_x + h^p b_x||v_y(h)N_y + h^p b_y|\}\,dx\,dy\,,$$

$$I_2(\delta,h) = \int\int_{T-U_\delta} E\Big(\Big[|v_x(h)N_x + h^p b_x| - |\{1-\rho_{xy}^2(h)\}^{\frac{1}{2}} v_x(h)N_y^o + h^p b_x|\Big] \times |v_y(h)N_y + h^p b_y|\Big)\,dx\,dy,$$

$$I_3(\delta,h) = \int\int_{T-U_\delta} E|v_y(h)N_y + h^p b_y| \times E\Big[|\{1-\rho_{xy}^2(h)\}^{\frac{1}{2}} v_x(h)N_y^o + h^p b_x| - |v_x(h)N_x + h^p b_x|\Big]\,dx\,dy,$$

and

$$I_4(\delta,h) = -\int\int_{U_\delta} \{E|v_x(h)N_x + h^p b_x|\}\{E|v_y(h)N_y + h^p b_y|\}\,dx\,dy.$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} I_1(\delta,h^*) &\le \sup_{x\in[-r,r]} E[\{v_x(h^*)N_x + (h^*)^p b_x\}^2] \int\int_{U_\delta} dx\,dy \\ &\le 5r\delta \sup_{x\in[-r,r]} [\mathrm{Var}\{f_n(x|h^*)\} + (h^*)^{2p} b_x^2] \\ &\le C\delta \sup_{x\in[-r,r]} \left\{\int_{-s}^{s} K(z)^2 f(x-h^*z)\,dz + b_x^2\right\} n^{-2p/(2p+1)} \\ &\le C_1 \delta n^{-2p/(2p+1)} \end{aligned} \tag{A.16}$$

for some constant $C_1 > 0$ not depending on either n or δ, by virtue of our assumptions on f and K. Noting that the random variable $\{v_x(h)N_x + h^p b_x\}\{v_y(h)N_y + h^p b_y\}$ has the same distribution as

$$\Big[\{1-\rho_{xy}^2(h)\}^{\frac{1}{2}} v_x(h)N_y^o + h^p b_x\Big]\{v_y(h)N_y + h^p b_y\} + \rho_{xy}(h)v_x(h)N_y\{v_y(h)N_y + h^p b_y\}$$

we obtain the bound

$$|I_2(\delta,h^*)| \le \int\int_{T-U_\delta} |\rho_{xy}(h^*)v_x(h^*)v_y(h^*)|\,dx\,dy + (2/\pi)^{\frac{1}{2}} \int\int_{T-U_\delta} |\rho_{xy}(h^*)v_x(h^*)(h^*)^p b_y|\,dx\,dy.$$

For large n the first integral on the right-hand side is equal to

$$\int\int_{T-U_\delta} |E\{f_n(x|h^*)\}E\{f_n(y|h^*)\}|\,dx\,dy\,n^{-1} \le C_2 n^{-1} \tag{A.17}$$

where $C_2 > 0$ does not depend on n. Also for large n, the second integral equals

$$(2/\pi)^{\frac{1}{2}} \int\int_{T-U_\delta} |E\{f_n(x|h^*)\}E\{f_n(y|h^*)\}(h^*)^p b_y v_y(h^*)^{-1}|\, dx\, dy\, n^{-1}$$
$$\leq C_3(h^*)^p \int_{-r}^{r} v_y(h^*)^{-1}\, dy\, n^{-1}$$
$$\leq C_4 n^{-1}$$

since $h^* = O\{n^{-1/(2p+1)}\}$ implies that $v_y(h^*) = O\{(h^*)^p\}$. Combining this with (A.17) we obtain for some $C_5 > 0$ and large n,

$$|I_2(\delta, h^*)| \leq C_5 n^{-1}. \tag{A.18}$$

Recalling that the function ψ introduced in Section 2.2 has the property $\psi(t) = E|N - t|$ for a N(0,1) random variable N we see that

$$|I_3(\delta, h^*)| \leq \int\int_{T-U_\delta} v_y(h^*)\psi\left(\frac{(h^*)^p b_y}{v_y(h^*)}\right)$$
$$\times \left| \{1 - \rho_{xy}^2(h^*)\}^{\frac{1}{2}} v_x(h^*)\psi\left(\frac{(h^*)^p b_x}{\{1 - \rho_{xy}^2(h^*)\}^{\frac{1}{2}} v_x(h^*)}\right) \right.$$
$$\left. - v_x(h^*)\psi\left(\frac{(h^*)^p b_x}{v_x(h^*)}\right) \right| dx\, dy$$
$$\leq (2/\pi)^{\frac{1}{2}} \int\int_{T-U_\delta} v_y(h^*)\psi\left(\frac{(h^*)^p b_y}{v_y(h^*)}\right) v_x(h^*)|1 - \{1 - \rho_{xy}^2(h^*)\}^{\frac{1}{2}}|\, dx\, dy$$
(by Lemma 2.2.1)
$$\leq (2/\pi)^{\frac{1}{2}} \left\{ \int\int_{T-U_\delta} v_y^2(h^*)\psi\left(\frac{(h^*)^p b_y}{v_y(h^*)}\right)^2 dx\, dy \right\}^{\frac{1}{2}}$$
$$\times \left(\int\int_{T-U_\delta} v_x^2(h^*)\left[1 - \{1 - \rho_{xy}^2(h^*)\}^{\frac{1}{2}}\right]^2 dx\, dy \right)^{\frac{1}{2}}.$$

By our assumptions on f and K, the second factor is bounded by

$$C_6 n^{-p/(2p+1)} \left(\int\int_{T-U_\delta} [1 - \{1 - \rho_{xy}^2(h^*)\}^{\frac{1}{2}}]^2\, dx\, dy \right)^{\frac{1}{2}}.$$

For all sufficiently large n,

$$\rho_{xy}^2(h^*) \leq C_7 n^{-2}[\mathrm{Var}\{f_n(x|h^*)\}\mathrm{Var}\{f_n(y|h^*)\}]^{-1} = o(1)$$

uniformly in $(x, y) \in T - U_\delta$, since $\mathrm{Var}\{f_n(x|h^*)\} = O\{n^{-2p/(2p+1)}\}$. It follows from this and the above bound that the second factor is $o\{n^{-p/(2p+1)}\}$ as $n \to \infty$. The first factor is no more than

$$(4r/\pi)^{\frac{1}{2}} \left\{ \int_{-r}^{r} v_y^2(h^*)\psi\left(\frac{(h^*)^p b_y}{v_y(h^*)}\right)^2 dy \right\}^{\frac{1}{2}}.$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ and writing $\sigma_y = \kappa_2 f(y)^{\frac{1}{2}}$ we obtain

$$\begin{aligned}
&\int_{-r}^{r} v_y^2(h^*)\psi\left(\frac{(h^*)^p b_y}{v_y(h^*)}\right)^2 dy \\
&\qquad \leq 2\int_{-r}^{r}\left\{v_y(h^*)\psi\left(\frac{(h^*)^p b_y}{v_y(h^*)}\right) - (nh^*)^{-\frac{1}{2}}\sigma_y\psi\left(\frac{(h^*)^p b_y}{(nh^*)^{-\frac{1}{2}}\sigma_y}\right)\right\}^2 dy \\
&\qquad\quad + 2c_1^{-1}n^{-2p/(2p+1)}\int_{-r}^{r}\sigma_y^2\psi\left(\frac{c_1^{p+\frac{1}{2}}b_y}{\sigma_y}\right)^2 dy \\
&\qquad \leq 2\int_{-r}^{r}\left\{(2/\pi)^{\frac{1}{2}}|v_y(h^*) - (nh^*)^{-\frac{1}{2}}\sigma_y|\right\}^2 dy \\
&\qquad\quad + 2c_1^{-1}n^{-2p/(2p+1)}\int_{-r}^{r}\sigma_y^2\psi\left(\frac{c_1^{p+\frac{1}{2}}b_y}{\sigma_y}\right)^2 dy \\
&\qquad = O\{n^{-2p/(2p+1)}\}
\end{aligned}$$

since $|v_y(h^*) - (nh^*)^{-\frac{1}{2}}\sigma_y| = o\{n^{-p/(2p+1)}\}$ uniformly in $y \in [-r,r]$. Therefore, on combining the orders of magnitude of both factors, we obtain

$$|I_3(\delta, h^*)| = o\{n^{-2p/(2p+1)}\}. \tag{A.19}$$

Using the same technique that was applied to $I_1(\delta, h^*)$ we may show that there is a constant $C_8 > 0$ such that for all $\delta > 0$,

$$|I_4(\delta, h^*)| \leq C_8\delta n^{-2p/(2p+1)}. \tag{A.20}$$

Combining (A.16), (A.18), (A.19) and (A.20) we obtain for every $\eta > 0$,

$$\limsup_{n\to\infty} n^{2p/(2p+1)}\mathrm{Var}\{G(h^*)\} \leq \eta,$$

which immediately leads to the required result. ∎

Proof of Theorem A.1.

With $b_x \equiv (\kappa_1/p!)f^{(p)}(x)$, $\sigma_x \equiv \kappa_2 f(x)^{\frac{1}{2}}$ and $\alpha > 0$ we shall put

$$D(f, K, \alpha) \equiv \alpha^{-\frac{1}{2}}\int_{-\infty}^{\infty}\sigma_x\psi\left(\frac{\alpha^{p+\frac{1}{2}}b_x}{\sigma_x}\right) dx$$

which is finite since f has compact support. Theorem 5.1 of Devroye and Györfi (1985, p.78) can easily be extended of the case of a pth order kernel to give

$$E\{J_n(h)\} = \int_{-\infty}^{\infty}(nh)^{-\frac{1}{2}}\sigma_x\psi\left(\frac{(nh^{2p+1})^{\frac{1}{2}}b_x}{\sigma_x}\right) dx + o\{h^p + (nh)^{-\frac{1}{2}}\},$$

producing the result

$$E\{J_n(h^*)\} = n^{-p/(2p+1)} D(f, K, c_1) + o\{n^{-p/(2p+1)}\}. \tag{A.21}$$

We shall first prove that

$$\lim_{n\to\infty} [J_n(h^*)/E\{J_n(h^*)\}] = 1 \tag{A.22}$$

in probability. In view of (A.21) this result will hold if and only if it can be shown that

$$|J_n(h^*) - E\{J_n(h^*)\}| = o_p\{n^{-p/(2p+1)}\}. \tag{A.23}$$

The left-hand side of (A.23) is dominated by

$$|J_n(h^*) - G(h^*)| + |G(h*) - E\{G(h^*)\}| + |E\{G(h^*)\} - E\{J_n(h^*)\}| \tag{A.24}$$

so it suffices to prove that the first two terms are each $o_p\{n^{-p/(2p+1)}\}$ and the third is $o\{n^{-p/(2p+1)}\}$.

Let $A(h)$ be given by

$$A(h) = \int_{-\infty}^{\infty} |(nh)^{-\frac{1}{2}} \xi(x|h) + b_n(x|h)|\, dx.$$

Lemma A.1 asserts that $|J_n(h^*) - A(h^*)| = o_p\{n^{-p/(2p+1)}\}$ since the elements of the sequence of Brownian bridges $\{W_n^o\}$ each have the same distribution as W^o used in the definition of $\xi(x|h)$. Therefore to prove that $|J_n(h^*) - G(h^*)| = o_p\{n^{-p/(2p+1)}\}$ it is enough to show that $|A(h^*) - G(h^*)| = o_p\{n^{-p/(2p+1)}\}$. Notice that

$$|A(h^*) - G(h^*)| \le \int_{-\infty}^{\infty} |b_n(x|h) - (h^*)^p b_x|\, dx$$

so by the compact support of f, the right-hand side is $o\{n^{-p/(2p+1)}\}$ by standard results for the asymptotic bias of $f_n(x|h)$ (see, e.g., Devroye and Györfi (1985 p.92)). For the second term we use Chebyshev's inequality to obtain for all $\epsilon > 0$,

$$P\{n^{p/(2p+1)}|G(h^*) - E\{G(h^*)\}| > \epsilon\} \le n^{2p/(2p+1)} \mathrm{Var}\{G(h^*)\}\epsilon^{-2}$$

$$= o(1)$$

from Lemma A.3. This proves that $|G(h^*) - E\{G(h^*)\}| = o_p\{n^{-p/(2p+1)}\}$. The last term of (A.24) is dominated by

$$\int_{-\infty}^{\infty} \left| (nh^*)^{-\frac{1}{2}} \sigma_x \psi \left(\frac{\{n(h^*)^{2p+1}\}^{\frac{1}{2}} b_x}{\sigma_x} \right) - v_x(h^*) \psi \left(\frac{(h^*)^p b_x}{v_x(h^*)} \right) \right| dx$$
$$+ |E\{J_n(h^*)\} - n^{-p/(2p+1)} D(f, K, c_1)|.$$

The second term is $o\{n^{-p/(2p+1)}\}$ since $E\{J_n(h^*)\} \sim n^{-p/(2p+1)} D(f, K, c_1)$. From Lemma 2.2.1 the first term is dominated by

$$(2/\pi)^{\frac{1}{2}} \int_{-\infty}^{\infty} |(nh^*)^{-\frac{1}{2}} \sigma_x - [\mathrm{Var}\{f_n(x|h^*)\}]^{\frac{1}{2}}| \, dx$$

which is also $o\{n^{-p/(2p+1)}\}$ since $[\mathrm{Var}\{f_n(x|h^*)\}]^{\frac{1}{2}} \sim c_1^{-\frac{1}{2}} n^{-p/(2p+1)} \sigma_x$ for all x and f has compact support. Hence $|E\{G(h^*)\} - E\{J_n(h^*)\}| = o\{n^{-p/(2p+1)}\}$ as had to be shown.

The result stated in the theorem follows from (A.22) and

$$\lim_{n \to \infty} [J_n(h_n^*)/E\{J_n(h^*)\}] = 1$$

in probability. This result can be proved in the same way as (A.22) using the fact that

$$\lim_{n \to \infty} n^{1/(2p+1)} h_n^* = c_1$$

in probability. ■

With some extra work our assumption of f having compact support can be weakened to the existence of a moment of order $2 + p^{-1} + \epsilon$ for some $\epsilon > 0$. This has not been done since it would produce a result which is still slightly weaker than that stated in Section 2.4.

Appendix B

LEAST-SQUARES CROSS-VALIDATION FOR NONPARAMETRIC ESTIMATION OF DENSITY DERIVATIVES

The problem of selecting a window-size for the estimation of density derivatives is briefly addressed in this appendix. Suppose we have a sample $X_1, \ldots, X_n$ of independent, real-valued random variables having common density f where f is r times differentiable. A kernel based estimator for $f^{(r)}(x)$ is

$$f_n^{(r)}(x|h) = n^{-1}h^{-r-1}\sum_{i=1}^{n} K^{(r)}\{(x - X_i)/h\}$$

(see Bhattacharya (1967)) where K is a r times differentiable kernel and h is the window-size. This estimator is, of course, obtained by differentiating the usual kernel density estimator r times. The window-size can be selected for the estimation of $f^{(r)}$ by generalising least squares cross-validation as follows. First note that minimisation of L_2 loss $M_n(h) = \int\{f_n^{(r)}(\cdot|h) - f^{(r)}\}^2$ is equivalent to minimisation of

$$M_n(h) - \int\{f^{(r)}\}^2 = \int f_n^{(r)}(\cdot|h)^2 - 2\int f_n^{(r)}(\cdot|h)f^{(r)}.$$

The first term on the right-hand side is known. However, the second involves the unknown function $f^{(r)}$. To overcome this we observe that

$$-2\int f_n^{(r)}(\cdot|h)f^{(r)} = -2(-1)^r\int f_n^{(2r)}(\cdot|h)f$$

from integration by parts, where it is now assumed that K has $2r$ derivatives available. The mean of the right-hand side can be estimated unbiasedly by

$$-2(-1)^r\{n(n-1)\}^{-1}h^{-2r-1}\sum\sum_{i\neq j}K^{(2r)}\{(X_i - X_j)/h\}$$

which leads to

$$\mathrm{CV}(h) = \int f_n^{(r)}(\cdot|h)^2 - 2(-1)^r n^{-2}h^{-2r-1}\sum\sum_{i\neq j}K^{(2r)}\{(X_i - X_j)/h\}$$

as the least-squares cross-validatory criterion to be minimised. The selected window-size is the value of h at which this minimum is attained; we shall call it $\hat{h}_n$. Under certain assumptions on the smoothness of f and K and the range of admissible values of h Härdle, Marron and Wand (1989) have shown that this selection rule is asymptotically optimal in terms of minimising L_2 loss.

The calculation of $\mathrm{CV}(h)$ in practice involves evaluation of $\int f_n^{(r)}(\cdot|h)^2$. Assuming that K is symmetric we obtain

$$\begin{aligned}\int f_n^{(r)}(\cdot|h)^2 &= n^{-2}h^{-2r-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K^{(r)}\{(x-X_i)/h\}K^{(r)}\{(x-X_j)/h\}\,dx \\ &= n^{-2}h^{-2r-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(K^{(r)}*K^{(r)})\{(X_i-X_j)/h\}.\end{aligned}$$

The evaluation of the convolution $K^{(r)}*K^{(r)}$ may be simplified by taking K to be ϕ where $\phi(z)=(2\pi)^{-\frac{1}{2}}e^{-z^2/2}$ is the Gaussian kernel. For $r=0,1,\ldots$ it may be readily established by induction that

$$(\phi^{(r)}*\phi^{(r)})(x) = 2^{-r-\frac{1}{2}}\phi^{(2r)}(x/2^{\frac{1}{2}}),$$

so that the cross-validatory criterion can be explicitly evaluated as

$$\begin{aligned}\mathrm{CV}(h) = n^{-2}h^{-2r-1}(-1)^r\Bigg[&2^{-r-\frac{1}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{(2r)}\{(X_i-X_j)/(2^{\frac{1}{2}}h)\} \\ &-2\sum\sum\nolimits_{i\neq j}\phi^{(2r)}\{(X_i-X_j)/h\}\Bigg].\end{aligned}$$

A small simulation was run to test the efficacy of the proposed window-size selection rule. Ten samples of size 500 were drawn from the extreme value density ($f(x)=e^xe^{-e^x}$) with the aim of estimating the first derivative of this density,

$$f'(x) = (1-e^x)e^xe^{-e^x},$$

using the Gaussian kernel estimator. The optimal window-size in this case is asymptotic to

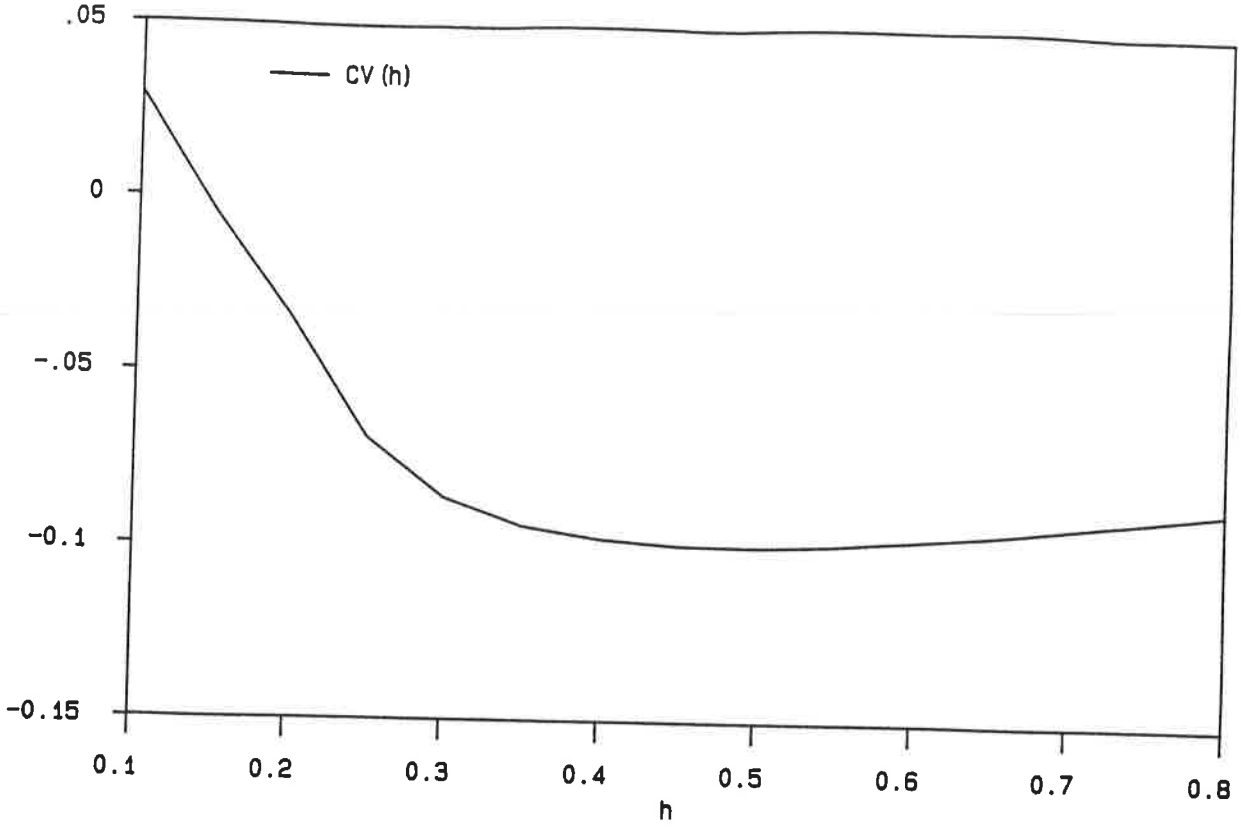
$$h^* = \left[\frac{12}{17\pi^{\frac{1}{2}}}\right]^{1/7}n^{-1/7}$$

which assumes the value $h^*\doteq 0.3608$ when $n=500$.

Table 1 displays the selected window-size and the L_2 error of the corresponding estimator for each replication. In Figure 1 we have plotted CV(h) and the estimate of f' based the selected window-size $\hat{h}_{500}$ for two of the samples from Table 1. Graphs (a) and (b) pertain to replication 5 while graphs (c) and (d) pertain to replication 9. These particular samples were chosen to depict "average case" performance of the density estimators, since they produced the fifth and sixth lowest realisations of $M_{500}(\hat{h}_{500})$ out of the ten replications.

Table 1: Values of $\hat{h}_{500}$ and $M_{500}(\hat{h}_{500})$ for extreme value data (10 replications).

Rep. no.	$\hat{h}_{500}$	$M_{500}(\hat{h}_{500})$
1	0.5499	0.0024
2	0.4826	0.0036
3	0.5237	0.0041
4	0.4277	0.0063
5	0.4808	0.0025
6	0.4550	0.0022
7	0.2821	0.0138
8	0.4650	0.0013
9	0.4740	0.0035
10	0.4860	0.0017



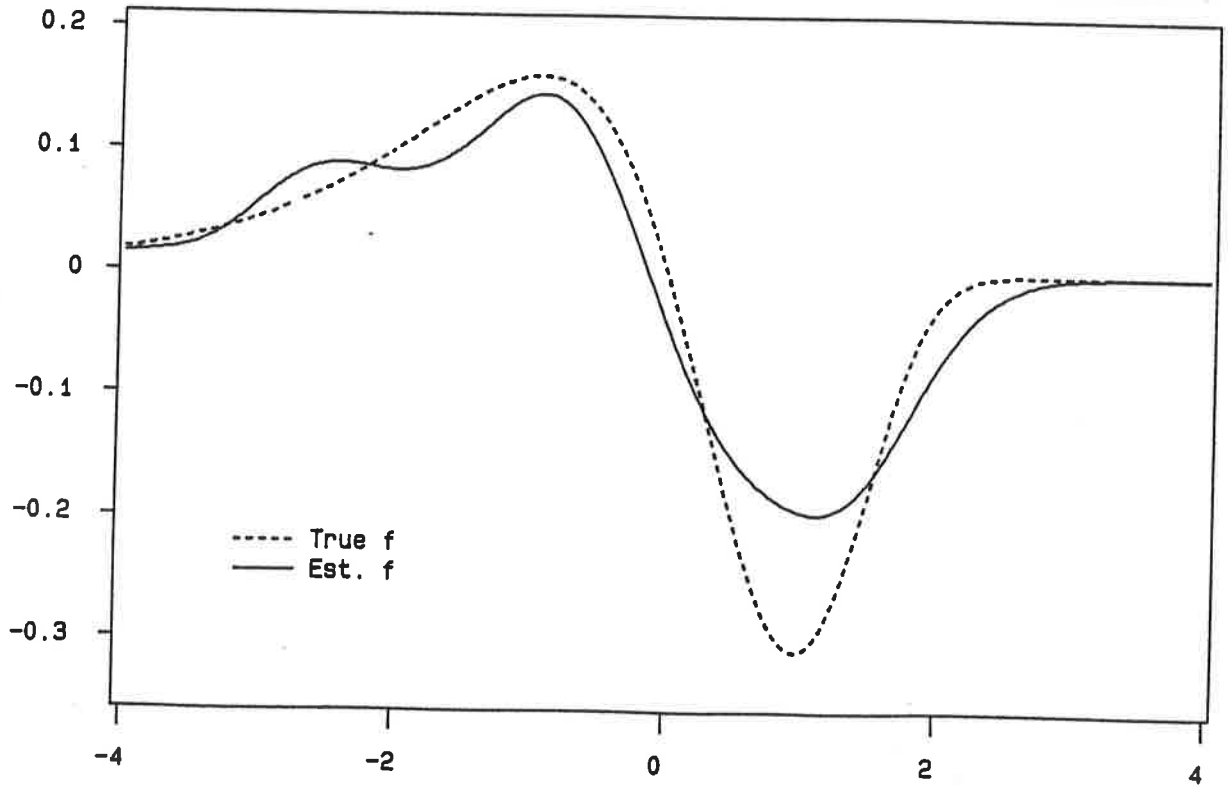
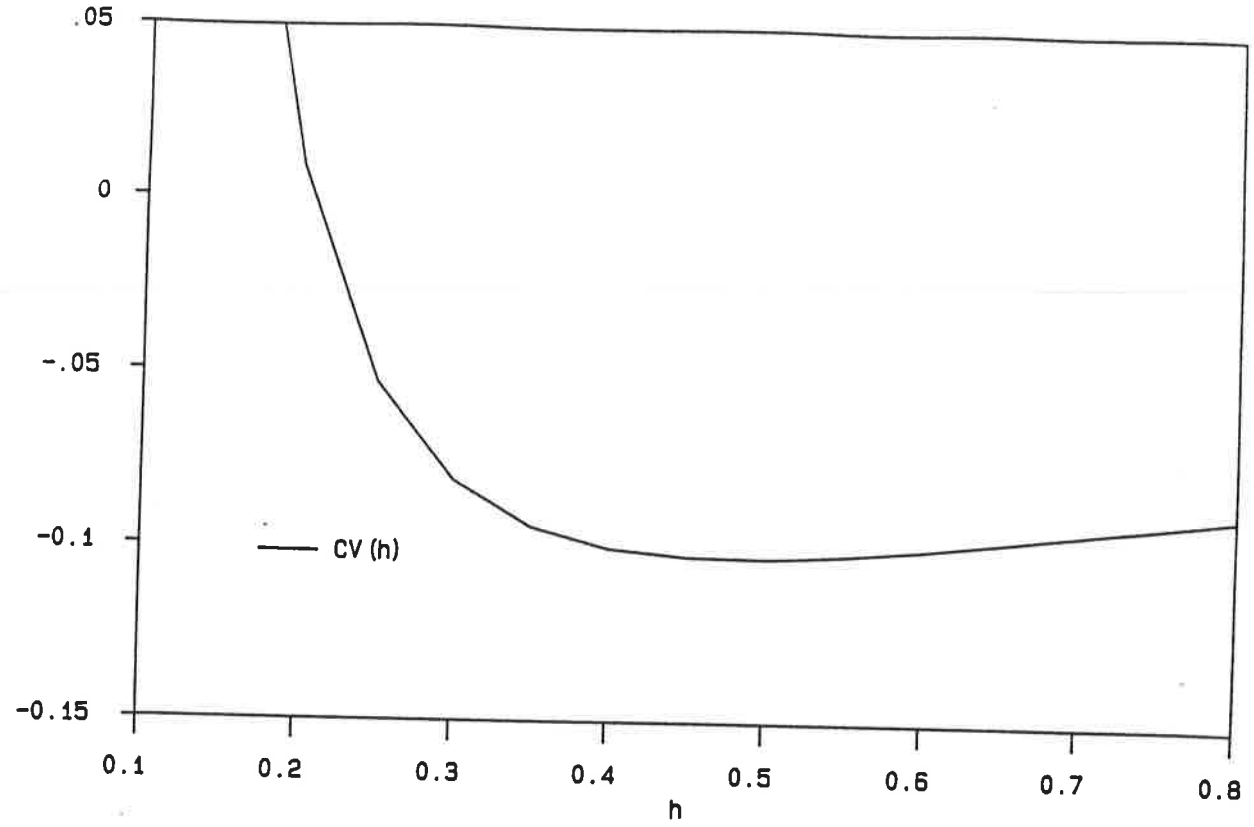


Figure 1 (a) and (b): Typical least squares cross-validatory score function $\mathrm{CV}(h)$ and estimate of the first derivative of the extreme value density with window-size chosen by minimising $\mathrm{CV}(h)$. The curve in (a) is $\mathrm{CV}(h)$ based on a sample of size 500 of extreme value data. In (b) the broken curve is f'; the unbroken curve is $f'_{500}(\cdot|\hat{h}_{500})$ where $\hat{h}_{500} = 0.4808$.



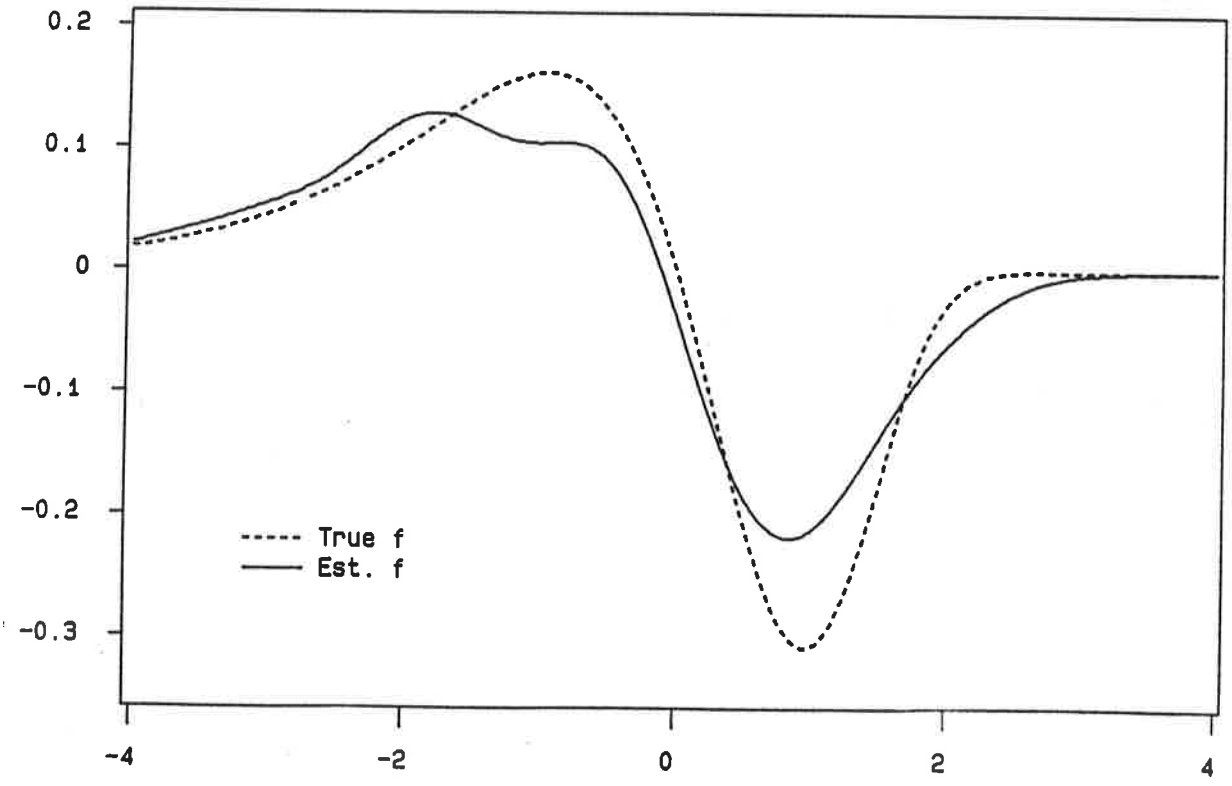


Figure 1 (c) and (d): Typical least squares cross-validatory score function CV(h) and estimate of the first derivative of the extreme value density with window-size chosen by minimising CV(h). The curve in (c) is CV(h) based on a sample of size 500 of extreme value data. In (d) the broken curve is f'; the unbroken curve is $f'_{500}(\cdot|\hat{h}_{500})$ where $\hat{h}_{500} = 0.4740$.

REFERENCES

AITCHISON, J. and AITKEN, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–20.

ANDERSON, J.A., WHALEY, K., WILLIAMSON, J. and BUCHANAN, W.W. (1972). A statistical aid to the diagnosis of keratoconjuctivitis sicca. *Quart. J. Med.*, **41**, 175–89.

ANDREWS, D.F. and HERZBERG, A.M. (1985). *Data.* Springer, New York.

BEAN S.J. and TSOKOS, C.P. (1980). Developments in nonparametric density estimation. *Int. Stat. Rev.*, **48**, 267–87.

BHATTACHARYA, P.K. (1967). Estimation of a probability density and its derivatives. *Sankhyā A*, **29**, 373–82.

BOWMAN, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–60.

BOWMAN, A.W., HALL, P. and TITTERINGTON, D.M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, **71**, 341–51.

BROWN, P.J. and RUNDELL, P.W.K. (1985). Kernel estimates for categorical data. *Technometrics*, **27**, 293–9.

COLLOMB, G. (1985). Nonparametric regression: An up-to-date bibliography. *Math. Oper. Statist.*, **16**, 297–307.

CÖRGŐ, M., CÖRGŐ, S., HORVÁTH, L. and MASON, D.M. (1986). Weighted empirical and quantile processes. *Ann. Prob.*, **14**, 31–85.

DEVROYE, L. (1988). The kernel estimate is relatively stable. *Prob. Theor. Related Fields*, **77**, 521–536.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

EDDY, W.F. (1980). Optimum kernel estimators of the mode. *Ann. Statist.*, **8**, 870–82.

FREEDMAN, D. and DIACONIS, P. (1981). On the histogram as a density estimator: L_2 Theory. *Z. Wahrsch. verw. Gebiete*, **57**, 453–76.

FRYER, M.J. (1977). A review of some nonparametric methods of density estimation. *J. Inst. Maths. Applics.*, **20**, 335–54.

GASSER, T. and MÜLLER, H.G. (1979). Kernel estimation of regression func-

tions. In *Smoothing Techniques for Curve Estimation* (ed T. Gasser and M. Rosenblatt), 23-68. Springer, Heidelberg.

HALL, P. and HEYDE, C.C. (1980). *Martingale Limit Theory and Its Application.* Academic, New York.

HALL, P. and WAND, M.P. (1988). On the minimization of absolute distance in kernel density estimation. *Statist. Prob. Lett.*, **6**, 311–4.

HAND, D.J. (1981). *Discrimination and Classification.* Wiley, Chichester.

HAND, D.J. (1982). *Kernel Discriminant Analysis.* Research Studies Press, Chichester.

HÄRDLE, W. and MARRON, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465–81.

HÄRDLE, W., MARRON, J.S. and WAND, M.P. (1989). Bandwidth choice for density derivatives. *J. R. Statist. Soc. B*, to appear.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13-30.

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent random variables, and the sample distribution function. *Z. Wahrsch. verw. Gebiete.*, **32**, 111–31.

MARRON, J.S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Econ.*, to appear.

NADARAYA, E.A. (1964). On estimating regression. *Theory Prob. App.*, **9**, 141–2.

PARZEN, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statist.*, **35**, 1065–76.

PRAKASA RAO, B.L.S. (1983). *Nonparametric Functional Estimation.* New York: Academic Press.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–7.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, **9**, 65–78.

SCHUCANY, W.R. (1989). Locally optimal window widths for kernel density estimation with large samples. *Statist. Prob. Lett.*, to appear.

SCOTT, D.W., TAPIA, R.A. and THOMPSON, J.R. (1977). Kernel density

estimation revisited. *Nonlinear Anal.*, **1**, 339–72.

SCOTT, D.W. (1979). On optimal and data-based histograms. *Biometrika*, **66**, 605–10.

SCOTT, D.W. (1985). Frequency polygons: Theory and application. *J. Amer. Statist. Assoc.*, **80**, 348–54.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285–97.

TAPIA, R.A. and THOMPSON, J.R. (1978). *Nonparametric Probability Density Estimation.* John Hopkins University Press, Baltimore.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā A*, **26**, 359–72.

WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.*, **41**, 1665–71.