Zero-Bias Locally Adaptive Density Estimators

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

September 14, 1995

SUMMARY: Strategies for improving fixed kernel estimators have focused on reducing the bias, either by locally adjusting the bandwidth or by utilizing properties of higher-order kernels. 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 usual parametric rate of $O(n^{-1})$. We investigate fully data-based algorithms that take advantage of this predicted theoretical behavior.

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

[1]Stephan R. Sain is Research Associate, Department of Statistical Science, Southern Methodist University, POB 750332, Dallas, TX 77275. David W. Scott is Professor, Department of Statistics, Rice University, POB 1892, Houston, TX 77251.

1. INTRODUCTION

The fixed kernel density estimator has been studied extensively in the literature, i.e.

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

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. The 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 of order $O(n^{-4/5})$. Recent surveys include Silverman (1986), Scott (1992), and Wand and Jones (1995).

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 a higher-order bias. A symmetric kernel is defined to be of order p (where p is assumed to be even) if $\int w^p K(w)\,dw \neq 0$ while $\int w^i K(w)\,dw = 0$ for $i = 1, \ldots, p-1$. Bartlett showed that the bias is of order $O(h^p f^{(p)}(x))$ with a mean squared error of order $O(n^{-2p/(2p+1)})$. Thus, if the density is sufficiently smooth, substantial improvement over the non-negative kernel (i.e. order $p = 2$) is asymptotically possible.

Current research has focused on varying the bandwidth of a fixed-order kernel (usually $p = 2$) 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 is the focus of this paper and 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.

A second variable kernel estimator was introduced by Breiman, Meisel, and Purcell (1977) who suggested varying the bandwidth at each sample (not estimation) 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 relationship of the bandwidth functions $h(\cdot)$ and $g(\cdot)$ is of great interest here.

Loftsgaarden and Quesenberry (1965) introduced the first practical local kernel density estimator in the form of the multivariate nearest neighbor estimator,

$$\hat{f}(\mathbf{x}) = \frac{k}{nV_d h_k(\mathbf{x})^d},$$

where $\mathbf{x} \in \Re^d$, $h_k(\mathbf{x})$ is the Euclidean distance from $\mathbf{x}$ to the kth nearest sample point, and V_d is the volume of the unit sphere in $\Re^d$. Note that in the univariate case, this estimator can be written as in (2) where K is the uniform kernel on $[-1, 1]$ and $h(x) = h_k(x)$. The same is true in the multivariate case if K is uniform on the unit sphere. 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 most popular implementation of the second 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 generated much excitement as the bias was shown to be of order $O(h^4)$, similar to the negative ($p = 4$) kernels of Bartlett (1963); see Abramson (1982), Silverman (1986), and Hall and Marron (1988) for details. However, Terrell and Scott (1992), Hall, Hu, and Marron (1994), and Sain and Scott (1995) discuss implementation problems as well as poor performance in certain situations. Recent research by Fan et al. (1993), Sain (1994), and Sain and Scott (1995) has focused on the practical issues associated with estimating optimal bandwidths for the general adaptive estimator.

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

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

Jones showed that some modest improvement over the fixed bandwidth case can be made. An example of $h^*(x)$ for the case of normal data is shown in Figure 1.

This work describes the surprising theoretical result that there exist certain regions in the support of an unknown density where a bandwidth can be found yielding exactly

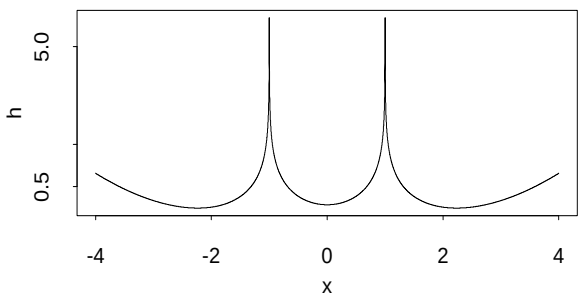

Figure 1: Optimal local bandwidth function, $h^*(x)$, for $N(0,1)$ kernel and data ($n = 100$).

zero bias pointwise, and that these zero-bias bandwidths vastly improve upon $h^*(x)$. Furthermore, a global estimation strategy utilizing these zero-bias bandwidths is shown theoretically to be vastly superior over the fixed bandwidth estimator as well as the simple strategy utilized by Jones (1990). Finally, a practical local cross-validation procedure extending work done by Hall and Schucany (1989) is shown to be effective for estimating these zero-bias bandwidths in practice. Finally, these ideas are expanded into a global estimation strategy similar to that discussed in Fan et al. (1993) and Sain (1994) for the adaptive sample point estimator.

2. EXISTENCE OF ZERO-BIAS BANDWIDTHS

Consider the univariate locally adaptive estimator 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 the bandwidth is taken to be some function of the estimation point, $h_x = h(x)$. Following the approach of Marron and Wand (1992), exact expressions for the pointwise bias and variance may be obtained when the true density is a normal mixture of the form

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

where $\sum w_i = 1$ and $\phi_\sigma(x-\mu)$ is a $N(\mu, \sigma^2)$ density. These exact bias/variance expressions

$$\text{Bias}[\hat{f}(x|h_x)] = \mu_1(x|h_x) - f(x) \tag{6}$$

and

$$\text{Var}[\hat{f}(x|h_x)] = \frac{1}{n}[\mu_2(x|h_x) - \mu_1(x|h_x)^2], \tag{7}$$

can be written as functions of the first two moments, μ_1 and μ_2, of the kernel, where

$$\begin{aligned} \mu_1(x|h) = E[K_h(x-\mathbf{X})] &= \int \phi_h(x-t) f(t)\, dt \\ &= \sum_{i=1}^{k} w_i \int \phi_h(x-t)\phi_{\sigma_i}(t-\mu_i)dt \\ &= \sum_{i=1}^{k} w_i \phi_{\sqrt{h^2+\sigma_i^2}}(x-\mu_i) \end{aligned} \tag{8}$$

and, similarly,

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

The pointwise MSE can be calculated by summing (7) and the square of (6).

Consider estimation of f at some point x in the "tail" of f. Then it will be shown in Theorem 2.1 that there exists a bandwidth, $h_0(x) > 0$, such that the bias is exactly zero, and that $h_0(x)$ corresponds to a local and possibly a global minima in the MSE. As the bias is not a function of the sample size, n, then $h_0(x)$ does not depend on n. We often shorten $h_0(x)$ to simply h_0.

An example of the existence of a zero-bias bandwidth, h_0, is given in Figure 2, where (5) is used to estimate f at $x = 2.5$ where $f(x)$ is $N(0,1)$ and $n = 100$. Note that the bias does in fact equal zero at $h_0 = 22.6$. Furthermore, there is a bandwidth corresponding to a global minima of the MSE that is slightly larger than h_0 (see Theorem 2.2). The local minima in the MSE located at $h = 0.345$ is consistent with the usual asymptotic theory exemplified by Equation (4) and represents the predicted best asymptotic local bias/variance tradeoff. One might ask why the MSE in Figure 2 is so much better near the zero-bias bandwidth, h_0. The answer is fairly straightforward. Usually, large bandwidths will improve variance, but only at the expense of a larger bias (note the rapid increase in the bias curve until $h = 5.0$). However, further increase in the bandwidth does not increase the bias as predicted by the asymptotic theory, but, surprisingly, reduces it, so that the bias vanishes at h_0 and finally converging to $-f(x)$. 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 conditions under which such bandwidths exist. First, the notion of *convexity* must be defined.

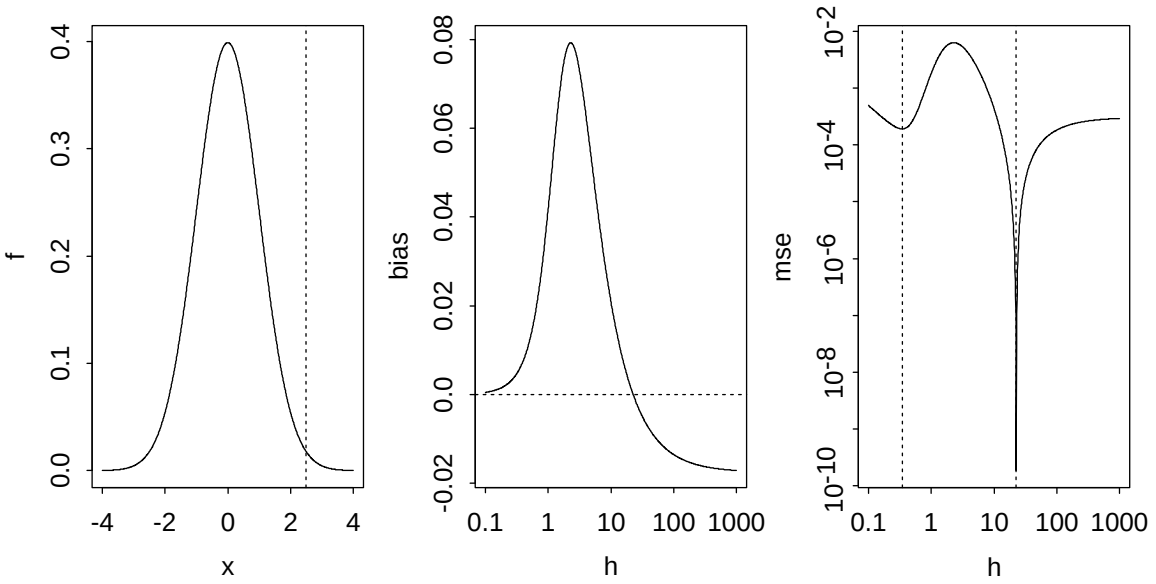

Figure 2: Bias and MSE curves for estimation at $x = 2.5$ when $f(x)$ is $N(0,1)$ and $n = 100$.

Definition 2.1 (Convexity) *A finite real-valued function, g, defined on some interval I, is said to be convex if for $a < b$, $a, b \in I$, and $\lambda \in [0,1]$,*

$$g(\lambda a + (1-\lambda)b) \leq \lambda g(a) + (1-\lambda)g(b).$$

Remark. Geometrically, a function is convex if all line segments connecting pairs of points of g in I lie above the function there. This definition is adopted as it includes finite-support densities by noting that $f(x) = 0$ beyond the support of f.

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 (6) at some $h > 0$. Suppose f is convex over the interval $(x-h, x+h)$. As in (8) and by an application of Jensen's inequality,

$$E[\hat{f}(x|h)] = E[K_h(x - \mathbf{X})] \quad = \quad \int_{-\infty}^{\infty} K_h(x-y) f(y)\, dy$$

$$
\begin{aligned}
&= \int_{x-h}^{x+h} K_h(x-y) f(y)\, dy \qquad (9)\\
&\geq f\left(\int_{x-h}^{x+h} y K_h(x-y)\, dy \right)\\
&= f(x).
\end{aligned}
$$

Therefore, $\text{Bias}[\hat{f}(x|h)] \geq 0$.

Clearly from (9), $E[\hat{f}(x|h)] = K_h * f(x)$, where "$*$" denotes convolution. 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 $\text{Bias}[\hat{f}(x|h)] \to -f(x) < 0$.

Now, the general behavior of the bias as a function of h can be established. Since the bias, defined as a convolution of continuous functions, is also a continuous function of h, there must be at least one finite $h_0 > 0$ where the bias function crosses zero. □

Remark 1. Note that the bias of $\hat{f}(x)$ does not involve the sample size as $\hat{f}(x)$ is the mean of n independent and identically distributed random variables. Hence, the value of $h_0(x)$ is a function solely of the kernel, the estimation point, and the underlying density as opposed to h_x^* which is also a function of the sample size. Unfortunately, the closed functional form of h_0 is elusive even with explicit knowledge of f and K. The standard Taylor's series arguments used to determine the functional form of h_x^* fail to be of any use in this case as h_0 does not go to zero for large n.

Remark 2. While a key section of the proof involves finite support of the kernel, the theorem holds for all K, including the normal kernel as was demonstrated in Figure 2. A heuristic argument can be given for a general kernel. Writing the bias as a convolution, $\text{Bias}[\hat{f}(x|h)] = (K_h * f - f)(x)$, it is clear that $\text{Bias}[\hat{f}(x|h)] < 0$ at the primary mode for all $h > 0$. Now, since $\int \text{Bias}[\hat{f}(x|h)]\, dx = 0$ for all $h > 0$, it follows that there must exist intervals over which the bias is positive in order for the integrated bias to equal zero. Hence, 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 as in the theorem.

Remark 3. The existence of such a bandwidth is not entirely unknown in the literature (see Hazelton, 1995). However, it is commonly held to be a theoretical curiosity and of no

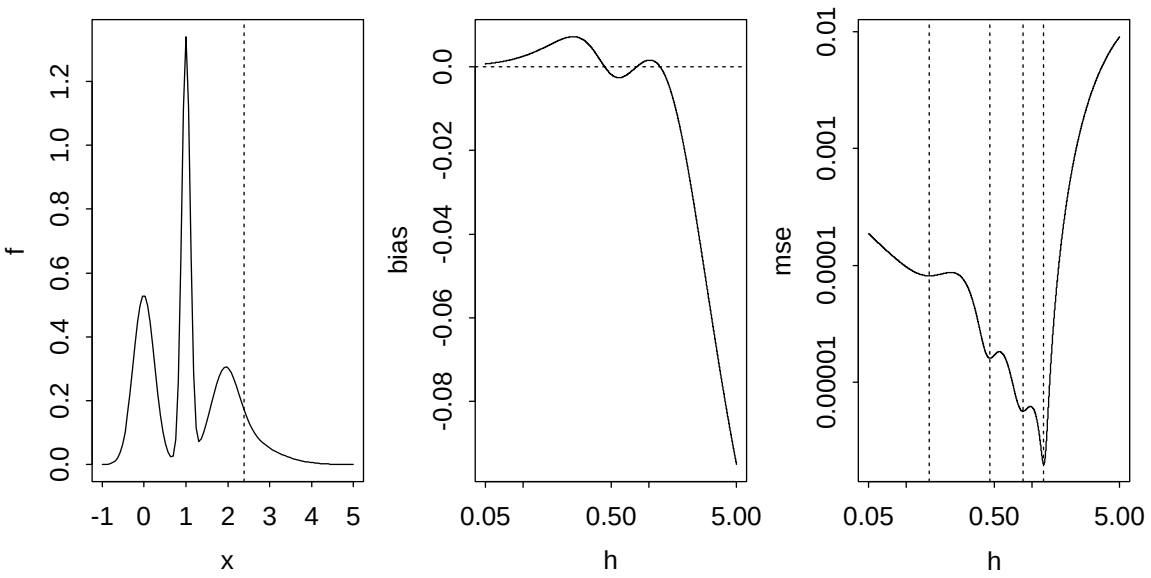

Figure 3: Bias and MSE curves for estimation at $x = 2.4$ when $f(x)$ is a normal mixture density and $n = 5000$.

practical utility. We shall show otherwise.

Remark 4. We give some examples of regions in $\Re^1$ where h_0 exists. For a normal density, $N(\mu, \sigma^2)$, h_0 exists for all x such that $|x - \mu| > \sigma$. For an exponential density, h_0 exists everywhere f is defined ($x > 0$). For the standard Cauchy density, h_0 exists for all $|x| > 1/\sqrt{3}$.

In the case of a multimodal density, the theorem still holds, but h_0 (if it exists) is not necessarily unique for each x. Figure 3 gives an example of just such a situation with a trimodal normal mixture density ($n = 5000$). The parameters of f are given in Table 1. At $x = 2.4$, f is convex and there are three bandwidths ($h = 0.464, 0.849, 1.24$) that have zero bias with corresponding local minima in the MSE. The largest of these bandwidths actually gives the global minimizer of the MSE, a value that is more than an order of magnitude smaller than the MSE at the local minima ($h = 0.153$) corresponding to h_x^*.

i	w_i	μ_i	σ_i
1	$0.\bar{3}$	0.0	0.25
2	$0.\bar{3}$	1.0	0.1
3	0.2	$1.91\bar{6}$	$0.\bar{3}$
4	$0.0\bar{6}$	2.2	0.5
5	$0.0\bar{6}$	2.7	$0.\bar{6}$

Table 1. Parameters of mixture density.

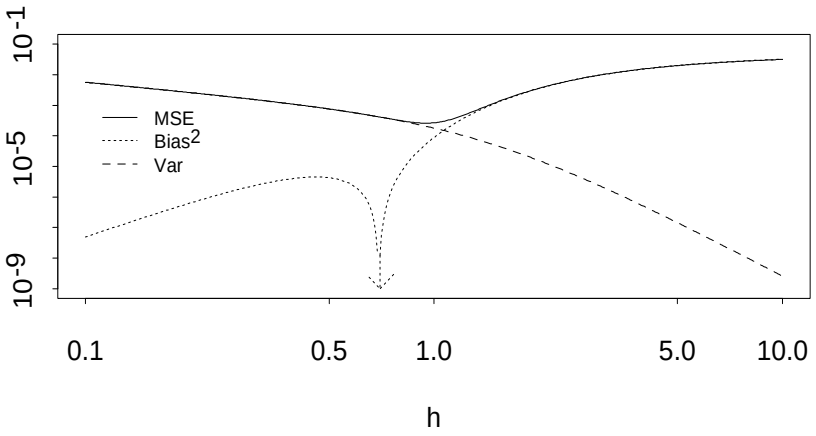

Figure 4: Bandwidth vs MSE for estimation at $x = 1.1$ where $f(x)$ is $N(0,1)$ and $n = 100$.

To understand this phenomenon, consider the effect on the bias of increasing h when f is multimodal. As h moves away from zero, the bias is positive due to contributions from the closest mode. As h gets even larger, the influence of the first mode is diminished and the bias decreases until the bandwidth gets large enough to emphasize the contribution of the next nearest peak, and so on. This effect continues until the bandwidth is large enough to wipe out all of the features of f. It is conceivable that a density that exhibits m modes could have up to $2m + 1$ bandwidths that yield zero bias at some point in a convex region of f.

In general, h_x^* and $h_0(x)$ will both result in a local and possibly global minima in the MSE as was seen in Figures 2 and 3, as a close examination of the MSE vs. h suggests. Given a sufficiently large sample size, the variance is negligible and the nature of the squared bias with respect to h dominates the MSE yielding a local minima near h_0. However, in some small-sample cases, the variance can dominate the bias with respect to the MSE. An example is shown in Figure 4 where the MSE is plotted as a function of h for estimation at $x = 1.1$ where $f(x)$ and K are $N(0,1)$ with $n = 100$. The zero-bias bandwidth is indicated by the arrow where the squared bias goes to zero near $h = 0.7$. However, with this sample size and with x being so close to the non-convex region of the density, the variance dominates the bias and any evidence of the local minima in the MSE near the zero-bias bandwidth is lost.

Figure 5 displays all of the local minima of the exact MSE as a function of both h and

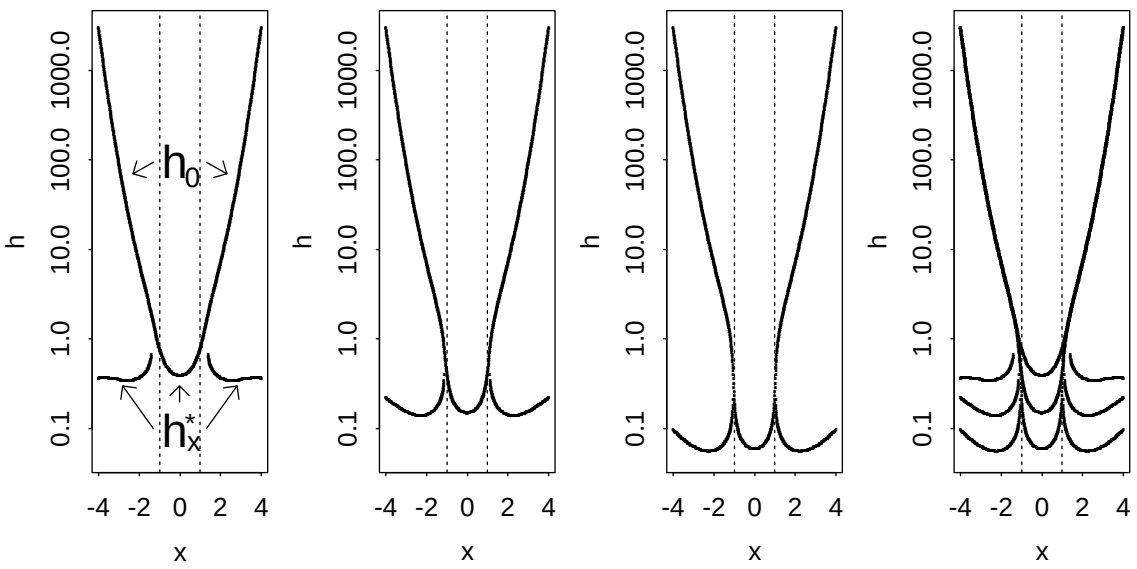

Figure 5: All locally optimal bandwidths for the pointwise local minima of the MSE for $N(0,1)$ data ($n = 10^2, 10^4, 10^6$, and all three combined).

the estimation point x for standard normal data with $n = 10^2$, 10^4, and 10^6. Note the bifurcation points ($x = \pm 1$) where the MSE begins to exhibit more than one local minima due to the phenomenon described above. Also note that the bandwidths in the far tails yielding zero bias are considerably larger than any previously considered in the literature. The rightmost plot overlays the first three plots and shows that these larger local minima in the MSE do not converge to zero as do those corresponding to h_x^*. In fact, as the following theorem shows, these larger local minima in the MSE converges, quite rapidly, to the zero-bias bandwidth, $h_0(x)$, which is constant with respect to n. An interesting exercise is to compare these plots with the asymptotic bandwidth function h_x^* shown in Figure 1. The asymptotic formula tracks the exact values near the mode and the smaller of the two bandwidths in the tails. However, h_x^* fails to describe the situation near ± 1 and misses the larger of the two bandwidths in the convex regions of f. This is due in part to the asymptotic theory failing to adequately describe the behavior of the bias for larger values of h.

Theorem 2.2 (Convergence of h_{mse} to h_0) *Assume the zero-bias bandwidth $h_0(x)$ exists and let $h_{mse}(x)$ be a local minima of the MSE$[\hat{f}(x)]$ in the neighborhood of $h_0(x)$. Then*

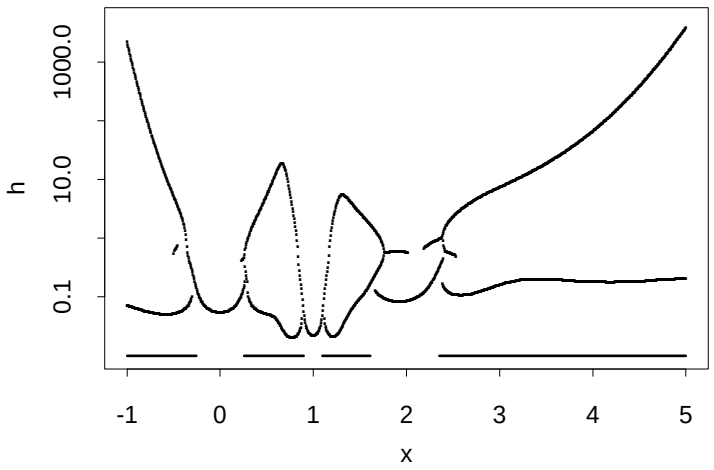

Figure 6: Pointwise local minima of the MSE for the normal mixture density ($n = 500$). The horizontal line segments just above the x-axis indicate regions where f is convex. (Note that the dotted lines should be connected; each dot represents a numerical search for a root.)

$h_{mse}(x) - h_0(x) = O(n^{-1})$. *Writing* $h_{mse}(x) = h_0(x) + \delta_{x,n} + o(n^{-1})$, *where*

$$\delta_{x,n} = \left[\frac{a}{2bh_0^2(x)}\right] n^{-1} > 0,$$

and

$$a = \int_{-\infty}^{\infty} K^2(w) f(x + h_0(x)w)\, dw, \qquad b = \left[\int_{-\infty}^{\infty} wK(w) f'(x + h_0(x)w)\, dw\right]^2.$$

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. In this example, the placement of $h_0(x)$ in the denominator only further limits the magnitude of δ_x since $h_0(x) \gg 0$ as $|x|$ gets larger. In addition, δ_x is a decreasing function of x, meaning the two bandwidths get closer as $|x|$ gets large. Finally, since a and b are both positive, if follows that $\delta_x \geq 0$ and $h_{mse} \geq h_0$.

Figure 6 displays all of the local minima of the pointwise MSE for the normal mixture shown in Figure 3. In a neighborhood of each peak, there is one local minima while in the tails and valleys there are as many as four local minima corresponding to zero-bias bandwidths. It is interesting to note that there are additional minima (for example, when

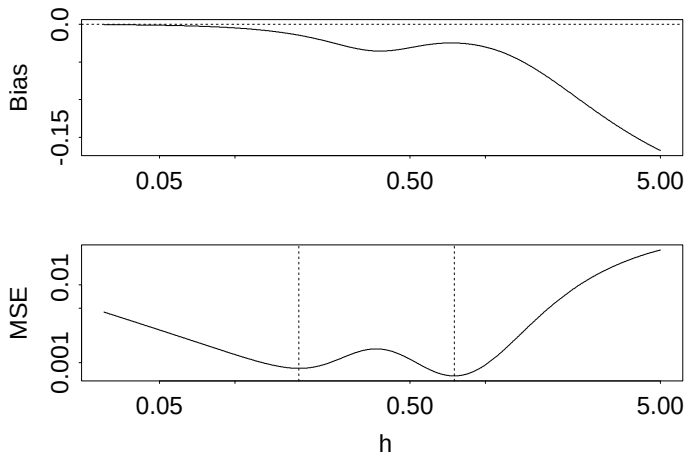

Figure 7: Bias and MSE curves for estimation of f at $x = 2.0$ for the mixture density.

$x = 2$; see Figure 7 for a detailed view at $x = 2$) that occur in the MSE in (non-convex) regions that are not explicitly covered under Theorem 2.1. While they are the direct result of bias behavior discussed previously, the minima are not caused by a zero-bias bandwidth defined under Theorem 2.1. However, such minima are the result of similar mechanics. As the bandwidth increases, the influence of additional modes causes fluctuations in the bias that bring the bias closer to zero resulting in a local minima in the MSE.

As a result of Theorem 2.2, the convergence rate of $\text{MSE}[\hat{f}(x|h_{mse})]$, where h_{mse} is the local minima 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 minima 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, $\text{MSE}[\hat{f}(x|h_{mse})] \leq \text{MSE}[\hat{f}(x|h_0)] = \text{Var}[\hat{f}(x|h_0)] = \frac{1}{n}\text{Var}[K_{h_0}(x)]$. □

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 represents a possible breakthrough in an important practical problem.

3. A COMPARISON OF OPTIMAL ESTIMATION STRATEGIES

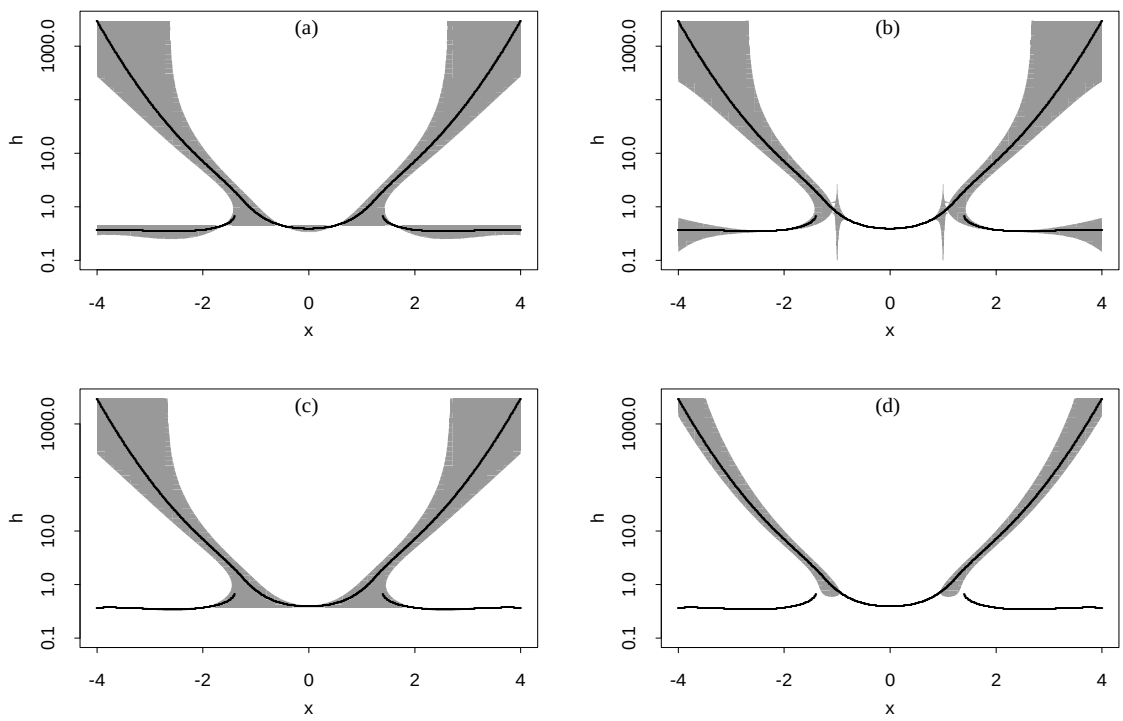

Figure 8: Pointwise local minima of the MSE for $N(0,1)$ data with $n = 100$. Shaded regions represent bandwidths that yield estimates with MSE at least as small as: (a) the optimal MISE fixed bandwidth approach, (b) a local approach using h_x^*, (c) a local approach using the smallest exact local minima of the MSE, and (d) direct parametric estimation.

Despite the fact that the estimator given in (5) does not necessarily integrate to one, balloon estimators have been used in global settings in the literature. In this section, the optimal global estimation strategies for choosing $h_x = h(x)$ utilizing $h_0(x)$ will be discussed and compared to other proposals, namely, the asymptotic bandwidth function, h_x^*. In particular, we address the sensitivity of the MSE to deviations from $h_0(x)$.

Figure 8 once again shows the local minima (bold lines) of the MSE for a random sample of size 100 of $N(0,1)$ data show in the first frame of Figure 5. The shaded region in $8(a)$ represents values of the bandwidth that will yield pointwise MSE smaller than the best MSE achieved with (1), where the bandwidth in (1) is chosen as the minimizer of the exact mean integrated squared error (MISE); see Marron and Wand (1992). There are two apparent approaches to local estimation in the tails that improve upon the fixed bandwidth approach. One could choose h_x to follow the smaller local minima in the MSE

(corresponding to the best tradeoff between variance and bias) or the larger local minima in the MSE (corresponding to the zero-bias bandwidth). While both approaches achieve gains over the fixed method in the tails, the improvement achieved by using the "tradeoff" approach are consistent with the marginal gains observed in the literature.

Many authors have suggested attempting to estimate the first local MSE minima using the asymptotic formula (4) for choosing h_x (Jones, 1990). However, Figure 8(b) shows that we can, in fact, do much better as evidenced by the large shaded regions surrounding the local minima in the MSE corresponding to the zero-bias estimates. Generally, h_x^* does yield gains over the fixed bandwidth in the neighborhood of the origin, but begins to fail far out in the tails (as indicated by the shaded regions near the smaller bandwidths beyond $|x| > 3$). The shaded regions around ± 1 are artifacts due to the divergence of the asymptotic formula around these points.

Part (c) of Figure 8 shows that the larger zero-bias bandwidths can achieve gains in MSE over the smaller local bandwidths associated with the asymptotic results. Finally, part (d) of Figure 8 shows the remarkable result that the zero-bias bandwidths can achieve a lower pointwise MSE in the tails than the maximum-likelihood parametric estimator, $\phi_s(x - \bar{x})$.

While not indicated by Figure 8, it is important to note the area covered by the shaded region near the larger local minima of the MSE grows as the sample size increases. This is due to the faster rate of convergence of the MSE near the zero-bias bandwidths.

Figure 9 displays estimates of 5 random samples of size $n = 100$ from a $N(0, 1)$ distribution. The bottom plot shows fixed kernel estimates using $h = 0.445$, the optimal MSE bandwidth. In the top plot, the first local minima (corresponding to h_x^*) of the MSE is chosen as h_x for the same data. Note that the estimates seem fairly consistent with the optimal fixed kernel bandwidth. There are some irregularities (jumps) in the estimates near ± 1, the bifurcation point of the local minima of the MSE; see Figure 5. Although this strategy might yield gains in the MSE over the optimal fixed bandwidth strategy, visually there is not much difference. In fact, the estimates are less appealing visually.

The estimates in the middle plot use bandwidths chosen along the "best path," which utilizes the zero-bias bandwidths in the tails of the distribution. This best path chooses

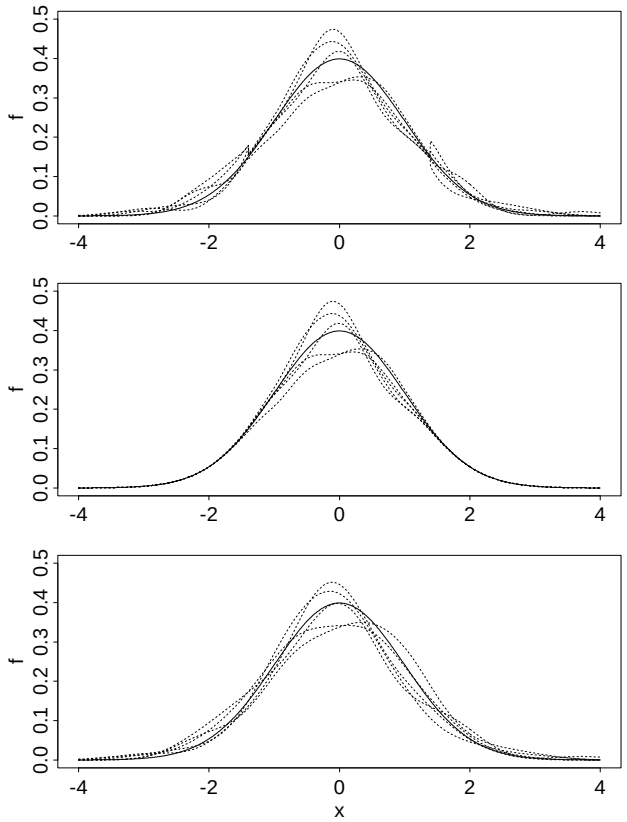

Figure 9: Estimates of samples of size $n = 100$ from $N(0,1)$. The true f is indicated by the solid line. The estimates $\hat{f}$ from top to bottom used h_x^* (first local minimum), "best path" h_x, and h fixed.

the h_x that corresponds to the global minima of the MSE at each estimation point. Note that in the neighborhood of the mode, there is not much difference between this estimator and the fixed kernel method. However, the gains in the tails are outstanding. One minor drawback is that the estimates do not integrate to one.

Figure 10 shows estimates of 5 random samples of size 500 from the normal mixture density. The same bandwidth strategies were employed as in Figure 9, and the results are similar. Of all these estimates, the estimates with the zero-bias bandwidths in the middle plot are clearly superior in regions where the zero-bias effect is available (almost 2/3 of the range of the data).

One of the original premises of this research was that there existed some relationship between optimal adaptive bandwidth functions $h(x)$ and $g(x_i)$. This idea is based on the supposed asymptotic relationship between the two bandwidth functions, so that, in

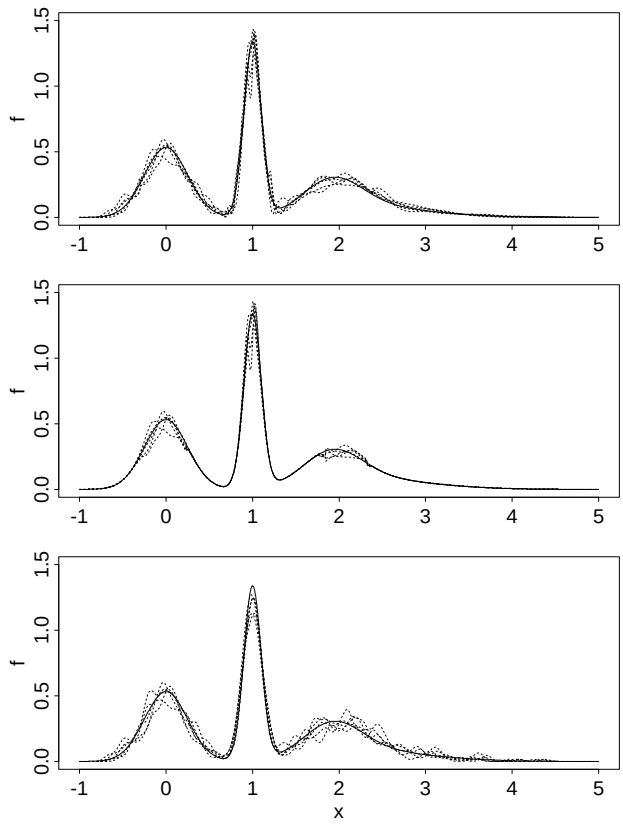

Figure 10: Estimates of samples of size $n = 500$ from normal mixture. The true f is the solid line. The estimates $\hat{f}$ from top to bottom used h_x^* (first local minimum), "best path" h_x, and $h = 0.0515$ fixed.

practice, one could use $g(x_i) = h^*(x_i)$, where $h^*(x)$ is given in Equation (4). However, after examining the pattern of optimal bandwidths in this case, our conclusion is that the relationship is truly asymptotic and not viable for small samples. In fact, even when using the pointwise zero-bias bandwidths that perform so well for $\hat{f}$, the results are not nearly as good as for the sample point estimator. Figure 11 shows estimates of the normal and normal mixture data where $g_i = g(x_i)$ is taken as the global minimizer of the MSE at x_i (the best-path approach). As can easily be seen, such a procedure leads to inferior estimates. In the normal case, the estimates tend to be kurtotic, exhibiting peakedness and heavy tails. In the multimodal case, this is again seen as the estimates at the modes tend to be too high and the valleys and tails too heavy.

4. A LOCAL CROSS-VALIDATION ALGORITHM

The results of the previous section suggest that an optimal estimation scheme should in-

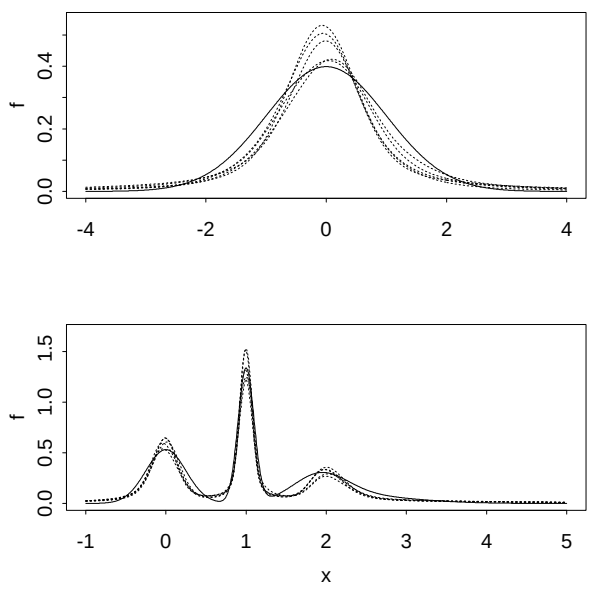

Figure 11: Estimates of $N(0,1)$ data ($n = 100$) and normal mixture data ($n = 500$) using the sample point estimator where g_i is taken as the global minimizer of the pointwise MSE.

corporate the zero-bias bandwidths whenever applicable. The question remains how these theoretical findings can be applied, if at all, in practice. In this section, a local cross-validation algorithm is presented to estimate h_0.

Rudemo (1982) and Bowman (1984) introduced the notion of least-squares cross-validation as a data-based method of estimating the global smoothing parameter for kernel density estimation. Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (1989) extend the global approach to a local procedure used to estimate $h^*(x)$ given in (4). The procedure seeks to minimize a criterion that is an estimate of the local integrated squared error (ISE) using the data in the neighborhood of x.

Consider the integrated squared error of an estimator calculated on some interval $[a, b]$,

$$\begin{aligned} \text{ISE}_{ab} &= \int [\hat{f}(x) - f(x)]^2 I_{[a \le x \le b]}(x)\, dx \\ &= \int_a^b [\hat{f}(x) - f(x)]^2 dx \\ &= \int_a^b \hat{f}(x)^2 dx - 2\int_a^b \hat{f}(x) f(x) dx + \int_a^b f(x)^2 dx, \end{aligned} \tag{10}$$

where I is the indicator function ($I_A(x) = 1\ \forall\ x \in A$, 0 otherwise). The final term in (10) is constant and can be ignored, while, letting K be $N(0,1)$, the first term can be calculated

explicitly as

$$\begin{aligned}\int_a^b \hat{f}(x)^2 dx &= \frac{1}{n^2}\sum_{i=1}^{n}\int_a^b \phi_h^2(x-x_i) + \frac{1}{n^2}\sum_{i\neq j}\int_a^b \phi_h(x-x_i)\phi_h(x-x_j) \\ &= \frac{1}{2n^2 h\sqrt{\pi}}\sum_{i=1}^{n}\left\{\Phi\left(\frac{\sqrt{2}(b-x_i)}{h}\right) - \Phi\left(\frac{\sqrt{2}(a-x_i)}{h}\right)\right\} \\ &+ \frac{1}{n^2}\sum_{i\neq j}\phi_{h\sqrt{2}}(x_i-x_j)\left\{\Phi\left(\frac{\sqrt{2}(b-x_{ij})}{h}\right) - \Phi\left(\frac{\sqrt{2}(a-x_{ij})}{h}\right)\right\},\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal density and $x_{ij} = (x_i + x_j)/2$. The cross-product term in (10) is calculated using the leave-one-out estimator

$$\hat{f}_{-i}(x_i) = \frac{1}{n-1}\sum_{j\neq i}\phi_h(x_i - x_j) \qquad x_i \in [a,b],$$

giving as the expression for the interval least-squares cross-validation criterion function

$$\mathrm{UCV}_{ab} = \int_a^b \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i) I_{[a\leq x_i\leq b]}(x_i). \tag{11}$$

Hall and Schucany (1989) show that such a procedure is unbiased in the sense that $E[\mathrm{UCV}_{ab}] + \int_a^b f(x)^2 dx = \mathrm{MISE}_{ab}$. They show that if $\hat{h}$ minimizes UCV_{ab}, then $\hat{h}/h_x^* \to 1$ almost surely, where h_x^* is the optimal asymptotic bandwidth given in (4) that minimizes the MSE. Hence, by using the data in a small interval around a point, x, an estimate of h_x^* can be found. In addition, they show that when using an interval $(a,b) = (x-\epsilon, x+\epsilon)$, ϵ should shrink to zero more slowly than $n^{-1/10}$, but that the choice of ϵ had little effect on the estimation of h_x^*.

While Hall and Schucany (1989) dealt solely with estimating the smaller bandwidth, h_x^*, with the local cross-validation procedure, it shall be shown here that the same procedure can be used to estimate h_{ab}, if it exists. Here, $h_{ab} = h_0(\eta)$, $a \leq \eta \leq b$, denotes the bandwidth minimizing the integrated squared bias over the interval $[a,b]$. In fact the primary result of this section shows that $\sigma(\hat{h}_{ab} - h_{ab}) = O(n^{-1/2})$, where $\hat{h}_{ab}$ is the minimizer of UCV_{ab} in the neighborhood h_{ab}. Note that the rate of convergence is slightly better than the observed rate of $O(n^{-3/10})$ achieved by the global version of UCV. By using UCV_{ab} to estimate h_{ab}, a reasonable estimator of h_0 at some point $a \leq \eta \leq b$ can be found by considering the data in some small interval around η.

Remark. While $O(n^{-1/2})$ may appear to represent only a modest improvement over the rate $O(n^{-3/10})$, recall $\hat{h}_{ab} = O(1)$ while $\hat{h}_{UCV} = O(n^{-1/5})$, so that the errors converge relatively at the rates $O(n^{-1/2})$ vs $O(n^{-1/10})$, respectively.

First, h_{ab} must be shown to exist and to yield a minima in MISE_{ab}.

Theorem 4.4 (Existence of h_{ab}) *Let $f(x)$ be a continuous probability density function such that f is convex on the interval $[a, b]$. Then there exists a bandwidth $h_{ab} = h_0(\eta)$, $a \leq \eta \leq b$ such that*

$$h_{ab} = \arg\min_h \int_a^b Bias^2[\hat{f}(x|h)]\, dx.$$

Proof: See Appendix.

It is clear that in regions where h_{ab} exists and for a sufficiently large sample size, a local minima in the MISE_{ab} exists. The form of the MISE_{ab} is given by,

$$\begin{aligned} \text{MISE}_{ab}[\hat{f}(x|h)] &= \int_a^b \text{Var}[\hat{f}(x|h)]dx + \int_a^b \text{Bias}^2[\hat{f}(x|h)]dx \\ &= \frac{1}{n}\int_a^b \text{Var}[K_h(x)]dx + \int_a^b \text{Bias}^2[\hat{f}(x|h)]dx. \end{aligned}$$

For large n and large h, the behavior of the MISE_{ab} is dominated by the bias. As before, a local minima of the MISE_{ab} occurs at a value slightly larger than h_{ab} and converges rapidly to h_{ab} as n gets large.

Now that the existence of h_{ab} and the behavior of MISE_{ab} in a neighborhood of h_{ab} has been established, the primary result of this section concerning the convergence of $\hat{h}_{ab}$ minimizing UCV_{ab} to h_{ab} can be stated and proved.

Theorem 4.5 (Convergence of $\hat{h}_{ab}$) *Let the conditions for existence of h_{ab} as stated in Theorem 4.4 be met. Let $\hat{h}_{ab}$ be the (possibly local) minimizer of UCV_{ab} in the neighborhood of h_{ab}. Then both $\sigma(\hat{h}_{ab} - h_{ab}) = O(n^{-1/2})$ and $\sigma((\hat{h}_{ab} - h_{ab})/h_{ab}) = O(n^{-1/2})$.*

Proof: See Appendix.

Figure 12 shows an example of $\text{UCV}_{ab}(h)$ for five samples of size $n = 1000$ from a $N(0, 1)$ distribution where $(a, b) = (2.25, 2.75)$ for estimation of the optimal bandwidths at $x = 2.5$. The dashed lines show the two local minima in the pointwise MSE for estimation of f at $x = 2.5$. All of the curves display at least two local minima distributed around the

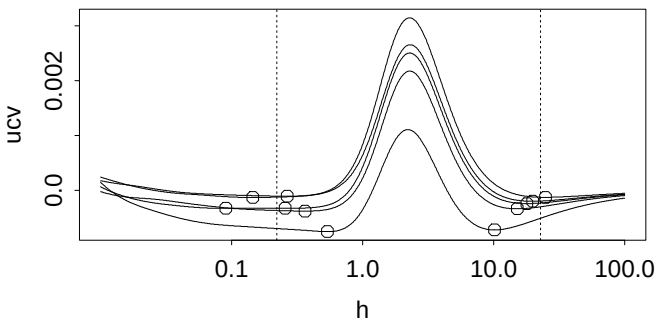

Figure 12: Five $\mathrm{UCV}_{ab}(h)$ functions for samples of size $n = 1000$ from $N(0,1)$ data with $(a,b) = (2.25, 2.75)$. Points represent local minima. Dashed lines represent the two local minima (h_x^* and h_{ab}) of the pointwise MSE at $x = 2.5$.

local minima of the MSE. One of the curves shows three local minima, with two of them around the smaller minima.

A small simulation study was performed to study the behavior of the local cross-validation algorithm and assess the algorithm's ability to find $h_{ab} = h_0(\eta)$. For each of intervals $(1.0, 2.0)$, $(1.5, 2.5)$, and $(2.0, 3.0)$, 100 samples of various sizes were generated from a $N(0,1)$ distribution and $\hat{h}_{ab}$ calculated. Such intervals would be used to find the optimal bandwidths for estimating f at the points $x = 1.5, 2.0,$ and 2.5. Estimates of h_{ab} were found by examining the values of the criterion over a range of possible h values and choosing three values of h that bracketed the location of the local minima. Then, Brent's method of inverse parabolic interpolation (Press et al, 1992) was used to find the minima of the criterion. In some cases that exhibited multiple minima, the estimate of h_{ab} was chosen as the largest local minima found.

The results of the simulation study support the results of Theorem 4.5. Figure 13 shows a plot of the standard deviation of $\hat{h}_{ab}$ as a function of sample size for each of three cases $(a,b) = (1.0, 2.0)$, $(1.5, 2.5)$, and $(2.0, 3.0)$. In each case the width of the interval was held constant to compare the standard deviations as the point at which the window is centered is changed. The dashed line in the plot represents convergence rate of $n^{-1/2}$. In each of the three cases, the rate of convergence of $\sigma(\hat{h}_{ab})$ is proportional to $n^{-1/2}$, although in the

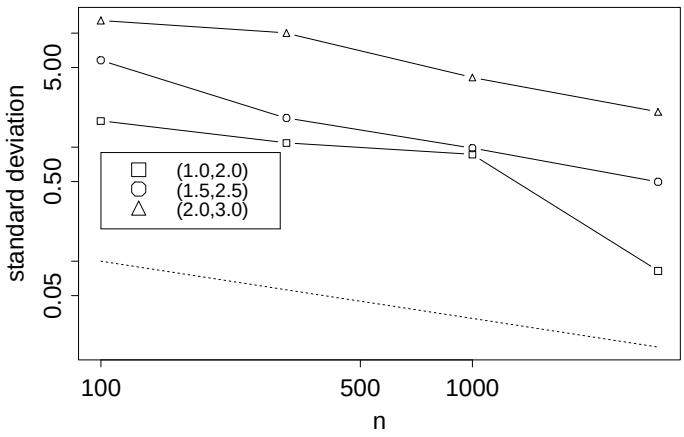

Figure 13: $\sigma(\hat{h} - h_{ab})$ for $N(0,1)$ samples of various sizes and several (a,b) intervals.

case of $(a,b) = (1.0, 2.0)$, the observed rate is somewhat erratic. Note that the standard deviations increase as the window is moved away from the origin.

The convergence of $\hat{h}_{ab}$ to h_{ab} is well-illustrated by Figure 14. However, a curiosity of the procedure becomes apparent. Especially in smaller samples, the distributions of $\hat{h}_{ab}$ exhibit an almost discrete nature. This behavior is also more pronounced as (a,b) gets farther out in the tails of f. An explanation is simple however, as the UCV_{ab} criterion is strongly dependent on the number of $x_i \in (a,b)$. The location of $\hat{h}_{ab}$ is such that samples with the same number of $x_i \in (a,b)$ have approximately the same value of $\hat{h}_{ab}$. Increasing or decreasing the number of $x_i \in (a,b)$ by only one point is enough to radically alter the value of $\hat{h}_{ab}$. Fortunately, this tendency is diminished as the sample size increases.

The erratic behavior of $\hat{h}_{ab}$ in the interval $(a,b) = (1.0, 2.0)$ is explained by a closer examination of the distribution of the optimal $\hat{h}_{ab}$ found. Figure 15 shows the distributions of $\hat{h}_{ab}$ for $(a,b) = (1.0, 2.0)$ and $n = 100, 316, 1000$. The dashed line in the histograms corresponds to $h_{ab} = 2.19$. Note the bimodality in the distributions. The interval is close to regions in which h_{ab} does not exist and the local cross-validation procedure is fooled by the large number and the distribution of points in the interval and chooses a optimal bandwidths corresponding to h_x^* rather than h_{ab}. Hall and Schucany (1989) note that, in their search for h_x^*, the width of the interval should decrease with n, but also note that in their simulation studies the interval width was not terribly important to the performance

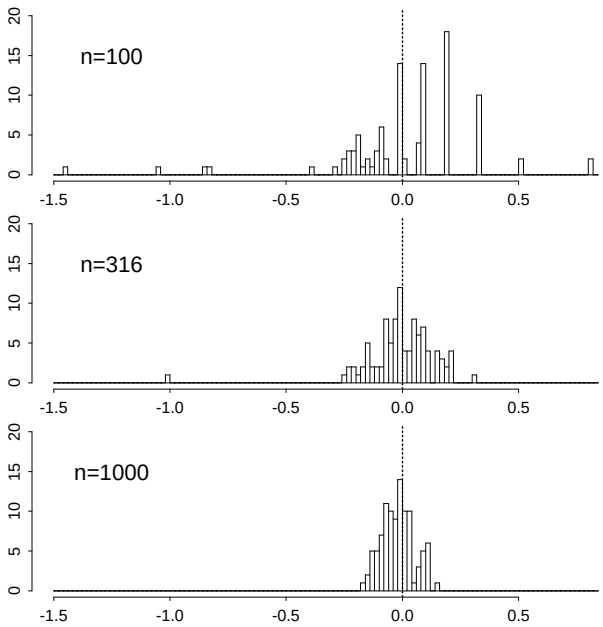

Figure 14: Histograms of $\log_{10}(\hat{h}_{ab}/h_{ab})$ for various sample sizes where $f(x) = N(0,1)$ and $(a,b) = (1.5, 2.5)$.

of the procedure. In this setting, it is apparent that such is not the case, especially near boundary regions. It is clear that the width of the interval should decrease as n increases and should be chosen to ensure the convexity of f over the entire interval. An example with $n = 316$ is shown in Figure 16 where $(a,b) = (1.0, 2.0)$ is reduced to $(a,b) = (1.25, 1.75)$. The bottom plot in the figure corresponds to the reduced window size. The top plot shows the distribution of $\hat{h}_{ab}$ for the wider interval for comparison. The dashed line in the plots again refers to h_{ab}, but in the second plot $h_{ab} = 2.35$, slightly different than that of the wider interval. Note that by choosing a smaller, more appropriate interval, the resulting distribution of $\hat{h}_{ab}$ is more closely centered about h_{ab} and exhibits none of the bimodality present in the top plot.

One final characteristic of the procedure must be noted. In some cases, particularly with smaller samples, no data will be present in a particular choice of interval rendering the UCV_{ab} procedure useless for estimating h_{ab} and h_0. For pointwise estimation, this problem can be overcome by merely increasing the width of the interval. However, due to the apparent discrete nature of the distribution of $\hat{h}_{ab}$ for smaller sample sizes, arbitrarily widening the size of an interval must be done with caution. Additional methods of dealing

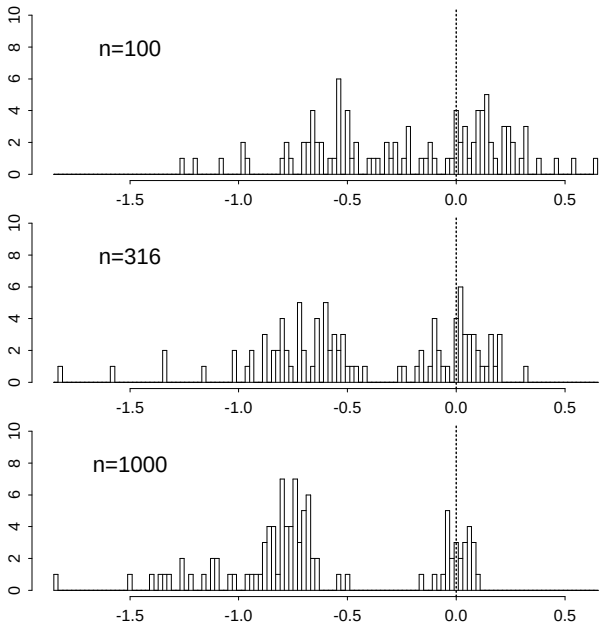

Figure 15: Histograms of $\log_{10}(\hat{h}_{ab}/h_{ab})$ for various sample sizes where $f(x) = N(0,1)$ and $(a,b) = (1.0, 2.0)$.

with this problem for global estimation will be discussed in the next section.

5. A GLOBAL ESTIMATION APPROACH

In this section, a global implementation procedure for estimating the bandwidth function $h(x)$ or h_x in (5) is developed. The procedure is similar in flavor to that discussed in Fan et al. (1993) and Sain (1994) for estimating the bandwidth function $g(x_i)$ in the sample point estimator (3). An outline of the new procedure is quite simple. The data are first binned according to an equally spaced mesh, where the mesh spacing is approximately $s \cdot n^{-1/10}$, where s is the sample standard deviation and n is the sample size. This spacing is consistent with the findings of Hall and Schucany (1989). The local cross-validation procedure is used to estimate bandwidths with the data in each non-empty bin. The estimated bandwidth is chosen as the largest bandwidth found in each non-empty bin.

A sequence of points of the form (t_i, h_i), $1 = 1, \ldots, m$, is obtained where t_i is the midpoint of the ith non-empty bin and h_i the estimated bandwidth for the ith non-empty bin. A continuous bandwidth function can be constructed from the m ordered pairs by using a natural cubic spline to interpolate the bandwidth values. Bandwidths for any given estimation point can then be taken directly from the spline.

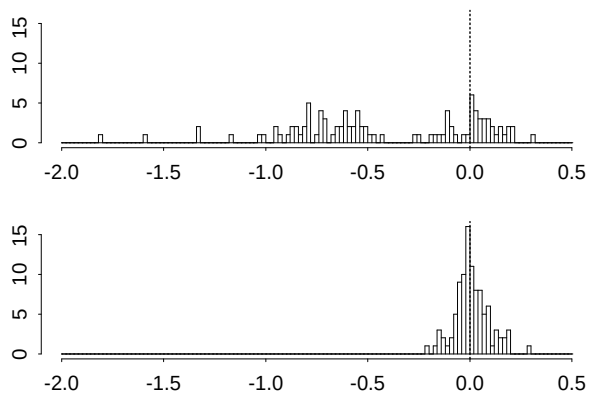

Figure 16: Histograms of $\log_{10}(\hat{h}_{ab}/h_{ab})$ for $(a,b) = (1.0, 2.0)$ and $(1.25, 1.75)$ for $f(x) = N(0,1)$ and $n = 316$.

An example using the suicide data ($n = 86$ and $s = 146.7$) studied extensively by Silverman (1986) is shown in Figure 17. The top plot shows a histogram of the data with the mesh spacing of 100 used to bin the data and the number of points in each bin. The middle plot shows the estimated bandwidths for each bin and the resulting spline function (solid line). Other local minima in the local cross-validation criterion are shown by the open circles. The unbiased cross-validation estimate for the fixed bandwidth estimator is shown by the dotted line.

Estimates using the local density estimator using the spline estimated bandwidth function (solid line) and the fixed bandwidth estimator using the unbiased cross-validation smoothing parameter (dotted line) are shown in the bottom plot. Note the two estimates agree near the mode where the estimated bandwidths are very similar. However, much of the noise in that tail of the fixed bandwidth estimate is not present in the local density estimate.

In the earlier sections, it was shown that the these zero-bias bandwidths can be useful not only in estimating the tails of a distribution, but also in emphasizing multiple modes by yielding better estimates of the density in the valleys between multiple modes. An example is given in Figure 18 using the Buffalo snowfall data ($n = 62$ and $s = 23.72$) which has been used extensively in the literature (see Scott (1992), Silverman (1986), and Parzen (1980)). The adaptive estimate does exhibit possible evidence of three bumps.

The top plot in Figure 18 shows the histogram of the data using a mesh spacing of 15.

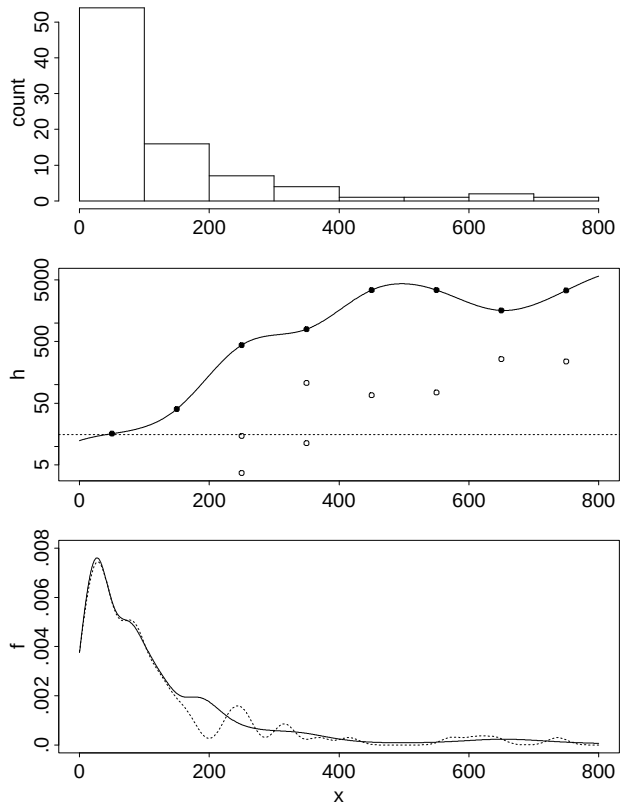

Figure 17: Example of the global estimation approach for the suicide data.

The estimated bandwidth functions are shown in the middle plot. Note that the bandwidth function estimated by the local cross-validation scheme does exhibit evidence suggestive of multimodality.

The resulting estimates are shown in the bottom plot. The local cross-validation procedure (solid line) does show more evidence of trimodality than the fixed bandwidth case, which shows what appears to be one mode and two shoulders.

6. DISCUSSION

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 the tails. A multivariate extension is being explored.

Rosenblatt (1956) first demonstrated the poor bias properties of kernel estimates. We have proved a partial counterexample. Furthermore, we have also shown a fully data-based

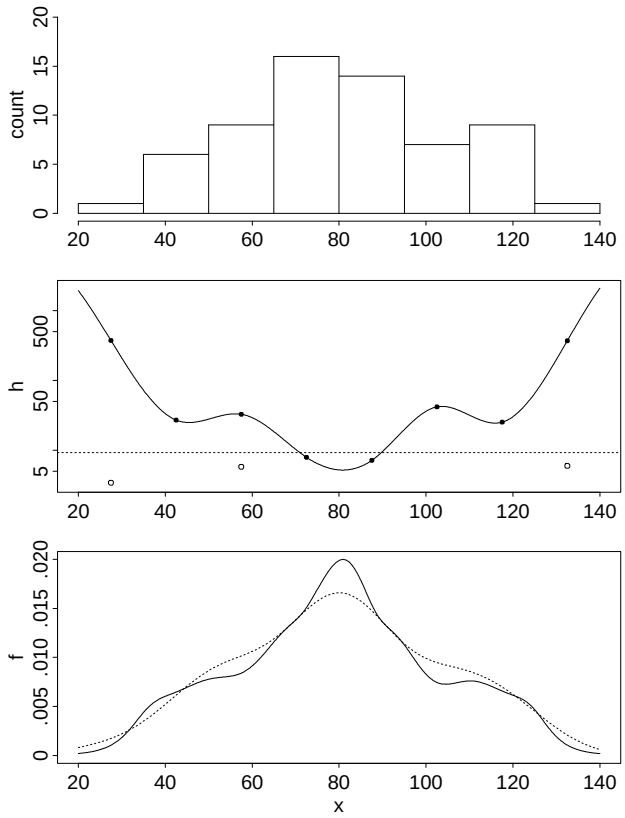

Figure 18: Example of the global estimation approach for the Buffalo snowfall data.

procedure to utilize these bias eliminating bandwidths in practice.

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. The effect of censoring on our results deserves study.

As an interesting aside, we have shown that the two basic adaptive strategies suggested in equations (2) and (3) are more dissimilar than many workers have imagined. For example, it has been assumed that if the optimal form of $h(x)$ in (2) was known, call it $h^*(x)$, then the optimal form for the sample adaptive bandwidth $g(x_i)$ could be well-approximated by the equation $g^*(x_i) = h^*(x_i)$. We have shown by example that this is not the case for either the asymptotic form h_x^* or for h_0.

ACKNOWLEDGMENTS

The authors would like to thank Keith Baggerly for his helpful comments. This research was supported in part by the National Science Foundation under grant DMS-9306658 and

the National Security Agency under grant MOD 9086-93.

APPENDIX

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

$$\begin{aligned} \text{Bias}[\hat{f}(x|h)] &= E[\hat{f}(x|h)] - f(x) \\ &= E[K_h(x)] - f(x), \end{aligned}$$

where

$$\begin{aligned} E[K_h(x)] &= \int \frac{1}{h_0+\delta} K\left(\frac{x-y}{h_0+\delta}\right) f(y)dy \\ &= \int K(w) f(x+(h_0+\delta)w)dw \\ &= \int K(w) f(x+h_0w+\delta w)dw. \end{aligned}$$

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

$$\begin{aligned} E[\hat{f}(x|h)] &= \int K(w)[f(x+h_0w) + \delta w f'(x+h_0w) \\ &\quad + \frac{1}{2}\delta^2 w^2 f''(x+h_0w) + \cdots]dw. \end{aligned}$$

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

$$\text{Bias}[\hat{f}(x|h)] = \delta \int wK(w)f'(x+h_0w)dw. \tag{12}$$

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

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

Expanding $f(x + h_0w + \delta w)$ as before, an asymptotic approximation of the variance of $\hat{f}(x|h)$ is given by

$$\text{Var}[\hat{f}(x|h)] = \frac{1}{n(h_0+\delta)}\int K^2(w) f(x+h_0w)dw. \tag{13}$$

Let $a = \int K^2(w) f(x + h_0 w) dw$ and $b = [\int w K(w) f'(x + h_0 w) dw]^2$, then combining (12) and (13), the MSE of $\hat{f}(x|h)$ is

$$\begin{aligned} \text{MSE}[\hat{f}(x|h)] &= \frac{a}{n(h_0 + \delta)} + \delta^2 b \\ &= \frac{a}{nh_0}\left(1 + \frac{\delta}{h_0}\right)^{-1} + \delta^2 b. \end{aligned}$$

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

$$\text{MSE}[\hat{f}(x|h)] = \frac{a}{nh_0} - \frac{a\delta}{nh_0^2} + \delta^2 b.$$

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

$$\frac{\partial}{\partial \delta}\text{MSE}[\hat{f}(x|h_{mse})] = -\frac{a}{nh_0^2} + 2\delta b = 0.$$

Solving for δ,

$$\delta = \delta_x = \frac{a}{2bnh_0^2}. \quad \square$$

Proof of Theorem 4.4: By the mean value theorem, the integrated squared bias can be written as

$$\int_a^b \text{Bias}^2[\hat{f}(x|h)]\, dx = (b - a)\, \text{Bias}^2[\hat{f}(\eta|h)], \tag{14}$$

where $a \leq \eta \leq b$. Taking the derivative of (14) and setting equal to zero yields

$$\begin{aligned} \frac{\partial}{\partial h}(b - a)\, \text{Bias}^2[\hat{f}(\eta|h)] &= \frac{\partial}{\partial h}(b - a)\left\{E[\hat{f}(\eta|h)] - f(\eta)\right\}^2 \\ &= 2\,(b - a)\left\{E[\hat{f}(\eta|h)] - f(\eta)\right\}\frac{\partial}{\partial h}E[\hat{f}(\eta|h)] = 0. \end{aligned} \tag{15}$$

Note that, by definition, η is in a convex interval of f, and setting $h = h_0(\eta)$ makes the bracketed term in (15) equal to zero by Theorem 2.1. Hence, $h = h_0(\eta)$ is a solution to (15) and a critical value. Similar arguments show that the second derivative of the integrated squared bias at $h = h_0(\eta)$ is positive. Thus, $h = h_0(\eta)$ is a minimizer. $\square$

Proof of Theorem 4.5: To capture the quadratic nature of MISE_{ab} around h_{ab}, expand MISE_{ab} in a second-order Taylor's series expansion about h_{ab},

$$\begin{aligned}\mathrm{MISE}_{ab}(h) &= c_0 + c_1(h - h_{ab}) + \frac{1}{2}c_2(h - h_{ab})^2 + \cdots \\ &= c_0 + \frac{1}{2}c_2(h - h_{ab})^2 + \cdots\end{aligned}$$

where

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

The linear term vanishes since h_{ab} is a local minima of MISE_{ab}; hence, $c_1 = 0$. Similarly for UCV_{ab},

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

where

$$\hat{c}_0 = c_0 - \int_a^b f(x)^2 dx + d_0 Z_0, \qquad \hat{c}_1 = d_1 Z_1, \qquad \hat{c}_2 = c_2 + d_2 Z_2,$$

with $d_i > 0$, $i = 0, 1, 2$, and Z_i, $i = 0, 1, 2$, are assumed to be random variables with mean zero and unit variance. Since UCV_{ab} is unbiased for MISE_{ab} (except for a constant level shift), then the Taylor's series for each must match term by term, on average; hence, $E\hat{c}_i = c_i$, or equivalently, $EZ_i = 0$.

Let $\hat{h}$ be the local minimizer of UCV_{ab}. Then,

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

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

$$\begin{aligned}\hat{h} - h_{ab} &= \frac{\hat{c}_1}{\hat{c}_2} = \frac{d_1 Z_1}{c_2 + d_2 Z_2} \\ &= \frac{d_1 Z_1}{c_2}\left(1 + \frac{d_2}{c_2}Z_2\right)^{-1}.\end{aligned}$$

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

$$\hat{h} - h_{ab} = \frac{d_1 Z_1}{c_2}\left(1 - \frac{d_2}{c_2}Z_2 + \left(\frac{d_2}{c_2}\right)^2 Z_2^2 + \ldots\right). \tag{16}$$

To establish the order of the rate of convergence of $\sigma(\hat{h} - h_{ab})$, the behavior of the terms d_1, d_2, and c_2 must be established. First, examining c_2,

$$\begin{aligned} c_2 &= \frac{\partial^2}{\partial h^2}\text{MISE}_{ab}(h_{ab}) \\ &= \frac{\partial^2}{\partial h^2}\left(\int_a^b \text{Var}[\hat{f}(x|h_{ab})]dx + \int_a^b \text{Bias}^2[\hat{f}(x|h_{ab})]dx\right) \\ &= \frac{1}{n}\frac{\partial^2}{\partial h^2}\int_a^b \left[E[K^2_{h_{ab}}(x)] - E[K_{h_{ab}}(x)]^2\right]dx + \frac{\partial^2}{\partial h^2}\int_a^b \left[E[K_{h_{ab}}(x)] - f(x)\right]^2 dx. \end{aligned}$$

Note that only the first term, corresponding to the variance, is a function of n. The second term is constant with respect to n. Therefore, c_2 can be rewritten as $c_2 = c_2^* + O(n^{-1})$, where c_2^* is a constant.

The next two terms, d_1 and d_2, are somewhat more complicated. Examining the variance of the derivative of UCV_{ab} with respect to h,

$$\begin{aligned} \text{Var}[\frac{\partial}{\partial h}\text{UCV}_{ab}(h)] &= \text{Var}[\hat{c}_1 + \hat{c}_2(h - h_{ab})] \\ &= \text{Var}[d_1 Z_1 + (c_2 + d_2 Z_2)(h - h_{ab})] \\ &= d_1^2 + d_2^2(h - h_{ab})^2. \end{aligned} \tag{17}$$

Taking the partial derivative with respect to h of the UCV_{ab} criterion given in (11) and rewriting,

$$\frac{\partial}{\partial h}\text{UCV}_{ab}(h) = \frac{1}{n^2}\sum_{i=1}^n \lambda_i + \frac{1}{n^2}\sum_{i\neq j}\gamma_{ij} - \frac{1}{n(n-1)}\sum_{i\neq j}\eta_{ij}$$

where

$$\begin{aligned} \lambda_i &= \frac{1}{2\sqrt{\pi}}\frac{\partial}{\partial h}\left\{\frac{1}{h}\left[\Phi\left(\frac{\sqrt{2}(b - x_i)}{h}\right) - \Phi\left(\frac{\sqrt{2}(a - x_i)}{h}\right)\right]\right\} \\ \gamma_{ij} &= \frac{\partial}{\partial h}\left\{\phi_{h\sqrt{2}}(x_i - x_j)\left[\Phi\left(\frac{\sqrt{2}(b - x_{ij})}{h}\right) - \Phi\left(\frac{\sqrt{2}(a - x_{ij})}{h}\right)\right]\right\} \\ \eta_{ij} &= 2\frac{\partial}{\partial h}\phi_h(x_i - x_j)I_{[a\leq x_i \leq b]}(x_i). \end{aligned}$$

Note that the indicator function in the third term of the $\frac{\partial}{\partial h}\text{UCV}_{ab}(h)$ does not reduce the number of terms in the double sum as the number of terms still increases as n^2. The variance of the derivative of UCV_{ab} can be written as

$$\text{Var}[\frac{\partial}{\partial h}\text{UCV}_{ab}(h)] = \frac{1}{n^4}\left\{\text{Var}\left[\sum_{i=1}^n \lambda_i\right]\right\} + \frac{1}{n^4}\left\{\text{Var}\left[\sum_{i\neq j}\gamma_{ij}\right]\right\}$$

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

$$+ \quad \frac{1}{n^3(n-1)}\left\{\mathrm{Cov}\left[\sum_{i=1}^{n}\lambda_{ij}, \sum_{i\neq j}\eta_{ij}\right]\right\} + \frac{1}{n^3(n-1)}\left\{\mathrm{Cov}\left[\sum_{i\neq j}\gamma_{ij}, \sum_{i\neq j}\eta_{ij}\right]\right\}.$$

The order of each term in the $\mathrm{Var}[\frac{\partial}{\partial h}\mathrm{UCV}_{ab}(h)]$ can now be examined. The first term can be expanded as

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

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 and the contribution to d_1^2 is $O(n^{-3})$. The second term in the variance is given by

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

where $i \neq j \neq k$. In this case, there are $n(n-1)(2n-3)$ nonzero terms in the variance giving a $O(n^{-1})$ contribution to d_1^2. Note that the behavior of the third term is exactly that of the previous term.

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

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

where the only nonzero terms are those where $i \neq j$. The next term involving λ_i and η_{kl} behaves in exactly the same manner. Finally, the last covariance term is

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

where the only nonzero terms are those where $i \neq j \neq k$.

Hence, $\mathrm{Var}[\frac{\partial}{\partial h}\mathrm{UCV}_{ab}(h)] = d_1^2 + d_2^2(h - h_{ab})^2 = O(n^{-1})$. A similar argument yields $d_2^2 = \mathrm{Var}[\frac{\partial^2}{\partial^2 h}\mathrm{UCV}_{ab}(h)] = O(n^{-1})$. All the terms in (17) except d_1^2 have all been shown to be $O(n^{-1})$; hence, d_1^2 can converge no slower than $O(n^{-1})$. Then, (16) can be rewritten as $\hat{h} - h_{ab} = d_1 Z_1/c_2$ and $\sigma(\hat{h} - h_{ab}) = d_1/c_2$ since the additional terms involving products of d_1 and d_2 are of higher order. Combining the previous results gives $d_1/c_2 = O(n^{-1/2})$. The remainder of the theorem follows since $h_{ab} = O(1)$. □

REFERENCES

Abramson, I. (1982a), "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.

Davis, K.B. (1975), "Mean Square Error Properties of Density Estimates," *Annals of Statistics*, **3**, 1025-1030.

Fan, J., Hall, P., Martin, M., and Patil, P. (1993), "On Local Smoothing of Nonparametric Curve Estimators," Technical Report No. 440, Deparment of Statistics, Stanford University.

Hall, P., Hu, T.C., and Marron, J.S. (1994), "Improved Variable Window Kernel Estimates of Probability Densities," *Annals of Statistics*, To appear.

Hall, P. and Marron, J.S. (1988), "Variable Window Width Kernel Estimates," *Probability Theory and Related Fields*, **80**, 37-49.

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

Hazelton, M. (1995), "Bandwidth Selection for Local Density Estimation," *Scandinavian Journal of Statistics*, To appear.

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

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.

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.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992), *Numerical Recipes in C*, Cambridge University Press.

Rosenblatt, M. (1956) "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, **27**, 832-837.

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. Manuscript, Department of Statistics, Rice University.

Sain, S.R. and Scott, D.W. (1995), "On Locally Adaptive Density Estimation," Submitted.

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

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

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

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