Zero-Bias Locally Adaptive Density Estimators

Stephan R. Sain[1] and David W. Scott[2]

May 1, 2001

SUMMARY: Strategies for improving fixed nonnegative kernel estimators have focused on reducing the bias, either by employing higher-order kernels or by locally adjusting the bandwidth. Intuitively, bandwidths in the tails should be relatively larger in order to reduce wiggles since there is less data available in the tails. We show that in regions where the density function is convex, it is theoretically possible to find local bandwidths such that the pointwise bias is exactly zero. The corresponding pointwise mean squared error converges at the parametric rate of $O(n^{-1})$ rather than the slower $O(n^{-4/5})$. These so-called zero-bias bandwidths are constant and are usually orders of magnitude larger than the optimal locally adaptive bandwidths predicted by asymptotic mean squared error analysis. We describe data-based algorithms for estimating zero-bias bandwidths over intervals where the density is convex. We find that our particular density estimator attains the usual $O(n^{-4/5})$ rate. However, we demonstrate that the algorithms can provide significant improvement in mean squared error, often clearly visually superior curves, and a new operating point in the usual bias-variance tradeoff.

KEY WORDS: Variable Bandwidth Kernel, Bias Reduction, Cross-Validation.

[1]Stephan R. Sain is Assistant Professor, Department of Statistical Science, Southern Methodist University, P.O. Box 750332, Dallas, TX 75275-0332 USA.

[2]David W. Scott is Professor, Department of Statistics MS-138, Rice University, 6100 Main Street, Houston, TX 77005-1892 USA. Supported in part by NSF DMS-9971797.

1. INTRODUCTION

The constant or fixed bandwidth kernel density estimator is given by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i), \tag{1}$$

where $K_h(x) = (1/h)K(x/h)$ and h is the smoothing parameter or bandwidth. The kernel $K(\cdot)$ is taken to be a nonnegative, symmetric function integrating to one. This nonnegative kernel estimator has pointwise bias of order $O(h^2 f''(x))$, pointwise variance of order $O(f(x)/nh)$, and resulting optimal mean squared error (MSE) of order $O(n^{-4/5})$. Recent surveys of kernel estimators include Silverman (1986), Thompson and Tapia (1990), Scott (1992), Wand and Jones (1995), and Simonoff (1996).

While the fixed kernel estimator represents a significant improvement over the histogram, many authors have sought to further improve upon this basic design by modifying the kernel or by varying the bandwidth in some fashion. Bartlett (1963) removed the nonnegativity restriction on the kernel, achieving estimators that exhibit bias of order $O(h^p f^{(p)}(x))$ with a mean squared error of order $O(n^{-2p/(2p+1)})$ for even order $p > 0$ kernels. Thus, if the density is sufficiently smooth, substantial improvement over the nonnegative kernel ($p = 2$) is asymptotically possible.

An alternative approach has focused on varying the bandwidth in some fashion in an attempt to better capture the "local" behavior exhibited by the underlying density. There are two formulations of primary interest. The first, which is the focus of this paper, is referred to as a *local density* or *balloon estimator*. The local density estimator varies the bandwidth with the estimation point, i.e.

$$\hat{f}_1(x) = \frac{1}{nh(x)}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)}\right) = \frac{1}{n}\sum_{i=1}^{n} K_{h(x)}(x - x_i). \tag{2}$$

Note that the bandwidth, $h(x)$, is written explicitly as a function of the estimation point. The first practical estimator of this form was the multivariate nearest neighbor estimator of Loftsgaarden and Quesenberry (1965). The nearest neighbor estimator has been criticized in the univariate setting (Silverman, 1986), but Terrell and Scott (1992) have shown that such an estimator performs nicely in dimensions beyond four.

The second variable kernel estimator was introduced by Breiman, Meisel, and Purcell (1977), who suggested varying the bandwidth at each sample point. The form of the *sample point adaptive estimator* is

$$\hat{f}_2(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{g(x_i)}K\left(\frac{x-x_i}{g(x_i)}\right) = \frac{1}{n}\sum_{i=1}^{n}K_{g(x_i)}(x-x_i). \tag{3}$$

Note that the bandwidth, $g(x_i)$, is now written explicitly as a function of the sample point. The most widely studied implementation of this type is the square-root rule introduced by Abramson (1982). The bandwidth in (3) is chosen so that $g(x_i) \propto f(x_i)^{-1/2}$, with f replaced by a pilot estimate. Abramson's proposal is of interest because the pointwise bias is of order $O(h^4)$, similar to the negative $p = 4$ kernels. However, Terrell and Scott (1992), Hall, Hu, and Marron (1995), and Sain and Scott (1996) discuss implementation problems as well as poor global performance in certain situations. Other work by Fan et al. (1993), Sain (1994), and Sain and Scott (1996) has focused on the theoretical details and practical issues associated with estimating optimal bandwidths for the general adaptive estimator with some success.

Jones (1990) gives a clear description of these two types of variable kernel estimators and compares them to the fixed bandwidth estimator. In a comparison of the fixed bandwidth estimator and the local density estimator using the bandwidth that minimizes the asymptotic pointwise mean-squared error (AMSE),

$$h^*(x) = \left[\frac{R(K)f(x)}{\sigma_K^4 f''(x)^2}\right]^{1/5} n^{-1/5}, \tag{4}$$

Jones (1990) showed that some modest improvement over the fixed bandwidth can be made.

This paper describes the surprising theoretical result that at points where the density is convex a local bandwidth can be found yielding exactly zero bias, and that these so-called "zero-bias bandwidths" vastly improve upon $h^*(x)$ (Section 2). Furthermore, no additional assumptions about the form of the underlying density are necessary. Estimation of the zero-bias bandwidth is discussed in Section 3 using a cross-validation procedure based on the work of Hall and Schucany (1989) and Mielniczuk et al. (1989). While the existence of the zero-bias bandwidth suggest kernel estimates with a higher-order MSE convergence rate in theory, the reality in practice, using a data-based estimate of the zero-bias bandwidth, is that error rate is no faster than what is traditional ($O(n^{-4/5})$). However, using the zero-bias bandwidth as a target, a different bias-variance trade-off is achieved that offers improved performance of estimates in the tails. This is demonstrated through the simulation study in Section 4 and the example in Section 5.

2. EXISTENCE OF ZERO-BIAS BANDWIDTHS

To illustrate the existence of the zero-bias bandwidth, consider the univariate locally adaptive estimator (2) with standard normal kernel

$$\hat{f}(x|h_x) = \frac{1}{n}\sum_{i=1}^{n} \phi_{h_x}(x - x_i), \tag{5}$$

where $\phi_\sigma(x-\mu)$ is a $N(\mu, \sigma^2)$ density and h_x denotes the bandwidth $h(x)$. Further, let the true density be a normal mixture of the form

$$f(x) = \sum_{i=1}^{k} w_i \phi_{\sigma_i}(x - \mu_i),$$

where $\sum w_i = 1$. Following Marron and Wand (1992), an exact expression for the pointwise MSE $=$ Bias^2 $+$ Var may be obtained using the formulae

$$\text{Bias}[\hat{f}(x|h_x)] = \mu_1(x|h_x) - f(x) \qquad \text{Var}[\hat{f}(x|h_x)] = \frac{1}{n}[\mu_2'(x|h_x) - \mu_1(x|h_x)^2], \tag{6}$$

where

$$\begin{aligned} \mu_1(x|h) = E[K_h(x-X)] &= \int \phi_h(x-t) f(t)\, dt \\ &= \sum_{i=1}^{k} w_i \int \phi_h(x-t)\phi_{\sigma_i}(t-\mu_i) dt \\ &= \sum_{i=1}^{k} w_i \phi_{\sqrt{h^2+\sigma_i^2}}(x-\mu_i) \end{aligned} \tag{7}$$

and

$$\mu_2'(x|h) = E[K_h^2(x-X)] = \frac{1}{2\sqrt{\pi}h}\sum_{i=1}^{k} w_i \phi_{\sqrt{\frac{1}{2}h^2+\sigma_i^2}}(x-\mu_i).$$

Consider estimation of f at some point x in the tail of f. It will be shown in Theorem 2.1 that there exists a bandwidth, $h_0(x) > 0$, such that the bias is exactly zero. As the

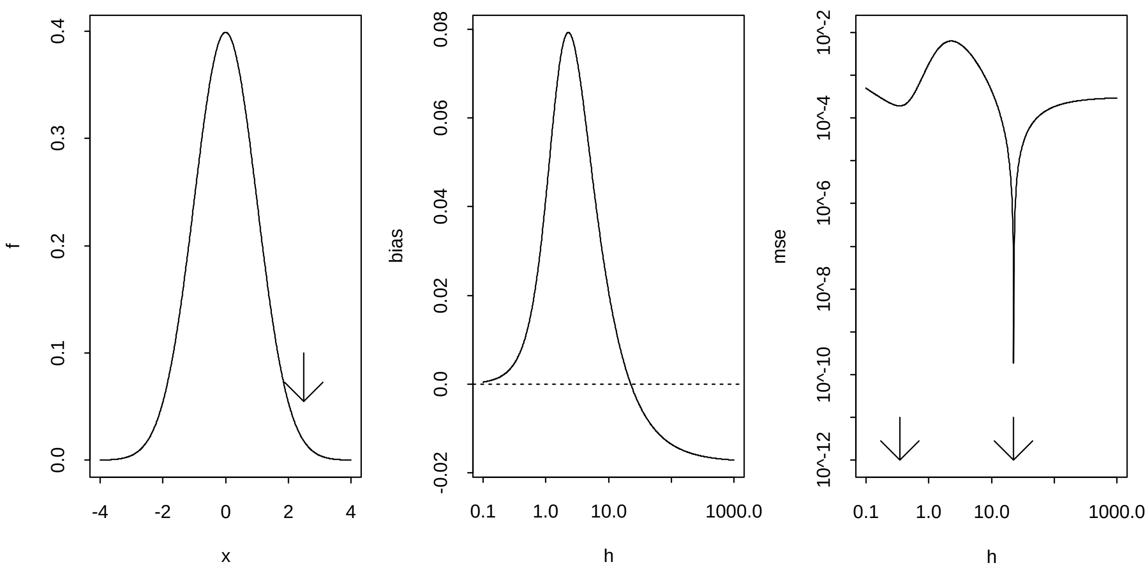

Figure 1: Exact bias and MSE curves for estimation at $x = 2.5$ when $f(x)$ and $K(x)$ are $N(0, 1)$ with $n = 100$.

bias is not a function of the sample size, n, then $h_0(x)$ also does not depend on n, but only on f, K, and x. We often shorten $h_0(x)$ to simply h_0.

An example of the existence a zero-bias bandwidth at $x = 2.5$ is given in Figure 1, where $f(x)$ and $K(x)$ are $N(0, 1)$, and $n = 100$. The value of the zero-bias bandwidth is $h_0 = 22.6$, but the global minimum of the MSE occurs at a slightly larger bandwidth (see Theorem 2.2). A local minimum in the MSE is located at $h = 0.345$, a value predicted by the usual asymptotic theory exemplified by Equation (4). One might ask why the MSE in Figure 1 is so much better near the zero-bias bandwidth, h_0. The answer is straightforward. Usually, large bandwidths will improve variance, but only at the expense of a larger bias. (Observe the rapid increase in the bias curve for $h > 0$.) However, increasing the bandwidth beyond $h = 2$ actually reduces the bias, which vanishes at $h_0 = 22.5$, and finally converges to $-f(2.5)$. The extent of the improvement is summarized in Theorem 2.3.

The following theorem not only demonstrates the existence of at least one zero-bias bandwidth for each x in the appropriate regions, but also gives a sufficient condition under which such bandwidths exist.

Theorem 2.1 (Existence of h_0) *Let $\hat{f}(x|h)$ be the locally adaptive density estimator given in Equation (2) with nonnegative, symmetric kernel supported on the interval* $(-1, 1)$.

Assume $f(x)$ is a continuous density function. Then, for every x in regions where f is convex there exists at least one bandwidth, $h_0 > 0$, such that

$$Bias[\hat{f}(x|h_0)] = 0.$$

Proof: The heart of the proof is an examination of the behavior of the bias as a function of h. Suppose h is small enough so that f is convex over the interval $(x-h, x+h)$. By an application of Jensen's inequality,

$$\begin{aligned} E[\hat{f}(x|h)] = E[K_h(x-X)] &= \int_{x-h}^{x+h} K_h(x-y) f(y)\, dy \\ &\geq f\left(\int_{x-h}^{x+h} y K_h(x-y)\, dy\right) = f(x). \end{aligned} \tag{8}$$

Therefore, $\text{Bias}[\hat{f}(x|h)] \geq 0$. Clearly from (8), $E[\hat{f}(x|h)] = K_h * f(x)$, where "$*$" denotes convolution, and is therefore a continuous function of h. As $h \to 0$, $K_h(w)$ acts like the Dirac delta function; hence, $E[\hat{f}(x|h)] \to f(x)$, and the $\text{Bias}[\hat{f}(x|h)] \to 0$. Similarly, as $h \to \infty$, $K_h \to 0$ everywhere; hence, the expectation vanishes, and the $\text{Bias}[\hat{f}(x|h)] \to -f(x) < 0$. Since the bias is a continuous function of h, there must be at least one finite $h_0 > 0$ where the bias function equals zero.

Remark 1. While a key section of the proof involves kernels with finite support, the theorem holds for all K. The following heuristic argument can be made rigorous. Writing the bias again as a convolution, $\text{Bias}[\hat{f}(x|h)] = (K_h * f - f)(x)$. Clearly $\int \text{Bias}[\hat{f}(x|h)]\, dx = 0$ for all $h > 0$. Hence, it follows that there must exist intervals over which the bias is positive in order for the integrated bias to equal zero. By continuity, there must exist at least one value of h at some point x at which the bias is exactly zero. Note that this argument does not give specific information about the location of such x.

Remark 2. Alternatively, for a general order $p = 2$ kernel, a sufficient condition on the existence of h_0 can be found through examination of the traditional Taylor's series bias approximation, $\text{Bias}[\hat{f}(x|h)] = \frac{1}{2}h^2\sigma_K^2 f''(x) + o(h^2)$. When f is convex, f'' is positive, and the bias is also positive for sufficiently small h. Hence, h_0 exists. This approach requires more assumptions about f and f'' than in Theorem 2.1.

Remark 3. The argument in the previous remark extends to both higher-order kernels and multivariate data. For higher-order kernels, $\text{Bias}[\hat{f}_p(x|h)] = \frac{1}{p!}h^p\mu_p f^{(p)}(x) + o(h^p)$, where

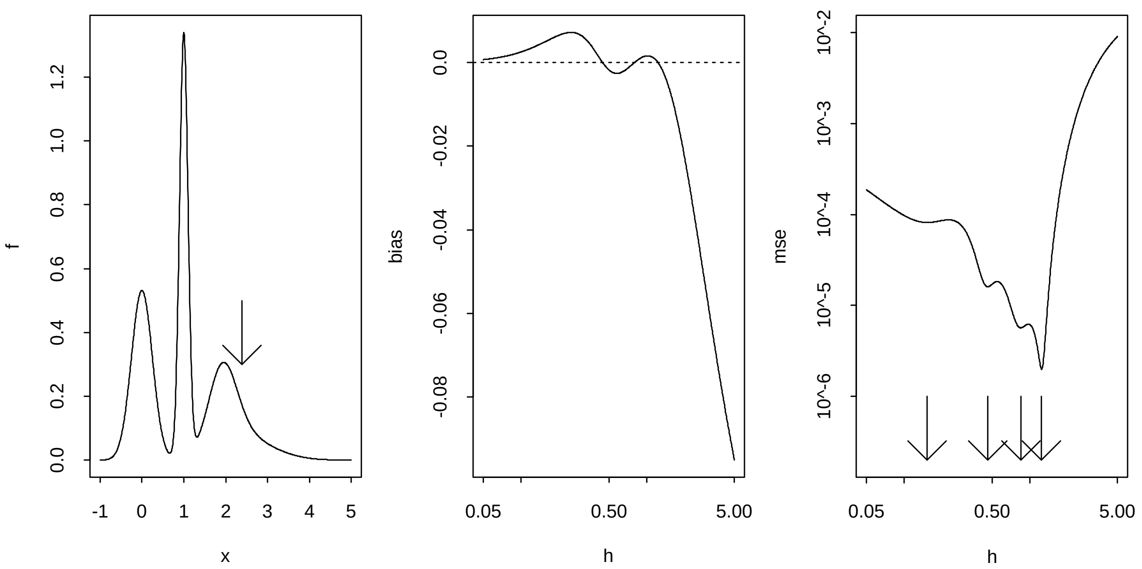

Figure 2: Exact bias and MSE curves at $x = 2.395$ for a Normal mixture with $n = 5000$.

$\mu_p = \int x^p K(x)\, dx$. Hence, for small h, the bias will be positive at x where $\mu_p f^{(p)}(x) > 0$ (implying a slightly different sufficient condition than is stated in Theorem 2.1 and Remark 2), and the same argument for existence of a zero-bias bandwidth holds. Finally, consider a bivariate product kernel estimator with single smoothing parameter, h, so that $\text{Bias}[\hat{f}(\mathbf{x}|h)] = \frac{1}{2}h^2\sigma_K^2[f_{11}(\mathbf{x}) + f_{22}(\mathbf{x})]$, where $\mathbf{x} = (x_1, x_2)^T$ and $f_{ii} = \partial^2 f/\partial x_i^2$. As long as $f_{11}(\mathbf{x}) + f_{22}(\mathbf{x}) > 0$, a zero-bias bandwidth exists for some h at $\mathbf{x}$. For the $N(\mathbf{0}, \mathbf{I}_2)$ density, such points exist when $x_1^2 + x_2^2 > 2$. See Sain (2001) for further discussion of such extensions.

Remark 4. The existence of the zero-bias bandwidth has been noted in the literature (see Scott, 1992, p. 137; Sain, 1994; Hazelton, 1998; Devroye and Lugosi, 1998). However, it is commonly held to be a theoretical curiosity and of no practical consequence. We now believe otherwise.

In the case of a multimodal density, h_0 (where it exists) is not necessarily unique for each x. Figure 2 gives an example of just such a situation with a trimodal normal mixture density ($n = 5000$). The parameters of f are given by

$$w = (1/3, 1/3, 1/5, 1/15, 1/15)', \quad \mu = (0, 1, 23/12, 2.2, 2.7)', \quad \sigma = (1/4, 1/10, 1/3, 1/2, 2/3)'$$

At $x = 2.395$, f is convex, and there are three zero-bias bandwidths ($h = 0.464, 0.849, 1.24$),

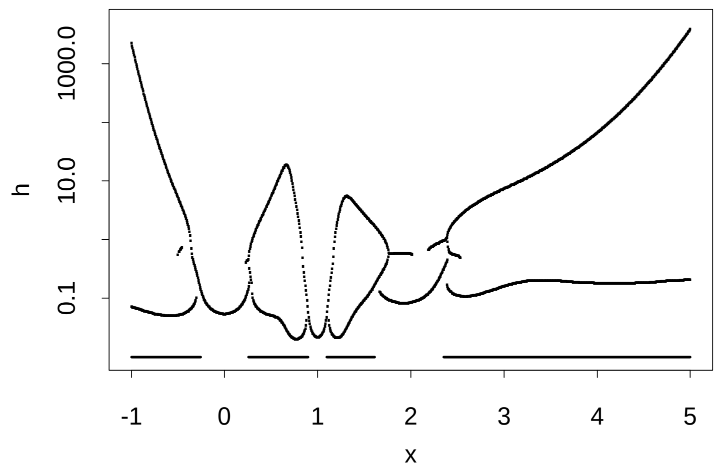

Figure 3: Bandwidths corresponding to local minima of the MSE for the normal mixture density ($n = 5000$). The horizontal line segments just above the x-axis indicate regions where f is convex. (Note that the dotted lines should be connected; each dot represents a numerical search for a root.)

each corresponding to a local minimum in the MSE. The largest of these bandwidths corresponds to the global minimizer of the MSE. All have smaller MSE than the usual asymptotic bandwidth at $h = 0.153$ by an order of magnitude. Figure 3 extends this example, displaying all locally optimal bandwidths for values of $x \in (-1, 5)$. The pattern in Figure 3 is complex and shows some unexpected behavior, which can be explained by a close examination of the behavior of the bias.

Sometimes, $h^*(x)$ and $h_0(x)$ are not sufficiently different to give two distinct minima in the MSE curve. Occasionally, the contributions of the bias from the various modes can also lead to local minima in the MSE that are not associated with a zero crossing of the bias. For mixtures, it is also possible to find zero-bias bandwidths not in convex regions (see Hazelton, 1998 and Sain, 2000).

For the standard Normal density, Figure 4 displays all of the local minima of the exact MSE as a function of both h and the estimation point x with $n = 10^2$, 10^4, and 10^6. Observe that bandwidths corresponding to the zero-bias bandwidths in the extreme tails are much larger than any previously considered in the literature. The rightmost plot contains an overlay of the first three plots and confirms that these larger optimal bandwidths do not vary with n (nor do they converge to zero as does $h^*(x)$). In fact, as the following theorem

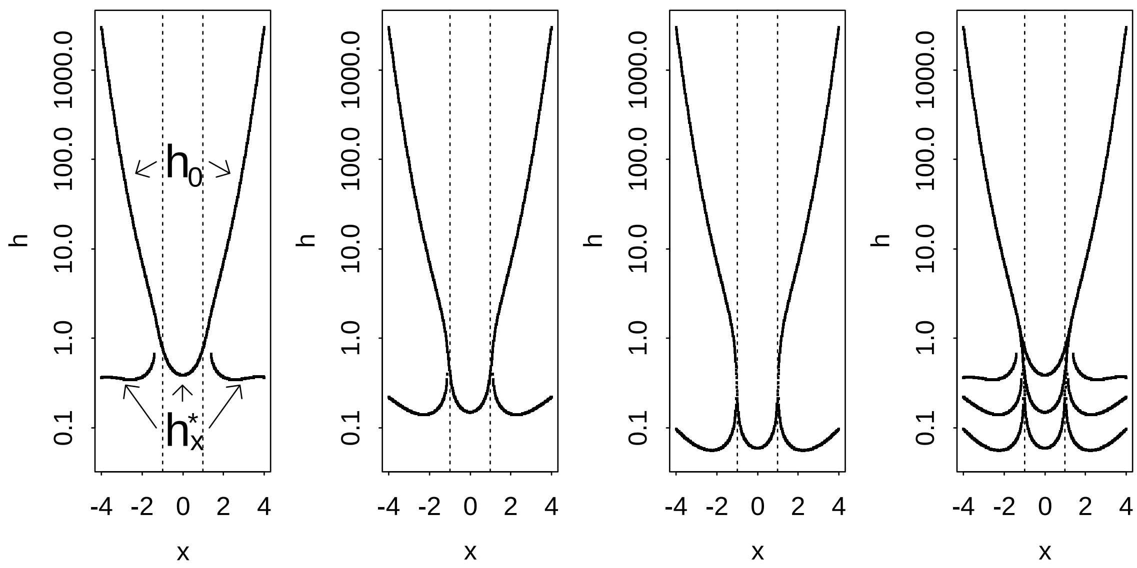

Figure 4: Pointwise local minima of the MSE for $N(0,1)$ data ($n = 10^2, 10^4, 10^6$, and all three combined).

shows, these bandwidths converge quite rapidly to the zero-bias bandwidths.

Theorem 2.2 (Convergence of h_{mse} to h_0) *Assume that $f(x)$ is a sufficiently differentiable density function and that the zero-bias bandwidth $h_0(x)$ exists. Let $h_{mse}(x)$ be a local minimum of the MSE$[\hat{f}(x)]$ in the neighborhood of $h_0(x)$. Then $h_{mse}(x) - h_0(x) = \delta_{x,n} = O(n^{-1})$, where*

$$\delta_{x,n} = \left[\frac{a_1 + h_0(x)a_2 - 2bh_0(x)f(x)}{2b^2h_0(x)^2}\right] n^{-1}, \tag{9}$$

with

$$a_1 = \int K^2(w)f(x - h_0(x)w)\, dw, \qquad a_2 = \int wK^2(w)f'(x - h_0(x)w)\, dw, \tag{10}$$

and

$$b \;=\; \int wK(w)f'(x - h_0(x)w)\, dw. \tag{11}$$

Proof: See Appendix.

Remark. For practical purposes, $h_{mse}(x)$ and $h_0(x)$ are identical. For example, when f is $N(0,1)$, $n = 100$, and $x = 2$, h_{mse} and h_0 differ only in the fourth decimal place. While it may not be clear from the theorem, $\delta \geq 0$ at $x = 2$. To see this consider a bandwidth slightly larger than h_0 and recall that the variance decreases as h increases.

As a result of Theorem 2.2, the convergence rate of MSE$[\hat{f}(x|h_{mse})]$, where h_{mse} is the local minimum corresponding to a zero-bias bandwidth, can be characterized.

Theorem 2.3 (MSE of $\hat{f}(x|h_{mse})$) *Let the conditions of Theorem 2.1 be met and denote the local minimum in the mean squared error corresponding to $h_0(x)$ by $h_{mse}(x)$. Then, MSE$[\hat{f}(x|h_{mse})] = O(n^{-1})$.*

Proof: Clearly, $\text{MSE}[\hat{f}(x|h_{mse})] \leq \text{MSE}[\hat{f}(x|h_0)] = \text{Var}[\hat{f}(x|h_0)] = \frac{1}{n}\text{Var}[K_{h_0}(x-X)]$.

Remark. Note that the rate $O(n^{-1})$ represents a significant improvement over the $O(n^{-4/5})$ rate achieved by the fixed bandwidth procedure using the optimal asymptotic bandwidth h_x^* and actually matches the rate achieved by parametric estimators. While the result is not global in nature, the existence of such a bandwidth yielding $O(n^{-1})$ convergence in the tails points towards a possible breakthrough in nonparametric estimation in regions of low density where data are sparse.

3. ESTIMATING THE ZERO-BIAS BANDWIDTH WITH DATA

The results of the previous sections suggest that an optimal estimation scheme should incorporate the zero-bias bandwidths over intervals in the convex regions of the underlying density . Hall and Schucany (1989) and Mielniczuk, et al. (1989) extend the idea of Rudemo (1982) and Bowman (1984) to consider a local form of least-squares cross-validation by estimating the integrated squared error (ISE) in a neighborhood of x. By minimizing this criterion, an estimator of h_x^* can be found. We show how it can also be used to find an approximation to $h_0(x)$.

Consider the integrated squared error on some closed interval $[x_0-\epsilon, x_0+\epsilon]$,

$$\begin{aligned}\text{ISE}_\epsilon(h) &= \int [\hat{f}(x|h)-f(x)]^2 I(x_0-\epsilon \leq x \leq x_0+\epsilon)\,dx = \int_{x_0-\epsilon}^{x_0+\epsilon} [\hat{f}(x|h)-f(x)]^2 dx \\ &= \int_{x_0-\epsilon}^{x_0+\epsilon} \hat{f}(x|h)^2 dx - 2\int_{x_0-\epsilon}^{x_0+\epsilon} \hat{f}(x|h)f(x)dx + \int_{x_0-\epsilon}^{x_0+\epsilon} f(x)^2 dx, \qquad (12)\end{aligned}$$

where I is the indicator function, i.e. $I(A) = 1$ if $x \in A$ and equals 0 otherwise. The final term in (12) is constant and can be ignored. The first term is solely a function of the kernel estimate and can be calculated directly. The cross-product term in (12) may be estimated

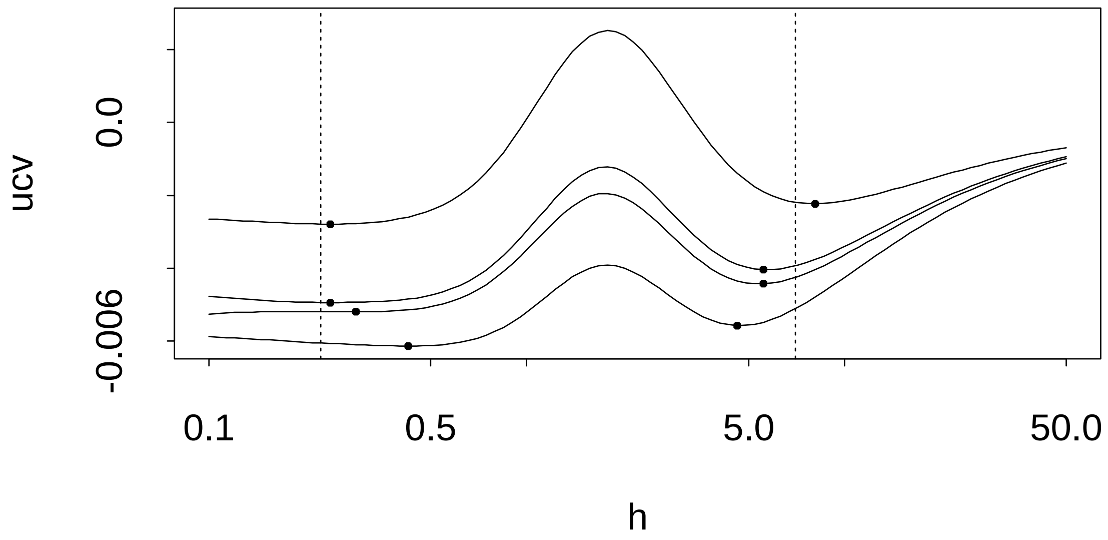

Figure 5: Cross-validation curves for four samples of size $n = 1000$ from a standard normal density with $\epsilon = 0.5$ and $x_0 = 2$. Local minima are indicated by solid dots. Optimal bandwidths are indicated by dotted vertical lines.

by -2 times the average of the leave-one-out estimator,

$$\hat{f}_{-i}(x_i|h) = \frac{1}{n-1}\sum_{j\neq i} K_h(x_i - x_j) \qquad x_i \in [x_0 - \epsilon, x_0 + \epsilon].$$

Putting these expressions together, the interval least-squares cross-validation criterion function is given by

$$\mathrm{UCV}_\epsilon(h) = \int_{x-\epsilon}^{x+\epsilon} \hat{f}(x|h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i|h) I(x_0 - \epsilon \le x_i \le x_0 + \epsilon). \tag{13}$$

Examining UCV_ϵ as a function of the bandwidth usually reveals two local minima, corresponding to $h^*_{x_0}$ and $h_0(x_0)$. Figure 5 shows an example with standard normal data and $n = 1000$. The figure displays cross-validation curves for four samples, using $\epsilon = 0.5$ to estimate the optimal bandwidth for $x_0 = 2$. The local minima in the curves are indicated as well as the two optimal values of the bandwidth, $h^*_{x_0}$ and $h_0(x_0)$, by the dotted vertical lines.

The choice of ϵ affects the behavior of the curves shown in Figure 5. Choosing $\epsilon > 1/2$ would further increase the stability of UCV_ϵ, but would provide a more biased estimate of the optimal bandwidth for $x_0 = 2$. Choosing $\epsilon << 1/2$ would reduce the bias but only a few data points would fall in $[x_0 - \epsilon, x_0 + \epsilon]$.

Let $\hat{h}_\epsilon$ be the local minimum of the UCV$_\epsilon$ criterion close to h_0. The primary result of this section suggests that by choosing ϵ on the order of $n^{-1/5}$, $\text{MSE}(\hat{h}_\epsilon) = E(\hat{h}_\epsilon - h_0)^2 = O(n^{-4/5})$. Before proceeding to that result, we first examine the behavior of h_ϵ, where

$$h_\epsilon = \arg\min_h \int_{x_0-\epsilon}^{x_0+\epsilon} \text{MSE}[\hat{f}(x|h)]\, dx. \tag{14}$$

Note that to study h_ϵ, it is only necessary to consider the bias since it dominates the $\text{MSE}[\hat{f}(x|h)]$ for larger n and h. The justification lies in an examination of the bias and variance as $h \to \infty$. As noted earlier, the bias goes to a constant $(-f(x_0))$ while the variance goes to zero.

Theorem 3.4 (Behavior of h_ϵ) *Let h_ϵ be defined as in (14). Then, for some estimation point x_0 in a convex region of f with corresponding zero-bias bandwidth h_0,*

$$h_\epsilon - h_0 = \beta_1 \epsilon^2 + o(\epsilon^2) \qquad \textit{and} \qquad \textit{MSE}(\hat{f}(x_0|h_\epsilon)) = O(1/n) + O(\epsilon^4),$$

where $\beta_1 = \alpha_3/\alpha_1$ and α_1 and α_3 are given in (18).

Proof: See Appendix.

Next we can establish the behavior of $\hat{h}_\epsilon$ with the following theorems.

Theorem 3.5 (Behavior of $\hat{h}_\epsilon$, Part I) *Let $\hat{h}_\epsilon$ be the local minimum of the cross-validation criterion defined in (13) in the neighborhood of h_ϵ. Then*

$$E(\hat{h}_\epsilon) - h_\epsilon = O(1/(\epsilon n)) \qquad \textit{and} \qquad \textit{Var}(\hat{h}_\epsilon - h_\epsilon) = \beta_2/(\epsilon n) + o(1/(\epsilon n)),$$

where $\beta_2 = f(x_0)/\alpha_1$ and α_1 is given in (18).

Proof: See Appendix.

The final task is to establish the behavior of $\hat{h}_\epsilon$ as an estimator of h_0.

Theorem 3.6 (Behavior of $\hat{h}_\epsilon$, Part II) *Let $\hat{h}_\epsilon$ be the local minimum of the cross-validation criterion defined in (13) in the neighborhood of h_ϵ. Then, by taking $\epsilon = O(n^{-1/5})$, $\textit{MSE}(\hat{h}_\epsilon)$ is $O(n^{-4/5})$.*

Proof: First, from Theorem 3.4 and 3.5, the bias of $\hat{h}_\epsilon$ can be examined, giving

$$E[\hat{h}_\epsilon] - h_0 = E[\hat{h}_\epsilon] - h_\epsilon + h_\epsilon - h_0 = O(\epsilon^2).$$

From Theorem 3.5, the variability of $\hat{h}_\epsilon$ is $O((\epsilon n)^{-1})$. Combining these results yields

$$\text{MSE}(\hat{h}_\epsilon) = \beta_1^2 \epsilon^4 + \frac{\beta_2}{\epsilon n}, \tag{15}$$

where β_1 and β_2 are given in Theorems 3.4 and 3.5, respectively. Differentiating with respect to ϵ, setting to zero, and solving for ϵ gives

$$\epsilon = \left[\frac{f(x_0)\alpha_1}{4\alpha_3^2}\right]^{1/5} n^{-1/5}.$$

where α_1 and α_3 are given in (18). Substituting into (15) gives the final result.

Remark 1. It is an interesting exercise to compare this result to that of the cross-validated estimates of the global fixed bandwidth. It is well-established that global cross-validation leads to estimates with a $\text{MSE}(\hat{h}_{UCV}) = O(n^{-3/5})$ and a coefficient of variation (CV) that is $O(n^{-1/10})$. The $\text{MSE}(\hat{h}_\epsilon)$ here is slightly faster and the relative rate is even more impressive, considering that the target (h_0) in this setting is actually constant with respect to n. Specifically,

$$\begin{aligned} \text{CV} &= \frac{\sqrt{\text{Var}(\hat{h}_\epsilon)}}{E(\hat{h}_\epsilon)} = \frac{c_1 n^{-4/10}}{h_\epsilon} = \frac{c_1 n^{-4/10}}{h_0 + c_2 n^{-2/5}} \\ &\approx \frac{c_1 n^{-4/10}}{h_0} \\ &= O(n^{-4/10}). \end{aligned}$$

This nearly achieves the best-possible rate of $O(n^{-1/2})$ for fixed bandwidth estimators (Hall and Marron, 1991; Fan and Marron, 1992; and references therein) and matches the best-possible rate of $O(n^{-2/5})$ for local bandwidth selectors (Hazelton, 1996).

Remark 2. The previous theorems discuss the behavior of $\hat{h}_\epsilon$. Taken a step further, the results suggest that $\text{MSE}(\hat{f}(x_0|\hat{h}_\epsilon)) = O(n^{-4/5})$, a rate that is consistent with that of data-based global and local bandwidth choices resulting from the more traditional bias-variance trade-off. To see this, consider the following expansion similar to that given in the Appendix for the proof of Theorem 3.4,

$$\begin{aligned} E(\hat{f}(x_0|\hat{h}_\epsilon)) &= EK(x_0, \hat{h}_\epsilon) \\ &= E\{K(x_0, h_0) + K_2(x_0, h_0)(\hat{h}_\epsilon - h_0) + \frac{1}{2}K_{2,2}(x_0, h_0)(\hat{h}_\epsilon - h_0)^2 + \ldots\}. \end{aligned}$$

Theorems 3.4 and 3.5, suggest that $\hat{h}_\epsilon$ can be written as $\hat{h}_\epsilon = h_0 + \beta_1\epsilon^2 + (\beta_2/\sqrt{\epsilon n})Z$ where Z is a random variable with $EZ = 0$ and $\mathrm{Var}(Z) = 1$. From Theorem 3.6, $\epsilon = O(n^{-1/5})$. Substituting these results into the expansion, it is seen that the $\mathrm{Bias}(\hat{f}(x_0|\hat{h}_\epsilon)) = O(n^{-2/5})$ which gives a bias contribution to the MSE of $O(n^{-4/5})$. A similar argument shows that the variance contribution is $O(n^{-1})$ after noting that h_0 is a constant.

In the next two sections, it will be shown through simulation and example that focusing on a data-based procedure that targets h_0 leads to a new bias-variance trade-off operating point that, while not giving an estimator with higher-order behavior, can lead to superior performance.

4. SIMULATION STUDY

In this section, results of a small simulation study are presented to verify the primary result in the previous section as well as to gain some further understanding of behavior of the cross-validation algorithm and our ability to successfully estimate the zero-bias bandwidths in practice. The behavior of the resulting density estimates will also be examined.

Samples of size ranging from 2000 to 8000 were drawn from three densities: $N(0,1)$, Beta(2,6), and Exp(1). These densities were chosen to cover a wide range of tail behaviors. The width (2ϵ) of the interval used in the cross-validation varied across a grid of values for ϵ. The sample sizes chosen allow a reasonable number of data points to fall inside the intervals.

Four hundred replications were drawn for each combination of sample size and density. For each replication and across a range of values of ϵ, the zero-bias bandwidth and the resulting density were estimated at $x_0 = 2$ for the standard normal, at $x_0 = 0.59$ for the beta density, and at $x_0 = 3.87$ for the exponential. Each of these estimation points represents the same approximate quantile (0.975).

For each value of ϵ, define the empirical squared bias of $\hat{h}_\epsilon$ as $b^2_{\hat{h}_\epsilon} = [(1/400)\sum_i \hat{h}_{\epsilon i} - h_0]^2$, the empirical variance of $\hat{h}_\epsilon$ as $s^2_{\hat{h}_\epsilon}$, and the empirical mean squared error as $\widehat{\mathrm{MSE}}(\hat{h}_\epsilon) = b^2_{\hat{h}_\epsilon} + s^2_{\hat{h}_\epsilon}$. Figure 6 displays these as a function of ϵ for the standard normal distribution. As expected, the plots of the squared bias do not show significant differences due to sample size. The $\widehat{\mathrm{MSE}}(\hat{h}_\epsilon)$ plots show clear minima with respect to ϵ. It is also clear that the

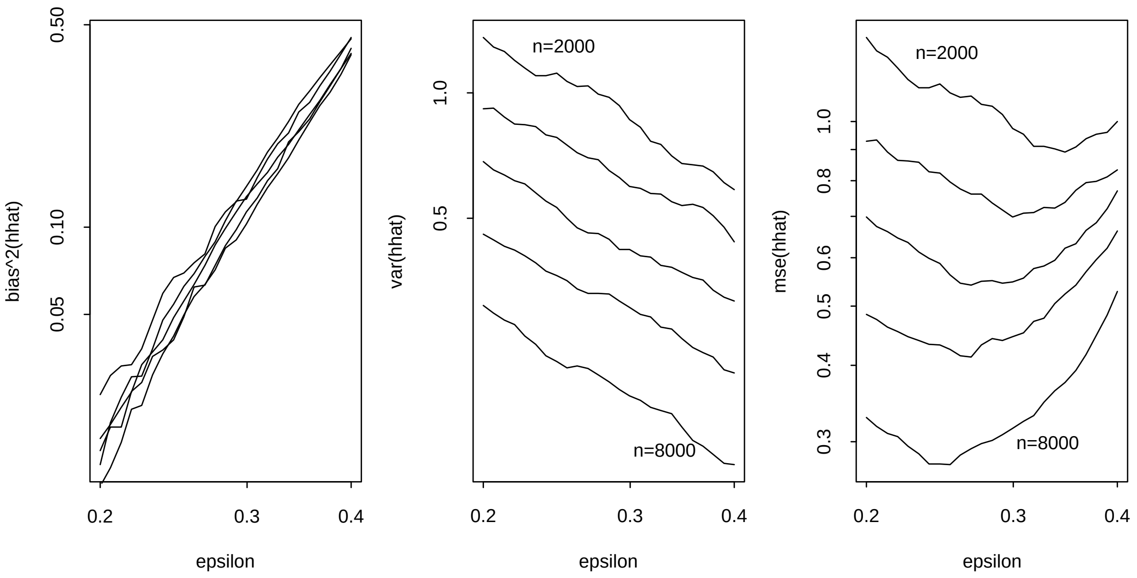

Figure 6: Plots of Bias^2, Var, and MSE of the estimated bandwidths as a function of ϵ for $N(0,1)$ samples of size $n = 2000$, 2828, 4000, 5657, and 8000 (top to bottom in each frame) and $x_0 = 2.0$.

optimal choice of ϵ is decreasing with sample size.

For each sample size, the minimum of the $\widehat{\mathrm{MSE}}(\hat{h}_\epsilon)$ with respect to ϵ was determined by naively picking the value of ϵ that minimizes the $\widehat{\mathrm{MSE}}(\hat{h}_\epsilon)$. Figure 7 shows these values of ϵ that minimize the $\widehat{\mathrm{MSE}}(\hat{h}_\epsilon)$ (denoted by $\hat{\epsilon}$) as a function of sample size for each of the three densities (left frame). These values of $\hat{\epsilon}$ track the theoretical rate of $n^{-1/5}$ derived in the previous section. The right frame in Figure 7 shows the $\widehat{\mathrm{MSE}}(\hat{h}_{\hat{\epsilon}})$ as a function of sample size. Even with such a crude approximation of the optimal value of ϵ, it is clear that the simulation results support the theory in the previous section since the $\widehat{\mathrm{MSE}}(\hat{h}_{\hat{\epsilon}})$ tracks the predicted rate of $n^{-4/5}$.

While Figures 6 and 7 verify the theoretical results concerning $\hat{h}_{\hat{\epsilon}}$, it is also of interest to study the actual density estimates using this local choice of bandwidth and compare them to actual density estimates computed using a global choice of bandwidth. Four hundred additional replications of each sample size and density were generated and the global least-squares cross-validation bandwidth ($\hat{h}_{UCV}$) determined. Density estimates were then calculated for both methods at the appropriate points ($x_0 = 2.0$ for the normal, $x_0 = 0.59$ for the beta, and $x_0 = 3.87$ for the exponential). For the zero-bias bandwidths, density

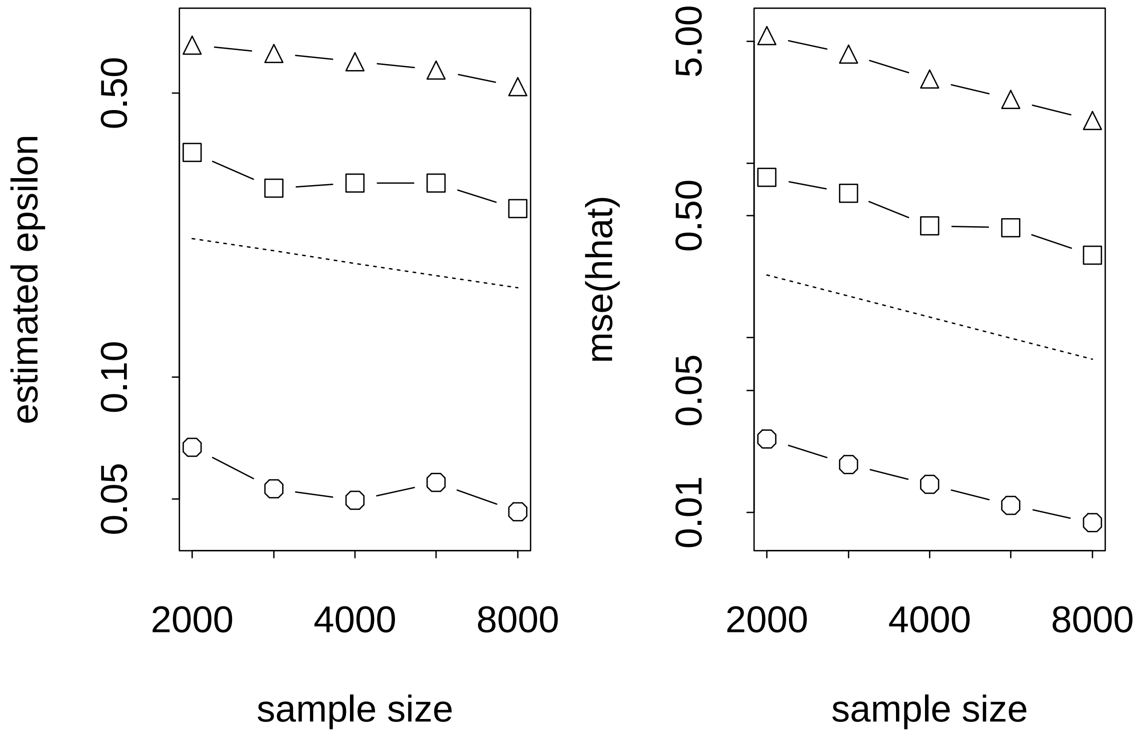

Figure 7: Left frame shows the optimal values of ϵ as a function of sample size for normal data (squares), beta data (circles), and exponential data (triangles). Dotted line represents $n^{-1/5}$. Right frame shows MSE of the estimated bandwidths computed with the optimal ϵ. Dotted line represents $n^{-4/5}$.

estimates were computed using bandwidths corresponding to $\hat{\epsilon}$. The $\widehat{\text{MSE}}(\hat{f})$ of the estimated densities was computed, and results for the normal and beta cases are summarized in Table 1.

In general, the local procedure based on estimates of the zero-bias bandwidths lead to substantial improvement in the MSE of the estimated density. The MSE is reduced by about 10% for normal data (except in one case), by 25-30% for beta data, and as much as 80% for exponential data (not shown).

However, the manner in which MSE is improved is considerably different for the two cases shown. For symmetric distributions such as the standard normal, the optimal global bandwidth will typically be larger than what is optimal for points in the tails, yielding pointwise estimates with relatively poor bias properties. Table 1 shows that the larger bandwidths of the local zero-bias procedure yield estimates with significantly better bias

Sample Size		N(0,1) 2000	2828	4000	5657	8000	Beta(2,6) 2000	2828	4000	5657	8000
	$\hat{h}_\epsilon$	1.72	.815	.594	.549	.271	31.9	14.0	21.9	11.7	10.2
$b^2(\hat{f}(x_0))$	$\hat{h}_{UCV}$	1.88	1.57	.951	.976	.795	8.25	8.41	2.82	1.95	.628
	$\hat{h}_\epsilon$	3.73	3.31	2.51	1.86	1.25	132.	105.	78.0	51.9	41.2
$s^2(\hat{f}(x_0))$	$\hat{h}_{UCV}$	4.32	2.84	2.11	1.66	1.23	206.	142.	105.	93.2	71.8
	$\hat{h}_\epsilon$	5.44	4.12	3.10	2.41	1.52	164.	119.	99.9	63.5	51.5
$\widehat{\text{MSE}}(\hat{f}(x_0))$	$\hat{h}_{UCV}$	6.20	4.41	3.06	2.64	2.02	214.	150.	108.	95.1	72.4

Table 1: MSE for density estimates at x_0 using both a global ($\hat{h}$) and zero-bias local ($\hat{h}_\epsilon$) procedures with normal ($x_0 = 2.0$) and beta ($x_0 = 0.59$) data. Values are reported as $\times 10^{-5}$.

properties and, in many cases, the larger bandwidths yield estimates with improved variance as well. For skewed distributions such as the beta used here, the optimal global bandwidth is much smaller to account for increased curvature near the mode. This yields estimates at points in the tails with better bias properties at the cost of substantially increased variance. While the zero-bias bandwidths are designed to overcome this problem, it appears from the simulations that this will only occur for even larger sample sizes than those used here. However, the MSE is dominated by the variance and the larger local bandwidths yield estimates with considerably better variability.

5. AN EXAMPLE

The results of the previous section deal with estimating the density at a point. Estimating the density over a particular region would require estimates of bandwidths for each point in that region, which, in turn, would require a different ϵ. In short, it is a daunting task. We consider a modification of the cross-validation function studied in the previous sections and replace the constant bandwidth by a bandwidth function, $h_x = h(x)$ defined over the region of interest. Hence, cross-validation is used to estimate the entire bandwidth function directly.

In practice, the logarithm of the bandwidth function h_x is parameterized as in Fan et al. (1993) and Sain (1994) by a cubic interpolating spline with the so-called "natural" end conditions. Knots are equally spaced along the x-axis in the region of interest. The natural

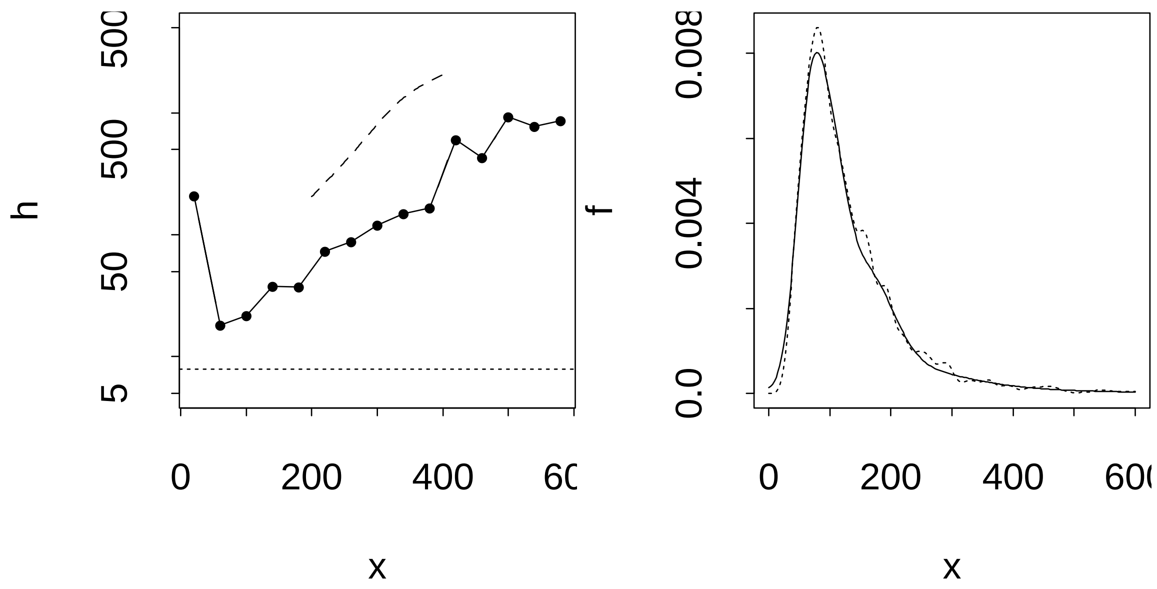

Figure 8: Estimated bandwidth functions and density estimates for the triglyceride data. The left plot shows $\hat{h}_{UCV} = 7.94$ (horizontal dotted line), $\hat{g}(x_i)$ (connected dots), and $\hat{h}_{ab}(x)$ (dashed line above the interval $[200, 400]$). The global density estimates for sample-point estimator using $\hat{g}(x_i)$ (solid line) and the fixed bandwidth estimator using $\hat{h}_{UCV}$ (dotted line) are shown in the right plot.

spline is completely determined by knowledge of the values at the knots; therefore, the cross-validation criterion may be minimized over the heights of the bandwidth function at these knot values. Any other values of the bandwidth function that are needed can be computed directly from the spline. Since only a small number of knot values is required (only four are used here), a considerable reduction in computational complexity is obtained.

The cross-validation function in (13) must be modified slightly and is given by

$$\mathrm{UCV}_{ab} = \int_a^b \hat{f}(x|h_x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i|h_{x_i}) I(a \le x_i \le b).$$

The first term cannot be computed explicitly, but it can easily be evaluated by numerical integration. Denote the spline-based cross-validation bandwidth function as $\hat{h}_{ab}(x)$.

To demonstrate, an example using real data is shown in Figures 8 and 9. The data are triglyceride levels of $n = 2116$ men in Istanbul, Turkey analyzed by Mahley, et al. (1995). The left frame in Figure 8 displays two locally adaptive bandwidth functions. The bandwidth function from a global approach using least-squares cross-validation and the binned sample point estimator ($\hat{g}(x_i)$) described in Sain and Scott (1996) is indicated by the connected solid circles. The dashed line indicates the estimated bandwidth function using the

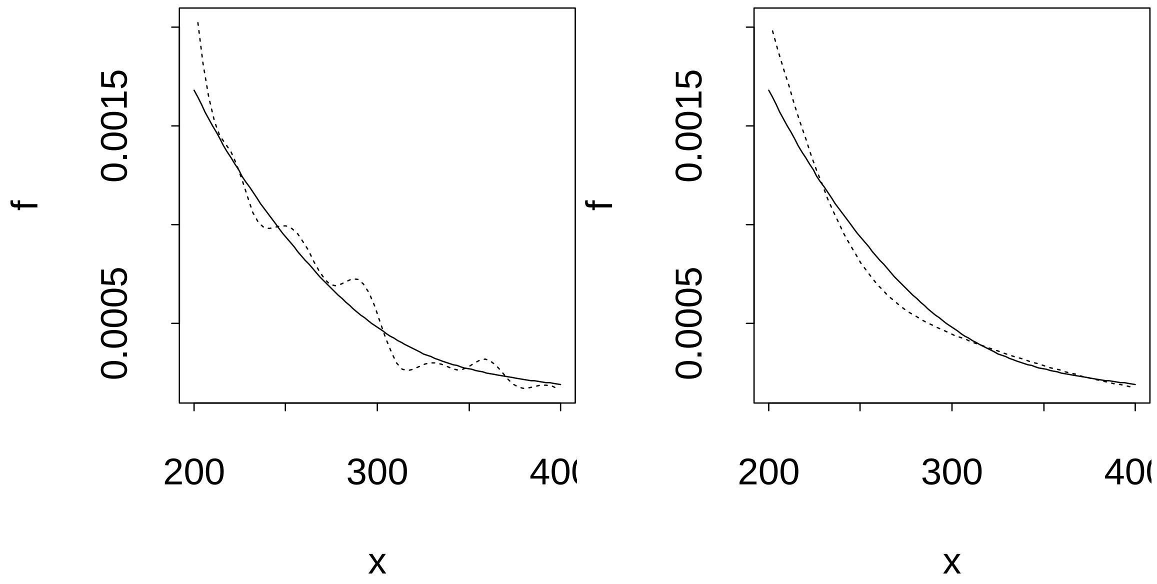

Figure 9: Density estimates in region of interest comparing the balloon estimator using $\hat{h}_{ab}(x)$ (solid line in both plots), the fixed bandwidth estimator using $\hat{h}_{UCV}$ (dotted line in left plot) and the sample-point estimator using $\hat{g}(x_i)$ (dotted line in right plot).

local cross-validation procedure for the balloon estimator ($\hat{h}_{ab}(x)$) for the interval $[200, 400]$ mg/dl (roughly the region defined by the 85th to 95th percentiles). For comparison, a fixed bandwidth estimator was also included with the smoothing parameter estimated by least-squares cross-validation ($\hat{h}_{UCV}$). The value of $\hat{h}_{UCV}$ found in this case was 7.94 (indicated by the dotted line), significantly smaller than bandwidths found by either adaptive approach.

One of the original goals of this research was to study the relationship between the balloon estimator and the sample point estimator. It was thought that, at least in an asymptotic sense, there would exist some fundamental relationship between the optimal bandwidth functions, $h(x)$ and $g(x_i)$, for the two approaches in Equations (2) and (3). While adapting to improve bias in the tails seems to drive both methods, this example shows that, if such a direct relationship exists, it is difficult to discern for finite samples. The bandwidths for the balloon estimator are much larger than those found using the sample point estimator and lead to significantly different estimates.

The right frame in Figure 8 shows the two global approaches with the sample point estimator using $\hat{g}(x_i)$ indicated by the solid line and the fixed bandwidth estimator using $\hat{h}_{UCV}$ with the dotted line. Note that the fixed bandwidth estimator appears to be undersmoothed for $x > 150$, as the small value of $\hat{h}_{UCV}$ tries to cope with the sharp mode at $x = 100$. The adaptive estimate seems to do a more than adequate job of smoothing over the whole range of the density.

Figure 9 shows a detail of the right tail, with the balloon estimator indicated by the solid line in both plots. A close examination reveals some interesting results. What is encouraging about the balloon estimator using $\hat{h}_{ab}(x)$ is that it tracks the fixed bandwidth estimator, except, of course, without the oscillations (left frame). Clearly $\hat{h}_{UCV}$ is too small; however, standard asymptotic theory suggests that a bandwidth that is too small leads to an estimate with better (average) bias properties. This would seem to support the results of the previous section and indicate that these larger zero-bias bandwidths incorporated by the balloon estimator yield an estimate that has considerably better bias properties while simultaneously having a smaller variance in a difficult region of the density to estimate.

To verify these results, another small simulation study was conducted to study these procedures and their ability to estimate the region between the 85 and 95 percentiles of the lognormal distribution. The zero-bias procedure was clearly superior, improving the empirical squared bias (integrated over the region) by approximately 50% and the empirical MSE (also integrated over the region) by approximately 40% over the two global methods (Sain, 2001).

6. DISCUSSION

For the past twenty years, a variety of attempts to create practical balloon and sample point estimators have met with increasing success. As we have shown, however, variable kernel estimation in convex regions is fundamentally different than near the modes. The extraordinarily large bandwidths predicted by the zero-bias theory in the tails where f is convex, have been shown to be estimable with data. While our practical implementation does not achieve higher-order convergence, it does appear to represent at least a constant factor improvement and has excellent bias and variance properties even in the extreme tails. We believe the large bandwidths found by the Sain and Scott (1996) global implementation of the sample point estimator reflect the zero-bias phenomenon. Better estimates just in the tails may be available with the current proposal.

It must be noted that the simulation studies in this work are not meant to be exhaustive as the intent is only to support the theoretical results and to facilitate some basic comparisons. Clearly there are other methods that would be viable competitors and may even be superior in certain situations. A comprehensive study along the lines of Jones and Signorini (1997), among others, would certainly be informative. It is, however, well beyond the scope of this work.

In the multivariate setting, Terrell and Scott (1992) noted that in regions where f is indefinite (i.e. Hessian has both positive and negative eigenvalues), the leading bias term could be eliminated even with a positive kernel. This work demonstrates that the entire bias can be eliminated in regions in the tails. The zero-bias idea may also be extended to nonparametric regression. For example, if the bias is proportional to $h^2[f(x)m(x)]''$, where m is the regression curve and f is the design density, then a zero-bias bandwidth exists at points where the initial bias moves the estimate away from the global mean. See Sain (2001).

Nonparametric estimates in tails have always been suspect, limiting application in important areas such as reliability and risk analysis. This work clearly paves a new path.

APPENDIX

Proof of Theorem 2.2: Let $h = h_0 + \delta$. The bias of $\hat{f}(x|h)$ is given by

$$\text{Bias}[\hat{f}(x|h)] = \int K(w)f(x - hw)\,dw - f(x) = \int K(w)f(x - h_0w - \delta w)\,dw - f(x).$$

Expanding $f(x - h_0w - \delta w)$ in a Taylor's series about $x - h_0w$,

$$E[\hat{f}(x|h)] = \int K(w)[f(x - h_0w) - \delta w f'(x - h_0w) + \cdots]\,dw.$$

Note that $\int K(w)f(x - h_0w)dw = E[\hat{f}(x|h_0)] = f(x)$ by definition of h_0. Ignoring the higher-order terms of δ, an asymptotic approximation of the bias of $\hat{f}(x|h)$ is given by

$$\text{Bias}[\hat{f}(x|h)] = -\delta b, \tag{16}$$

where b is given in (11).

A similar argument can be used to derive an expression for the variance of $\hat{f}(x|h)$:

$$\begin{aligned}\text{Var}[\hat{f}(x|h)] &= \frac{1}{n}\left[E[K_h^2(x-X)] - [E\,K_h(x-X)]^2\right] \\ &= \frac{1}{n}\int \frac{1}{(h_0+\delta)^2}K^2\left(\frac{x-y}{h_0+\delta}\right) f(y)\,dy - \frac{1}{n}\left[f(x)-b\delta\right]^2 \\ &= \frac{1}{n(h_0+\delta)}\int K^2(w) f(x-h_0w-\delta w)\,dw - \frac{1}{n}\left[f(x)-b\delta\right]^2.\end{aligned}$$

Expanding $f(x-h_0w-\delta w)$ as before and ignoring the higher order terms in δ/n, the asymptotic variance of $\hat{f}(x|h)$ is given by

$$\text{Var}[\hat{f}(x|h)] = \frac{a_1-\delta a_2}{n(h_0+\delta)} - \frac{f(x)^2}{n} + \frac{2\delta b f(x)}{n}, \tag{17}$$

where a_1 and a_2 are given in (10). Hence, the MSE of $\hat{f}(x|h)$ is

$$\text{MSE}[\hat{f}(x|h)] = \frac{a_1-\delta a_2}{n(h_0+\delta)} - \frac{f(x)^2}{n} + \frac{2\delta b f(x)}{n} + \delta^2 b^2.$$

Expanding $(1+\delta/h_0)^{-1}$ in a Taylor's series and ignoring the higher-order terms of δ/n,

$$\text{MSE}[\hat{f}(x|h)] = \frac{1}{n}\left[\frac{a_1}{h_0} - f(x)^2\right] - \frac{\delta}{n}\left[\frac{a_1}{h_0^2} + \frac{a_2}{h_0} - 2bf(x)\right] + \delta^2 b^2.$$

Since $h = h_{mse}$ is a local minimizer of the MSE and h_0 is a constant, the partial derivative of the MSE with respect to δ equals zero, yielding

$$\frac{\partial}{\partial\delta}\text{MSE}[\hat{f}(x|h_{mse})] = 2\delta b^2 - \frac{1}{n}\left[\frac{a_1}{h_0^2} + \frac{a_2}{h_0} - 2bf(x)\right] = 0.$$

Solving for δ and rearranging yields (9), completing the proof.

Proof of Theorem 3.4: Consider expanding $\text{Bias}^2[\hat{f}(x|h)] = (EK_h(x-X) - f(y))^2$ in a Taylor's series around $x = x_0$ and $h = h_0$. Write $EK_h(x-X) = EK(x,h)$ and indicate derivatives as subscripts, e.g. $EK_{1,2}(x,h)$ indicates the mixed partial derivative of the expectation evaluated at x and h. Noting that $\text{Bias}[\hat{f}(x_0|h_0)] = 0$, the expansion yields

$$\begin{aligned}\text{Bias}^2[\hat{f}(x,h)] &= \left(EK_1(x_0,h_0) - f'(x_0)\right)^2 (x-x_0)^2 + 2EK_2(x_0,h_0)\left(EK_1(x_0,h_0)\right. \\ &- \left. f'(x_0)\right)(h-h_0)(x-x_0) + EK_2(x_0,h_0)^2 (h-h_0)^2 \\ &+ \left(EK_1(x_0,h_0) - f'(x_0)\right)\left(EK_{1,1}(x_0,h_0) - f''(x_0)\right)(x-x_0)^3 \\ &+ \left(EK_2(x_0,h_0)\left(EK_{1,1}(x_0,h_0) - f''(x_0)\right)\right.\end{aligned}$$

$$
\begin{aligned}
&+ \quad 2\left(EK_1(x_0,h_0) - f'(x_0)\right) EK_{1,2}(x_0,h_0)) \left(h-h_0\right)\left(x-x_0\right)^2 \\
&+ \quad \left(2EK_2(x_0,h_0)EK_{1,2}(x_0,h_0) + \left(EK_1(x_0,h_0) - f'(x_0)\right)\right) \\
&\times \quad EK_{2,2}(x_0,h_0)\left(h-h_0\right)^2\left(x-x_0\right) + EK_2(x_0,h_0)EK_{2,2}(x_0,h_0)\left(h-h_0\right)^3 \\
&+ \quad \ldots .
\end{aligned}
$$

Evaluating the integral in (14) is straightforward, and we can easily take the partial derivative with respect to h. This yields

$$[\alpha_1(h-h_0) + \alpha_2(h-h_0)^2 + O((h-h_0)^3)]\epsilon + [\alpha_3 + O((h-h_0))]\epsilon^3 + O(\epsilon^5),$$

where

$$
\begin{aligned}
\alpha_1 &= 4EK_2(x_0,h_0)^2 \qquad \alpha_2 = 6EK_2(x_0,h_0)EK_{2,2}(x_0,h_0) \\
\alpha_3 &= (2/3)\left(2EK_{1,2}(x_0,h_0)(EK_1(x_0,h_0) - f'(x_0)\right) \\
&+ \quad EK_2(x_0,h_0)(EK_{1,1}(x_0,h_0) - f''(x_0))) . \qquad (18)
\end{aligned}
$$

Ignoring the higher-order terms involved ϵ and $h-h_0$ and setting this derivative equal to zero yields an expression of the form

$$\alpha_2(h-h_0)^2 + \alpha_1(h-h_0) + \alpha_3\epsilon^2 = 0. \qquad (19)$$

Since $h = h_\epsilon$ is a solution to (19), we can write $h_\epsilon - h_0 = \alpha_3/\alpha_1\epsilon^2 + o(\epsilon^2)$.

The MSE$[\hat{f}(x_0|h_\epsilon)]$ can now evaluated. Write $h_\epsilon = h_0 + \eta$ where $\eta = O(\epsilon^2)$. Then, as in the proof of Thereom 2.2,

$$\text{Bias}[\hat{f}(x|h_\epsilon)] = \eta \int wK(w)f'(x+h_0w)dw + \ldots$$

Hence, Bias $= O(\eta)$. A similar computation shows that the variance is $O(1/n + \eta/n)$. Therefore MSE$[\hat{f}(x_0|h_\epsilon)] = O(1/n) + O(\eta^2)$. Since $\eta = O(\epsilon^2)$, MSE$[\hat{f}(x_0|h_\epsilon)] = O(1/n) + O(\epsilon^4)$.

Proof of Theorem 3.5: To capture the quadratic nature of $\text{MISE}_\epsilon(h) = \int_{x_0-\epsilon}^{x_0+\epsilon} \text{MSE}(x,h)dx$ around h_ϵ, expand MISE_ϵ in a second-order Taylor's series expansion about h_ϵ,

$$\text{MISE}_\epsilon(h) = c_0 + c_1(h-h_\epsilon) + \frac{1}{2}c_2(h-h_\epsilon)^2 + O((h-h_\epsilon)^3),$$

where

$$c_0 = \mathrm{MISE}_\epsilon(h_\epsilon), \qquad c_1 = \frac{\partial}{\partial h}\mathrm{MISE}_\epsilon(h)\bigg|_{h_\epsilon}, \qquad \text{and} \qquad c_2 = \frac{\partial^2}{\partial h^2}\mathrm{MISE}_\epsilon(h)\bigg|_{h_\epsilon}.$$

Since h_ϵ corresponds to a local minimum of MISE_ϵ, $c_1 = 0$.

Similarly for UCV_ϵ,

$$\mathrm{UCV}_\epsilon(h) = \hat{c}_0 + \hat{c}_1(h-h_\epsilon) + \frac{1}{2}\hat{c}_2(h-h_\epsilon)^2 + O((h-h_\epsilon)^3),$$

where

$$\hat{c}_0 = \mathrm{UCV}_\epsilon(h_\epsilon) = c_0 + d_0 Z_0, \qquad \hat{c}_1 = \frac{\partial}{\partial h}\mathrm{UCV}_\epsilon(h)\bigg|_{h_\epsilon} = c_1 + d_1 Z_1 = d_1 Z_1$$

$$\hat{c}_2 = \frac{\partial^2}{\partial h^2}\mathrm{UCV}_\epsilon(h)\bigg|_{h_\epsilon} = c_2 + d_2 Z_2,$$

and $Z_i,\ i = 0,1,2$ are assumed to be random variables with mean zero and unit variance.

Let $\hat{h}$ be the local minimizer of UCV_ϵ. Then, ignoring higher-order terms,

$$\frac{\partial}{\partial h}\mathrm{UCV}_\epsilon(h)\bigg|_{\hat{h}} = \hat{c}_1 + \hat{c}_2(\hat{h} - h_\epsilon) = 0.$$

Solving for $(\hat{h} - h_\epsilon)$,

$$\hat{h} - h_\epsilon \;=\; -\frac{\hat{c}_1}{\hat{c}_2} = -\frac{d_1 Z_1}{c_2 + d_2 Z_2} = -\frac{d_1 Z_1}{c_2}\left(1 + \frac{d_2}{c_2}Z_2\right)^{-1}.$$

Expanding $(1 + (d_2/c_2)Z_2)^{-1}$ in a Taylor's series expansion and matching leading powers of Z,

$$\hat{h} - h_\epsilon \;=\; -\frac{d_1 Z_1}{c_2}\left(1 - \frac{d_2}{c_2}Z_2 + \left(\frac{d_2}{c_2}\right)^2 Z_2^2 + \ldots\right) \approx -\frac{d_1 Z_1}{c_2}. \tag{20}$$

To establish the order of the bias and variance of $\hat{h}$, the behavior of the terms d_1, d_2, and c_2 must be established. First, examining c_2,

$$\begin{aligned} c_2 &= \frac{\partial^2}{\partial h^2}\mathrm{MISE}_\epsilon(h)\bigg|_{h_\epsilon} \\ &\approx \frac{\partial^2}{\partial h^2}\int_{x_0-\epsilon}^{x_0+\epsilon} \left[E[K_h(x)] - f(x)\right]^2 dx\bigg|_{h_\epsilon} = 4EK_2(x_0,h_0)^2\epsilon + O(\epsilon^2). \end{aligned} \tag{21}$$

This result follows by expanding the bias term as in Theorem 3.4, evaluating the integral, and plugging in $h_\epsilon = h_0 + O(\epsilon^2)$.

The second term, d_1, is somewhat more complicated. By definition,

$$d_1^2 = \mathrm{Var}[\frac{\partial}{\partial h}\mathrm{UCV}_\epsilon(h)\Big|_{h_\epsilon}].$$

Taking the partial derivative with respect to h of the UCV_ϵ criterion and rewriting yields,

$$\frac{\partial}{\partial h}\mathrm{UCV}_\epsilon(h)\Big|_{h_\epsilon} = \frac{1}{n^2}\sum_{i=1}^{n}\lambda_i + \frac{1}{n^2}\sum_{i\neq j}\gamma_{ij} - \frac{1}{n^2}\sum_{i\neq j}\eta_{ij}$$

where

$$\lambda_i = \tfrac{\partial}{\partial h}\int_{x_0-\epsilon}^{x_0+\epsilon} K_h^2(x-x_i)dx\Big|_{h_\epsilon} \qquad \gamma_{ij} = \tfrac{\partial}{\partial h}\int_{x_0-\epsilon}^{x_0+\epsilon} K_h(x-x_i)K_h(x-x_j)dx\Big|_{h_\epsilon}$$
$$\eta_{ij} = I(x_0-\epsilon \le x_i \le x_0+\epsilon)\ \tfrac{\partial}{\partial h}K_h(x_i-x_j)\Big|_{h_\epsilon}.$$

The variance of the derivative can be written as

$$\begin{aligned}\mathrm{Var}[\frac{\partial}{\partial h}\mathrm{UCV}_\epsilon\Big|_{h_\epsilon}] &= \frac{1}{n^4}\left\{\mathrm{Var}\left[\sum_i \lambda_i\right]\right\} + \frac{1}{n^4}\left\{\mathrm{Var}\left[\sum_{i\neq j}\gamma_{ij}\right]\right\} + \frac{1}{n^4}\left\{\mathrm{Var}\left[\sum_{i\neq j}\eta_{ij}\right]\right\}\\ &+ \frac{2}{n^4}\left\{\mathrm{Cov}\left[\sum_i \lambda_i, \sum_{i\neq j}\gamma_{ij}\right]\right\} - \frac{2}{n^4}\left\{\mathrm{Cov}\left[\sum_{i=1}^{n}\lambda_i, \sum_{i\neq j}\eta_{ij}\right]\right\}\\ &- \frac{2}{n^4}\left\{\mathrm{Cov}\left[\sum_{i\neq j}\gamma_{ij}, \sum_{i\neq j}\eta_{ij}\right]\right\}.\end{aligned}$$

First, the order of each term with respect to n is examined. The first term can be expanded as

$$\frac{1}{n^4}\mathrm{Var}\left[\sum_i \lambda_i\right] = \frac{1}{n^4}\left[\sum_i \mathrm{Var}\lambda_i + \sum_{i\neq j}\mathrm{Cov}[\lambda_i,\lambda_j]\right] = \frac{1}{n^3}\mathrm{Var}\lambda_i = O(n^{-3}).$$

Since x_i is independent of x_j, the covariance terms are zero and there are only n nonzero terms in the variance of the sum.

The second term in the variance is given by

$$\frac{1}{n^4}\mathrm{Var}\left[\sum_{i\neq j}\gamma_{ij}\right] = \frac{1}{n^4}\{2n(n-1)\mathrm{Var}\gamma_{ij} + 4n(n-1)(n-2)\mathrm{Cov}[\gamma_{ij},\gamma_{ik}]\} = O(n^{-1}),$$

where $i \neq j \neq k$.

For the third term,

$$\begin{aligned}
\frac{1}{n^4}\mathrm{Var}\left[\sum_{i\neq j}\eta_{ij}\right] &= \frac{1}{n^4}\left\{n(n-1)\mathrm{Var}\eta_{ij} + n(n-1)(n-2)\mathrm{Cov}[\eta_{ij},\eta_{ik}] + n(n-1)^2\mathrm{Cov}[\eta_{ij},\eta_{jk}]\right.\\
&+ \left. n(n-1)(n-2)\mathrm{Cov}[\eta_{ij},\eta_{ki}] + n(n-1)(n-2)\mathrm{Cov}[\eta_{ij},\eta_{kj}]\right\}\\
&= O(n^{-1}).
\end{aligned}$$

Note that some care must be taken for the third term since it is not always symmetric, i.e. η_{ij} does not always equal η_{ji}.

The covariance terms are a bit more complicated, but still just involve counting the number of nonzero terms. The first covariance term is given as

$$\frac{2}{n^4}\mathrm{Cov}\left[\sum_i \lambda_i, \sum_{i\neq j}\gamma_{ij}\right] = \frac{2}{n^4}\left\{2n(n-1)\mathrm{Cov}\left[\lambda_i,\gamma_{ij}\right]\right\} = O(n^{-2}).$$

The second covariance term is similar.

Finally, the last covariance term is

$$\begin{aligned}
\frac{2}{n^4}\mathrm{Cov}\left[\sum_{i\neq j}\gamma_{ij}, \sum_{i\neq j}\eta_{ij}\right] &= \frac{2}{n^4}\left\{2n(n-1)\mathrm{Cov}\left[\gamma_{ij},\eta_{ij}\right] + 4n(n-1)(n-2)\mathrm{Cov}\left[\gamma_{ij},\eta_{ik}\right]\right\}\\
&= O(n^{-1}).
\end{aligned}$$

Writing out the dominating terms with respect to n gives

$$\begin{aligned}
d_1^2 = \mathrm{Var}[\frac{\partial}{\partial h}\mathrm{UCV}_\epsilon\bigg|_{h_\epsilon}] &= \frac{2}{n}\left\{4\mathrm{Cov}[\gamma_{ij},\gamma_{ik}] + \mathrm{Cov}[\eta_{ij},\eta_{ik}] + \mathrm{Cov}[\eta_{ij},\eta_{jk}]\right.\\
&+ \left.\mathrm{Cov}[\eta_{ij},\eta_{ki}] + \mathrm{Cov}[\eta_{ij},\eta_{kj}] - 4\mathrm{Cov}[\gamma_{ij},\eta_{ik}]\right\}.
\end{aligned}$$

Since these covariance terms are all functions of ϵ, a closer examination is necessary. The first can be written as

$$\mathrm{Cov}[\gamma_{ij},\gamma_{ik}] = E[\gamma_{ij}\gamma_{ik}] - E[\gamma_{ij}]E[\gamma_{ik}].$$

Noting that γ_{ij} can be written as

$$\gamma_{ij} = 2\epsilon\,\frac{\partial}{\partial h}K_h(x_0-x_i)K_h(x_0-x_j)\bigg|_{h_0} + O(\epsilon^3)$$

the covariance is easily seen to be $O(\epsilon^2)$.

The next covariance terms involve η's are somewhat more complicated. The first term can be written as

$$\begin{aligned}
\mathrm{Cov}[\eta_{ij}, \eta_{ik}] &= E[I(x_0 - \epsilon \le x_i \le x_0 + \epsilon)^2 \, \frac{\partial}{\partial h} K_h(x_i - x_j)\Big|_{h_\epsilon} \, \frac{\partial}{\partial h} K_h(x_i - x_k)\Big|_{h_\epsilon}] \\
&- E[I(x_0 - \epsilon \le x_i \le x_0 + \epsilon) \, \frac{\partial}{\partial h} K_h(x_i - x_j)\Big|_{h_\epsilon}] \\
&\times E[I(x_0 - \epsilon \le x_i \le x_0 + \epsilon) \, \frac{\partial}{\partial h} K_h(x_i - x_k)\Big|_{h_\epsilon}].
\end{aligned}$$

Noting that $I(x_0 - \epsilon \le x_i \le x_0 + \epsilon)^2 = I(x_0 - \epsilon \le x_i \le x_0 + \epsilon)$, the expectations can be rewritten as

$$\begin{aligned}
\mathrm{Cov}[\eta_{ij}, \eta_{ik}] &= \int_{x_0-\epsilon}^{x_0+\epsilon} \int\int \frac{\partial}{\partial h} K_h(x-y) \frac{\partial}{\partial h} K_h(x-z) f(x) f(y) f(z) dx dy dz\Big|_{h_\epsilon} \\
&- [\int_{x_0-\epsilon}^{x_0+\epsilon} \int \frac{\partial}{\partial h} K_h(x-y) f(x) f(y) dx dy dz\Big|_{h_\epsilon}]^2 \\
&= 2f(x_0) E K_2(x_0, h_0)^2 + O(\epsilon^2).
\end{aligned}$$

The other covariance terms involving η in which the first subscript are not the same will be $O(\epsilon^2)$ by similar arguments. The last covariance term involves both γ and η and is easily shown to be $O(\epsilon^2)$.

Hence, $d_1^2 = 4f(x_0)EK_2(x_0, h_0)^2\epsilon/n + O(\epsilon^2/n)$. Note that a similar result can be shown for d_2^2, further justifying the approximation in (20). Combining this result with the previous result for c_2 in (21) gives $\mathrm{E}(\hat{h}) - h_\epsilon = d_1 d_2 \mathrm{E}(Z_1 Z_2)/c_2^2 = O\left((\epsilon\, n)^{-1}\right)$ and $\mathrm{Var}(\hat{h} - h_\epsilon) = d_1^2/c_2^2 = f(x_0)/(4EK_2(x_0, h_0)^2 \epsilon n) = f(x_0)/(\alpha_1 \epsilon n) = O\left((\epsilon\, n)^{-1}\right)$, where α_1 is given in (18).

REFERENCES

Abramson, I. (1982), "On Bandwidth Variation in Kernel Estimates -A Square Root Law," *The Annals of Statistics*, **10**, 1217-1223.

Bartlett, M.S. (1963), "Statistical Estimation of Density Functions," *Sankhya Series A*, **25**, 245-254.

Bowman, A.W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, **71**, 353-360.

Breiman, L., Meisel, W., and Purcell, E. (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, **19**, 135-144.

Devroye, L. and Lugosi, G. (1998), "Variable Kernel Estimates: On the Impossibility of Tunning the Parameters," Unpublished Manuscript.

Fan, J. and Marron, J.S. (1992), "Best Possible Constant for Bandwidth Selection," *Annals of Statistics*, **20**, 2057-2070.

Fan, J., Hall, P., Martin, M., and Patil, P. (1993), "On Local Smoothing of Nonparametric Curve Estimators," *Journal of the American Statistical Association*, **91**, 258-266.

Hall, P., Hu, T.C., and Marron, J.S. (1995), "Improved Variable Window Kernel Estimates of Probability Densities," *Annals of Statistics*, **23**, 1-10.

Hall, P. and Marron, J.S. (1991), "Lower Bounds for Bandwidth Selection in Density Estimation," *Probability Theory and Related Fields*, **90**, 149-173.

Hall, P. and Schucany, W.R. (1989), "A Local Cross-Validation Algorithm," *Statistics & Probability Letters*, **8**, 109-117.

Hazelton, M. (1996), "Optimal Rates for Local Bandwidth Selection," *Nonparametric Statistics*, **7**, 57-66.

Hazelton, M. (1998), "Bias Annihilating for Local Density Estimates," *Statitics and Probability Letters*, **38**, 305-309.

Jones, M.C. (1990), "Variable Kernel Density Estimates and Variable Kernel Density Estimates," *Australian Journal of Statistics*, **32**, 361-371.

Jones, M.C. and Signorini, D.F. (1997), "A Comparison of Higher-Order Bias Kernel Density Estimators," *Journal of the American Statistical Association*, **92**,1063-1073.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," *Annals of Mathematical Statistics*, **36**, 1049-1051.

Marron, J.S. and Wand, M.P. (1992), "Exact Mean Integrated Squared Error," *Annals of Statistics*, **20**, 712-736.

Mahley, R.W., Palaoglu, K.E., Atak, Z., Dawson-Pepin, J., Langlois, A., Cheung, V., Onat, H., Fulks, P., Mahley, L.L., Vakar, F., Ozbayrakci, S., Gokdemir, O., and Winkler, W. (1995), "Turkish Heart Study: Lipids, Lipoproteins, and Apolipoproteins," *Journal of Lipid Research*, **36**, 839-859.

Mielniczuk, J., Sarda, P., and Vieu, P. (1989), "Local Data-Driven Bandwidth Choice for Density Estimation," *Journal of Statistical Planning and Inference*, **23**, 53-69.

Rudemo, M. (1982), "Empirical Choice of Histogram and Kernel Density Estimators," *Scandinavian Journal of Statistics*, **9**, 65-78.

Sain, S.R. (1994), "Adaptive Kernel Density Estimation," Unpublished Ph.D. Dissertation, Department of Statistics, Rice University.

Sain, S.R. (2000), "A New Characterization and Estimation of the Zero-Bias Bandwidth," Technical Report, Department of Statistical Science, Southern Methodist University.

Sain, S.R. (2001), "Bias Reduction and Elimination with Kernel Estimators," *Communications in Statistics*, To appear.

Sain, S.R. and Scott, D.W. (1996), "On Locally Adaptive Density Estimation," *Journal of the American Statistical Association*, **91**, 1525-1534.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

Sheather, S.J. (1986), “An Improved Data-Based Algorithm for Choosing the Window Width when Estimating the Density at a Point,” *Computational Statistics and Data Analysis*, **4**, 61-65.

Sheather, S.J. and Jones, M.C. (1991), “A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation”, *Journal of the Royal Statistical Society*, **53**, 683- 690.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Terrell, G.R. and Scott, D.W. (1992), “Variable Kernel Density Estimation,” *The Annals of Statistics*, **20**, 1236-1265.

Thompson, J.R. and Tapia, R.A. (1990), *Nonparametric Function Estimation, Modeling and Simulation*, Philadelphia: Society for Industrial and Applied Mathematics.

Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.