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Sain, S.R. (1994), "Adaptive Kernel Density Estimation," Unpublished Ph.D. Dissertation, Department of Statistics, Rice University.

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.

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.

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

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. (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. (1996), "Bandwidth Selection for Local Density Estimation," *Scandinavian Journal of Statistics*, **23**, 221-232.

Hjort, L.N. and Jones, M.C. (1996), "Locally Parametric Nonparametric Density Estimation," *Annals of Statistics*, **24**, 1619-1647.

Hyndman, R.J., Bashtannyk, D.M., and Grunwald, G.K. (1996), "Estimating and Visualizing Conditional Densities," *Journal of Computational and Graphical Statistics*, **5**, 315-336.

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.

Loader, C.R. (1996), "Local Likelihood Density Estimation," *Annals of Statistics*, **24**, 1602-1618.

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.

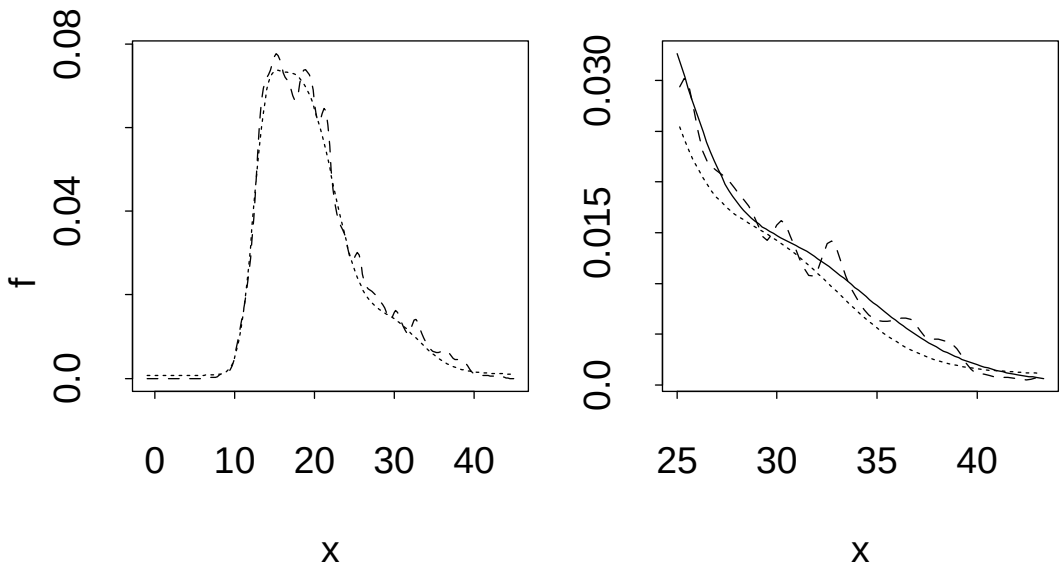

Figure 15: Density estimates using three methods: balloon estimator (solid line, right frame only); sample point estimator (dotted line); and the fixed bandwidth estimator (dashed line).

portant areas such as reliability and risk analysis. This work clearly paves a new path.

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-9626187 and the Department of Defense under contract MDA 904-95-C-2203.

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.

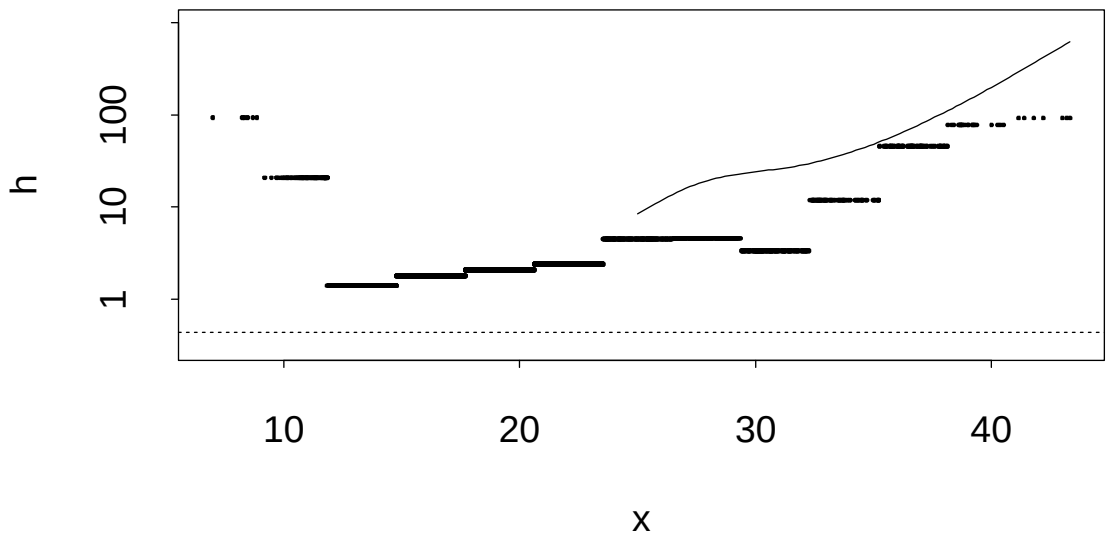

Figure 14: Estimated bandwidth functions for the Melbourne temperature data. The solid circles indicate the estimated bandwidths using the sample point estimator while the solid line show the estimated bandwidth function for the balloon estimator. The dotted line is the value of the estimated fixed bandwidth.

the tails where f is convex. While our practical implementation does not achieve $O(n^{-1})$ convergence, it does appear to represent at least a constant factor improvement and have excellent bias properties even in the extreme tails. We believe the large bandwidths found by the Sain and Scott (1996) implementation of the sample point estimator reflect the zero-bias phenomenon and give an excellent global estimate. Better tail estimates may be available with the current proposal.

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 multivariate extension indicated earlier is being explored.

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 design density, then a zero-bias bandwidth exists at points where the initial bias moves the estimate away from the global mean.

Nonparametric estimates in tails have always been suspect, limiting application in im-

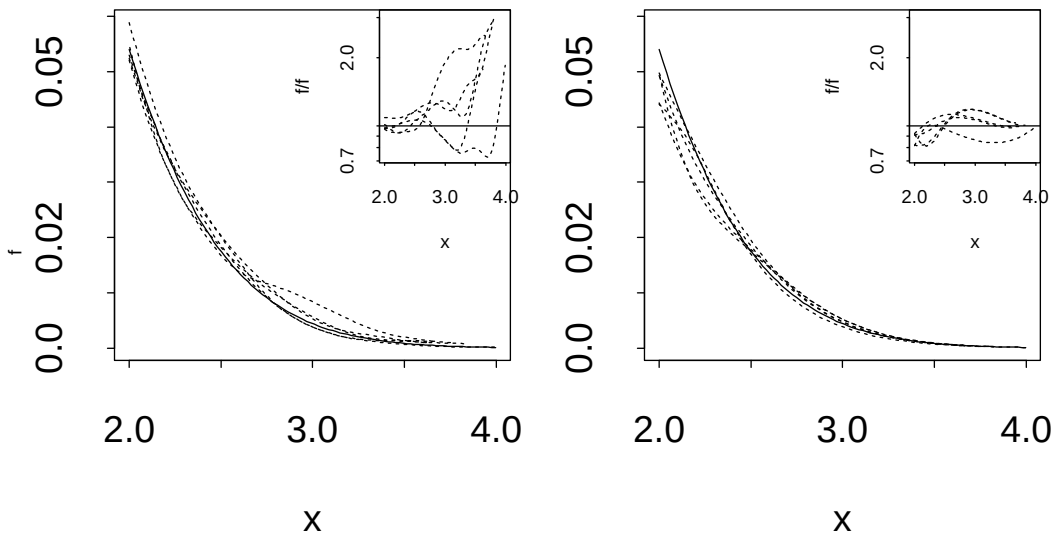

Figure 13: Estimated densities using the best global MISE bandwidth (left) and the estimated $h_0(x)$ (right) for $N(0,1)$ data (n=10000). Shown in the detail are $\log_{10}(\hat{f}/f)$.

dicated by the solid line). A close examination reveals some interesting results. While the sample point estimator does a good job globally, at least in comparison with the fixed estimator, the estimate is smaller than the balloon estimator in the right tail. What is encouraging about the balloon estimator is that it tracks the fixed bandwidth estimator, except, of course, without the variability. Clearly the estimated fixed bandwidth is too small, but the 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 also having a much smaller variance in a difficult region of the density to estimate.

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, variable kernel estimation in convex regions is fundamentally different than near the modes. The extraordinarily large bandwidths predicted by the zero-bias theory have been shown to be estimable in

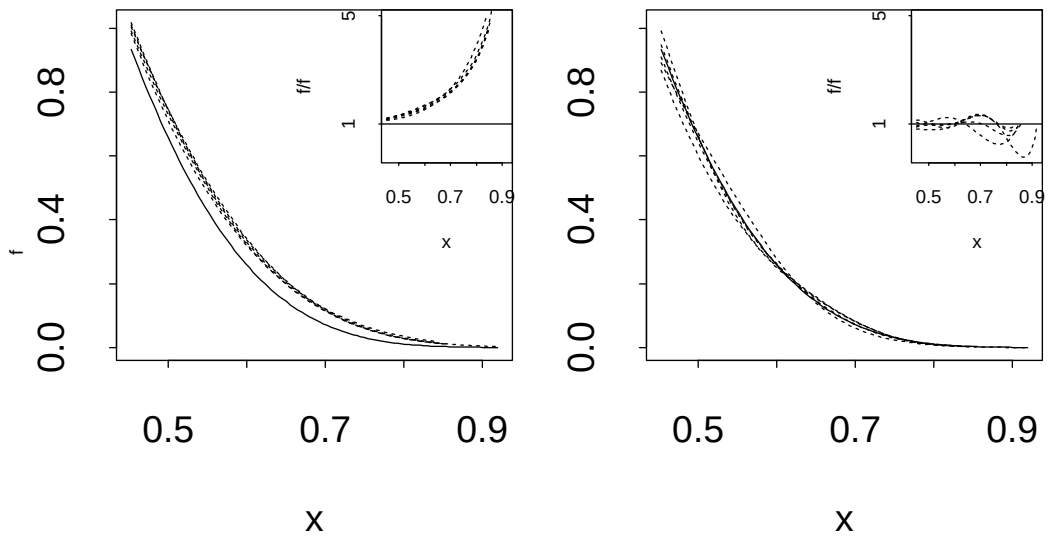

Figure 12: Estimated densities using the best global MISE bandwidth (left) and the estimated $h_0(x)$ (right) for beta(2,6) data (n=10000). Shown in the detail are $\log_{10}(\hat{f}/f)$.

by either adaptive approach.

One of the original premises 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. Furthermore, it was hoped that this relationship could be exploited in developing practical algorithms for locally adaptive density estimation. This example shows that, if such a relationship exists, it is truly asymptotic in nature. Note the scales for the two bandwidth functions. The bandwidths for the balloon estimator are much larger than those found using the sample point estimator and lead to significantly different estimates.

Figure 15 shows the estimated densities. In the first plot, the two global approaches are shown with the sample point estimator indicated by the dotted line and the fixed bandwidth estimator with the dashed line. Note that the fixed bandwidth estimator appears to be undersmoothed for $x > 15$, as the small value of the least-squares cross-validation bandwidth tries to cope with the sharp rise at $x = 10$. The adaptive estimate seems to do a more than adequate job of smoothing over the whole range of the density.

The second plot shows a detail of the right tail, including the balloon estimator (in-

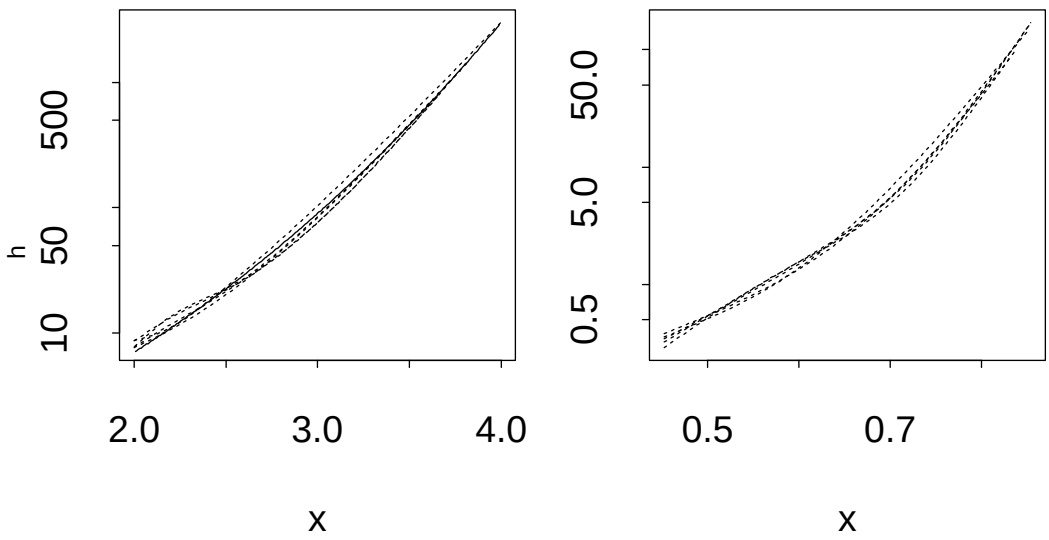

Figure 11: Estimated bandwidth functions for the two test cases: $N(0,1)$, and $B(2,6)$.

over the fixed estimator in the tail, particularly taking into consideration that the adaptive density estimates are data-based while the others use the theoretically optimal bandwidth. Note all the estimates using h_{MISE} are biased upward, as is to be expected in a convex region, while $\hat{h}(x)$'s bias behaves better.

A second example with standard normal data is shown in Figure 13 and shows similarly promising results.

5. AN EXAMPLE

Finally, an example using real data is shown in Figures 14 and 15. The data are daily maximum temperatures in Melbourne, Australia, from 1981-1990 ($n = 3650$) taken from Hyndman et al. (1996). Figure 14 displays two locally adaptive bandwidth functions. The bandwidth function from the global approach using least-squares cross-validation and the binned sample point estimator described in Sain and Scott (1996) is indicated by the solid circles. The solid line indicates the estimated bandwidth function using the local cross-validation procedure for the balloon estimator outlined in the previous section for the right tail of the density. For comparison, a fixed bandwidth estimator was also included with the smoothing parameter estimated by least-squares cross-validation. The value of $\hat{h}$ found in this case was 0.437 (indicated by the dotted line), significantly smaller than those found

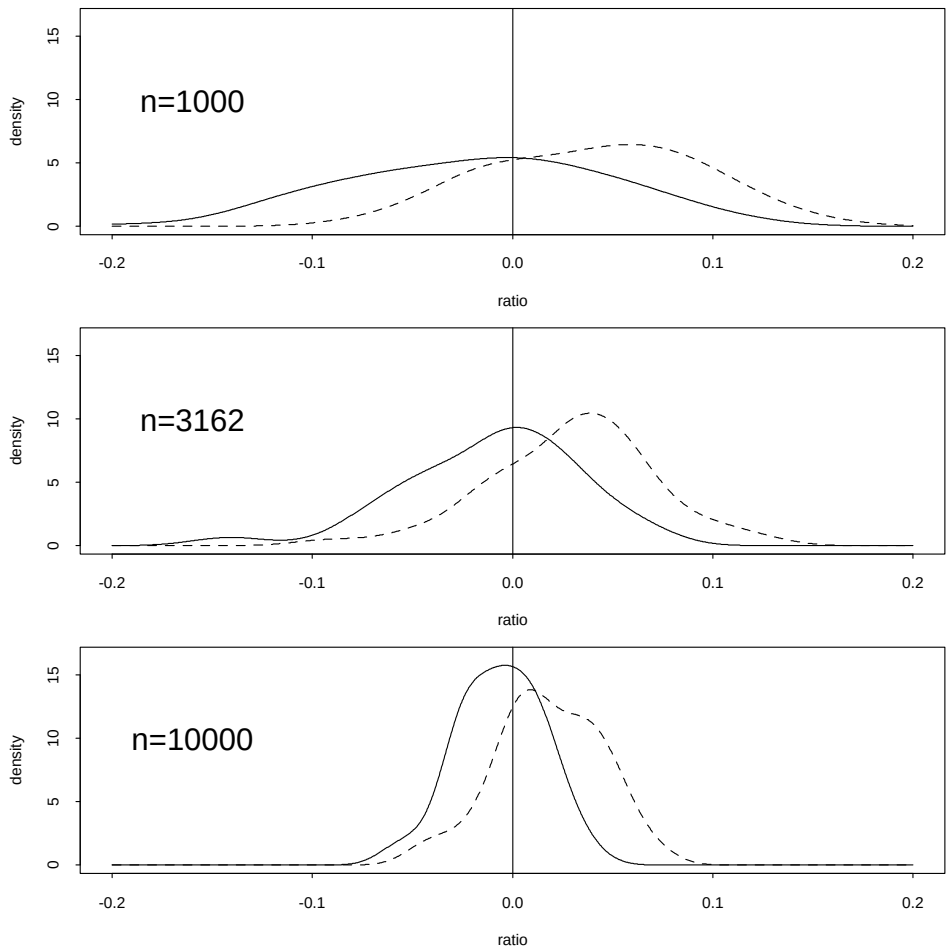

Figure 10: Kernel estimates of $\log_{10}(\hat{f}(2)/f(2))$ for the two estimation schemes outlined in the text. Solid line indicates estimates using $\hat{h}_0(x)$ and the dashed line indicates estimates using $\hat{h}_{\text{MISE}}$. The sample sizes for the three plots (from top to bottom) are $n = 1000, 3162, 10000$.

convex region (right tail) of these densities. Four knot locations were used in each case, at equal probability quantiles for the normal density and at equally spaced intervals for the beta density. The cross-validation criterion was then minimized for the heights of the estimated bandwidth functions at the knots. The estimated bandwidth functions (dotted lines) are shown in Figure 11. Note that in each case, the cubic spline does an excellent job of estimating the true bandwidth function (solid line).

The next two plots (Figures 12 and 13) show five estimated densities using $\hat{h}_0(x)$ and the theoretical bandwidth that minimizes the MISE. The first plot (Figure 12) shows the estimates for the beta density. It is clear that the adaptive estimator is clearly superior

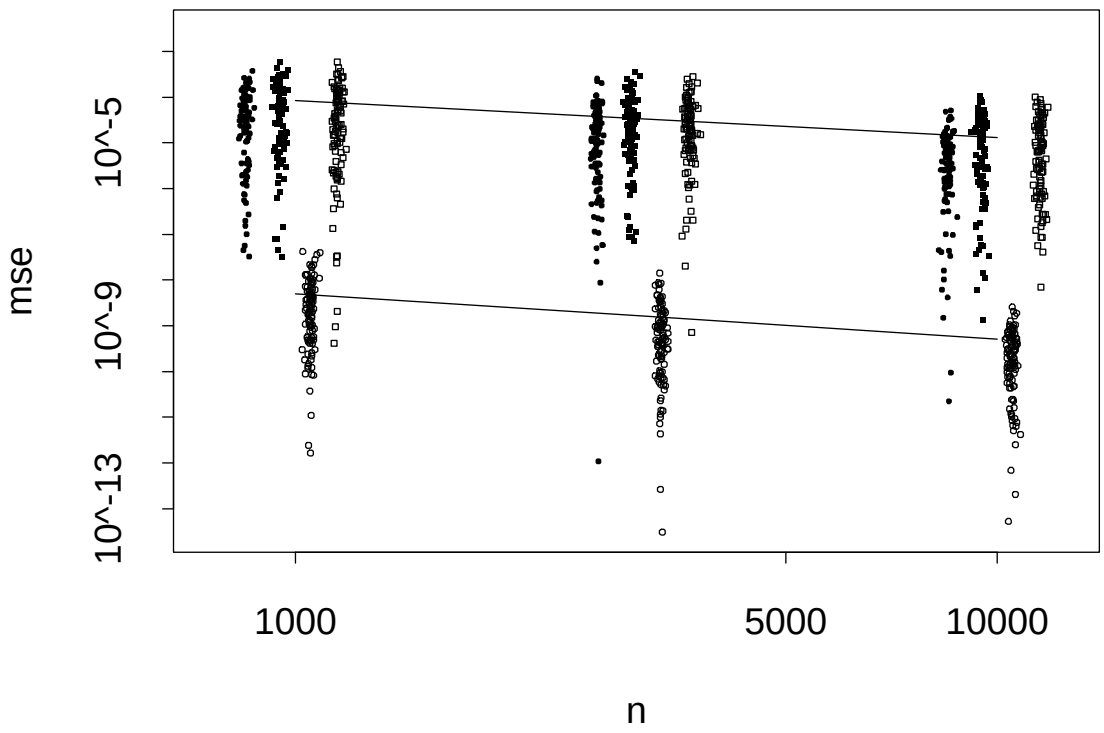

Figure 9: Plot of MSE versus sample size showing the results from a simulation study. For each of three sample sizes, four jittered dot plots are shown corresponding (from left to right): $\hat{h}_0$, $\hat{h}$, h_0, h_{MISE}. See text for details.

obtained. Figure 10 shows a more detailed view of the two estimation procedures. Shown in this figure are kernel estimates of $\log_{10}(\hat{f}(2)/f(2))$ for the two data-based bandwidth estimation schemes. The solid line indicates the distribution of the ratio using $\hat{h}_0(x)$ and the dashed line indicates the distribution using $\hat{h}_{\text{MISE}}$. While the variability of the two procedures seem similar, it is clear that the estimation scheme targeting the zero-bias bandwidths has better bias properties than the global procedure and appears to yield estimates that are superior to the global procedure.

While it may not be possible to achieve the $O(n^{-1})$ behavior in practice, these zero-bias bandwidths still hold great promise for estimation in the tails where fixed bandwidth methods have traditionally had difficulties. To demonstrate behavior of the estimation scheme outlined at the beginning of the section with different types of tail behavior, five samples were drawn from the standard normal distribution ($n = 10000$) and the beta(2,6) distribution ($n = 10000$). Bandwidth functions were estimated using the cubic spline in a

product term in (15) is estimated by -2 times the average of the leave-one-out estimator,

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

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

$$\mathrm{UCV}_{ab} = \int_a^b \hat{f}(x|h_x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i|h_{x_i}) I(a \leq x_i \leq b). \tag{16}$$

It is interesting to note that the estimated cross-product term is the same whether the sample point estimator with $g(x_i)$ or the balloon estimator with $h(x)$ is used.

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 (although other schemes for choosing these end conditions can be used). Knots are placed along the x-axis in the region of interest. The cross-validation criterion given in (16) is then minimized over the heights of the bandwidth function at these knot values.

A small simulation study was performed in an attempt to examine convergence properties of the estimation scheme outlined above. One hundred samples of various sizes were drawn from a standard normal distribution with the intent to estimate the density at a single point, namely $x = 2$. The exact zero-bias bandwidth at $x = 2$ was used as well as the estimated zero-bias bandwidth obtained using the local cross-validation procedure outlined above. However, a linear spline approximation to the logarithm of the bandwidth function was used, and the interval of interest was chosen to be $[1.5, 2.5]$. A global approach using the exact minimizer of the MISE, h_{MISE}, as well as an least-squares cross-validation estimate of h_{MISE} were also used for comparison.

Plotted in Figure 9 are the MSEs at $x = 2$ for h_0 and h_{MISE} as n increases. These are indicated by the solid lines and represent the convergence rates of n^{-1} and $n^{-4/5}$. Also plotted are $(\hat{f}(2) - \phi(2))^2$ for each of four cases at $n = 1000$, 3162, and 10000. Circles are used for estimates obtained using h_0 while squares are used for estimates obtained using h_{MISE}. Open shapes correspond to theoretical results while filled shapes correspond to results with estimated bandwidths.

It is clear from the figure that despite the $O(n^{-1})$ MSE convergence rate achieved by using $h_0(x)$, the same is not true when using our estimate of $h_0(x)$. However, some gain is

and the fixed kernel method. However, the gains in the tails are outstanding as can clearly be seen from the second column of plots focusing on the tails. One drawback to this global approach is that the estimates do not integrate to one but it certainly shows promise when estimating the density in convex regions of f, namely in the tails of f.

4. A CROSS-VALIDATION ALGORITHM

In spite the common use of the balloon estimator in practice, the fact the estimator does not integrate to one makes it somewhat less attractive as a global estimator, especially since recent study of the sample point estimator has yielded promising algorithms for choosing bandwidths (see, for example, Fan et al., 1993 and Sain and Scott, 1996). However, the results of the previous section suggest that an optimal estimation scheme should incorporate the zero-bias bandwidths whenever possible and that algorithms for estimation of the zero-bias bandwidths over intervals in the convex regions of the underlying density should be investigated. Here, we extend the work of Hall and Schucany (1989) and use a device introduced by Fan et al. (1993) and Sain (1994) and consider a local least-squares cross-validation algorithm to estimate $h_0(x)$ in such regions.

Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (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 $h_0(x)$.

The modification considered here is to replace the constant bandwidth used by Hall and Schucany (1989) by a bandwidth function, $h_x = h(x)$ defined over the interval of interest. Consider the integrated squared error on some closed interval $[a, b]$,

$$\begin{aligned} \mathrm{ISE}_{ab} &= \int [\hat{f}(x|h_x) - f(x)]^2 I(a \le x \le b)\, dx = \int_a^b [\hat{f}(x|h_x) - f(x)]^2 dx \\ &= \int_a^b \hat{f}(x|h_x)^2 dx - 2\int_a^b \hat{f}(x|h_x) f(x) dx + \int_a^b f(x)^2 dx. \end{aligned} \tag{15}$$

The final term in (15) is constant and can be ignored. The first term is solely a function of the kernel estimate and can be calculated directly using numerical integration The cross-

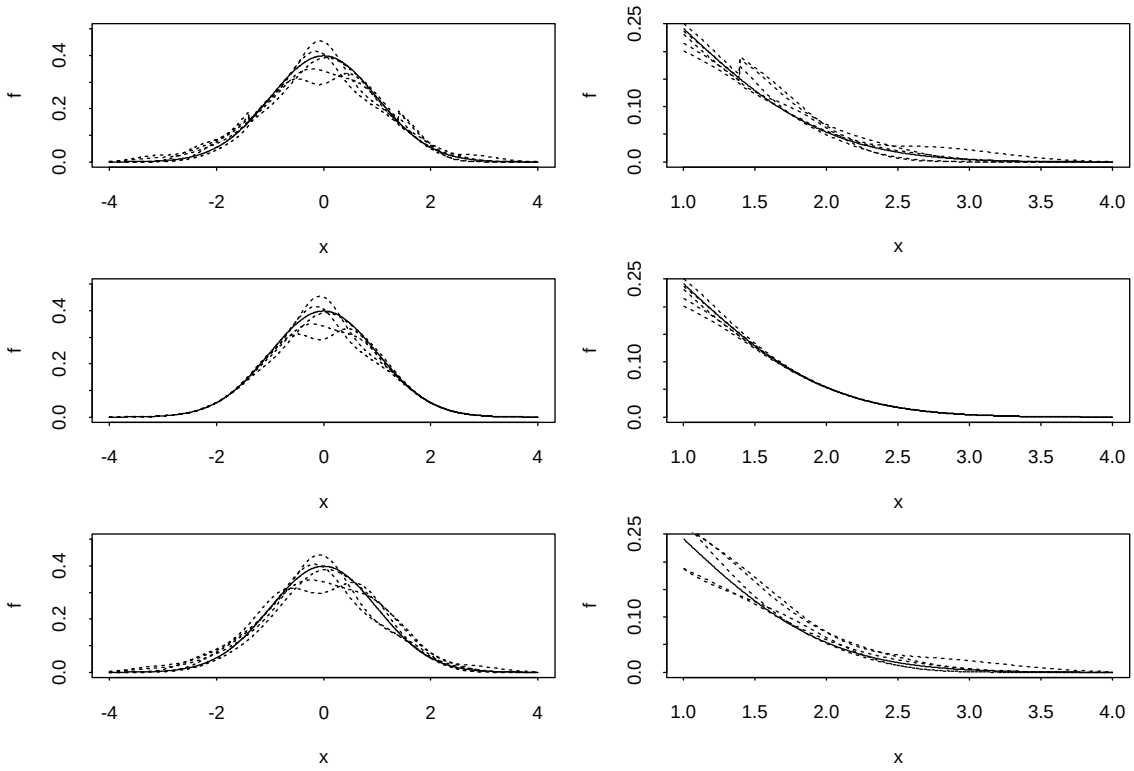

Figure 8: Five estimates with samples of size $n = 100$ from $N(0,1)$. The true f is indicated by the solid line. The estimates $\hat{f}$ from top to bottom used h_x^* (first local minimum), "best path" h_x using $h_0(x)$ in the tails, and h fixed. The second column shows a blowup of the right tail.

bution. The bottom plot shows fixed kernel estimates using $h = 0.445$, the optimal MISE bandwidth. In the top plot, the first local minimum (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 minimum of the MSE; see Figure 1 and 6. Although this strategy might yield gains in the MSE (and MISE) over the optimal fixed bandwidth strategy, visually there is not much difference. In fact, the estimates are less appealing visually due to the jumps introduced.

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 the h_x that corresponds to the global minimum of the MSE at each estimation point. Note that in the neighborhood of the mode, there is not any difference between this estimator

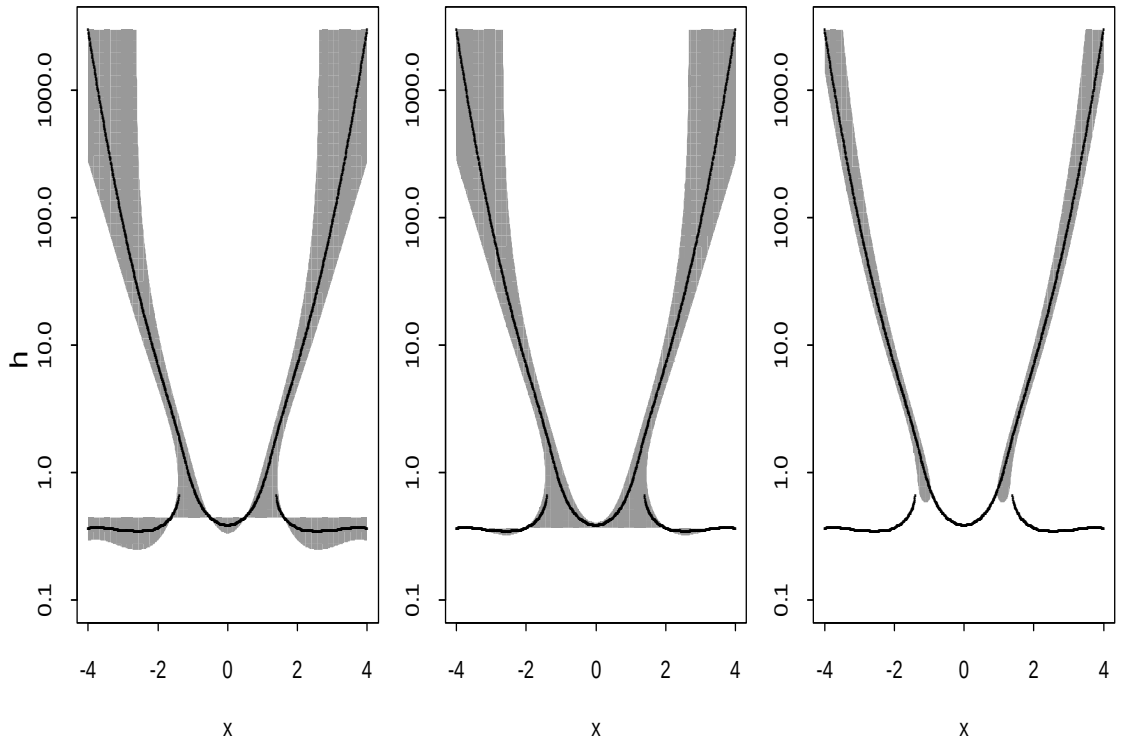

Figure 7: Pointwise local minima of the MSE for $N(0,1)$ data with $n = 100$. Shaded regions represent bandwidths that yield estimates with MSE at least as small as: (a) the optimal MISE fixed bandwidth approach, (b) a local approach using the exact local minimum of the MSE (h_x^*), and (c) direct parametric estimation.

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 minimum in the MSE (corresponding to h_x^*) or the larger local minimum in the MSE (corresponding to $h_0(x)$). 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. Figure 7(b) shows that we can, in fact, do better as evidenced by the large shaded regions surrounding the local minima in the MSE corresponding to the zero-bias estimates. Finally, part (c) of Figure 7 shows the amusing result that the zero-bias bandwidths can achieve a lower pointwise MSE in the tails than the maximum-likelihood parametric estimator, $\phi_s(x - \bar{x})$.

Figure 8 displays estimates of 5 random samples of size $n = 100$ from a $N(0,1)$ distri-

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 (10), completing the proof. □

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, $\delta \geq 0$. To see this consider a bandwidth slightly larger than h_0. This bandwidth will have a smaller variance than h_0, and the bias will still be negligible (at least when compared to the variance). Hence, the MSE will also be smaller.

As a result of Theorem 2.2, the convergence rate of $\text{MSE}[\hat{f}(x|h_{mse})]$, 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 represents a possible breakthrough in an important practical problem.

3. A COMPARISON OF OPTIMAL ESTIMATION STRATEGIES

In this section, optimal 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 7 once again shows the local minima (bold lines) of the MSE for a random sample of size 100 of $N(0,1)$ data. The shaded region in $7(a)$ represents the range of values of the bandwidth that will yield pointwise MSE smaller than when the bandwidth

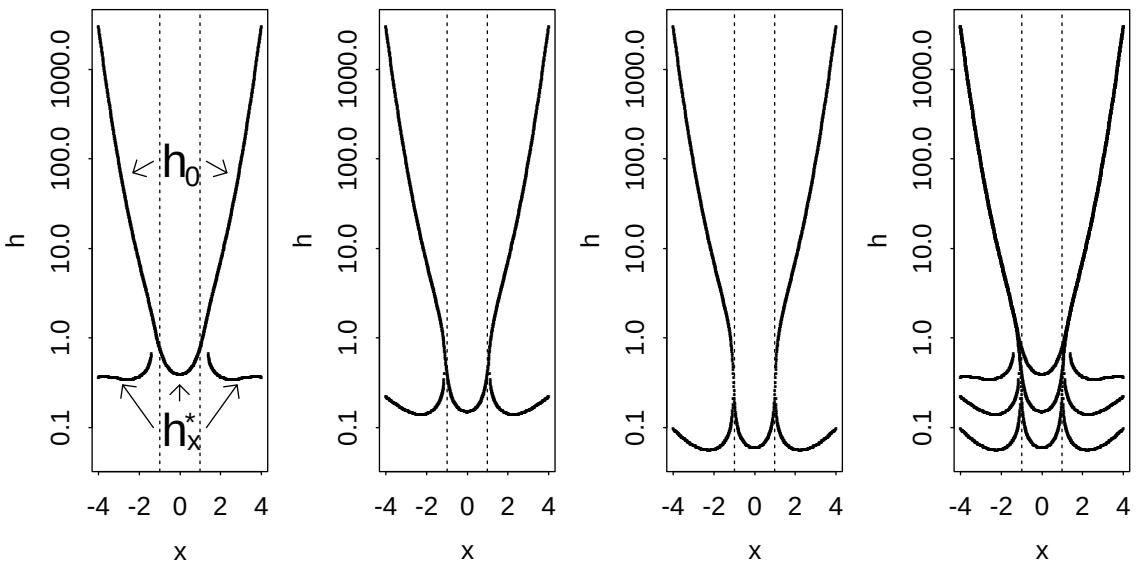

Figure 6: All locally optimal bandwidths for the pointwise local minima of the MSE for $N(0,1)$ data ($n = 10^2, 10^4, 10^6$, and all three combined).

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

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

where b again is given in (12). Expanding $f(x - h_0 w - \delta w)$ as before and ignoring the higher order terms in δ, the asymptotic variance of $\hat{f}(x|h)$ is given by

$$\mathrm{Var}[\hat{f}(x|h)] = \frac{a_1 - \delta a_2}{n(h_0+\delta)} - \frac{f(x)^2}{n} + \frac{2\delta b f(x)}{n}, \tag{14}$$

where a_1 and a_2 are also given in (11). Combining (13) and (14), the MSE of $\hat{f}(x|h)$ is

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

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

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

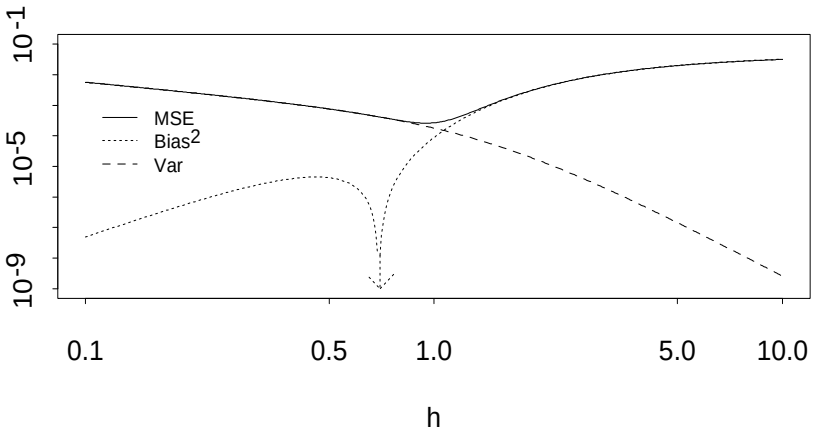

Figure 5: Bandwidth vs MSE for estimation at $x = 1.1$ where $f(x)$ is $N(0,1)$ and $n = 100$.

with

$$a_1 = \int K^2(w) f(x - h_0(x)w)\, dw, \qquad a_2 = \int wK^2(w) f'(x - h_0(x)w)\, dw, \tag{11}$$

and

$$b \;=\; \int wK(w) f'(x - h_0(x)w)\, dw. \tag{12}$$

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

$$\begin{aligned} \text{Bias}[\hat{f}(x|h)] &= E[K_h(x - X)] - f(x) \\ &= \int K(w) f(x - hw)\, dw - f(x) \\ &= \int K(w) f(x - h_0 w - \delta w)\, dw - f(x). \end{aligned}$$

Expanding $f(x - h_0 w - \delta w)$ in a Taylor's series about $x - h_0 w$,

$$E[\hat{f}(x|h)] = \int K(w)[f(x - h_0 w) - \delta w f'(x - h_0 w) + \cdots]\, dw.$$

Note that $\int K(w) f(x - h_0 w) dw = E[\hat{f}(x|h_0)] = f(x)$ 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

$$\text{Bias}[\hat{f}(x|h)] \;=\; -\delta b, \tag{13}$$

where b is given in (12).

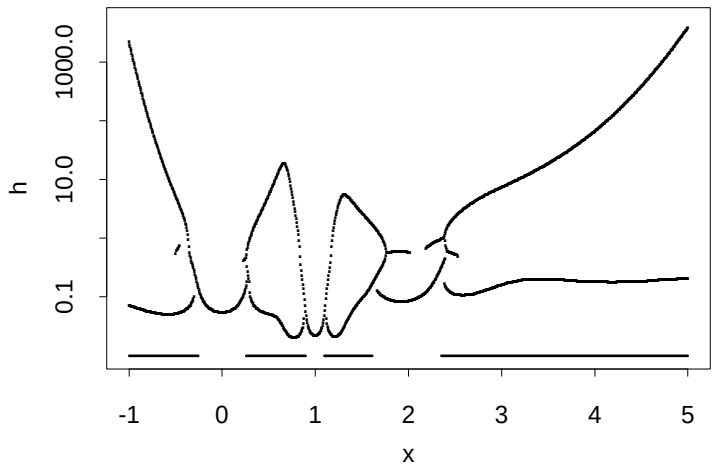

Figure 4: Pointwise local minima of the MSE for the normal mixture density ($n = 5000$). The horizontal line segments just above the x-axis indicate regions where f is convex. (Note that the dotted lines should be connected; each dot represents a numerical search for a root.)

5 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 $h = 0.7$.

For the standard Normal density, Figure 6 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. Beyond the inflection points ($x = \pm 1$), there are two "optimal" bandwidths. Note that the greater bandwidths in the tails are considerably larger than any previously considered in the literature. The rightmost plot contains an overlay of the first three plots and demonstrates that these larger optimal bandwidths in the MSE do not vary with n (nor do they converge to zero as do those corresponding to $h^*(x)$). In fact, as the following theorem shows, these larger local minima in the MSE converge, quite rapidly, to the zero-bias bandwidths.

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 minimum of the MSE$[\hat{f}(x)]$ in the neighborhood of $h_0(x)$. Then $h_{mse}(x) - h_0(x) = O(n^{-1})$. Writing $h_{mse}(x) = h_0(x) + \delta_{x,n}$, then*

$$\delta_{x,n} = \left[\frac{a_1 + h_0(x) a_2 - 2b h_0(x) f(x)}{2b^2 h_0(x)^2} \right] n^{-1}, \tag{10}$$

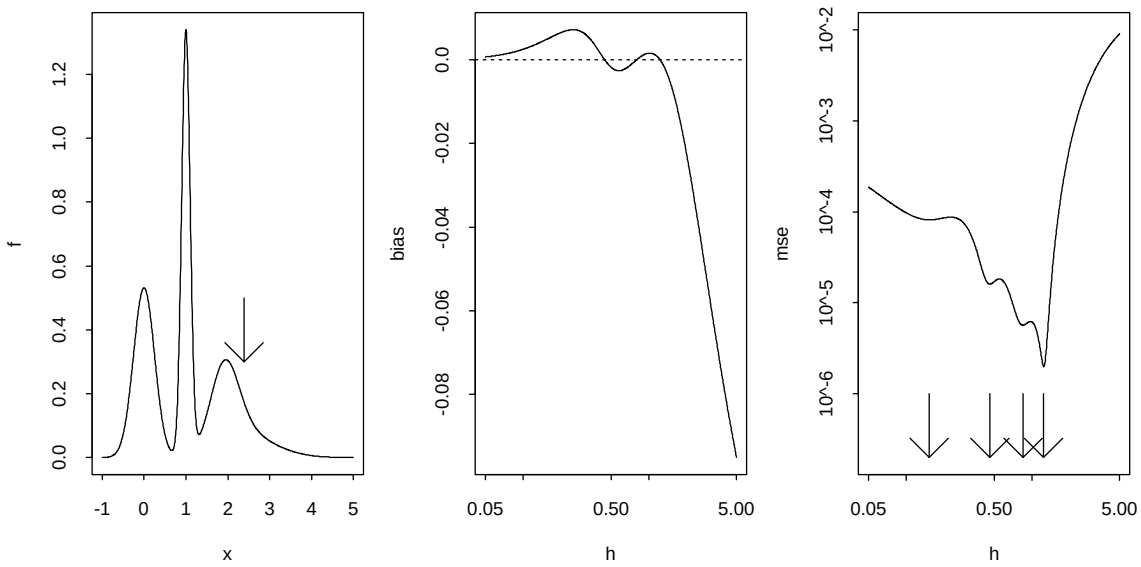

Figure 3: Exact bias and MSE curves at $x = 2.395$ for a Normal mixture with $n = 5000$.

to h_x^*.

Figure 4 displays all of the locally optimal bandwidths of the pointwise MSE for the normal mixture shown in Figure 3. In a neighborhood of each peak, there is one local minimum while in the tails and valleys there are as many as four optimal bandwidths.

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 the Normal mixture density displayed in Figure 3.

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 increases further, more valleys and peaks are introduced into the convolution computation. It is conceivable that a density with m sufficiently separated modes could have up to $2m - 1$ zero-bias bandwidths for some x.

In general, $h^*(x)$ and $h_0(x)$ will both result in a local minimum in the MSE as was seen in Figures 2 and 3. However, in some small-sample cases, $h^*(x)$ and $h_0(x)$ may not be sufficiently different to give two minima in the MSE curve. An example is shown in Figure

it is clear that $\text{Bias}[\hat{f}(x|h)] < 0$ at the largest mode for all $h > 0$. Now, $\int \text{Bias}[\hat{f}(x|h)]\,dx = 0$ 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 as does the theorem.

Remark 3. Alternatively, for general kernel, a sufficient condition on the existence of a zero-bias bandwidth 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, by the argument given in the above proof, there must exist a bandwidth yielding zero bias.

Remark 4. The argument in the previous remark extends to higher-order kernels as well as to the multivariate case. 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 $\mu_p = \int x^p K(x)\,dx$. Hence, for small h, the bias will be positive at x where $\mu_p f^{(p)}(x) > 0$, and the same argument for existence of a zero-bias bandwidth holds. Finally, consider a bivariate product kernel estimator with single smoothing parameter, h. In that case, $\text{Bias}[\hat{f}(\mathbf{x}|h)] = \frac{1}{2}h^2\sigma_K^2[f_{11}(\mathbf{x}) + f_{22}(\mathbf{x})]$, where $\mathbf{x} = (x_1, x_2)^T$ and $f_{ii} = \partial^2 f/\partial x_i^2$. Again, as long as $f_{11}(\mathbf{x}) + f_{22}(\mathbf{x}) > 0$, a zero-bias bandwidth exists for some h at $\mathbf{x}$. It is interesting to note that for the $N(\mathbf{0}, \mathbf{I}_2)$ density, such points exist when $x_1^2 + x_2^2 > 2$.

Remark 5. The existence of the zero-bias bandwidth is not entirely unknown in the literature (see Scott, 1992, p. 137; Hazelton, 1996). However, it is commonly held to be a theoretical curiosity and of no practical consequence. We now believe otherwise.

Remark 6. We give some examples of regions where h_0 exists: $|x - \mu| > \sigma$ for a normal density; $x > 0$ for an exponential density; and $|x| > 1/\sqrt{3}$ for a standard Cauchy density.

In the case of a multimodal density, the theorem still holds, but h_0 (where 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 $x = 2.395$, f is convex, but there are three bandwidths ($h = 0.464, 0.849, 1.24$) that have zero bias with corresponding local minimum 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 minimum ($h = 0.153$) corresponding

en 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) as a function of h. Suppose h is small enough so that f is convex over the interval $(x-h, x+h)$. As in (8) and by an application of Jensen's inequality,

$$\begin{aligned} E[\hat{f}(x|h)] = E[K_h(x - \mathbf{X})] &= \int_{-\infty}^{\infty} K_h(x-y) f(y)\, dy \\ &= \int_{x-h}^{x+h} K_h(x-y) f(y)\, dy \\ &\geq f\left(\int_{x-h}^{x+h} y\, K_h(x-y)\, dy \right) \\ &= f(x). \end{aligned} \tag{9}$$

Therefore, Bias$[\hat{f}(x|h)] \geq 0$. Clearly from (9), $E[\hat{f}(x|h)] = K_h * f(x)$, 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 Bias$[\hat{f}(x|h)] \to 0$. Similarly, as $h \to \infty$, $K_h \to 0$ everywhere; hence, the expectation vanishes, and the Bias$[\hat{f}(x|h)] \to -f(x) < 0$. Now, the general behavior of the bias as a function of h has been established. Since the bias is a continuous function of h, and 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 derive $h^*(x)$ 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. A heuristic argument which can be made rigorous can be given for a general kernel. Writing the bias again as a convolution, Bias$[\hat{f}(x|h)] = (K_h * f - f)(x)$,

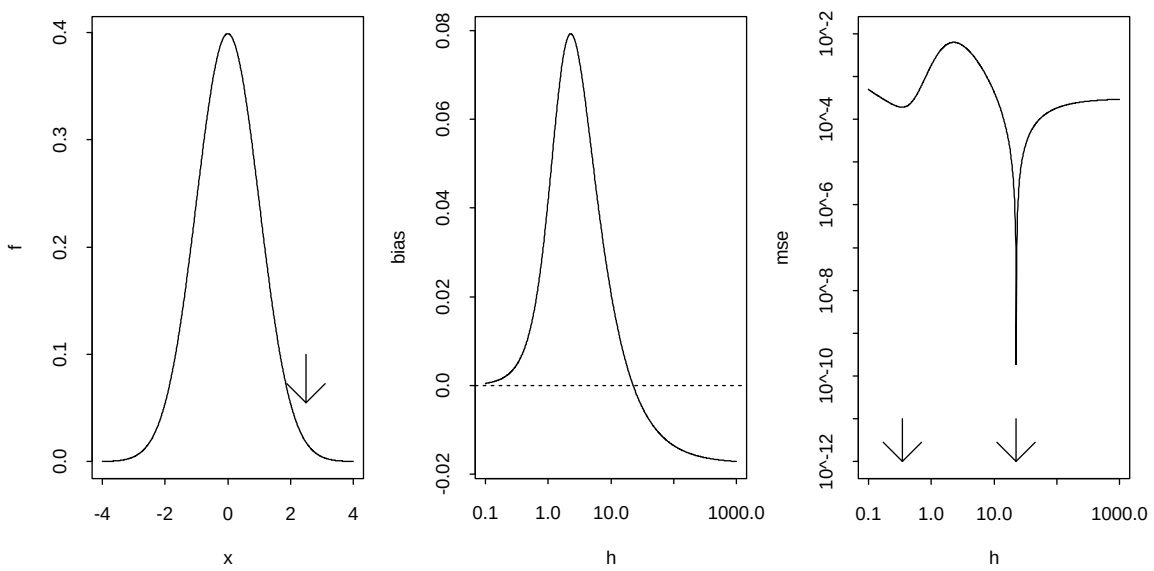

Figure 2: Exact bias and MSE curves for estimation at $x = 2.5$ when $f(x)$ and $K(x)$ are $N(0,1)$ with $n = 100$.

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 $f(x)$ and $K(x)$ are $N(0,1)$ with $n = 100$. Note that the bias does in fact equal zero at $h_0 = 22.6$. Furthermore, there is a bandwidth corresponding to a global minimum of the MSE that is slightly larger than h_0 (see Theorem 2.2). The local minimum 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 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 $h = 2.0$. 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 $-f(2.5)$. 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 conditions under which such bandwidths exist.

Theorem 2.1 (Existence of h_0) *Let $\hat{f}(x|h)$ be the locally adaptive density estimator giv-*

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. The exact bias/variance expressions

$$\begin{aligned} \text{Bias}[\hat{f}(x|h_x)] &= \mu_1(x|h_x) - f(x) & (6)\\ \text{Var}[\hat{f}(x|h_x)] &= \frac{1}{n}[\mu_2'(x|h_x) - \mu_1(x|h_x)^2], & (7) \end{aligned}$$

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

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

and, similarly, the (noncentral) second moment

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

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

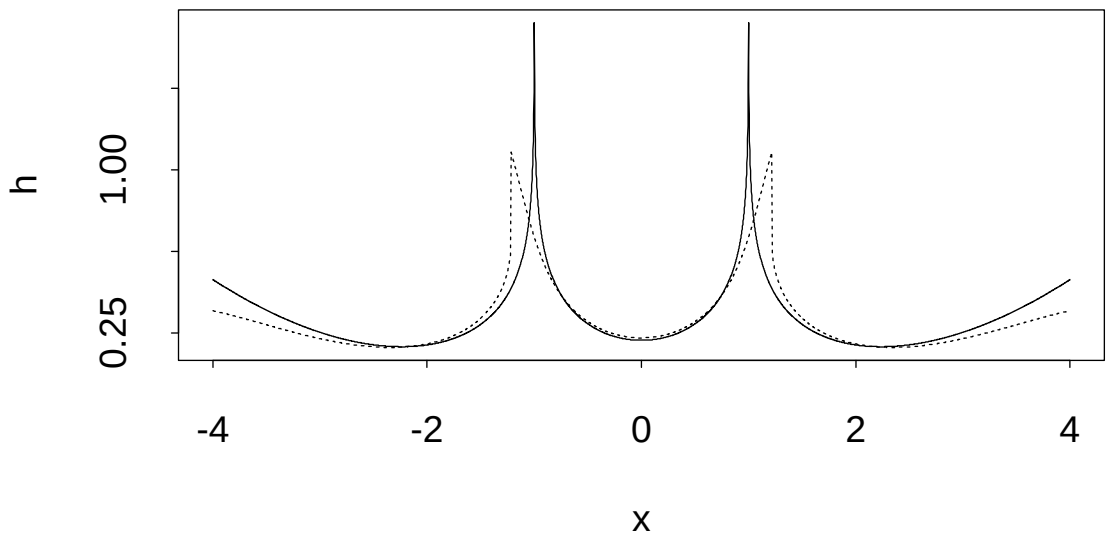

Figure 1: The asymptotically optimal local bandwidth function, $h^*(x)$, for $N(0,1)$ kernel and data ($n = 1000$). The dashed line shows the local bandwidth function that minimizes the exact mean squared error point-by-point.

displayed in this figure is the exact MSE minimizing bandwidth curve, which differs from $h^*(x)$ in two interesting ways. The large bandwidths predicted at the inflection points $x = \pm 1$ actually occur at larger values of $|x|$ for finite n. Of more interest is that the wider bandwidths expected in the tails are not much different than the optimal bandwidth at the mode. Our result below is that much wider bandwidths in the tail can make sense.

This work describes the surprising theoretical result that at points where the unknown density is convex a bandwidth can be found yielding exactly zero bias pointwise, and that these so-called "zero-bias bandwidths" vastly improve upon $h^*(x)$. Furthermore, an estimation strategy utilizing these zero-bias bandwidths is shown theoretically to be superior over the fixed bandwidth estimator as well as the simple strategy utilized by Jones (1990). Finally, a practical cross-validation procedure based on the work of Hall and Schucany (1989) and extending that of Fan et al. (1993) and Sain (1994) is shown to be promising for estimating these zero-bias bandwidths.

Note that the bandwidth, $g(x_i)$, is now written explicitly as a function of the sample point.

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 widely studied 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 is of interest because the pointwise bias is 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 (1995), and Sain and Scott (1996) discuss implementation problems as well as poor global performance in certain situations.

Recent research by Fan et al. (1993), Sain (1994), and Sain and Scott (1996) has focused on the practical issues associated with estimating optimal bandwidths for the general adaptive estimator. Related theory and implementation of local-likelihood-based nonparametric density estimation has been described by Loader (1996) and Hjort and Jones (1996).

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),

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

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

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 of kernel estimators include Silverman (1986), 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 a higher-order bias. A kernel is defined to be of order p 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 point. The form of the *sample point adaptive estimator* is

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

Zero-Bias Locally Adaptive Density Estimators

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

September 5, 1997

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 parametric rate of $O(n^{-1})$ rather than the slower $O(n^{-4/5})$. These so-called zero-bias bandwidths are usually orders of magnitude larger than the optimal locally adaptive bandwidths predicted by asymptotic mean squared error analysis. This finding helps explain some of the large bandwidths found by least-squares cross-validation using the sample point kernel estimator (Sain and Scott, *JASA*, 91:1525–1534). We design an algorithm for estimating the density only in an interval where the density is convex (typically, in the tail) and investigate a fully data-based algorithm to see if we can take advantage of the predicted theoretical behavior. We find that our particular estimator and implementation does not attain $O(n^{-1})$ convergence, but does provide some improvement in mean squared error and often clearly superior visually pleasing curves.

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

[1]Stephan R. Sain is Assistant Professor, Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332 USA. David W. Scott is Professor, Department of Statistics MS-138, Rice University, 6100 Main Street, Houston, TX 77005-1892 USA.