JOURNAL OF MULTIVARIATE ANALYSIS 26, 59–88 (1988)

Minimizing L_1 Distance in Nonparametric Density Estimation

PETER HALL AND MATTHEW P. WAND

Australian National University, Canberra, Australia

Communicated by P. Révész

We construct a simple algorithm, based on Newton's method, which permits asymptotic minimization of L_1 distance for nonparametric density estimators. The technique is applicable to multivariate kernel estimators, multivariate histogram estimators, and smoothed histogram estimators such as frequency polygons. It has an "adaptive" or "data-driven" version. We show theoretically that both theoretical and adaptive forms of the algorithm do indeed minimize asymptotic L_1 distance. Then we apply the algorithm to derive concise formulae for asymptotically optimal smoothing parameters. We also give numerical examples of applications of the adaptive algorithm.

1. INTRODUCTION

The "L_1 view" of nonparametric density estimation has recently received considerable attention, due in large part to the monograph by Devroye and Györfi [5]. These authors give a particularly lucid exposition of the mathematical attractions of L_1 distance: it is always well-defined as a metric on the space of density functions; it is invariant under monotone transformations; and it is proportional to the total variation metric. Devroye and Györfi point out that there are technical difficulties associated with L_1 optimality, and they circumvent these by working with upper and lower bounds to L_1 distance. In the present paper we work directly with an asymptotic formula for L_1 distance. We produce a simple, rapidly converging, iterative algorithm which permits minimization of L_1 distance, in both theoretical and adaptive ("data-driven") approaches to the problem. This leads to a practical procedure for asymptotic minimization of L_1 loss, which we illustrate numerically.

Received February 13, 1987; revised September 7, 1987.

Key words and phrases: asymptotic optimality, histogram estimator, kernel estimator, L_1 distance, nonparametric density estimator.

0047-259X/88 $3.00

In more detail, our main contributions are as follows.

(i) We derive a general asymptotic expression for the L_1 distance between a density f and its estimate $\hat{f}$, and show that minimization of this formula is equivalent to solving an equation $L(v) = 0$ $(v > 0)$, where L is a readily computable, strictly increasing, continuous function with $L(0) = -\infty$ and $L(+\infty) > 0$. The equation is easily solved by Newton's method, and so minimization of L_1 distance reduces to a simple, rapidly converging iteration (Section 2).

(ii) We illustrate the versatility of result (i) by giving versions of the function L for general kernel estimators, for histogram estimators, and for smoothed histogram estimators in $d \geqslant 1$ dimensions. Other cases, such as histospline estimators, may be treated in an identical manner (Section 2). Applying the iteration argument in (i), we derive formulae for theoretical asymptotically L_1-optimal window sizes in the case of kernel estimators (Section 4).

(iii) We develop an adaptive, strongly consistent version of the procedure described in (i). This amounts to constructing an "estimate" $\hat{L}$ of the function L, involving only the data and not the unknown density f. The estimate $\hat{L}$ has the same basic properties as L, and so the equation $\hat{L}(v) = 0$ is easily solved via a rapidly converging iteration. We prove that the resulting adaptive density estimator asymptotically minimizes L_1 distance (Section 3). A numerical illustration shows that the procedure is practicable (Section 4).

Section 2 describes and discusses our general approach to minimizing L_1 distance, Section 3 introduces adaptive techniques, and Section 4 summarizes numerical results. All proofs are deferred to Section 5. We know of no other work which computes explicit formulae for minimum L_1 distance or for asymptotically optimal smoothing parameters. However, formulae are available in the case of L_2 loss; see Prakasa Rao [11, Chaps. 2 and 3]. Our adaptive methods for minimizing L_1 distance are distantly related to those suggested by Woodroofe [16], in that both are based on "plug-in" rules. A further account of minimization of L_1 distance, including application to histogram estimators such as that studied by Scott [13], appears in Wand [15].

The case where the smoothing parameter is allowed to depend on location, x, is beyond the scope of this paper, although it may be treated in a somewhat similar manner. In particular, the equation "$L(v) = 0$" (see (i) above) takes the form "$L_x(v) = 0$" in the location-dependent case, and has an adaptive version. Brief details and examples will be given in [9].

We close this section with a little notation. A random n-sample from the distribution with unknown density f (in $d \geqslant 1$ dimensions) will be represen-

ted by $X_1, ..., X_n$. The *univariate* standard Normal density and distribution functions will be denoted by ϕ and Φ, respectively. Unqualified integrals will be over either $\mathbb{R}^d$ or $\mathbb{R}$; the case will be clear from context.

2. Asymptotic Mimization of L_1 Distance

In this section we show how to minimize asymptotic formulae for L_1 distance in the case of kernel estimators, histogram estimators, and smoothed histogram estimators (i.e., first-order histosplines). Our initial exposition is tailored to the case of kernel estimators, where relatively simple formulae are available for bias and variance. Immediately after that work we illustrate our argument with two examples treating general kernel estimators. Then we give two examples dealing with general histogram estimators, and there we show that only minor modifications to the earlier argument are necessary to handle the histogram case.

Suppose the estimate $\hat{f}$ of the d-variate density f is so constructed that bias and standard deviation are of the same order of magnitude, roughly equal to δ say. If $\hat{f}$ is asymptotically Normally distributed then we may write

$$\hat{f} - f = \delta(b - \sigma Z), \tag{2.1}$$

where δb is asymptotic to the bias of $\hat{f}$, $\delta\sigma$ is asymptotic to the standard deviation, and $Z = Z(x)$ is asymptotically Normal $N(0, 1)$. Of course, this representation is far from being unique. For kernel estimators we may choose δ, b, and σ so that δ depends only on n, and b and σ depend only on x. In the work below it is convenient to think of the representation as having this form.

In view of (2.1),

$$\delta^{-1} \int_{\mathbb{R}^d} E\,|\hat{f}(x) - f(x)|\,dx = \int_{\mathbb{R}^d} dx \int_{-\infty}^{\infty} |b(x) - \sigma(x)z| \times \phi(z)\,dz + o(1) \tag{2.2}$$

as $n \to \infty$. Since bias integrates to zero then $\int b(x)\,dx = 0$, and so the right-hand side of (2.2) (excluding the $o(1)$ term) equals

$$\begin{aligned}
&\int_{\mathbb{R}^d} dx \int_{-\infty}^{b(x)/\sigma(x)} \{b(x) - \sigma(x)z\}\,\phi(z)\,dz \\
&\qquad - \int_{\mathbb{R}^d} dx \int_{b(x)/\sigma(x)}^{\infty} \{b(x) - \sigma(x)z\}\,\phi(z)\,dz \\
&\quad = 2 \int_{\mathbb{R}^d} dx \int_{-\infty}^{b(x)/\sigma(x)} \{b(x) - \sigma(x)z\}\,\phi(z)\,dz \\
&\quad = 2 \int_{\mathbb{R}^d} \sigma(x)\,dx \int_{-\infty}^{b(x)/\sigma(x)} \Phi(z)\,dz.
\end{aligned}$$

Therefore asymptotic L_1 loss is given by $\delta\lambda$, where

$$\lambda \equiv 2 \int_{\mathbb{R}^d} \sigma(x)\, dx \int_{-\infty}^{b(x)/\sigma(x)} \Phi(z)\, dz.$$

In most cases of practical interest, bias and standard deviation can be balanced against one another to achieve the "optimum." This means that for some $r>0$, $b(x) \equiv u^r b_0(x)$ and $\sigma(x) \equiv u^{-1}\sigma_0(x)$, where $u>0$ is an adjustable parameter not depending on x, and b_0 and σ_0 are fixed functions of x, not depending on u. Thus, λ is really a function of u. See the end of this section for examples. We show next how to find the value u^* which minimizes

$$\lambda(u) \equiv 2 \int_{\mathbb{R}^d} \sigma_0(x)\, dx \int_{-\infty}^{u^r b_0(x)/\sigma_0(x)} \Phi(uz)\, dz. \tag{2.3}$$

Noting that $\int_{z<y} z\phi(z)\, dz = -\phi(y)$, we see that

$$\begin{aligned}\frac{1}{2}\lambda'(u) &= \int_{\mathbb{R}^d} \left[\int_{-\infty}^{u^r b_0(x)/\sigma_0(x)} z\phi(uz)\, dz + ru^{r-1}\{b_0(x)/\sigma_0(x)\} \right. \\ &\qquad \left. \times \Phi\{u^{r+1} b_0(x)/\sigma_0(x)\} \right] \sigma_0(x)\, dx \\ &= u^{-2} \Lambda(u^{r+1}),\end{aligned}$$

where

$$\Lambda(v) \equiv \int_{\mathbb{R}^d} \left[rvb_0(x)\, \Phi\{vb_0(x)/\sigma_0(x)\} - \sigma_0(x)\, \phi\{vb_0(x)/\sigma_0(x)\} \right] dx.$$

The "optimal" value u^* of u is a solution of the equation $\Lambda(u^{r+1}) = 0$. A notable feature of the function Λ is that it involves only one integration; the function λ involved two.

We show now that the equation $\Lambda(v) = 0$ has a unique positive solution. Put $L(v) \equiv v^{-1}\Lambda(v)$, and observe that

$$\begin{aligned}L'(v) &= \int_{\mathbb{R}^d} \{rb_0^2\sigma_0^{-1}\phi(vb_0/\sigma_0) + \sigma_0 v^{-2}\phi(vb_0/\sigma_0) + b_0^2\sigma_0^{-1}\sigma(vb_0/\sigma_0)\}\, dx \\ &= \int_{\mathbb{R}^d} \{(r+1)\, b_0^2\sigma_0^{-1} + \sigma_0 v^{-2}\}\, \phi(vb_0/\sigma_0)\, dx.\end{aligned}$$

The right-hand side is assuredly positive, proving that $L(v)$ is continuous and strictly increasing. Also, $L(0) = -\infty$, and as $v \to \infty$,

$$\begin{aligned}L(v) &= -\int_{\mathbb{R}^d} \left[rb_0\{1 - \Phi(vb_0/\sigma_0)\} + v^{-1}\sigma_0\phi(vb_0/\sigma_0) \right] dx \\ &\to -\int_{\mathbb{R}^d} rb_0 I(b_0 < 0)\, dx > 0.\end{aligned}$$

Therefore the equation $L(v)=0$, and so also $\Lambda(u^{r+1})=0$, has a unique positive solution.

In practice, the equation $L(v)=0$ may be solved using Newton's method, as follows. Let

$$H(v) \equiv L(v)/L'(v) = \left[\int_{\mathbb{R}^d} \{rb_0\Phi(vb_0/\sigma_0) - v^{-1}\sigma_0\phi(vb_0/\sigma_0)\}\, dx\right] \times \left[\int_{\mathbb{R}^d} \{(r+1)\, b_0^2\sigma_0^{-1} + \sigma_0 v^{-2}\}\, \phi(vb_0/\sigma_0)\, dx\right]^{-1}.$$

If v' is an approximation to the solution of $L(v)=0$ then $v''=v'-H(v')$ is a better approximation, and the approximations converge rapidly on iteration. This method of minimizing L_1 loss will be used repeatedly in the numerical work of Section 4.

The function Λ may be written in more homogeneous form as

$$\Lambda(v) = \int_{\mathbb{R}^d} \{r(vb_0/\sigma_0)\, \Phi(vb_0/\sigma_0) - \phi(vb_0/\sigma_0)\}\, \sigma_0\, dx.$$

Let v^* denote the (unique) solution of $\Lambda(v)=0$. If we are able to alter the construction of our estimator $\hat{f}$ in such a way that b_0 changes to $a_1 b_0$ and σ_0 changes to $a_2\sigma_0$, for *constants* a_1 and a_2, then we see from the above representation of Λ that v^* changes to $v^* a_2/a_1$. Since the value u^* of u which minimizes $\Lambda(u)$ is just the solution of $\Lambda(u^{r+1})=0$, then u^* changes to $u^*(a_2/a_1)^{1/(r+1)}$ under the transformation. This trite observation is important in the case of kernel estimators. It means that once we have derived the value of u^* for a particular kernel, we can easily find its value for all other kernels of the same "order," as will be shown in Examples 2.1 and 2.2 below.

We now give four examples which illustrate the forms which b_0, σ_0, r, and δ can take. Example 2.1 discusses regularity conditions which are sufficient for a rigorous proof of result (2.2).

EXAMPLE 2.1. *General kernel estimator in $d=1$ dimension.* Put

$$\hat{f}(x|h) \equiv (nh)^{-1} \sum_{j=1}^{n} K\{(x-X_j)/h\},$$

where K is a pth-order kernel—that is, $\int |z^p K(z)|\, dz < \infty$, K is bounded, and

$$\int_{-\infty}^{\infty} z^j K(z)\, dz = \begin{cases} 1 & \text{if } j=0 \\ 0 & \text{if } 1 \leqslant j \leqslant p-1 \\ (-1)^p c_1 \neq 0 & \text{if } j=p, \end{cases}$$

where $p \geqslant 1$ is an integer. If f is bounded and if $f^{(p)}$ is bounded and continuous then

$$\hat{f}(x|h) - f(x) = (c_1/p!)\, h^p f^{(p)}(x) + (nh)^{-1/2} c_2 f(x)^{1/2} Z_1 + o(h^p) \qquad (2.4)$$

as $n \to \infty$, $h = h(n) \to 0$, and $nh \to \infty$, where $c_2 \equiv (\int K^2)^{1/2}$ and $Z_1 = Z_1(x)$ is asymptotically Normal $N(0, 1)$. (See Prakasa Rao [11, p. 44ff] or Rosenblatt [12].) Take $h \equiv n^{-1/(2p+1)} u^2$, $r \equiv 2p$, $\delta \equiv n^{-p/(2p+1)}$, $b_0(x) \equiv (c_1/p!) f^{(p)}(x)$, and $\sigma_0(x) \equiv c_2 f(x)^{1/2}$. Then (2.1) and (2.2) hold with $b \equiv u^r b_0$ and $\sigma \equiv u^{-1}\sigma_0$.

Let u_0^* denote the value of u which minimizes $\lambda(u)$ (defined at (2.3), with $d = 1$ and $r = 2p$) for a particular pth-order kernel K_0. Let $c_{0,1}$ and $c_{0,2}$ be the versions of c_1 and c_2, respectively, in the case of this kernel. Then if u^* is the value of u which minimizes $\lambda(u)$ for any other pth-order kernel,

$$u^* = u_0^* \{(c_{0,1} c_2)/(c_1 c_{0,2})\}^{1/(2p+1)}, \qquad (2.5)$$

where c_1 and c_2 are computed for the kernel K. Therefore, once we know the value of u^* for a particular kernel, we can easily derive it for all other kernels of the same order.

The effect of changing from one pth-order kernel to another is only to alter the values of c_1 and c_2. Suppose that after such a change, $c_1 \mapsto a_1 c_1$ and $c_2 \mapsto a_1 c_2$ for constants a_1 and a_2. Then L_q distance, which is asymptotic to $\delta \lambda_q(u)^{1/q}$, where

$$\lambda_q(u) \equiv \int dx \int_{-\infty}^{\infty} |u^r b_0(x) - u^{-1}\sigma_0(x) z|^q \, \phi(z)\, dz$$

(compare (2.2)), changes to $\delta \lambda_q^\dagger(u)^{1/q}$, where

$$\begin{aligned} \lambda_q^\dagger(u) &\equiv \int dx \int_{-\infty}^{\infty} |u^r a_1 b_0(x) - u^{-1} a_2 \sigma_0(x) z|^q \, \phi(z)\, dz \\ &= (a_1 a_2^r)^{q/(r+1)} \lambda_q\{(a_1/a_2)^{1/(r+1)} u\}. \end{aligned}$$

Notice that $\inf_u \lambda_q^\dagger(u) = (a_1 a_2^r)^{q/(r+1)} \inf_u \lambda_q(u)$, implying that *no matter what the value of* q, the optimal kernel is the one which minimizes $a_1 a_2^r$. For example, the Barlett–Epanechnikov kernel [2, 6], which is known to be optimal in the sense of minimizing L_2 distance when $p = 2$, is also optimal in any L_q metric ($1 \leqslant q < \infty$) when $p = 2$. In particular it is optimal in the L_1 metric. The argument above applies without change to d-dimensional kernel estimators, and so the d-dimensional version of the Bartlett–Epanechnikov kernel [3] is optimal in any L_q metric.

The easiest way to give a rigorous proof of (2.2), here and in the other examples, is to establish "pointwise convergence" of the integrand, that is,

$$\delta^{-1} E\,|\hat{f}(x) - f(x)| \to \int_{-\infty}^{\infty} |b(x) - \sigma(x) z|\, \phi(z)\, dz, \qquad \text{for all } x, \quad (2.6)$$

and then prove convergence of the integral by applying a version of the dominated convergence theorem. Pointwise convergence is easily proved in all four of our examples, under the assumption that f and the derivatives of f appearing in b are bounded and continuous. Result (2.2) follows readily from the pointwise convergence in (2.6), if f vanishes outside a compact set (and, in Examples 2.1 and 2.2, if K is compactly supported). The case of compact support is paid particular attention by Devroye and Györfi [5] in their study of the L_1 metric. More general situations can also be handled, as we show in the next paragraph.

The convergence in (2.2) is usually uniform, in the following sense. (We treat only the case of Example 2.1.) Define $h_u \equiv n^{-1/(2p+1)}u^2$, and let $\lambda(u)$ be as in (2.3). Then we have:

THEOREM 2.1 ($d=1$). *If K is a compactly supported pth-order kernel for some $p \geqslant 1$; if $E(|X_1|^{1+\varepsilon}) < \infty$ for some $\varepsilon > 0$; if f is bounded; and if $f^{(p)}$ is bounded, continuous, and integrable; then*

$$n^{p/(2p+1)} \int_{-\infty}^{\infty} E\,|\hat{f}(x|h_u) - f(x)|\,dx = \lambda(u) + o(1) \tag{2.7}$$

uniformly in $u \in [C^{-1}, C]$, for each $C > 1$. Furthermore,

$$\inf_{h>0} \int_{-\infty}^{\infty} E\,|\hat{f}(x|h) - f(x)|\,dx \sim n^{-p/(2p+1)}\,\lambda(u^*), \tag{2.8}$$

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

A proof of Theorem 2.1 is given in Section 5. The techniques are not specific to $d=1$ dimension, and the theorem may be readily generalized to multivariate cases. The condition $E(|X_1|^{1+\varepsilon}) < \infty$ used in the theorem is close to being necessary for results of this type, since the function $\lambda(u)$ is not even well-defined for densities such as the Cauchy which have $E(|X_1| = \infty)$. (The reason is that $\int f^{1/2} = \infty$.)

EXAMPLE 2.2: *Nonnegative kernel estimator in $d \geqslant 1$ dimensions.* Put

$$\hat{f}(x|h) \equiv (nh^d)^{-1} \sum_{j=1}^{n} K\{(x - X_j)/h\},$$

where K is a bounded d-variate probability density satisfying $\int \|z\|^2 K(z)\,dz < \infty$ and $\int zK(z)\,dz = 0$. Let

$$c_1 \equiv \int_{\mathbb{R}^d} z_j^2 K(z)\,dz \qquad \text{and} \qquad c_2 \equiv \left\{ \int_{\mathbb{R}^d} K^2(z)\,dz \right\}^{1/2},$$

where z_j denotes the jth component of z and it is assumed that the integral defining c_1 does not depend on j. If f is bounded and if all the second derivatives of f are bounded and continuous then

$$\hat{f}(x|h) - f(x) = (c_1/2)h^2 \nabla^2 f(x) + (nh^d)^{-1/2} c_2 f(x)^{1/2} Z_1 + o(h^2) \quad (2.9)$$

as $n \to \infty$, $h = h(n) \to 0$, and $nh^d \to \infty$, where $\nabla^2 f = \sum_j (\partial/\partial x_j)^2 f$ is the Laplacian and $Z_1 = Z_1(x)$ is asymptotically Normal $N(0, 1)$. (See Prakasa Rao [11, p. 182ff] or Rosenblatt [12].) Take $h \equiv n^{-1/(d+4)} u^{2/d}$, $r \equiv 4/d$, $\delta \equiv n^{-2/(d+4)}$, $b_0(x) \equiv (c_1/2) \nabla^2 f(x)$, and $\sigma_0(x) \equiv c_2 f(x)^{1/2}$. Then (2.1) and (2.2) hold with $b \equiv u^r b_0$ and $\sigma \equiv u^{-1} \sigma_0$. An analogue of formula (2.5) describes the effect of changing from one kernel to another.

EXAMPLE 2.3: *Histogram estimator in $d \geqslant 1$ dimensions.* Divide all of $\mathbb{R}^d$ into a lattice of cubes with side length h. Given $x \in \mathbb{R}^d$, let $A(x) \equiv \prod_{1 \leqslant j \leqslant d} (a_j - \frac{1}{2}h, a_j + \frac{1}{2}h]$ be that cube in the lattice containing x, and write $N(x)$ for the number of observations from the sample which fall into $A(x)$. The histogram estimator of $f(x)$ is

$$\hat{f}(x) \equiv N(x)/(nh^d).$$

If f is bounded and if the first derivatives of f are bounded and continuous then

$$\hat{f}(x) - f(x) = \sum_{j=1}^{d} (a_j - x_j) f_j(x) + (nh^d)^{-1/2} f(x)^{1/2} Z_1 + o(h), \quad (2.10)$$

as $n \to \infty$, $h \to 0$, and $nh^d \to \infty$, where x_j denotes the jth component of x, $f_j(x) \equiv (\partial/\partial x_j) f(x)$, and Z_1 is asymptotically Normal $N(0, 1)$. (See Section 5 for a sketch of the proof.) Take $h \equiv n^{-1/(d+2)} u^{2/d}$, $r \equiv 2/d$, $\delta \equiv n^{-1/(d+2)}$, $b_0(x) \equiv h^{-1} \sum_j (a_j - x_j) f_j(x)$, and $\sigma_0(x) \equiv f(x)^{1/2}$. Then (2.1) and (2.2) hold with $b \equiv u^r b_0$ and $\sigma \equiv u^{-1} \sigma_0$.

This example differs from the previous two in that b_0 depends on h. However, that dependence turns out to be unimportant, as we now show. Remember that our general asymptotic expression for L_1 distance is $\delta \lambda(u)$, where

$$\lambda(u) \equiv \int_{\mathbb{R}^d} dx \int_{-\infty}^{\infty} |u^r b_0(x) - u^{-1} \sigma_0(x) z| \, \phi(z) \, dz; \quad (2.11)$$

see (2.2). In the present example,

$$\lambda(u) = \int_{\mathbb{R}^d} dx \int_{-\infty}^{\infty} \left| u^r \sum_{j=1}^{d} \eta_j f_j(x) - u^{-1} \sigma_0(x) z \right| \phi(z) \, dz,$$

where $\eta_j \equiv h^{-1}(a_j - x_j) \in (-\frac{1}{2}, \frac{1}{2})$. As $h \to 0$, $\lambda(u)$ converges to

$$\int_{\mathbb{R}^d \times (-1/2,1/2)^d} dx\, dy \int_{-\infty}^{\infty} |u^r b_0(x, y) - u^{-1}\sigma_0(x) z|\, \phi(z)\, dz,$$

where $y \equiv (y_1, ..., y_d)$ and $b_0(x, y) \equiv \sum_{1 \leqslant j \leqslant d} y_j f_j(x)$. This has the same form as the right-hand side of (2.11), provided we replace $b_0(x)$ by $b_0(x, y)$ and the integral over $\mathbb{R}^d$ by an integral over $\mathbb{R}^d \times (-\frac{1}{2}, \frac{1}{2})^d$. Of course, neither $b_0(x, y)$ nor $\sigma_0(x)$ depends on n or u.

With these trivial changes, the argument given in the first part of this section goes through as before. In particular, if we define

$$\Lambda(v) \equiv \int_{\mathbb{R}^d \times (-1/2,1/2)^d} [rvb_0(x, y)\, \Phi\{vb_0(x, y)/\sigma_0(x)\} - \sigma_0(x)\, \phi\{vb_0(x, y)/\sigma_0(x)\}]\, dx\, dy$$

and $L(v) \equiv v^{-1}\Lambda(v)$, then L is continuous and strictly increasing from $-\infty$ to a positive number, and the value of u such that $h \equiv n^{-1/(d+2)} u^{2/d}$ asymptotically minimizes L_1 distance is the unique positive solution of the equation $L(u^{r+1}) = 0$ (or equivalently, of $\Lambda(u^{r+1}) = 0$). The solution may be found rapidly by using the iteration argument given earlier in this section.

Reiterating, the only change we need make to our earlier theory to treat the case of a histogram estimator is to replace the bias term $b_0(x)$ by a function $b_0(x, y)$ of y as well as x, and to integrate over y as well as x. The next example shows that to treat the case of a *smoothed* histogram estimator, or histospline, or frequency polygon, the only requisite change is to replace both $b_0(x)$ and $\sigma_0(x)$ by functions of y as well as x, and to integrate over x and y.

EXAMPLE 2.4: *Frequency polygon estimator in $d = 1$ dimension.* Divide $\mathbb{R}$ into a lattice of segments, each of length h and having the form $(t - \frac{1}{2}h, t + \frac{1}{2}h]$ for some t. Any $x \in \mathbb{R}$ may be expressed uniquely in the form $x = a + \eta h$, where $0 < \eta \leqslant 1$ and a is the midpoint of one of the segments. Let $A_1(x) \equiv (a - \frac{1}{2}h, a + \frac{1}{2}h]$ and $A_2(x) \equiv (a + \frac{1}{2}h, a + \frac{3}{2}h]$, and write $N_j(x)$ for the number of sample values lying in $A_j(x)$. The frequency polygon estimator of $f(x)$ is

$$\hat{f}(x) = \{(1 - \eta)\, N_1(x) + \eta N_2(x)\}(nh)^{-1}.$$

If f is bounded and f'' is bounded and continuous then

$$\hat{f}(x) - f(x) = \tfrac{1}{6} h^2 f''(x)(3\eta - 3\eta^2 + \tfrac{1}{4}) + (nh)^{-1/2} f(x)^{1/2} (1 - 2\eta + 2\eta^2)^{1/2} Z_1 + o(h^2), \quad (2.12)$$

as $n \to \infty$, $h \to 0$, and $nh \to \infty$, where Z_1 is asymptotically Normal $N(0, 1)$. (See Section 5 for a sketch of the proof.) Take $h \equiv n^{-1/5}u^2$, $r \equiv 4$, $\delta \equiv n^{-2/5}$, $b_0(x) \equiv \frac{1}{6} f''(x)(3\eta - 3\eta^2 + \frac{1}{4})$, and $\sigma_0(x) \equiv f(x)^{1/2}(1 - 2\eta + 2\eta^2)^{1/2}$. Then (2.1) and (2.2) hold with $b \equiv u^r b_0$ and $\sigma \equiv u^{-1}\sigma_0$.

As in the previous example, the dependence of b_0 and σ_0 on h is unimportant. To appreciate why, observe that asymptotic L_1 distance equals $\delta\lambda(u)$, where $\lambda(u)$ is given by (2.11), and just as in Example 2.3, $\lambda(u)$ converges to

$$\int_{\mathbb{R} \times (0,1)} dx\, dy \int_{-\infty}^{\infty} |u^r b_0(x, y) - u^{-1}\sigma_0(x, y)z|\, \phi(z)\, dz$$

as $n \to \infty$, where $b_0(x, y) \equiv \frac{1}{6} f''(x)(3y - 3y^2 + \frac{1}{4})$ and $\sigma_0(x, y) \equiv f(x)^{1/2}(1 - 2y + 2y^2)^{1/2}$. Neither $b_0(x, y)$ nor $\sigma_0(x, y)$ depends on n or u. If we replace $b_0(x)$ by $b_0(x, y)$ and $\sigma_0(x)$ by $\sigma_0(x, y)$, and integrate over x and y rather than just over x, then the theory in the first part of this section goes through without changes.

3. Adaptive Minimization of L_1 Distance

In this section we concentrate on kernel estimators. Variants of our arguments may be used with other estimators, such as histogram estimators. We shall introduce an adaptive, data-driven method for minimizing L_1 distance, and prove that it does indeed minimize L_1 loss in an asymptotic sense.

Suppose the data are in $d \geqslant 1$ dimensions, and define the kernel estimator

$$\hat{f}(x) = \hat{f}(x|h) \equiv (nh^d)^{-1} \sum_{j=1}^{n} K\{(x - X_j)/h\}.$$

Examples 2.1 and 2.2 in the previous section discussed such estimators in detail, and gave instances of the decomposition of $\hat{f} - f$ into bias and variance components; see (2.4) and (2.9). Let

$$J(h) \equiv \int_{\mathbb{R}^d} E\,|\hat{f}(x|h) - f(x)|\, dx$$

denote L_1 distance. As we showed in Section 2, if $h_u \equiv n^{-1/\{d(r+1)\}}u^{2/d}$ for a correctly chosen $r > 0$ (depending on the kernel K), then as $n \to \infty$,

$$J(h_u) \sim n^{-r/\{2(r+1)\}}\lambda(u) \tag{3.1}$$

uniformly in $u \in [C^{-1}, C]$ for any $C > 1$, where

$$\lambda(u) \equiv \int_{\mathbb{R}^d} dx \int_{-\infty}^{\infty} |u^r b_0(x) - u^{-1}\sigma_0(x)z| \, \phi(z)\, dz,$$

$\sigma_0(x)$ is proportional to $f(x)^{1/2}$, $b_0(x)$ is proportional to a linear combination of derivatives of f, and $u > 0$ is a variable parameter. Furthermore,

$$\inf_{h>0} J(h) \sim n^{-r/\{2(r+1)\}} \lambda(u^*), \tag{3.2}$$

where u^* is the unique value of u which minimizes $\lambda(u)$. (Theorem 2.1 gives regularity conditions sufficient for (3.1) and (3.2).) Of course, both b_0 and σ_0 are unknown, and so any attempt at minimizing λ using only information in the data must involve explicit or implicit estimation of b_0 and σ_0. In the present section we discuss the explicit approach to this problem.

Let $\hat{b}_0$ and $\hat{\sigma}_0$ be L_1 consistent estimators of b_0 and σ_0, respectively. That is,

$$\int_{\mathbb{R}^d} (|\hat{b}_0 - b_0| + |\hat{\sigma}_0 - \sigma_0|) \to 0 \tag{3.3}$$

almost surely as $n \to \infty$. Assume too that $\int \hat{b}_0 = 0$. Later in this section we shall discuss candidates which satisfy these conditions. Put

$$\hat{\lambda}(u) \equiv \int_{\mathbb{R}^d} dx \int_{-\infty}^{\infty} |u^r \hat{b}_0(x) - u^{-1}\hat{\sigma}_0(x)z| \, \phi(z)\, dz. \tag{3.4}$$

The argument given in the early part of Section 2 shows that there is a unique $\hat{u}^* > 0$ which minimizes $\hat{\lambda}(u)$, and that $\hat{u}^*$ may be found by iteration. Indeed, if we define a sequence $v_0, v_1, v_2, \ldots$ by

$$v_{j+1} = v_j - \hat{H}(v_j), \qquad j \geq 0,$$

where

$$\hat{H}(v) \equiv \left[\int_{\mathbb{R}^d} \{ r\hat{b}_0 \Phi(v\hat{b}_0/\hat{\sigma}_0) - v^{-1}\hat{\sigma}_0 \phi(v\hat{b}_0/\hat{\sigma}_0) \} \, dx \right]$$
$$\times \left[\int_{\mathbb{R}^d} \{ (r+1)\, \hat{b}_0^2 \hat{\sigma}_0^{-1} + \hat{\sigma}_0 v^{-2} \} \, \phi(v\hat{b}_0/\hat{\sigma}_0)\, dx \right]^{-1},$$

then the sequence $\{v_j\}$ converges to that value $\hat{v}^*$, $0 < \hat{v}^* < \infty$, which is such that $\hat{u}^* \equiv (\hat{v}^*)^{1/(r+1)}$.

Remember that u^* is the unique u minimizing $\lambda(u)$. In view of (3.3) we have, for any $C > 1$,

$$\sup_{C^{-1} \leq u \leq C} |\hat{\lambda}(u) - \lambda(u)| \to 0 \tag{3.5}$$

almost surely, and also $\hat{u}^* \to u^*$ almost surely and $\lambda(\hat{u}^*) \to \lambda(u^*)$ almost surely. We shall define our adaptive, data-driven window to be

$$\hat{h}^* \equiv n^{-1/\{d(r+1)\}}(\hat{u}^*)^{2/d}. \tag{3.6}$$

Let $h^* \equiv n^{-1/\{d(r+1)\}}(u^*)^{2/d}$ be the deterministic, asymptotically optimal window. Then $\hat{h}^*/h^* \to 1$ almost surely and $J(\hat{h}^*)/J(h^*) \to 1$ almost surely. (For the latter, use (3.1) and (3.5).) Indeed, noting the asymptotic relation (3.2), we have

$$J(\hat{h}^*)/\inf_{h>0} J(h) \to 1 \tag{3.7}$$

almost surely. In this sense, the adaptive window $\hat{h}^*$ provides asymptotic minimization of L_1 distance.

Observe that $J(\hat{h}^*)$ is not the same as L_1 distance computed for $\hat{f}(x|\hat{h}^*)$. The latter would be

$$\int_{\mathbb{R}^d} E\,|\hat{f}(x|\hat{h}^*) - f(x)|\,dx,$$

whereas

$$J(\hat{h}^*) = \int_{\mathbb{R}^d} \{E\,|\hat{f}(x|h) - f(x)|\}_{h=\hat{h}^*}\,dx.$$

It would be more in keeping with the fact that $\hat{h}^*$ is a random variable to examine $\hat{h}^*$ in the context of minimizing *raw* L_1 distance, defined by

$$\hat{J}(h) \equiv \int_{\mathbb{R}^d} |\hat{f}(x|h) - f(x)|\,dx.$$

(Of course, $J = E(\hat{J})$.) A natural question to ask is whether $\hat{h}^*$ is asymptotically as good as the window which minimizes $\hat{J}$; that is, whether

$$\hat{J}(\hat{h}^*)/\inf_{h>0} \hat{J}(h) \to 1 \tag{3.8}$$

almost surely. This result is analogous to (3.7), and if it were true it would provide another sense in which the adaptive window $\hat{h}^*$ was asymptotically optimal. In fact, (3.8) *is* true under appropriate regularity conditions. It follows from (3.7) above and (3.9) below.

THEOREM 3.1 ($d \geqslant 1$). *If K is compactly supported and Hölder continuous, and if f is bounded, then*

$$\{\inf_{h>0} \hat{J}(h)\}/\{\inf_{h>0} J(h)\} \to 1 \quad \text{and} \quad \hat{J}(\hat{h}^*)/J(\hat{h}^*) \to 1 \tag{3.9}$$

almost surely as $n \to \infty$.

We now give examples of the estimators $\hat{b}_0$ and $\hat{\sigma}_0$. For notational convenience we concentrate on the case of $d=1$ dimension, which was the subject of Example 2.1. In this circumstance, Theorem 2.1 gives regularity conditions sufficient for (3.1) and (3.2). The techniques used to derive Theorem 2.1 and Theorem 3.2 below are not specific to $d=1$ dimension, and our results are readily extended to multivariate cases.

Define

$$\hat{f}(x|h) \equiv (nh)^{-1} \sum_{j=1}^{n} K\{(x-X_j)/h\},$$

where K is a pth-order kernel for some $p \geqslant 1$. The versions of b_0 and σ_0 appropriate to this context are $b_0(x) \equiv (c_1/p!) f^{(p)}(x)$ and $\sigma_0(x) \equiv c_2 f(x)^{1/2}$, where $c_1 \equiv \int z^p K(z)\, dz \neq 0$ and $c_2 \equiv (\int K^2)^{1/2}$; see Example 2.1. As our estimates of $f^{(p)}$ and $f^{1/2}$ we shall take

$$\hat{f}_1^{(p)}(x) \equiv (nh_1^{p+1})^{-1} \sum_{j=1}^{n} K_1^{(p)}\{(x-X_j)/h_1\}$$

and

$$\hat{f}_2(x)^{1/2} \equiv \left[(nh_2)^{-1} \sum_{j=1}^{n} K_2\{(x-X_j)/h_2\} \right]^{1/2},$$

where K_1 and K_2 are kernels (possibly the same as K), K_2 is nonnegative, and it is assumed that $K_1^{(p)}$ is well-defined. Of course, $\hat{f}_1^{(p)}$ is just the pth derivative of an ordinary kernel estimator. Its existence as a numerical quantity does not require existence of $f^{(p)}$, but for convergence we do need to assume that the pth derivative is well-defined and finite. Our estimates of b_0 and σ_0 are

$$\hat{b}_0(x) \equiv (c_1/p!)\, \hat{f}_1^{(p)}(x), \qquad \hat{\sigma}_0(x) \equiv c_2 \hat{f}_2(x)^{1/2}. \tag{3.10}$$

Notice that $\int \hat{b}_0 = 0$. The only other property required of $\hat{b}_0$ and $\hat{\sigma}_0$ is the L_1 convergence described by (3.3), and that follows from the following theorem.

THEOREM 3.2 ($d=1$). *Assume K_1 and K_2 are bounded, compactly supported, and integrate to unity; $K_1^{(p)}$ is well-defined and bounded; K_2 is nonnegative; $E(|X_1|^{1+\varepsilon}) < \infty$ for some $\varepsilon > 0$; f is bounded; $f^{(p)}$ is bounded, continuous, and integrable; and $h_1, h_2 \to 0$, $nh_1^{2p+1}/\log n \to \infty$, and $nh_2 \to \infty$. Then*

$$\int_{-\infty}^{\infty} |\hat{f}_1^{(p)} - f^{(p)}| \to 0 \tag{3.11}$$

and

$$\int_{-\infty}^{\infty} |\hat{f}_2^{1/2} - f^{1/2}| \to 0 \tag{3.12}$$

almost surely.

The condition $E(|X_1|^{1+\varepsilon}) < \infty$ was imposed also in Theorem 2.1, and as explained there it is needed to exclude f's such as the Cauchy density, for which the theorem fails. The condition $nh_1^{2p+1} \to \infty$ is *necessary* for weak pointwise consistency of $\hat{f}_1^{(p)}$, and our assumption that $nh_1^{2p+1}/\log n \to \infty$ is only slightly more restrictive. Nonnegativity of K_2 is needed to ensure that $\hat{f}_2^{1/2}$ is real-valued.

In conclusion, we point out that if f, K_1, and K_2 satisfy the conditions in Theorem 3.2, if the pth-order kernel K satisfies the conditions in Theorem 3.1, if $\hat{b}_0$ and $\hat{\sigma}_0$ are defined by (3.10), if $\hat{u}^*$ denotes the value of u minimizing $\hat{\lambda}(u)$ (defined at (3.4), with $d=1$ and $r=2p$), and if $\hat{h}^*$ is given by (3.6), then results (3.7) and (3.8) hold:

$$J(\hat{h}^*)/\inf_{h>0} J(h) \to 1 \qquad \text{and} \qquad \hat{J}(\hat{h}^*)/\inf_{h>0} \hat{J}(h) \to 1$$

almost surely. In these two senses, our adaptive, data-driven window $\hat{h}^*$ provides asymptotic minimization of L_1 distance. The next section describes numerical applications of this idea.

4. Numerical Results

In this section we confine attention to $d=1$ dimension, and to symmetric, nonnegative kernels. This amounts to taking $p=2$ in Example 2.1, and means that the "asymptotically optimal" window in the sense of minimizing L_1 distance is $h^* = n^{-1/5}(u^*)^2$, where u^* is that value of u which minimizes

$$\lambda(u) \equiv \int_{-\infty}^{\infty} dx \frac{1}{2} \int_{-\infty}^{\infty} \left| u^4 \frac{1}{2} c_1 f''(x) - u^{-1} c_2 f(x)^{1/2} z \right| \phi(z)\, dz.$$

(Here $c_1 \equiv \int z^2 K(z)\, dz$ and $c_2 \equiv (\int K^2)^{1/2}$. See Example 2.1 for details.) Work in this section falls naturally into two parts. First, we discuss numerical values of the constant $C^{(1)} = (u^*)^2$ in the formula $h^* = C^{(1)} n^{-1/5}$, for the case where f is either Normal or a mixture of two Normals. Then we show how to implement the adaptive method described in Section 3, and illustrate those ideas by applying our techniques to simulated data. Throughout the section we stress differences between L_1 and L_2 minimization.

Put $g(x) \equiv c_1(2c_2)^{-1} f''(x)/f(x)^{1/2}$, and define

$$H(v) \equiv v \int_{-\infty}^{\infty} [4vg(x)\, \Phi\{vg(x)\} - \phi\{vg(x)\}]\, f(x)^{1/2}\, dx$$

$$\times \left(\int_{-\infty}^{\infty} [5\{vg(x)\}^2 + 1]\, \phi\{vg(x)\}\, f(x)^{1/2}\, dx \right)^{-1}.$$

Define the sequence $v_0, v_1, v_2, \ldots$ by $v_{j+1} = v_j - H(v_j)$, $j \geq 0$, where $v_0 > 0$ is arbitrary. We showed in Section 2 that this sequence converges to the number v^* such that $u^* = (v^*)^{1/5}$. Thus, $h^* = C^{(1)} n^{-1/5}$, where $C^{(1)} = (v^*)^{2/5}$. (If v_0 is chosen much larger than v^* then v_1 may be negative, due to the fact that L is concave. This difficulty is easily overcome by using a smaller value of v_0.)

Once we know the value of h^* for a particular symmetric, nonnegative kernel K_0, we can easily derive it for all other kernels of this type. Indeed, let h_0^* be the version of h^* for K_0, and let $c_{0,1}$ and $c_{0,2}$ be the corresponding versions of c_1 and c_2. Then the value of h^* for the kernel K is

$$h^* = h_0^* \{(c_{0,1} c_2)/(c_1 c_{0,2})\}^{2/5};$$

see (2.5). In the next three paragraphs we work with the Bartlett–Epanechnikov kernel, $K(x) \equiv \frac{3}{4}(1 - x^2)$ if $|x| \leq 1$, 0 if $|x| > 1$. This K is bounded, compactly supported, and Hölder continuous.

When f is the standard Normal density and K is the Bartlett–Epanechnikov kernel, the constant $C^{(1)}$ in the formula $h^* = C^{(1)} n^{-1/5}$ is $C^{(1)} = 2.279$. By way of comparison, the window $C^{(2)} n^{-1/5}$ which is asymptotically optimal in the sense of minimizing L_2 distance has $C^{(2)} = 2.345$. Since $C^{(1)} < C^{(2)}$ then minimizing L_1 distance provides slightly less smoothing than minimizing L_2 distance. However, the two constants are remarkably close. In the case of L_2 distance, it is sometimes suggested that when the data distribution is unknown, the window be chosen as though the data were Normal, resulting in $h = 2.345 \hat{\sigma} n^{-1/5}$, where $\hat{\sigma}$ is sample standard deviation. The analogue of this proposal in the case of L_1 distance is of course $h = 2.279 \hat{\sigma} n^{-1/5}$, provided the Bartlett–Epanechnikov kernel is in use.

Recall from Section 2 that minimum asymptotic L_1 loss is $n^{-2/5} \inf_u \lambda(u)$, where $\lambda(u)$ is defined at (2.3). Devroye and Györfi [5, pp. 78–79] give bounds for $\inf_u \lambda(u)$. In the case where f is standard Normal and K is Bartlett–Epanechnikov, these bounds (together with the exact value) are

$$1.002 < \inf_u \lambda(u) = 1.022 < 1.341.$$

The lower bound in particular is remarkably accurate.

Devroye and Györfi [5, p. 107] also suggest an approximation to the asymptotically optimal window: $h \sim 1.664 n^{-1/5}$, compared to the actual value $2.279 n^{-1/5}$. The discrepancy here is due to the fact that the function λ is quite flat in the vicinity of the minimizing value, $u^* = 1.510$ $(= 2.279^{1/2})$. Indeed,

$$|\lambda(1.45) - \lambda(u^*)|/\lambda(u^*) = 0.006, \quad |\lambda(1.55) - \lambda(u^*)|/\lambda(u^*) = 0.003.$$

It is much faster to derive u^* using the iterative argument described in the second paragraph of this section, than to find it by a direct attempt at minimization.

Table 4.1 lists values of $C^{(1)}$, $C^{(2)}$, and the ratio $C^{(1)}/C^{(2)}$ for several equal-proportion, two-component Normal mixtures, with densities

$$f_\sigma(x) = \tfrac{1}{2}(2\pi\sigma^2)^{-1/2} \exp\{-(x+1)^2/2\sigma^2\} + \tfrac{1}{2}(2\pi\sigma^2)^{-1/2} \exp\{-(x-1)^2/2\sigma^2\}.$$

(The only variable parameter governing f is the common variance, σ^2.) In all these cases the ratio $C^{(1)}/C^{(2)}$ is less than one. This occurrence seems to be more common than $C^{(1)}/C^{(2)} > 1$, although the latter can arise. A case in point is that of equal-proportion, two-component Normal mixtures with means $(1, -1)$ and variances $(1, 1/10)$, where $C^{(1)} = 1.03$ and $C^{(2)} = 0.98$.

The closeness of the ratio $C^{(1)}/C^{(2)}$ to unity in many cases of interest means that from a practical viewpoint there is often little to choose between a density estimate which has been optimised in an L_1 sense and one which has been L_2 optimised. Graphs of $\hat{f}(x|h)$ for $h = C^{(1)}n^{-1/5}$ and $h = C^{(2)}n^{-1/5}$ are virtually indistinguishable when f is a Normal mixture.

We conducted a series of experiments using the adaptive window selection rule suggested in Section 3. We took f to be the standard Normal density, and K, K_1, K_2 all to be the standard Normal kernel. (Since this K does not have compact support then, strictly speaking, results in Section 3 do not apply to it. That may be remedied by using arguments from [7].) We selected the window h_2, needed in the construction of $\hat{f}_2^{1/2}$, by squared-error cross-validation. Thus, $\hat{f}(\cdot|h_2)$ asymptotically minimizes L_2 loss. We took h_1, needed for $\hat{f}_1''$, to be simply $h_2^{9/5}$, in the knowledge that a window of size $n^{-1/5}$ is optimal for estimating f, whereas a window of size $n^{-1/9}$ is optimal for estimating f''. Constructing $\hat{u}^*$ in the manner described in Section 3, we

TABLE 4.1

Values of $C^{(1)}$, $C^{(2)}$ for Bartlett–Epanechnikov Kernel K and for Equal-Proportion, Two-Component Normal Mixture Density f with Means $(1, -1)$ and Variances (σ^2, σ^2). Windows $C^{(1)}n^{-1/5}$, $C^{(2)}n^{-1/5}$ Are Asymptotically Optimal for L_1, L_2 Loss, Respectively.

σ^2	$C^{(1)}$	$C^{(2)}$	$C^{(1)}/C^{(2)}$
∞	$\sim 2.279\sigma$	$\sim 2.345\sigma$	0.972
5^2	11.6	12.0	0.972
5^1	5.60	5.80	0.966
5^0	3.01	3.26	0.925
5^{-1}	1.15	1.18	0.969
5^{-2}	0.524	0.539	0.972
0	$\sim 2.279\sigma$	$\sim 2.345\sigma$	0.972

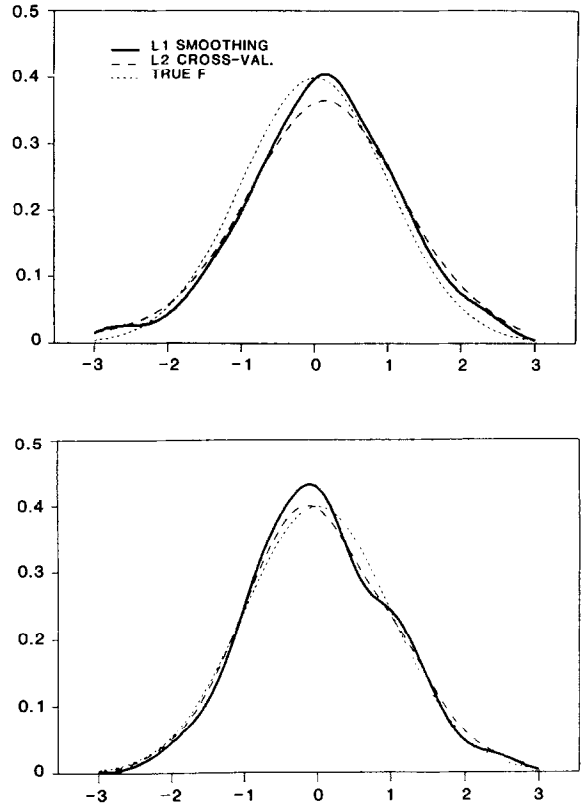

FIG. 4.1. Numerical examples of adaptive procedure discussed in Section 3, with f, K, K_1, K_2 all equal to standard Normal density ϕ. Solid curve is $\hat{f}(\cdot|\hat{h}^*)$ (asymptotically L_1 optimal), dashed curve is $\hat{f}(\cdot|h_2)$ (h_2 found by squared-error cross-validation; asymptotically L_2 optimal), and dotted curve is true f. Sample size is $n = 100$.

defined $\hat{h}^* \equiv n^{-1/5}(\hat{u}^*)^2$. (The quantity $\hat{C}^{(1)} \equiv (\hat{u}^*)^2$ is a consistent estimator of $C^{(1)}$.) We compared graphs of $\hat{f}(\cdot|\hat{h}^*)$ and $\hat{f}(\cdot|h_2)$. For large n ($n \geq 200$) there was little difference between the true curves, although $\hat{h}^*$ was a little more robust than h_2 against sampling fluctuations. This appears to be due to the fact that $\hat{h}^*/(C^{(1)}n^{-1/5})$ converges to unity at rate $n^{-2/9}$ (the rate of consistency of $\hat{f}''(\cdot|h_1)$ for f''), whereas $h_2/(C^{(2)}n^{-1/5})$ converges to unity only at rate $n^{-1/10}$ (see [8]). For $n = 100$ there was a tendency for $\hat{C}^{(1)}$ to underestimate $C^{(1)}$ and for $n^{1/5}h_2$ to overestimate $C^{(2)}$, but neither $\hat{h}^*$ nor h_2 gave curves which were closer, on average, to the true density. Figure 4.1 depicts two typical results.

5. Proofs

Throughout our proofs the symbols C, C_1, C_2, ... denote positive generic constants, possibly different at different appearances.

Proof of Theorem 2.1. We show first that for any $C \geqslant 1$ and $h \leqslant 1$,

$$I(n, h, C) = \int_{|x| > C} E\,|\hat{f}(x|h) - f(x)|\,dx \leqslant g(C)\{(nh)^{-1/2} + h^p\}, \quad (5.1)$$

where g does not depend on n or h and converges to zero as $C \to \infty$.

Notice that

$$(p-1)!\,|E\hat{f}(x|h) - f(x)| = \left| h^p \int_{-\infty}^{\infty} K(z)\,dz \times \int_0^1 f^{(p)}(x - thz)(1-t)^{p-1}\,dt \right|, \quad (5.2)$$

and that if $|hz| \leqslant C$ and $0 < t < 1$, $\{x: |x| > 2C\} \subseteq \{x: |x - thz| > C\}$. Therefore if $h \leqslant 1$,

$$\begin{aligned} I_1(n, h, 2C) &\equiv \int_{|x| > 2C} |E\hat{f}(x|h) - f(x)|\,dx \\ &\leqslant h^p \left\{ \int_{-\infty}^{\infty} |K(z)|\,dz \int_{|y| > C} |f^{(p)}(y)|\,dy \right. \\ &\quad \left. + \int_{|hz| > C} |K(z)|\,dz \int_{-\infty}^{\infty} |f^{(p)}(y)|\,dy \right\} \\ &\leqslant g_1(2C) h^p, \end{aligned}$$

where

$$\begin{aligned} g_1(2C) &\equiv \int_{-\infty}^{\infty} |K(z)|\,dz \int_{|y| > C} |f^{(p)}(y)|\,dy \\ &\quad + C^{-p} \int_{-\infty}^{\infty} |z^p K(z)|\,dz \int_{-\infty}^{\infty} |f^{(p)}(y)|\,dy. \end{aligned}$$

Also, $\operatorname{var}\{\hat{f}(x|h)\} \leqslant (nh)^{-1} \int K^2(z)\, f(x - hz)\,dz$. Given $\alpha > 1$, put

$$g_2(C) \equiv \left\{ \int_{|x| > C} (1 + |x|^{\alpha})^{-1}\,dx \right\}^{1/2}.$$

Notice that $1+|x|^\alpha \leqslant 2^\alpha(1+|x-hz|^\alpha+|hz|^\alpha)$. Therefore if $h \leqslant 1$,

$$\begin{aligned} I_2(n,h,C) &\equiv \int_{|x|>C} [\operatorname{var}\{\hat{f}(x|h)\}]^{1/2}dx \\ &\leqslant g_2(C)\left[\int_{-\infty}^{\infty} \operatorname{var}\{\hat{f}(x|h)\}(1+|x|^\alpha)\,dx\right]^{1/2} \\ &\leqslant g_3(C)(nh)^{-1/2}, \end{aligned}$$

where

$$\begin{aligned} g_3(C) \equiv g_2(C)\,2^{\alpha/2}\Bigg[&\int_{-\infty}^{\infty} K^2(z)\,dz \int_{-\infty}^{\infty} (1+|x|^\alpha)\,f(x)\,dx \\ &+\int_{-\infty}^{\infty} |z|^\alpha\,K(z)^2\,dz\Bigg]^{1/2}. \end{aligned}$$

This quantity is finite if α is sufficiently close to unity. The desired result (5.1), with $g \equiv g_1 + g_3$, follows from the estimates in this paragraph and the fact that $I \leqslant I_1 + I_2$.

Let b_0 and σ_0 be the functions defined in Example 2.1. Put

$$\lambda(u,C) \equiv \int_{|x|\leqslant C} dx \int_{-\infty}^{\infty} |u^r b_0(x) - u^{-1}\sigma_0(x)z|\,\phi(z)\,dz,$$

where $r \equiv 2p$. Then for any $C_1 > 1$,

$$\lim_{C_2\to\infty} \sup_{u\in[C_1^{-1},C_1]} |\lambda(u,C_2)-\lambda(u)| = 0. \tag{5.3}$$

Techniques used to prove Theorem 1 on p. 78 of [5] are readily adapted to show that for any $C_2 > 0$,

$$\sup_{u\in[C_1^{-1},C_1]} \left| n^{p/(2p+1)} \int_{|x|\leqslant C_2} E\,|\hat{f}(x\,|\,h_u) - f(x)|\,dx - \lambda(u,C_2)\right| \to 0 \tag{5.4}$$

as $n\to\infty$. Result (2.7), with uniform convergence, follows from (5.1), (5.3), and (5.4).

We showed in Section 2 that the continuous function λ has a unique minimum, occurring at the point u^*. Result (2.8) will follow from this fact and (2.7) if we prove that for some $C > 0$,

$$J(h) \geqslant C[\min\{(nh)^{-1/2}, 1\} + \min(h^p, 1)] \tag{5.5}$$

whenever $n \geqslant 1$ and $h > 0$. Now,

$$3E\,|\hat{f} - f| \geqslant E\,|\hat{f} - E\hat{f} + E\hat{f} - f| + 2\,|E\hat{f} - f| \geqslant E\,|\hat{f} - E\hat{f}| + |E\hat{f} - f|,$$

and so

$$3J \geqslant J_1 + J_2, \tag{5.6}$$

where $J_1 \equiv \int E\,|\hat{f} - E\hat{f}|$ and $J_2 \equiv \int |E\hat{f} - f|$. We may show from (5.2) that

$$\liminf_{h \to 0} h^{-p} \int_{-\infty}^{\infty} |E\hat{f}(x|h) - f(x)|\,dx > 0,$$

and so there exist C_1, $C_2 > 0$ such that $J_2(h) \geqslant C_1 h^p$ whenever $0 \leqslant h \leqslant C_2$. The inequality $J_2(h) \geqslant C_3$ for $h > C_2$ is easily established separately. Therefore $J_2(h) \geqslant C_4 \min(h^p, 1)$, for all $h > 0$. The desired result (5.5) is a consequence of this inequality, (5.6), and the lemma below. (We state and prove the lemma for d-dimensional data, since that form will be needed in the proof of Theorem 3.1.)

LEMMA 5.1 ($d \geqslant 1$). *If the d-variate kernel K is bounded, vanishes outside a compact set, and integrates to unity, if f is bounded, and if the d-variate kernel density estimator $\hat{f}$ is based on K, then for a constant $C > 0$ not depending on n or h, such that for $n \geqslant 1$ and $0 < h \leqslant 1$,*

$$\int_{\mathbb{R}^d} E\,|\hat{f}(x|h) - E\hat{f}(x|h)|\,dx > C \min\{(nh^d)^{-1/2}, 1\}.$$

Proof of Lemma 5.1. By an inequality for moments of sums of independent random variables [1; 5, p. 90],

$$|\,E|\,\hat{f}(x|h) - E\hat{f}(x|h)| - (2/\pi)^{1/2}\,\{E\,|\hat{f}(x|h) - E\hat{f}(x|h)|^2\}^{1/2}| \leqslant C_1(nh^d)^{-1},$$

where C_1 does not depend on x, n, or h. Simple calculations show that for $h^d \geqslant n^{-1}$ and for some bounded region $\mathscr{R}$,

$$\int_{\mathscr{R}} (E\,|\hat{f} - E\hat{f}|^2)^{1/2} \geqslant C_2(nh^d)^{-1/2}.$$

Therefore if $h^d \geqslant C_3 n^{-1}$ and C_3 is sufficiently large,

$$\int_{\mathscr{R}} E\,|\hat{f} - E\hat{f}| \geqslant \tfrac{1}{2}\,C_2(nh^d)^{-1/2}.$$

This proves the lemma for $h^d \geqslant C_3 n^{-1}$.

To treat the case $h^d \leqslant C_3 n^{-1}$, suppose K vanishes outside a ball of radius s centred at the origin. Then $\hat{f}(x|h) = 0$ if $|x - X_j| > sh$ for each j, $1 \leqslant j \leqslant n$. Therefore if $h^d \leqslant C_3 n^{-1}$ and n is large then the chance that $\hat{f}(x|h)$ equals zero exceeds

$$p(x; n) \equiv \{P(|x - X| > sh)\}^n \geqslant (1 - Bv_d s^d h^d)^n$$
$$\geqslant C_4 \exp(-nBv_d s^d h^d) \geqslant C_5 > 0,$$

where B is an upper bound to f, v_d equals the content of the d-dimensional ball of unit radius, and C_5 does not depend on x, n, or h. Hence

$$\int_{\mathbb{R}^d} E\,|\hat{f} - E\hat{f}| \geqslant \int_{\mathbb{R}^d} P(x; n)\,|E\hat{f}(x|h)|\,dx$$
$$\geqslant C_5 \left| \int_{\mathbb{R}^d} E\hat{f}(x|h)\,dx \right| = C_5,$$

which completes the proof of Lemma 5.1.

The condition that f be bounded, imposed in Theorems 2.1 and 3.1 and in Lemma 5.1, may be relaxed. We do not pursue such generalizations here, because the condition of boundedness is mild, natural, and commonly imposed in work of this type—see, e.g., [8, 14].

Proof of Theorem 3.1. Our proof uses a very powerful result due to Devroye [4]. We state it here, for convenience.

LEMMA 5.2 ($d \geqslant 1$). *If K is bounded and compactly supported then there exist positive constants C_1, C_2, and C_3, depending on K but not on f or n, such that*

$$\sup_{h>0} P\{|\hat{J}(h) - J(h)| > \varepsilon\} \leqslant C_1 \exp(-C_2 n\varepsilon^2)$$

whenever $C_3 n^{-1/2} \leqslant \varepsilon \leqslant 1$.

Lemma 5.2 is a corollary of Devroye's Theorem 1 [4]. In fact, Devroye shows that we may take $C_1 = 2$ and $C_2 = (32 \int |K|)^{-1}$.

Let h_0, $\hat{h}_0$ be the values of h which minimize J, $\hat{J}$, respectively. It is easy to see that for $a > 0$ sufficiently large we have $n^{-a} \leqslant h_0 \leqslant n^a$ for all large n, and also

$$P\{n^{-a} \leqslant \hat{h}_0(n), \hat{h}^* \leqslant n^a, \text{ all } n \geqslant n'\} \to 1$$

as $n' \to \infty$. Given $c > 0$, let $\mathscr{H} = \mathscr{H}(a, c) = \{h_1, h_2, \ldots\}$ be the nonrandom sequence defined by $n^{-a} = h_1 < h_2 < \cdots < h_{m-1} \leqslant n^a < h_m < \cdots$ and

$h_{i+1} - h_i = n^{-c}$, $i \geqslant 1$. For each $h \in \mathscr{I} \equiv [n^{-a}, n^a]$, let $H(h)$ be a value in $\mathscr{H}$ which minimizes $|h - H(h)|$. By Hölder continuity and compact support of K, we may choose $c = c(a)$ so large that for some $C > 0$,

$$\sup_{h \in \mathscr{I}} |\hat{J}(h) - \hat{J}\{H(h)\}| \leqslant Cn^{-1},$$

no matter what the sample $\mathscr{X} = \{X_1, ..., X_n\}$ or the value of n. This inequality entails $|J(h) - J\{H(h)\}| \leqslant Cn^{-1}$ for all $h \in \mathscr{I}$, and so with $\Delta \equiv \hat{J} - J$ we have

$$\sup_{h \in \mathscr{I}} |\Delta(h) - \Delta\{H(h)\}| \leqslant 2Cn^{-1} \tag{5.7}$$

uniformly in samples $\mathscr{X}$.

Suppose we prove that for some $\eta > 0$, $C_0 > 0$, and all sufficiently large n,

$$\inf_{h \in \mathscr{I}} J(h) \geqslant C_0 n^{-1/2+\eta}. \tag{5.8}$$

Taking $\varepsilon = n^{-(1-\eta)/2}$ in the lemma we see that for large n,

$$\begin{aligned} P\{\sup_{1 \leqslant j \leqslant m} |\Delta(h_j)| > n^{-(1-\eta)/2}\} &\leqslant \sum_{j=1}^{m} P\{|\Delta(h_j)| > n^{-(1-\eta)/2}\} \\ &\leqslant C_1 m \exp(-C_2 n^{\eta}). \end{aligned}$$

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

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

implying (by the Borel–Cantelli lemma) that

$$n^{(1-\eta)/2} \sup_{1 \leqslant j \leqslant m} |\Delta(h_j)| \to 0$$

almost surely. In view of (5.7) this entails

$$n^{(1-\eta)/2} \sup_{h \in \mathscr{I}} |\Delta(h)| \to 0$$

almost surely, and together with (5.8) this gives

$$\{\inf_{h \in \mathscr{I}} \hat{J}(h)\}/\{\inf_{h \in \mathscr{I}} J(h)\} \to 1 \qquad \text{and} \qquad \hat{J}(\hat{h}^*)/J(\hat{h}^*) \to 1$$

almost surely. Theorem 3.1 follows: note the property mentioned in the second sentence of the previous paragraph.

It remains to prove (5.8). Recall result (5.6): $3J \geq J_1 + J_2$, where $J_1 \equiv \int E\,|\hat{f} - E\hat{f}|$ and $J_2 \equiv \int |E\hat{f} - f|$. We know from Lemma 5.1 that for $h \leq 1$, $J_1 \geq C \min\{(nh^d)^{-1/2}, 1\}$. Inequality (5.8) follows from these estimates and the fact that for some $\xi > 0$,

$$J_2 \geq C \min(h^\xi, 1). \tag{5.9}$$

We conclude by proving (5.9).

Since K is bounded then for each $\varepsilon > 0$ there exists $h' > 0$ such that $|\hat{f}| \leq \varepsilon$ for all $n \geq 1$, all x, all samples $\mathscr{X}$, and all $h \geq h'$. Therefore we trivially have (5.9) for sufficiently large h—say for $h \geq h''$. If $h < h''$ then

$$|E\hat{f} - f| \leq C \equiv 2(\sup f) \int_{\mathbb{R}^d} |K|,$$

so that

$$J_2 \geq C^{-1} \int_{\mathbb{R}^d} |E\hat{f} - f|^2 \geq C' h^\xi$$

for some $\xi > 0$, the second inequality following from Lemma 1 of Stone [14].

Proof of Theorem 3.2. (i) *Proof of* (3.11). Result (3.11) follows from

$$\int_{-\infty}^{\infty} |E\hat{f}_1^{(p)} - f^{(p)}| \to 0, \tag{5.10}$$

$$\int_{-\infty}^{\infty} |\hat{f}_1^{(p)} - E\hat{f}_1^{(p)}| \to 0 \qquad \text{almost surely,} \tag{5.11}$$

and we prove these limit theorems separately.

(i.a) *Proof of* (5.10). Observe that

$$E\hat{f}_1^{(p)}(x) - f^{(p)}(x) = \int_{-\infty}^{\infty} K_1(z)\{f^{(p)}(x - h_1 z) - f^{(p)}(x)\}\, dz,$$

and so by continuity of $f^{(p)}$ and compact support of K_1,

$$\sup_{|x| \leq C} |E\hat{f}_1^{(p)}(x) - f^{(p)}(x)| \to 0$$

for each $C > 0$. If K_1 vanishes outside the interval $[-s, s]$, and if $h_1 s \leq \frac{1}{2} C$, then

$$\int_{|x| \leq C} |E\hat{f}_1^{(p)} - f^{(p)}| \leq 2 \left(\int_{-\infty}^{\infty} |K_1| \right) \int_{|x| > (1/2)C} |f^{(p)}|.$$

Result (5.10) follows from the last two displayed estimates.

(i.b) *Proof of* (5.11). We begin by stating a version of Bernstein's inequality (see, e.g., Hoeffding [10, p. 17]), which we need on three occasions.

LEMMA 5.3. *If $Y_1, ..., Y_n$ are independent and identically distributed with zero mean and variance σ^2, and if each $|Y_j| \leqslant c$, then*

$$P\left(\left|\sum_{j=1}^{n} Y_j\right| > t\right) \leqslant 2\exp\{-\tfrac{1}{2}t^2(n\sigma^2 + ct)^{-1}\}, \qquad \text{all} \quad t > 0.$$

For any $\xi > 0$, the integral on the left-hand side of (5.11) is dominated by

$$\int_{|x| \leqslant \xi} |\hat{f}_1^{(p)} - E\hat{f}_1^{(p)}| + \int_{|x| > \xi} |\hat{f}_1^{(p)}|$$

$$+ \int_{|x| > \xi} |f^{(p)}| + \int_{-\infty}^{\infty} |E\hat{f}_1^{(p)} - f^{(p)}|.$$

Therefore it suffices to show that for some sequence $\xi = \xi(n)$ diverging to $+\infty$, we have

$$\int_{|x| > \xi} |\hat{f}_1^{(p)}| \to 0 \qquad \text{almost surely,} \tag{5.12}$$

$$\int_{|x| \leqslant \xi} |\hat{f}_1^{(p)} - E\hat{f}_1^{(p)}| \to 0 \qquad \text{almost surely.} \tag{5.13}$$

If the support of K_1 is confined to $[-s, s]$, and if h_1 is so small and ξ so large that $h_1 s \leqslant \frac{1}{2}\xi$, then the left-hand side of (5.12) is dominated by

$$(nh_1^{p+1})^{-1} \sum_{j=1}^{n} \int_{|x| > \xi} |K_1^{(p)}\{(x - X_j)/h_1\}|\, dx$$

$$\leqslant C_1(nh_1^p)^{-1} \sum_{j=1}^{n} \int_{|X_j + h_1 x| > \xi; |x| \leqslant s} dx$$

$$\leqslant 2C_1 s(nh_1^p)^{-1} \sum_{j=1}^{n} I(|X_j| > \tfrac{1}{2}\xi), \tag{5.14}$$

where $C_1 \equiv \sup |K_1^{(p)}|$. Suppose $E(|X_1|^\alpha) < \infty$, where $\alpha > 1$. Then $\pi \equiv P(|X_1| > \frac{1}{2}\xi) \leqslant C_2 \xi^{-\alpha}$, and so if we take $\xi \equiv h_1^{-p/\beta}$, where $(2\alpha)/(\alpha + 1) < \beta < \alpha$, we have

$$E\{(nh_1^p)^{-1} \sum_{j=1}^{n} I(|X_j| > \tfrac{1}{2}\xi)\} \leqslant C_2 h_1^{p(\alpha - \beta)/\beta} \to 0.$$

Furthermore, for each $\varepsilon > 0$ we have by Lemma 5.3,

$$q \equiv P\left[\left|\sum_{j=1}^{n} \{I(|X_j| > \tfrac{1}{2}\xi) - \pi\}\right| > \varepsilon n h_1^p\right]$$
$$\leqslant 2\exp[-\tfrac{1}{2}(\varepsilon n h_1^p)^2 \{n\pi(1-\pi) + \varepsilon n h_1^p\}^{-1}].$$

Now, $\pi(1-\pi) \leqslant C_2 h_1^{p\alpha/\beta} \leqslant C_2 h_1^p$, and so

$$q \leqslant 2\exp\{-C_3(\varepsilon) n h_1^p\} = O(n^{-k})$$

for all $k > 0$, since $n h_1^{2p+1} \to \infty$. Therefore, by the Borel–Cantelli lemma,

$$(n h_1^p)^{-1} \sum_{j=1}^{n} I(|X_j| > \tfrac{1}{2}\xi) \to 0$$

almost surely, which together with (5.14) proves (5.12).

To establish (5.13), put $\tau^2(x) \equiv \max\{\tau_1^2(x), (1+|x|^{2\alpha})^{-1}\}$, where

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

and let $\mathscr{S}_c$ denote the set of values of $x \in (0, c)$ such that $(1+|x|^\alpha)\tau^2(x) > 2$. We shall prove separately that

$$\int_{|x| \leqslant \xi; x \in \mathscr{S}_\infty} |\hat{f}_1^{(p)} - E\hat{f}_1^{(p)}| \to 0 \qquad \text{almost surely,} \tag{5.15}$$

$$\int_{|x| \leqslant \xi; x \notin \mathscr{S}_\infty} |\hat{f}_1^{(p)} - E\hat{f}_1^{(p)}| \to 0 \qquad \text{almost surely.} \tag{5.16}$$

As a prelude to deriving (5.15) we show that the Lebesgue measure of $\mathscr{S}_\infty$, which we denote by $\mathscr{L}(\mathscr{S}_\infty)$, is bounded. To prove this, let Y_c have the uniform distribution on $(0, c)$, and observe that $(1+|x|^\alpha)\tau^2(x) > 2$ if and only if $(1+|x|^\alpha)\tau_1^2(x) > 2$. By Markov's inequality,

$$c^{-1}\mathscr{L}(\mathscr{S}_c) \leqslant \tfrac{1}{2} E\{(1+|Y_c|^\alpha)\tau_1^2(Y_c)\},$$

and so

$$\mathscr{L}(\mathscr{S}_\infty) \leqslant \frac{1}{2}\int_0^\infty (1+|x|^\alpha)\tau_1^2(x)\, dx < \infty$$

uniformly in $h_1 \leqslant 1$, since $E(|X_1|^\alpha) < \infty$. For each $\varepsilon > 0$, the left-hand side of (5.15) is dominated by $\varepsilon\mathscr{L}(\mathscr{S}_\infty) + h_1^{-(p+1)}(2\sup|K_1^{(p)}|)M_1$, where

$$M_1 \equiv \int_{\mathscr{S}_\infty} I\left[\left|\sum_{j=1}^{n} \{K_1^{(p)}((x-X_j)/h_1) - EK_1^{(p)}((x-X_j)/h_1)\}\right| > \varepsilon n h_1^{p+1}\right] dx.$$

The desired result (5.15) will follow from this observation via Markov's inequality and the Borel–Cantelli lemma if we show that for all $\varepsilon > 0$ and $k > 0$, $E(M_1) = O(n^{-k})$. To establish the latter bound, take

$$Y_j \equiv K_1^{(p)}\{(x - X_j)/h_1\} - EK_1^{(p)}\{(x - X_j)/h_1\},$$
$$c \equiv 2 \sup |K_1^{(p)}|, \qquad t \equiv \varepsilon n h_1^{p+1}$$

in Lemma 5.3. Then $\sigma^2 \leqslant h_1 \tau_1^2(x) \leqslant C_1 h_1$, the latter inequality holding since f and $K_1^{(p)}$ are bounded. Therefore

$$\begin{aligned} \tfrac{1}{2} t^2 (n\sigma^2 + ct)^{-1} &\geqslant C_2(\varepsilon)(nh_1^{p+1})^2 \{nh_1 \tau_1^2(x) + nh_1^{p+1}\}^{-1} \\ &\geqslant C_3(\varepsilon)\, nh_1^{2p+1}, \end{aligned}$$

and so, since $nh_1^{2p+1}/\log n \to \infty$,

$$\begin{aligned} E(M_1) &\leqslant 2 \int_{\mathscr{S}_\infty} \exp\{-C_3(\varepsilon)\, nh_1^{2p+1}\}\, dx \\ &= 2\mathscr{L}(\mathscr{S}_\infty) \exp\{-C_3(\varepsilon)\, nh_1^{2p+1}\} \\ &= O(n^{-k}) \end{aligned}$$

for all $k > 0$, as required.

Finally, we derive (5.16). For each $\varepsilon > 0$, the left-hand side of (5.16) is dominated by

$$\varepsilon \int_{-\infty}^{\infty} \tau(x)^{\beta/\alpha}\, dx + h_1^{-(p+1)}(2 \sup |K_1^{(p)}|)\, M_2,$$

where

$$M_2 \equiv \int_{|x| \leqslant \xi; x \notin \mathscr{S}_\infty} I\left\{\left|\sum_{j=1}^{n} Y_j\right| > \varepsilon n h_1^{p+1} \tau(x)^{\beta/\alpha}\right\} dx$$

and Y_j is as defined in the previous paragraph. Now, by Hölder's inequality,

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

uniformly in $h_1 \leqslant 1$, using the fact that $E(|X_1|^{\alpha}) < \infty$ and $\alpha\beta > 2\alpha - \beta$. Therefore (5.16) will follow via Markov's inequality and the Borel–Cantelli

lemma if we show that $E(M_2) = O(n^{-k})$ for all $\varepsilon > 0$ and $k > 0$. Apply Lemma 5.3 once more, with the same values of Y_j and c but this time with $t \equiv \varepsilon n h_1^{p+1} \tau(x)^{\beta/\alpha}$. Since $\sigma^2 \leqslant h_1 \tau^2(x)$ then

$$\begin{aligned} T = T(x) &\equiv \tfrac{1}{2} t^2 (n\sigma^2 + ct)^{-1} \\ &\geqslant C_4(\varepsilon)(nh_1^{p+1})^2 \tau(x)^{2\beta/\alpha} \\ &\quad \times \{nh_1 \tau(x)^2 + nh_1^{p+1} \tau(x)^{\beta/\alpha}\}^{-1}. \end{aligned} \tag{5.17}$$

To bound the right-hand side of (5.17) we treat two cases separately. First of all, suppose $h_1^p \leqslant \tau(x)^{2-(\beta/\alpha)}$. Then

$$\begin{aligned} T &\geqslant \tfrac{1}{2} C_4(\varepsilon)\, nh_1^{2p+1} \tau(x)^{-2(\alpha-\beta)/\alpha} \\ &\geqslant 2^{(\beta/\alpha)-2} C_4(\varepsilon)\, nh_1^{2p+1} (1 + |x|^\alpha)^{(\alpha-\beta)/\alpha}, \end{aligned}$$

the second inequality holding when $x \notin \mathscr{S}_\infty$, since the latter statement means $\tau(x)^{-2} > \frac{1}{2}(1 + |x|^\alpha)$. Second, suppose $h_1^p > \tau(x)^{2-(\beta/\alpha)}$. Then

$$\begin{aligned} T &\geqslant \tfrac{1}{2} C_4(\varepsilon)\, nh_1^{p+1} \tau(x)^{\beta/\alpha} \geqslant \tfrac{1}{2} C_4(\varepsilon)\, nh_1^{p+1} (1 + |x|^{2\alpha})^{-\beta/2\alpha} \\ &\geqslant C_5(\varepsilon)\, nh_1^{2p+1}, \end{aligned}$$

the second inequality holding because $\tau(x)^2 \geqslant (1 + |x|^{2\alpha})^{-1}$ and the third holding when $|x| \leqslant \xi = h_1^{-p/\beta}$. Combining both these bounds we conclude that $T \geqslant C_6(\varepsilon)\, nh_1^{2p+1}$ when $|x| \leqslant \xi$ and $x \notin \mathscr{S}_\infty$. Therefore

$$\begin{aligned} E(M_2) &\leqslant 2 \int_{|x| \leqslant \xi; x \notin \mathscr{S}_\infty} e^{-T(x)}\, dx \\ &\leqslant 4\xi \exp\{-C_6(\varepsilon)\, nh_1^{2p+1}\} = O(n^{-k}) \end{aligned}$$

for all $k > 0$, as had to be shown.

(ii) *Proof of* (3.12). Observe that

$$\begin{aligned} \int_{-\infty}^{\infty} |\hat{f}_2^{1/2} - f^{1/2}| &\leqslant \int_{-\infty}^{\infty} |\hat{f}_2 - f|^{1/2} \\ &\leqslant \left\{ \int_{-\infty}^{\infty} |\hat{f}_2(x) - f(x)|\, (1 + |x|^\alpha)\, dx \right\}^{1/2} \\ &\quad \times \left\{ \int_{-\infty}^{\infty} (1 + |x|^\alpha)^{-1}\, dx \right\}^{1/2}, \end{aligned}$$

where (as before) $\alpha > 1$ is chosen so that $E(|X_1|^\alpha) < \infty$. It is well known that under the conditions of Theorem 3.2, $\int |\hat{f}_2 - f| \to 0$ almost surely.

(See, e.g., Theorem 1, p. 12 of [5].) Therefore it suffices to prove that with probability one,

$$\lim_{C \to \infty} \limsup_{n \to \infty} \int_{|x| > C} \hat{f}_2(x)(1 + |x|^\alpha)\, dx = 0. \tag{5.18}$$

Suppose K_2 vanishes outside the interval $[-s, s]$, and let h_2 be so small that $h_2 s \leqslant \frac{1}{2} C$. The integral on the left-hand side of (5.18) equals

$$\begin{aligned} n^{-1} \sum_{j=1}^{n} & \int_{|X_j + h_2 y| > C} K_2(y)(1 + |X_j + h_2 y|^\alpha)\, dy \\ & \leqslant 2^\alpha (\sup |K_2|)\, n^{-1} \sum_{j=1}^{n} (1 + |X_j|^\alpha)\, I(|X_j| > \tfrac{1}{2} C) \\ & \quad + (2h_2)^\alpha \int_{-\infty}^{\infty} |y|^\alpha K_2(y)\, dy. \end{aligned}$$

Result (5.18) follows from this inequality via the strong law of large numbers.

Proof of (2.10). The variable $N(x)$ appearing in the definition of $\hat{f}(x)$ is Binomial, and from this fact it follows easily that $\hat{f}(x) - E\hat{f}(x) = (nh^d)^{-1/2} f(x)^{1/2} Z(x, n)$, where $Z(x, n)$ is asymptotically Normal $N(0, 1)$. Notice too that for $y \in A(x)$,

$$f(y) = f(x) + \sum_{j=1}^{d} (y - x)_j f_j(x) + o(h),$$

so that

$$\begin{aligned} n^{-1} E\{N(x)\} &= \int_{A(x)} f(y)\, dy \\ &= h^d f(x) + \sum_{j=1}^{d} h^{d-1} f_j(x) \int_{a_j - (1/2)h}^{a_j + (1/2)h} (t - x_j)\, dt + o(h^{d+1}) \\ &= h^d f(x) + h^d \sum_{j=1}^{d} f_j(x)(a_j - x_j) + o(h^{d+1}). \end{aligned}$$

Therefore $E\hat{f}(x) - f(x) = \sum_j (a_j - x_j) f_j(x) + o(h)$.

Proof of (2.12). The variables $(N_1(x), N_2(x), n - N_1(x) - N_2(x))$ have a joint multinomial distribution, and $N_1(x)$ and $N_2(x)$ are asymptotically independent and identically Normally distributed. From this fact it follows that $\hat{f}(x) - E\hat{f}(x) = (nh)^{-1/2} f(x)^{1/2} \{(1 - \eta)^2 + \eta^2\}^{1/2} Z(x, n)$, where $Z(x, n)$

is asymptotically Normal $N(0, 1)$. Also, writing $a_1 \equiv a$ and $a_2 \equiv a + h$, we have

$$\begin{aligned} n^{-1}E\{N_j(x)\} &= \int_{A_j(x)} f(y)\,dy \\ &= \int_{a_j-(1/2)h}^{a_j+(1/2)h} \left\{ f(x) + (y-x)\,f'(x) \right. \\ &\qquad \left. + \frac{1}{2}(y-x)^2 f''(x) \right\} dy + o(h^3) \\ &= h\left[f(x) + (a_j - x)\,f'(x) \right. \\ &\qquad \left. + \frac{1}{6}\left\{ 3(a_j-x)^2 + \frac{1}{4}h^2 \right\} f''(x) \right] + o(h^3). \end{aligned}$$

Therefore, noting that $x - a = \eta h$,

$$\begin{aligned} E\{\hat{f}(x)\} &= (nh)^{-1}\,E\{(1-\eta)\,N_1(x) + \eta N_2(x)\} \\ &= f(x) + \tfrac{1}{6}h^2 f''(x)(3\eta - 3\eta^2 + \tfrac{1}{4}) + o(h^2), \end{aligned}$$

after a little algebra.

ACKNOWLEDGMENT

We are grateful to Luc Devroye for supplying a copy of [4] prior to publication.

REFERENCES

[1] VON BAHR, B. (1965). On the convergence of moments in the central limit theorem. *Ann. Math. Statist.* **36** 808–818.

[2] BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A* **25** 245–254.

[3] DEHEUVELS, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl.* **25** 5–42.

[4] DEVROYE, L. (1986). The kernel estimate is relatively stable. Preprint.

[5] DEVROYE, L., AND GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

[6] EPANECHNIKOV, V. A. (1969). Nonparametric estimates of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

[7] HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In *Proceedings, Sixth International Symposium on Multivariate Analysis* (P. R. Krishnaiah, Ed.), pp. 289–309, North-Holland, Amsterdam.

[8] HALL, P., AND MARRON, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581.

[9] HALL, P., AND WAND, M. P. (1987). On the minimization of absolute distance in kernel density estimation. *Statist. Probab. Lett.*, in press.

[10] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[11] PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation.* Academic Press, New York.

[12] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

[13] SCOTT, D. W. (1985). Frequency polygons: Theory and application. *J. Amer. Statist. Assoc.* **80** 348–354.

[14] STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

[15] WAND, M. P. (1988). *Nonparametric Density Estimation and Discrimination.* Ph.D. thesis, Australian National University.

[16] WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.* **41** 1665–1671.