The Annals of Statistics
1992, Vol. 20, No. 2, 712–736

EXACT MEAN INTEGRATED SQUARED ERROR[1]

BY J. S. MARRON AND M. P. WAND

University of North Carolina and Rice University

An exact and easily computable expression for the mean integrated squared error (MISE) for the kernel estimator of a general normal mixture density, is given for Gaussian kernels of arbitrary order. This provides a powerful new way of understanding density estimation which complements the usual tools of simulation and asymptotic analysis. The family of normal mixture densities is very flexible and the formulae derived allow simple exact analysis for a wide variety of density shapes. A number of applications of this method giving important new insights into kernel density estimation are presented. Among these is the discovery that the usual asymptotic approximations to the MISE can be quite inaccurate, especially when the underlying density contains substantial fine structure and also strong evidence that the practical importance of higher order kernels is surprisingly small for moderate sample sizes.

1. Introduction. Substantial research has been devoted to kernel density estimation. This is because it provides a simple, yet appealing, context in which to study problems and issues that arise in all types of nonparametric curve estimation. This includes regression, spectral density and hazard estimation, and also a variety of other estimators, including histograms, splines and orthogonal series.

Three important and useful tools for understanding the behavior of nonparametric curve estimators are asymptotic analysis, simulation and numerical calculation of error criteria.

Each of these methods provides many useful insights into the complicated structure present in the study of curve estimation. However, each has its limitations as well.

The strength of asymptotic analysis is that it frequently allows simultaneous study of many different specific examples, through general results applying to entire classes of settings. The weakness of asymptotics is that they only describe behavior in the limit. This is still very useful in many situations because the asymptotics describe the actual situation quite well. However, it is less useful when the asymptotics have not yet kicked in (that is, in studying situations where the asymptotically dominant effect has not taken over yet). Perhaps the biggest drawback to asymptotics is that it is very difficult to determine in a given situation which of these possibilities is occurring.

Received July 1990; revised March 1991.

[1]Research partially supported by the Mathematical Sciences Institute at Cornell University. Research of the first author was also partially supported by NSF Grant DMS-89-02973. Research of the second author was commenced while in the Department of Statistics, Texas A & M University.

AMS 1980 *subject classifications*. Primary 62G05; secondary 65D30.

Key words and phrases. Gaussian-based kernel, integrated squared error, kernel estimator, nonparametric density estimation, normal mixture, window width.

The strength of simulation is that one can clearly understand any setting, most especially those where asymptotics are clearly not appropriate (e.g., small samples). Furthermore, the only level of approximation, the Monte Carlo variability, can be made as small as desired, and also can usually be precisely understood. Despite these important strengths, the weakness of simulation should also be recognized. This is that the lessons are limited to only the set of examples that can be studied (versus the wide classes that can often be analyzed via asymptotic methods). These limits are of practical importance, because very substantial effort is required, in terms of both programming and also CPU time, to do even a moderate scale simulation study.

Numerical calculation of error criteria does something to overcome the time problems entailed in simulation. In short, at least in simple cases, one can look at a much broader base of examples with the same amount of research effort. However, numerical methods have the weakness that in complicated problems, they, too, can involve very substantial effort on the part of both the researcher and his equipment, which again limits the number of examples that can be considered. Furthermore, understanding the errors involved in this type of approximation is often much trickier than in simulation.

In this paper a variation of the numerical approximation technique is presented, which we believe is worth separate consideration in its own right. The main idea is exact calculation of error criteria for special classes of examples, which make the calculation tractable, but at the same time represent a broad base of examples. In our opinion, such classes will turn out to be rather generally available, especially in curve estimation. The strength of this approach to problems is that many more examples can be considered than when the same effort is devoted to simulation (most especially large samples require no additional CPU time) or to complicated numerical work. An additional payoff is that there are no errors to be concerned with, no negligible term as found in asymptotics, no Monte Carlo variability as in simulation, no numerical error as in that type of approximation. Of course, it still needs to be kept in mind that this method has the weakness of being limited only to studying individual examples.

Most of the paper involves application of this idea to studying simple kernel density estimation. The points made above are borne out by the fact that the paper studies many more different aspects of this topic than do most other papers. Furthermore, we see many other areas where one can gain much through analysis by this method, including important variations on the simple density estimation considered here such as the multivariate case and derivatives. Also, there is much to be gained from the study of related curve estimation problems, including regression estimation.

Fryer (1976) and Deheuvels (1977) first showed that the mean integrated squared error (MISE) could be calculated exactly when both the underlying density and the kernel function are Gaussian. This is because the MISE can be written in terms of convolutions, which are simply evaluated in the Gaussian case.

In Section 2 of this paper, this idea is extended in two directions. First, we note that the simple convolution property still holds when the underlying

density is a normal mixture. Second, we allow the kernel to be of arbitrary order. This is accomplished by use of the Gaussian-based kernels of Wand and Schucany (1990), which retain the convolution property.

Exact MISE calculations for the Fourier integral density estimator were performed by Davis (1981) and Hart (1984). Also, see Gasser, Müller and Mammitzsch (1985) for a different but related type of exact calculation in nonparametric regression.

In Section 3, examples are given which show that the class of normal mixture densities is a very broad one, allowing easy study of many different types of problems that arise in density estimation. These same specific examples are also of interest because they provide an interesting test set for simulation study of data-based window width selection. Simple evaluation of the performance of these methods provides an additional application of our exact MISE formulae.

The error inherent to the usual asymptotic analysis of the MISE is studied in Section 4. An especially surprising aspect of this for us is that the MISE can have local minima. The closely related problem of how the MISE optimal window width relates to its usual asymptotic minimizer is studied in Section 5. It is seen in these sections that many of the key ideas often still hold up, but some of them may require prohibitively large sample sizes before the asymptotic effects provide a reasonable description of the actual situation.

The practical effectiveness of higher order kernels is investigated in Section 6. While these are known to be always superior in the limit (if enough smoothness is assumed on the underlying density), they are not used much in practice, because they lose some of the intuitive appeal of the nonnegative kernels. It is seen that sample sizes in the hundreds are needed for higher order kernels to be worthwhile for very simple densities, and far more, even into the millions are required for densities with more complicated structure. Proofs of our results are in Section 7.

The approach taken throughout this paper concerns estimation of the entire curve, which is what we feel is most relevant for density estimation. However, the basic ideas are easily adaptable to estimating a density at a point, through calculation of the pointwise mean squared error. An extension which does not appear to be simple is to the case of other norms, such as L_1 or L_∞. For these we know of no analogs of our methods, because we do not know how to express them simply in terms of convolutions.

2. Exact MISE for normal mixture densities. Let $X_1, \ldots, X_n$ be a set of independent $\mathbb{R}$-valued random variables each having density f. For a given window width h, the kernel estimator of $f(x)$ is given by

$$f_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{2.1}$$

where $K_h(u) = h^{-1}K(u/h)$ and K is a symmetric function satisfying $\int K = 1$

and

$$\int x^j K(x)\,dx = \begin{cases} 0, & j = 1,\ldots,p-1, \\ \mu_p(K) \neq 0, & j = p, \end{cases}$$

where this defines $\mu_j(K)$. The integer p is called the order of the kernel. Note that the symmetry of K implies that p is an even integer. The integer $p/2$ will be denoted throughout by r.

The MISE of $f_n(\cdot; h)$ is given by

$$\text{MISE}(h) = E\int \{f_n(x; h) - f(x)\}^2\,dx,$$

however, simple manipulation leads to

$$\text{MISE}(h) = n^{-1}h^{-1}\int K^2 + (1 - n^{-1})\int (K_h * f)^2 - 2\int (K_h * f) f + \int f^2. \tag{2.2}$$

Fryer (1976) and Deheuvels (1977) observed that if f is normal and K is the Gaussian (standard normal) kernel, then closed form expressions are available for these convolutions and their integrals and exact MISE calculations are possible. However, this result applies to only one special case of the density estimation problem, so its use is rather limited. In this work we exploit the tractability of the normal density much further by taking f to be a general normal mixture density and K to be a pth-order Gaussian-based kernel, both of which we now define. First, let ϕ denote the standard normal density and put $\phi_\sigma(x) = \sigma^{-1}\phi(x/\sigma)$. Let $(w_1, \ldots, w_k)$ be a vector with positive entries summing to unity and set

$$f(x) = \sum_{l=1}^{k} w_l \phi_{\sigma_l}(x - \mu_l), \tag{2.3}$$

where $-\infty < \mu_l < \infty$ and $\sigma_l > 0$ for $l = 1, \ldots, k$. We will say that f has a normal k-mixture density with parameters $\{(w_l, \mu_l, \sigma_l^2)\colon\ l = 1, \ldots, k\}$. The Gaussian-based kernel of order $p = 2r$ studied here is

$$G_{2r}(x) = \frac{(-1)^r \phi^{(2r-1)}(x)}{2^{r-1}(r-1)!x} = \sum_{s=0}^{r-1} \frac{(-1)^s}{2^s s!}\phi^{(2s)}(x) \tag{2.4}$$

and can be viewed as the natural extension of the Gaussian second-order kernel to higher-order kernels [Wand and Schucany (1990)]. The second representation of G_{2r} is used extensively in our calculations and follows from the first via the recurrence formula for Hermite polynomials.

In what follows, the notation for derivatives of scaled versions of ϕ will be

$$\phi_\sigma^{(s)}(x) = (d^s/dx^s)\phi_\sigma(x) = \sigma^{-(s+1)}\phi^{(s)}(x/\sigma).$$

For a given value of r the constant $\mathscr{C}_1(r)$ is defined by

$$\mathscr{C}_1(r) = \frac{1}{\pi^{1/2}} \sum_{s=0}^{r-1} \sum_{s'=0}^{r-1} \frac{(2s + 2s')!}{2^{3s+3s'+1} s!s'!(s + s')!}.$$

We then have:

THEOREM 2.1. *Let f be the normal mixture density defined by* (2.3) *and K be the $(2r)$th-order Gaussian-based kernel defined by* (2.4). *Then*

$$\text{MISE}(h) = \frac{\mathscr{C}_1(r)}{nh} + \left(1 - \frac{1}{n}\right) \sum_{s=0}^{r-1} \sum_{s'=0}^{r-1} \frac{(-1)^{s+s'}}{2^{s+s'} s!s'!} \mathscr{U}(h; s + s', 2)$$

$$- 2 \sum_{s=0}^{r-1} \frac{(-1)^s}{2^s s!} \mathscr{U}(h; s, 1) + \mathscr{U}(h; 0, 0),$$

where

$$\mathscr{U}(h; s, q) = \sum_{l=1}^{k} \sum_{l'=1}^{k} w_l w_{l'} h^{2s} \phi_{ll'q}^{(2s)}(\mu_l - \mu_{l'})$$

and $\sigma_{ll'q} = (\sigma_l^2 + \sigma_{l'}^2 + qh^2)^{1/2}$.

The proof of this result is given in Section 7.

3. Examples of normal mixture densities. One way of seeing that the class of normal mixture densities is a very broad one comes from the fact that any density can be approximated arbitrarily closely in various senses by a normal mixture. This idea is made visually clear in Figure 1. These fifteen densities have been carefully chosen because they typify many different types of challenges to curve estimators. The first five represent different types of problems that can arise for unimodal densities. The rest of the densities are multimodal. Densities number 6 to 9 are mildly multimodal and one might hope to be able to estimate them fairly well with a data set of moderate size. The remaining densities are strongly multimodal and will be very hard to recover in full with a moderate sample size, but still are well worth studying, because the issue of just how much of them can be recovered is an important one. We believe these densities effectively model many real data situations.

The Gaussian density, #1, occupies a special place in curve estimation, because it is close to being the easiest possible density to estimate, in a sense discovered by Terrell and Scott (1985). The skewed unimodal density, #2, is not far from the Gaussian in shape, being only mildly skewed and was chosen to resemble the extreme value density in appearance.

The next three densities are much farther from the Gaussian. The strongly skewed density, #3, departs in the direction of skewness and was chosen to

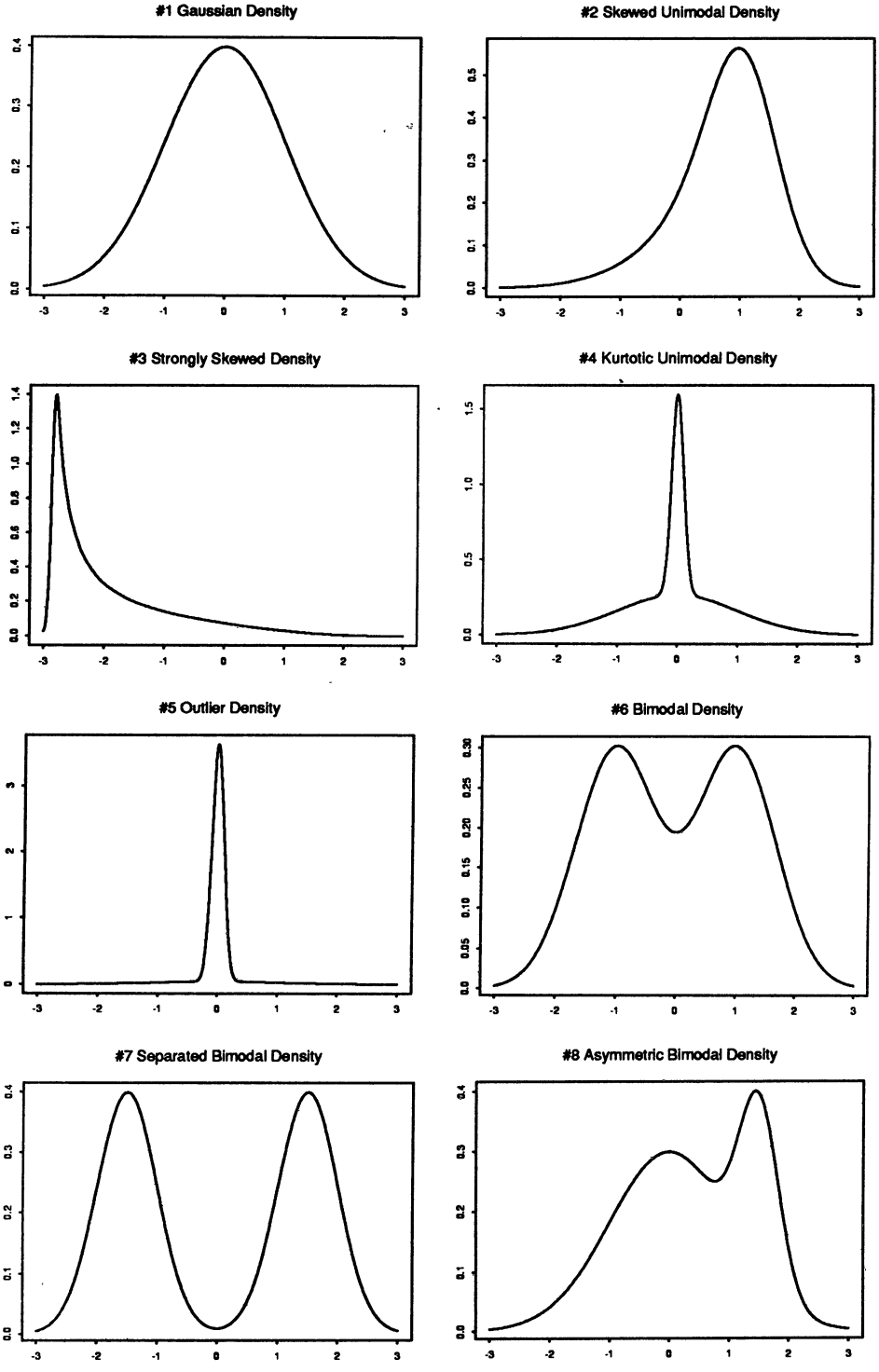


FIG. 1. *Normal mixture densities.*

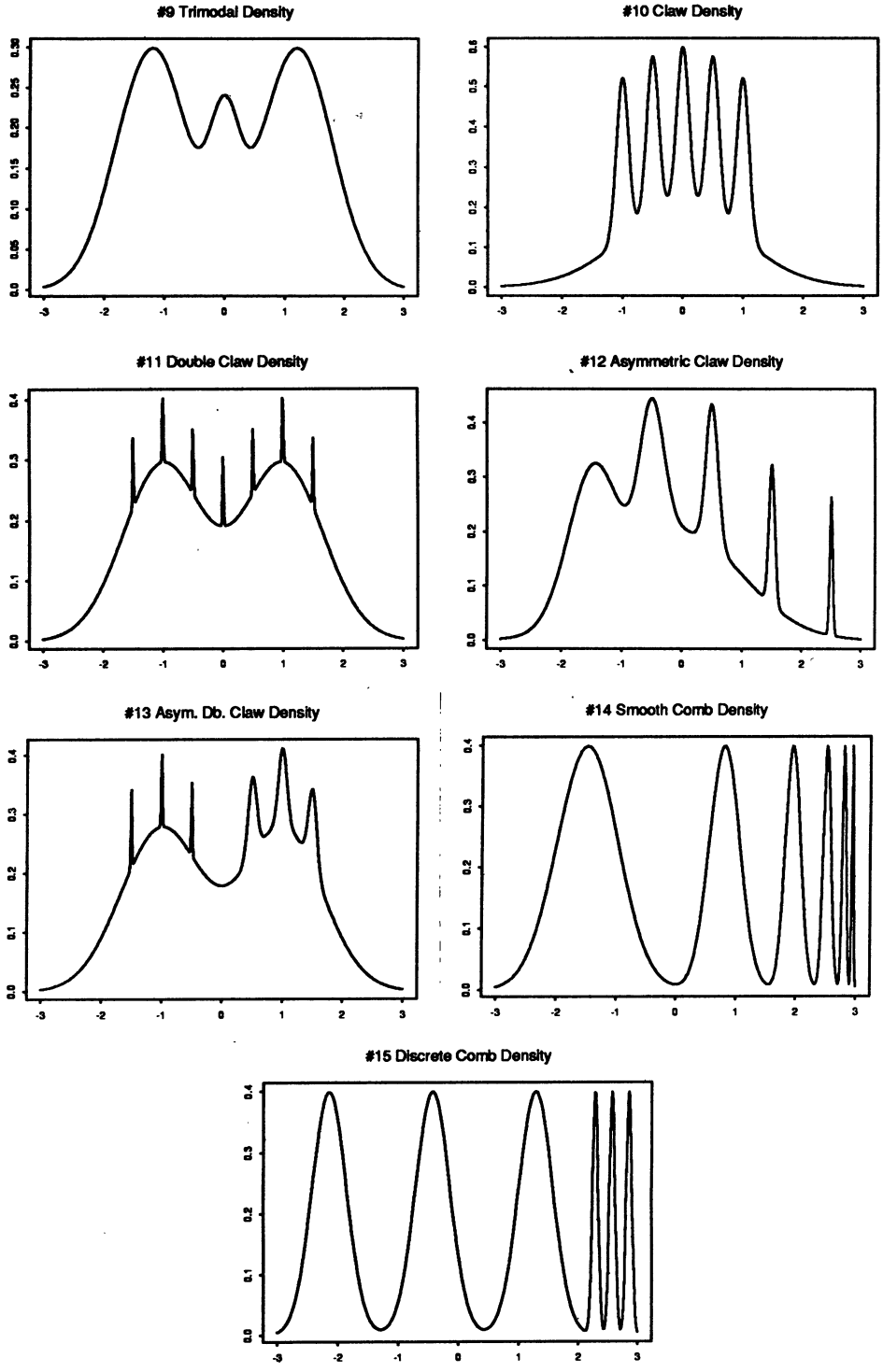


FIG. 1. *Continued.*

resemble to lognormal. The kurtotic unimodal density, #4, is a departure in the direction of heavy kurtosis. This also provides a reasonable approximation to a density with a couple of discontinuities, because of the very sharp rises where the peak meets the two shoulders. The outlier density, #5, has a shape similar to the Gaussian (although it appears different because of the scale chosen), except that 10% of the observations are multiplied by 10, that is, are strong outliers.

The bimodal densities, #6, #7 and #8 and the trimodal density, #9, represent important but simple departures from the unimodal. The claw density, #10, is of special interest because this is where the surprising result of local minima in the MISE occurs. The double claw density, #11, is essentially the same as #6, except that approximately 2% of the probability mass appears in the spikes. Hence for small sample sizes, there is essentially no practical difference between these, although the asymptotic approximations are far different. Also of interest here is to study the point at which there are enough data that the spikes are of practical importance.

The asymmetric claw and double claw densities, #12 and #13, are modifications of #10 and #11, respectively. The smooth and discrete comb densities, #14 and #15, are enhancements of the basic idea of #7. Both of these are shown here because they have much different Fourier transform properties, since #15 has two strong periodic components, while #14 has essentially no periodic tendencies. Also for any given sample size, some of the peaks in #14 can be well recovered, some only marginally recovered and others not recovered at all. On the other hand, this happens in a more blockwise fashion for #15.

The values for the parameters are given in Table 1. For ease in plotting, these have been chosen so that

$$\min_l(\mu_l - 3\sigma_l) = -3 \quad \text{and} \quad \max_l(\mu_l + 3\sigma_l) = 3.$$

4. Comparison of MISE and its asymptotic representation. While the form of MISE is given at (2.2) may seem simple, as compared to being an $n + 1$-fold integral as one might at first expect (and we do indeed seem to have for the L_1 or L_∞ norms), it is still sufficiently complicated that important features of the estimation problem are not plainly visible. For example, as demonstrated by Figures 2.4 and 2.5 in Silverman (1986), the window width h is crucial to the performance of the estimator, as the resulting curve estimate is too wiggly when h is too small and too smooth for h large. A simple asymptotic analysis of MISE makes easy understanding of issues such as this available. For f having a continuous pth derivative, the asymptotic mean integrated squared error (AMISE) is given by

$$\text{AMISE}(h) = n^{-1}h^{-1}\int K^2 + h^{2p}\{\mu_p(K)/p!\}^2\int\{f^{(p)}\}^2.$$

Assuming $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, it is true that $\text{MISE}(h) = \text{AMISE}(h) + o(n^{-1}h^{-1} + h^{2p})$ [see Silverman (1986), pages 38–39]. The value

TABLE 1
Parameters for 15 example normal mixture densities

Density	$w_1N(\mu_1,\sigma_1^2) + \cdots + w_kN(\mu_k,\sigma_k^2)$
#1 Gaussian	$N(0,1)$
#2 Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{12},(\frac{5}{9})^2)$
#3 Strongly skewed	$\sum_{l=0}^{7} \frac{1}{8}N(3\{(\frac{2}{3})^l - 1\},(\frac{2}{3})^{2l})$
#4 Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
#5 Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2)$
#6 Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
#7 Separated bimodal	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2) + \frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
#8 Skewed bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$
#9 Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$
#10 Claw	$\frac{1}{2}N(0,1) + \sum_{l=0}^{4} \frac{1}{10}N(l/2 - 1,(\frac{1}{10})^2)$
#11 Double claw	$\frac{49}{100}N(-1,(\frac{2}{3})^2) + \frac{49}{100}N(1,(\frac{2}{3})^2) + \sum_{l=0}^{6} \frac{1}{350}N((l-3)/2,(\frac{1}{100})^2)$
#12 Asymmetric claw	$\frac{1}{2}N(0,1) + \sum_{l=-2}^{2} (2^{1-l}/31)N(l + \frac{1}{2},(2^{-l}/10)^2)$
#13 Asymmetric double claw	$\sum_{l=0}^{1} \frac{46}{100}N(2l - 1,(\frac{2}{3})^2) + \sum_{l=1}^{3} \frac{1}{300}N(-l/2,(\frac{1}{100})^2) + \sum_{l=1}^{3} \frac{7}{300}N(l/2,(\frac{7}{100})^2)$
#14 Smooth comb	$\sum_{l=0}^{5} (2^{5-l}/63)N(\{65 - 96(\frac{1}{2})^l\}/21,(\frac{32}{63})^2/2^{2l})$
#15 Discrete comb	$\sum_{l=0}^{2} \frac{2}{7}N((12l - 15)/7,(\frac{2}{7})^2) + \sum_{l=8}^{10} \frac{1}{21}N(2l/7,(\frac{1}{21})^2)$

of h which minimizes AMISE(h) is

$$h_{\text{AMISE}} = \left[\frac{(p!)^2 \int K^2}{2p\mu_p^2(K)\int\{f^{(p)}\}^2 n}\right]^{1/(2p+1)}$$

which is an approximation to h_{MISE}, the minimizer of MISE(h). The corre-

sponding minimum AMISE is

$$\inf_{h>0} \mathrm{AMISE}(h) = \left(\frac{2p+1}{2p}\right)\left[\frac{2p\mu_p^2(K)\int\{f^{(p)}\}^2\left(\int K^2\right)^{2p}}{(p!)^2 n^{2p}}\right]^{1/(2p+1)}.$$

For comparison between MISE(h) and AMISE(h) and their respective minimizers in the case of normal mixture densities and Gaussian-based kernels, we will need:

THEOREM 4.1. *For f as in Theorem* 2.1, *we have for* $s = 0, 1, \ldots,$

$$\int\{f^{(s)}\}^2 = (-1)^s \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \phi^{(2s)}_{\sigma_{ll'}}(\mu_l - \mu_{l'}),$$

where $\sigma_{ll'} = (\sigma_l^2 + \sigma_{l'}^2)^{1/2}$. *Also for K as in Theorem* 2.1,

$$\mu_p(K) = \int x^{2r} K(x)\,dx = \frac{(-1)^{r+1}(2r)!}{2^r r!} \quad \text{and} \quad \int K^2 = \mathscr{C}_1(r).$$

The required formulae involving AMISE follow from this.

COROLLARY 4.1. *Let f and K be as in Theorem* 2.1. *Define*

$$\mathscr{C}_2(r) = \frac{1}{2^{2r}(r!)^2}\sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \phi^{(4r)}_{\sigma_{ll'}}(\mu_l - \mu_{l'}),$$

where $\sigma_{ll'}$ is as in Theorem 4.1. *For the kernel estimator* (2.1), *we have*

$$\mathrm{AMISE}(h) = n^{-1}h^{-1}\mathscr{C}_1(r) + h^{4r}\mathscr{C}_2(r),$$

$$h_{\mathrm{AMISE}} = \left\{\frac{\mathscr{C}_1(r)}{4r\mathscr{C}_2(r)n}\right\}^{1/(4r+1)}$$

and

$$\inf_{h>0} \mathrm{AMISE}(h) = \left(\frac{4r+1}{4r}\right)\left\{\frac{4r\mathscr{C}_1(r)\mathscr{C}_2(r)^{4r}}{n^{4r}}\right\}^{1/(4r+1)}.$$

The proof of Theorem 4.1 is not given, because it follows very easily from Corollaries 4.1 of 5.2 of Aldershof, Marron, Park and Wand (1991).

We define the integrated variance, the integrated squared bias and their corresponding asymptotic forms by

$$\mathrm{IV}(h) = \int \mathrm{Var}\{f_n(x;h)\}\,dx, \qquad \mathrm{AIV}(h) = n^{-1}h^{-1}\int K^2,$$

$$\mathrm{ISB}(h) = \int\{Ef_n(x;h) - f(x)\}^2\,dx$$

and

$$\mathrm{AISB}(h) = h^{2p}\{\mu_p(K)/p!\}^2\int\{f^{(p)}\}^2.$$

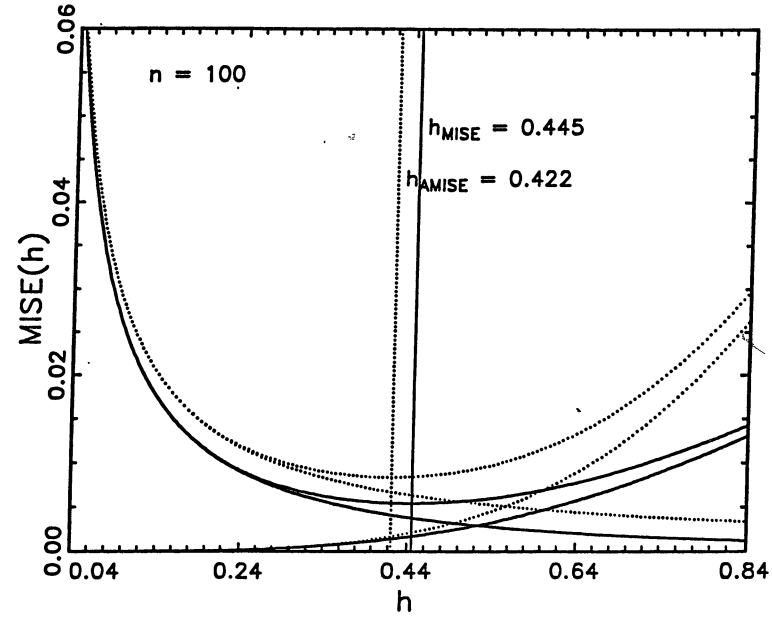
The asymptotic representation AMISE(h) is particularly attractive because it enables easy understanding of the trade-off involved in choice of h. In particular, for h too small, recall $f_n(\cdot; h)$ is then too wiggly, which is quantified by the term AIV(h) being large [the fact that AIV(h) is inversely proportional to the number of observations in a typical window has a definite intuitive appeal]. On the other hand, when h is too large, features of the underlying density are smoothed away by $f_n(\cdot; h)$, which is nicely quantified by the term AISB(h) [in particular note the curvature of f plays an important and intuitively sensible role].

Figures 2a and 2b show pictorially how MISE and AMISE depend on h for the Gaussian density, #1, when the sample size is $n = 100$, for the standard nonnegative Gaussian kernel, which has order $p = 2$. Figure 2a shows h on the ordinary scale, with Figure 2b showing the same thing, but with h on a $\log_{10}$ scale. We feel that the $\log_{10}$ scale is the more informative way to present such pictures (not intuitively surprising, since the window width is a scale factor and works in a multiplicative fashion), so this convention will be followed for the rest of this paper. Note that the log scale is not only more appropriate for making pictures of this type, but also provides a more efficient design when choosing equally space grids of window widths for simulation studies (because the ordinary scale will waste observations by putting essentially too few on the left side and too many on the right).

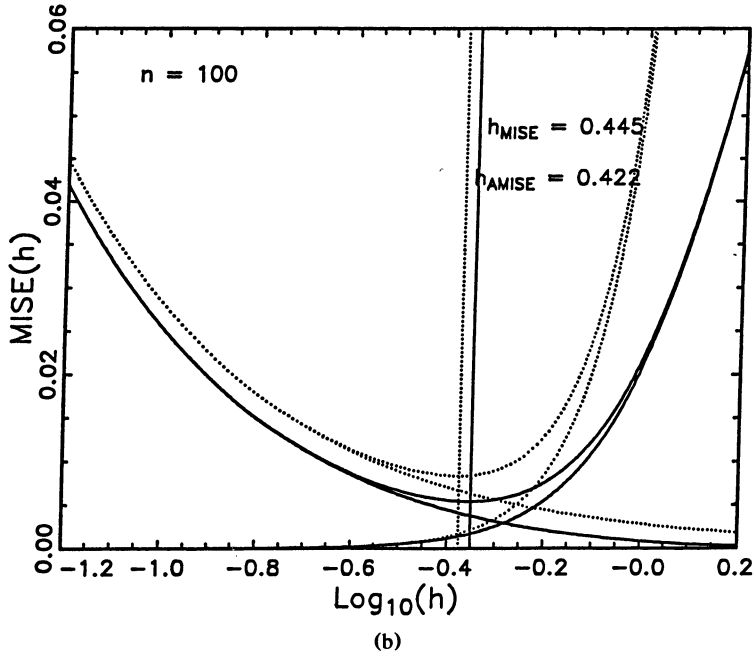
The curves IV(h) and AIV(h) are very large for h small and, as expected, tend to 0 as h grows. Again as intuitively expected, the curves ISB(h) and AISB(h) increase in h, and tend to 0 as $h \to 0$. Note that the IV(h) $\approx$ AIV(h) approximation is quite good and uniform in the sense that the vertical distances between these stay fairly constant. On the other hand, the ISB(h) $\approx$ AISB(h) approximation is comparatively worse and quite nonuniform in character, as the curves come together for h small, but diverge widely for large h. We have made such plots for all fifteen densities (but do not show them here because the main ideas are the same) and this type of behavior is very typical, although the bias approximation can be far worse, as demonstrated in Figure 3 below.

Figure 2 also shows the curves MISE(h) and AMISE(h) which are the sums of the respective variance and squared bias components. Note that the approximation of these curves is good for small h, but poor for large h.

Figure 3 shows the same setup as Figure 2b, but the underlying density is now the double claw, #11. Note that as in Figure 2, the IV(h) $\approx$ AIV(h) approximation is quite good, but the ISB(h) $\approx$ AISB(h) approximation is terrible. The reason the bias approximation is so poor is that AISB(h) is proportional to $\int (f'')^2$, which is very large because of the spikes apparent in the true f. However, while these spikes have an enormous effect on AISB(h), they are not really present in any practical sense, because as indicated in Table 1, the probability mass of these seven spikes represents approximately 2% of the total probability mass, which corresponds to about two observations in this case. Indeed the actual MISE(h) is very nearly the same as for the bimodal density, #6; see Figure 4 below.



(a)



(b)

FIG. 2. *Comparisons of* MISE(h) *and* AMISE(h) *for the Gaussian density, #1, Gaussian kernel, plotted on the ordinary h scale (part* a) *and the* $\log_{10}(h)$ *scale (part* b). *Solid curves are* MISE(h), IV(h), ISB(h) *and dotted curves are* AMISE(h), AIV(h), AISB(h). *The minimizers of* MISE(h) *and* AMISE(h) *are* h_{MISE} *and* h_{AMISE}, *respectively.*

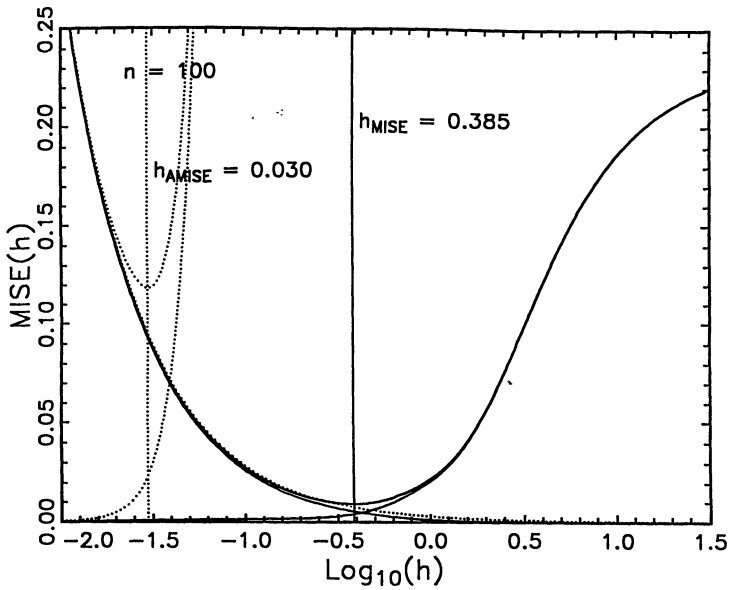


FIG. 3. MISE(h) *and* AMISE(h) *for the double claw density*, #11, $n = 100$, *Gaussian kernel. Solid curves are* MISE(h), IV(h), ISB(h) *and dotted curves are* AMISE(h), AIV(h), AISB(h).

Another interesting feature is that while AMISE(h) $\to \infty$ as $h \to \infty$ (recall it is proportional to h^4), MISE(h) seems to level off. This is in fact always the case, but it usually requires a large vertical scale to see this (for example, it is not visible using the much different scales in Figure 2). Indeed, it is straightforward to show that $\lim_{h\to\infty} \text{MISE}(h) = \int f^2$, which is not so surprising because for large h, $f_n(\cdot; h)$ tends to the function which is identically 0.

These considerations motivate the question of how well AMISE(h) approximates MISE(h) if one includes higher order terms in the Taylor expansion represented by AISB(h). In particular, assuming f is sufficiently smooth and that the kernel is second order, define for $j = 1, 2, \ldots$, the j term improvement of AMISE by

$$\text{AMISE}_j(h) = \frac{1}{nh}\int K^2 + \sum_{s=1}^{j} h^{2s}\left\{\frac{\mu_s(K)}{s!}\right\}^2 \int \{f^{(s)}\}^2.$$

Figure 4 shows how well the first few of these approximate MISE(h) for the bimodal density, #6, and $n = 100$, using the Gaussian kernel. Note that this type of approximation gains very little for h large. This may not be too surprising, because the real gains for this type of expansion occur in the limit as $h \to 0$. On the other hand, the Taylor series is convergent pointwise in h, but Figure 4 shows that this convergence is highly nonuniform. This is perhaps the worst case we have seen in this study of asymptotic results being very far from accurately indicating what is actually happening. We conclude

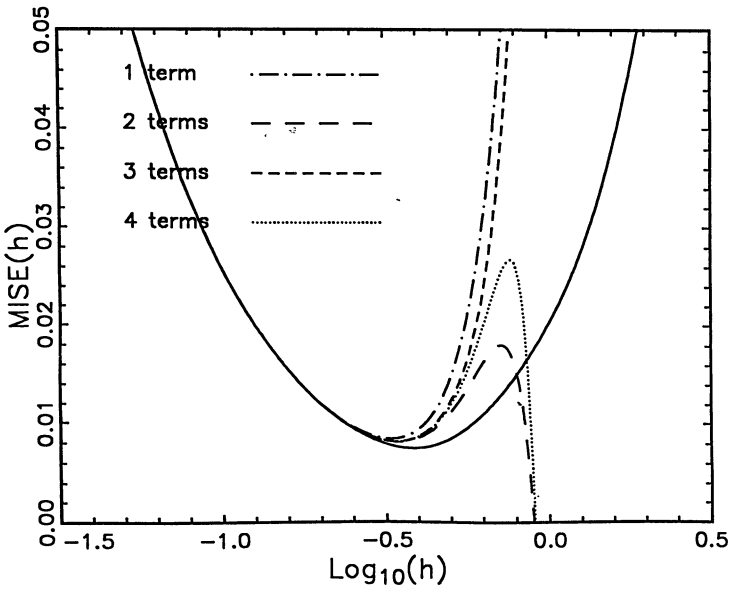


FIG. 4. MISE(h) *and several orders of* AMISE(h) *approximation for the bimodal density, #6,* $n = 100$, *Gaussian kernel. The number of terms is that used in the Taylor approximation of* ISB(h) *by* AISB(h), *that is, the number* j *in* AMISE(h).

that very careful interpretation (and even healthy skepticism) is clearly indicated for arguments based on approximation of ISB(h) by higher order terms in this sense. Note that in all of these pictures, MISE(h) < AMISE(h). This is also true for all of the other pictures we have made, and is in fact true in general, at least for nonnegative kernels (we speculate that it is true in general, but do not have a proof).

THEOREM 4.2. *For* f'' *continuous and square-integrable and* K *nonnegative,*

$$\text{MISE}(h) < \text{AMISE}(h) \quad \textit{for all } h > 0.$$

One consequence of Theorem 4.2 is that the AMISE assessment of the density estimation problem always appears harder than it actually is. Our pictures lead us to conjecture that there may be an analogous theorem available to the effect that $h_{\text{AMISE}} < h_{\text{MISE}}$, but we have not found a simple proof.

5. The minimizer of MISE. In this section the minimizers h_{MISE} and h_{AMISE} of MISE(h) and AMISE(h) are discussed. Details of calculations are postponed to the end of the section.

In view of the fact that MISE(h) so intuitively quantifies the trade-off between undersmoothing and oversmoothing, we were quite surprised to

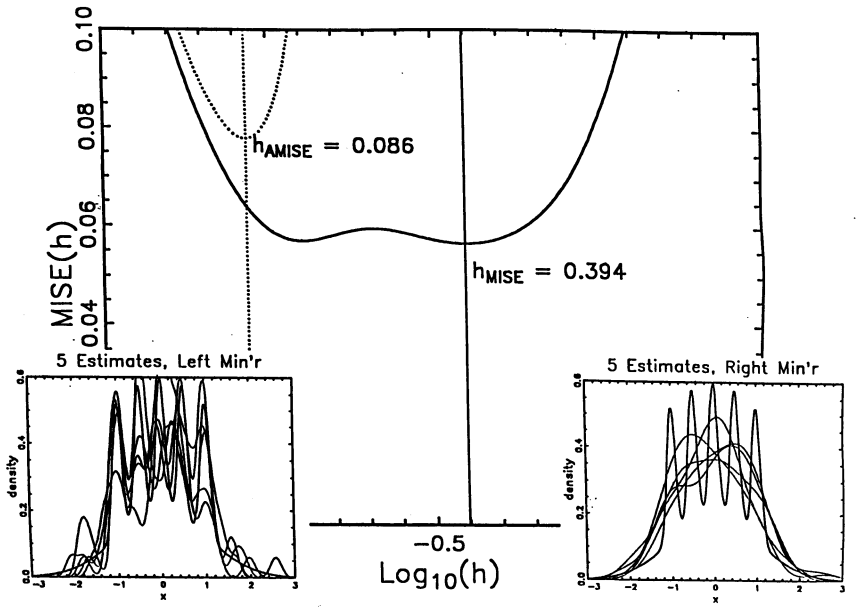


FIG. 5. MISE(h) *and* AMISE(h) *for the claw density,* #10, $n = 53$, *Gaussian kernel, solid curve is* MISE(h), *dotted curve is* AMISE(h). *Small plots show density estimates for five simulated data sets, using as window widths the left and right minimizers, respectively.*

discover that h_{MISE} is not always uniquely defined. This comes about because MISE(h) can have local minima, as shown in Figure 5, which shows AMISE(h) and MISE(h) for the claw density, #10, for $n = 53$, and the Gaussian kernel. The reason that the unusual value $n = 53$ was chosen here is that for $n = 54$, the right side of the MISE(h) curve in the figure stays essentially the same [recall ISB(h) is independent of n], but the left side moves downward [since IV(h) is proportional to n^{-1}], so the overall minimizer h_{MISE} suddenly moves all the way over to the left minimum. An interpretation of this phenomena is that for samples of size $n \leq 53$, a larger window width which completely ignores the fingers in the claw density is more appropriate in the sense of MISE(h), while for $n \geq 54$, a good MISE window width will be a much smaller one which attempts to estimate well the fingers. This interpretation is verified by the small plots in Figure 5 which show density estimates at these window widths and sample sizes for five simulated data sets.

Another interesting viewpoint on the behavior of h_{MISE} comes from studying it as a function of n. Figure 6 shows plots of $\log_{10}(h_{\text{MISE}})$ and $\log_{10}(h_{\text{AMISE}})$ as functions of $\log_{10}(n)$ for the Gaussian kernel and the bimodal, #6, and the double claw, #11, densities. The log-log scale is very informative, because optimal window widths are usually thought of in terms of powers of n. Note that h_{AMISE} is in fact exactly a linear function of n on this scale and the lines are parallel both having slope $-1/5$ [but far different intercepts, because of

the different values of $\int(f'')^2]$. For #6, h_{AMISE} begins to provide a decent approximation to h_{MISE} for sample sizes between 100 and 1000. This shape and behavior were quite typical for all the densities #1–#9. However, for the remaining densities, it takes much larger sample sizes for this approximation to be acceptable. For example, in the case of #11, n needs to be closer to one million for good approximation. As remarked above, since the spikes in #11 represent only about 2% of the probability mass, for small sample sizes these densities are practically the same. This shows up nicely in the way the two h_{AMISE} curves are the same up until about $n = 1000$. It is only for larger sample sizes that the effect of the peaks is enough, that is pays in the MISE sense to use a smaller window width for #11. The corresponding curves for the densities #10–#15 exhibited a variety of interesting features related to those shown here for #11, however there are too many to show here. One feature which is quite different is for density #10, where there is a discontinuity in the h_{MISE} curve between $n = 53$ and $n = 54$ for the reasons given at the beginning of this section.

Figure 6 shows how the best window width behaves as a function of n, but an even more interesting question is how well one does using this window width. This is addressed in Figure 7, which shows, in the same setup $\log_{10}\{\inf_{h>0} \mathrm{MISE}(h)\}$ and $\log_{10}\{\inf_{h>0} \mathrm{AMISE}(h)\}$ versus $\log_{10}(n)$. Again the AMISE curves are linear, both having slope $-4/5$, but with much different intercepts. As expected, the $\mathrm{AMISE}(h) \approx \mathrm{MISE}(h)$ approximation gets visibly

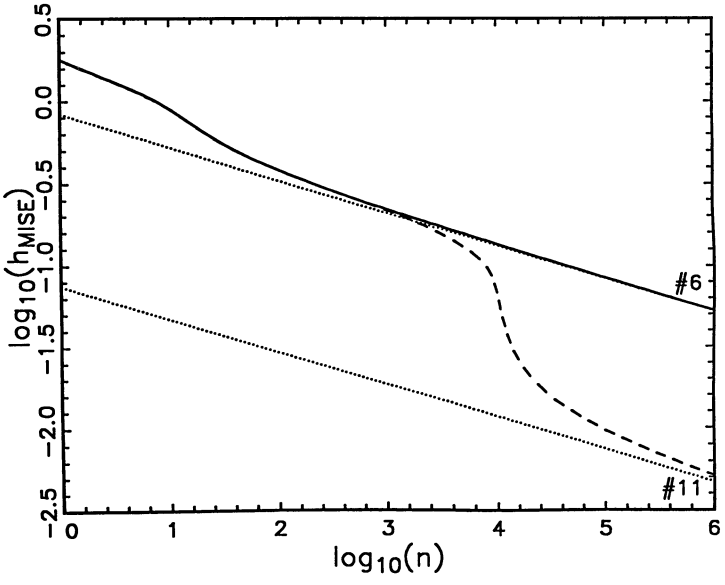


FIG. 6. $\log_{10}(h_{\mathrm{MISE}})$ *(solid and dashed curves) and* $\log_{10}(h_{\mathrm{AMISE}})$ *(dotted lines) as a function of* $\log_{10}(n)$ *for the bimodal density, #6 (solid) and the double claw density, #11 (dashed), Gaussian kernel.*

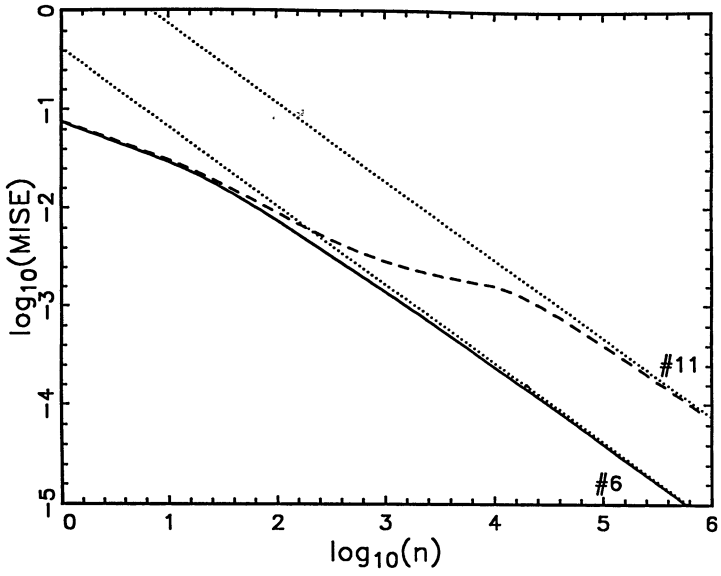


FIG. 7. $\log_{10}\{\inf_{h>0} \mathrm{MISE}(h)\}$ (*solid and dashed curves*) *and* $\log_{10}\{\inf_{h>0} AIMSE(h)\}$ (*dotted lines*) *as a function of* $\log_{10}(n)$ *for the bimodal density,* #6 (*solid*) *and the double claw density,* #11 (*dashed*), *Gaussian kernel.*

better for larger n. Again for smaller n, the two MISE curves are practically the same, although they are now visibly separate for sample sizes substantially less than 100. The reason that this separation occurs much earlier than in Figure 6 seems to be that the true MISE tends to be influenced strongly by the largest deviation between $f_n(\cdot; h)$ and f.

The minimizer, h_{AMISE}, of AMISE(h) is found from straightforward calculus and is given in Corollary 4.1. The minimizer, h_{MISE}, of MISE(h) is not so simple, being only implicitly defined, and as shown in Figure 5 is not necessarily unique. We have developed an effective numerical algorithm for finding this, details are available from the authors. It is a modified Newton's method, which requires formulae for the derivatives. These are given in the next theorem.

THEOREM 5.1. *Let f and K be as in Theorem* 2.1. *For the kernel estimator* (2.1), *we have*

$$\frac{d}{dh}\mathrm{MISE}(h) = -\frac{\mathscr{C}_1(r)}{nh^2} + \left(1 - \frac{1}{n}\right)\sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(-1)^{s+s'}}{2^{s+s'}s!s'!}\mathscr{U}'(h; s+s', 2)$$

$$- 2\sum_{s=0}^{r-1}\frac{(-1)^s}{2^s s!}\mathscr{U}'(h; s, 1),$$

where

$$\mathscr{U}'(h;s,q) = \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'}\left\{2sh^{2s-1}\phi_{ll'q}^{(2s)}(\mu_l - \mu_{l'}) + qh^{2s+1}\phi_{\sigma_{ll'q}}^{(2s+2)}(\mu_l - \mu_{l'})\right\}$$

and $\sigma_{ll'q} = (\sigma_l^2 + \sigma_{l'}^2 + qh^2)^{1/2}$ *is as in Theorem* 2.1. *In addition,*

$$\frac{d^2}{dh^2}\mathrm{MISE}(h) = \frac{2\mathscr{C}_1(r)}{nh^3} + \left(1 - \frac{1}{n}\right)\sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(-1)^{s+s'}}{2^{s+s'}s!s'!}\mathscr{U}''(h;s+s',2)$$

$$- 2\sum_{s=0}^{r-1}\frac{(-1)^s}{2^s s!}\mathscr{U}''(h;s,1),$$

where

$$\mathscr{U}''(h;s,q) = \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'}\left\{2s(2s-1)h^{2s-2}\phi_{\sigma_{ll'q}}^{(2s)}(\mu_l - \mu_{l'})\right.$$

$$\left. + q(4s+1)h^{2s}\phi_{\sigma_{ll'q}}^{(2s+2)}(\mu_l - \mu_{l'}) + q^2h^{2s+2}\phi_{\sigma_{ll'q}}^{(2s+4)}(\mu_l - \mu_{l'})\right\}.$$

The proof of Theorem 5.1 involves straightforward differentiation of the MISE(h) expression in Theorem 2.1.

6. Effect of higher order kernels. A very important application of our exact MISE ideas is to the problem of understanding when higher order kernels are actually more effective. For a sufficiently smooth underlying density, their faster rate of convergence means that higher order kernels will always be more effective for large enough sample sizes, but this says nothing about the crucial issues of:

1. How large an n is needed?
2. What happens before the asymptotics take effect?

These are especially important in practice because there is a definite price to be paid, in terms of plausibility and also interpretability, by higher order kernels. This is because they take on negative values and thus miss out on much of the beautiful and simple intuition available for nonnegative kernels.

Figure 8 is an analog of Figure 2b, for density #6 and kernel orders $p = 2$, 4, 6, 8, except that IV(h), ISB(h) and the asymptotic versions of everything are now removed to keep the plot from being too cluttered. Since the main point of higher order kernels is reduction of bias, it is not surprising that when the kernel order increases, the bias decreases for each h. As pointed out by Härdle (1986), the higher order kernels do however pay a price in terms of increased variance. Both of these effects mean that h_{MISE} moves to the right for higher kernel order. Note that the magnitude of these effects decreases with increasing kernel order, that is, there is a law of diminishing returns, which reflects the fact that higher order corrections become increasingly fine.

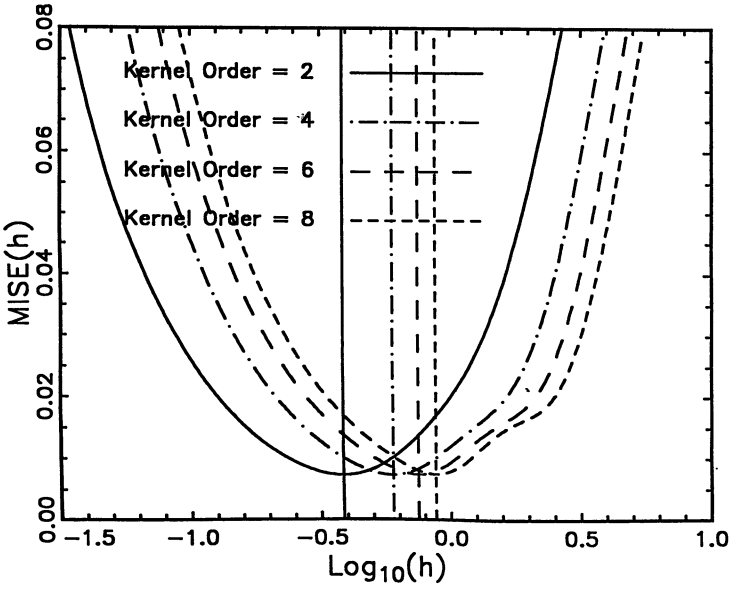


FIG. 8. MISE(h) *for the bimodal density, #6,* $n = 100$, *Gaussian based kernels of orders* $p = 2$, 4, 6 *and* 8.

We have looked at AMISE(h) as well, but it does not seem worth including a plot because the essential ideas are much the same as in Figure 2b. Once again, IV(h) is reasonably well-approximated by AIV(h), but the approximation of ISB(h) by AISB(h) is even worse for higher kernel orders and this time the effect does not decrease with higher kernel order. Similarly, the approximation of h_{MISE} by h_{AMISE} tends to get worse for higher order kernels.

It is interesting that in this case, the heights of the minima of these curves are all about the same. This shows that for $n = 100$, there is no practical gain from higher order kernels, that is, there is not enough data to take effective advantage of the higher order bias improvement. The same picture for $n = 1000$, on the other hand, does show marked improvement for the kernel of order $p = 4$, but little more for orders $p = 6$ and 8.

A more powerful way of viewing this is to look at analogs of Figures 6 and 7. We do not show the higher order version of Figure 6 to save space, since the lessons are much the same as in that picture, although of course the asymptotic window widths have a gentler slope (reflecting their slower rates of convergence). Another interesting feature is that for higher order kernels, it takes substantially large sample sizes to get the same degree of approximation of h_{MISE} by h_{AMISE}. Another point we had not expected is that discontinuities in the h_{MISE} curve, caused by local minima in MISE(h) as in Figure 5, appear very frequently for higher kernel orders. It is the prevalence of these local minima that makes the modification of Newton's method discussed at the end of Section 5 so important.

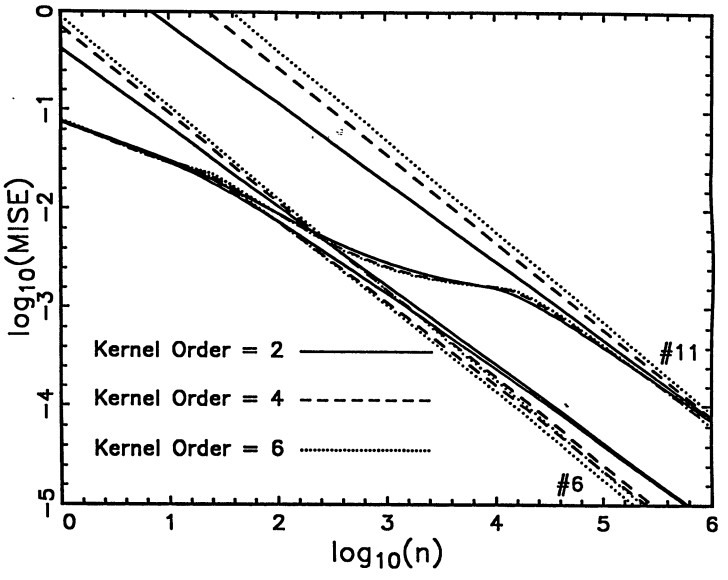


FIG. 9. $\log_{10}\{\inf_{h>0} \text{MISE}(h)\}$ *and* $\log_{10}\{\inf_{h>0} \text{AMISE}(h)\}$ *as a function of* $\log_{10}(n)$ *for the bimodal density, #6, and the double claw density, #11, and Gaussian-based kernel orders* $p = 2$ *(solid), 4 (dashed), 6 (dotted). The* AMISE's *are represented by straight lines. The corresponding* MISE's *are curves that are always below, but asymptotic to, the* AMISE's. *All density #11 curves are above the corresponding density #6 curves.*

Figure 9 is an analog of Figure 7 for higher order kernels. Note that for density #6, the performance for all the kernels orders is about the same near $n = 100$, and the higher order kernels are at least slightly better after that. However, we do not feel this effect is strong enough to overcome the lack of intuition inherent to higher order kernels until around $n = 1000$, and then only for order 4, with much less improvement in going from 4 to 6. In the other direction, it is not easy to see on this plot, but in fact for around $n = 20$, the higher order kernels give MISE(h) actually worse than that for the nonnegative one. Such behavior was typical for all of the densities #1 to #9, although the point where the higher order kernels become dominant increases with the complexity (this is only an intuitive notion reflecting a rough idea of distance from the Gaussian) of the density. The curves for h_{AMISE} also seem to converge roughly at a single point, which is substantially larger than the trade-off point for the true h_{MISE} curves. This shows it is not really sufficient to study this trade-off by looking at only AMISE, but instead our exact MISE calculations are needed. Again, as expected from the discussion around Figures 6 and 7, for small samples sizes, the curves for density #11 are very close to those for #6. However for larger sample sizes, they diverge to values that are far worse, because the spikes in #11 make it much harder to estimate. Note that which kernel order is best oscillates substantially with increasing n, and indeed the higher order kernels still have not become finally dominant even for

TABLE 2
Relative efficiency of kernel order 2 with respect to kernel order 4

	Threshold					
	MISE			AMISE		
Density	**100%**	**90%**	**75%**	**100%**	**90%**	**75%**
#1	2	9	88	85	245	1516
#2	5	25	268	176	506	3131
#3	345	1951	24100	5153	14779	91506
#4	96	383	3742	768	2202	13634
#5	5	19	170	105	302	1871
#6	88	246	1635	334	959	5935
#7	12	42	471	168	481	2981
#8	203	722	6454	1456	4176	25855
#9	275	1816	21635	5245	15042	93138
#10	193	596	5422	1281	3674	22747
#11	205822	$\geq 10^6$	$\geq 10^6$	$\geq 10^6$	$\geq 10^6$	$\geq 10^6$
#12	18235	111287	$\geq 10^6$	245389	703768	$\geq 10^6$
#13	345519	$\geq 10^6$	$\geq 10^6$	$\geq 10^6$	$\geq 10^6$	$\geq 10^6$
#14	13559	107241	$\geq 10^6$	249359	715155	$\geq 10^6$
#15	1069	5143	66175	12138	34811	215538

n = 1,000,000. This behavior is typical for the difficult-to-estimate densities #10–#15. Of course our formulae make it very easy to extend these pictures to larger n, but this seems pointless in view of current computing limitations (on actually working with such large data sets, and in particular computing kernel estimates), so we have not done so.

From looking at such pictures for all of our fifteen densities, we believe it is fair to say that in situations where the true density has features that make their presence felt, but cannot be well recovered, the higher order kernels have no advantage over the nonnegative kernel. This is seen in Figure 9, for samples sizes roughly between 10 and 100 for both densities #6 and #11, and for sample sizes roughly between 6,000 and 200,000 for density #11. On the other hand, when all features can be reasonably well recovered, the higher order kernel tends to be superior, although usually only marginally so.

A way of quantifying this idea for the densities shown in Figure 1 is given in Table 2. This is motivated by classical notions of efficiency. However as pointed out in Section 3.3.2 of Silverman (1986), it is tricky to do this directly in terms of MISE, because the usual efficiency interpretation of how much additional data is needed does not hold due to the $n^{-4/5}$ dependence of MISE on n. Instead of using $\text{MISE}^{5/4}$ as done there, we exploit the much deeper information available to us here by using an exact version. This is especially appropriate for those situations, as seen in Figures 7 and 9, where the slope of the MISE curve is less than its limiting value. Given two kernel orders $p < p'$ and given a sample size n, define the relative efficiency of the kernel of order p with respect to the kernel order p', to be the proportion of the sample (for the

lower order kernel) that would be needed for the same performance when using the higher order kernel, that is,

$$\mathrm{RE}_{p,p'}(n) = n'/n,$$

where n' is chosen so that $\inf_{h>0} \mathrm{MISE}(h; p, n) = \inf_{h>0} \mathrm{MISE}(h; p', n')$. For example, $\mathrm{RE}_{p,p'}(n) = 1/2$ means that the kernel of order p' requires only $n/2$ observations to have the same MISE as the kernel of order p using n observations. For n large enough, $\mathrm{RE}_{p,p'}(n)$ decreases in n. Instead of presenting graphs of these curves, we have chosen to summarize them by finding those values of n where the curves last cross below the efficiencies of 100%, 90% and 75%. These values shows when the asymptotic advantage of higher order kernels have taken effect as quantified by these three ways. They are listed in Table 2 for $p = 2$ and $p' = 4$. Values in the table larger than 1,000,000 are omitted because the only really important fact is that they are above this large value.

Our personal choice is to call the higher order kernel dominant (that is, increased efficiency overcomes better intuitive properties) when the efficiency of the lower order kernel falls below 75%. One reason for this is that a study of plots of $\mathrm{RE}_{p,p'}(n)$ versus $\log_{10}(n)$ shows that in those regions where the kernels are not comparable, the efficiencies tend to oscillate roughly between 80% and 125%.

Analogs of Table 2 in the cases of $p = 2$, $p' = 6$ and $p = 4$, $p' = 6$ are available from the authors. One main lesson from these tables is that $p' = 6$ usually dominates $p = 2$ for smaller n than is required for $p' = 4$ to dominate $p = 2$. These tables also show that much larger values of n are needed for $p' = 6$ to dominate $p = 4$ than for $p' = 4$ to dominate $p = 2$.

Note that these thresholds for the efficiencies are much worse when measured in terms of AMISE. The reason for this is that the straight lines tend to cross later than the corresponding curves in Figure 9. An important consequence is that it is not enough to study this kernel choice problem through AMISE, but instead our exact MISE calculations are required.

Our personal conclusion from this is that we cannot recommend the use of higher order kernels in practice. The situations in which they are sufficiently dominant require sample sizes much larger than those that many people work with. For those who do work with such large samples, the higher order kernel is only going to be much better in cases where the underlying density does not have much in the way of interesting features.

7. Proofs.

PROOF OF THEOREM 2.1. Our proof makes use of the following two results from Aldershof, Marron, Park and Wand (1991). For $\sigma, \sigma' > 0$ and $r, r' = 0, 1, 2, \ldots,$

$$\int \phi_\sigma^{(r)}(x - \mu)\phi_{\sigma'}^{(r')}(x - \mu')\,dx = (-1)^r \phi_{\tilde{\sigma}}^{(r+r')}(\mu - \mu'), \tag{7.1}$$

where $\tilde{\sigma} = (\sigma^2 + \sigma'^2)^{1/2}$ and

$$\phi_\sigma^{(2r)}(0) = (-1)^r 2^{-(2r+1)/2}\pi^{-1/2}(2r)!(r!)^{-1}\sigma^{-(r+1)}. \tag{7.2}$$

It is easily shown that

$$\operatorname{Var}\{f_n(x;h)\} = n^{-1}\operatorname{Var}\{K_h(x - X_1)\} = n^{-1}\left\{(K_h^2 * f)(x) - (K_h * f)^2(x)\right\}.$$

However, for $K = G_{2r}$,

$$\begin{aligned}
\int (K_h^2 * f) &= \int\int K_h^2(x-u)\,dx f(u)\,du = \int K_h^2(x)\,dx \\
&= \frac{1}{h}\sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(-1)^{s+s'}}{2^{s+s'}s!s'!}\int \phi^{(2s)}(x)\phi^{(2s')}(x)\,dx \\
&= \frac{1}{h}\sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(-1)^{s+s'}}{2^{2s+2s'+1/2}s!s'!}\phi^{(2s+2s')}(0) \\
&= \frac{1}{h\pi^{1/2}}\sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(2s+2s')!}{2^{3s+3s'+1}s!s'!(s+s')!} = \frac{\mathscr{C}_1(r)}{h},
\end{aligned}$$

which uses the second representation of G_{2r} given by (2.4), along with (7.1) and (7.2). Appealing again to (7.1) (here and in all following arguments),

$$\begin{aligned}
(K_h * f)(x) &= \int K(u) f(x - hu)\,du \\
&= \sum_{s=0}^{r-1}\sum_{l=1}^{k}\frac{(-1)^s w_l}{2^s s!}\int \phi^{(2s)}(u)\phi_{\sigma_l}(hu - x + \mu_l)\,du \\
&= \sum_{s=0}^{r-1}\sum_{l=1}^{k}\frac{(-1)^s w_l}{2^s s! h}\int \phi^{(2s)}(u)\phi_{\sigma_l/h}\{u - (x-\mu_l)/h\}\,du \\
&= \sum_{s=0}^{r-1}\sum_{l=1}^{k}\frac{(-1)^s w_l}{2^s s! h}\phi^{(2s)}_{\{1+(\sigma_l/h)^2\}^{1/2}}\{(x-\mu_l)/h\} \\
&= \sum_{s=0}^{r-1}\sum_{l=1}^{k}\frac{(-1)^s w_l h^{2s}}{2^s s!}\phi^{(2s)}_{(h^2+\sigma_l^2)^{1/2}}(x - \mu_l).
\end{aligned}$$

Thus,

$$\begin{aligned}
\int (K_h * f)^2 &= \sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\sum_{l=1}^{k}\sum_{l'=1}^{k}\frac{(-1)^{s+s'}}{2^{s+s'}s!s'!}w_l w_{l'} h^{2s+2s'} \\
&\qquad \times \int \phi^{(2s)}_{(h^2+\sigma_l^2)^{1/2}}(x-\mu_l)\phi^{(2s')}_{(h^2+\sigma_{l'}^2)^{1/2}}(x-\mu_{l'})\,dx \qquad (7.3) \\
&= \sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(-1)^{s+s'}}{2^{s+s'}s!s'!}\mathscr{U}(h; s+s', 2)
\end{aligned}$$

and so

$$\int \mathrm{Var}\{f_n(x;h)\}\,dx$$

$$= \frac{\mathscr{C}_1(r)}{nh} - \frac{1}{n}\sum_{s=0}^{r-1}\sum_{s'=0}^{r-1}\frac{(-1)^{s+s'}}{2^{s+s'}s!s'!}\mathscr{U}(h;s+s',2). \tag{7.4}$$

Next observe that

$$\int\{Ef_n(x;h) - f(x)\}^2\,dx = \int(K_h * f)^2 - 2\int(K_h * f)f + \int f^2. \tag{7.5}$$

The first term on the right-hand side of (7.5) is given by (7.3). The second term is

$$\int(K_h * f)f = \sum_{l=1}^{k}\sum_{l'=1}^{k}\sum_{s=0}^{r-1}\frac{(-1)^s}{2^s s!}w_l w_{l'}h^{2s}$$

$$\times\int\phi^{(2s)}_{(h^2+\sigma_l^2)^{1/2}}(x-\mu_l)\phi_{\sigma_l}(x-\mu_{l'})\,dx \tag{7.6}$$

$$= \sum_{s=0}^{r-1}\frac{(-1)^s}{2^s s!}\mathscr{U}(h;s,1)$$

and the third term is

$$\int f^2 = \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'}\phi_{(\sigma_l^2+\sigma_{l'}^2)^{1/2}}(\mu_l - \mu_{l'}) = \mathscr{U}(h;0,0). \tag{7.7}$$

Noting that

$$\mathrm{MISE}(h) = \int \mathrm{Var}\{f_n(x;h)\}\,dx + \int\{Ef_n(x;h) - f(x)\}^2\,dx$$

and combining (7.3), (7.4), (7.5), (7.6) and (7.7), we obtain the required result. □

PROOF OF THEOREM 4.2. Since

$$\mathrm{IV}(h) = \mathrm{AIV}(h) - n^{-1}\int(K_h * f)^2,$$

it is enough to show that

$$\mathrm{ISB}(h) \le \mathrm{AISB}(h).$$

For this, use Taylor's theorem with integral remainder to obtain

$$\begin{aligned}
\mathrm{ISB}(h) &= \int \left\{ h^2 \int\int_0^1 u^2 K(u)(1-t)^2 f''(x-hut)\,dt\,du \right\} \\
&\qquad \times \left\{ h^2 \int\int_0^1 u'^2 K(u')(1-t')^2 f''(x-hu't')\,dt'\,du' \right\} dx \\
&= h^4 \int\int\int_0^1\int_0^1 u^2 K(u)(1-t)^2 u'^2 K(u')(1-t')^2 \\
&\qquad\qquad \times \int f''(x-hut)\, f''(x-hu't')\,dx\,dt\,dt'\,du\,du' \\
&\le h^4 \int\int\int_0^1\int_0^1 u^2 K(u)(1-t)^2 u'^2 K(u')(1-t')^2 \\
&\qquad\qquad \times \left\{ \int f''(x-hut)^2\,dx \int f''(x-hu't')^2\,dx \right\}^{1/2} dt\,dt'\,du\,du' \\
&= h^4 \left\{ \mu_p(K)/2 \right\}^2 \int \{f''\}^2 = \mathrm{AISB}(h),
\end{aligned}$$

where we have used the Cauchy–Schwarz inequality and $\int_0^1 (1-t)^2 = 1/2$. □

REFERENCES

ALDERSHOF, B., MARRON, J. S., PARK, B. U. and WAND, M. P. (1991). Facts about the Gaussian probability density function. Unpublished manuscript.

DAVIS, K. B. (1981). Estimation of the scaling parameter for a kernel-type density estimate. *J. Amer. Statist. Assoc.* **76** 632–636.

DEHEUVELS, P. (1977). Estimation nonparametrique de la densité par histogrammes generalisés. *Rev. Statist. Appl.* **25** 5–42.

FRYER, M. J. (1976). Some errors associated with the nonparametric estimation of density functions. *J. Inst. Math. Appl.* **18** 371–380.

GASSER, TH., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

HÄRDLE, W. (1986). A note on jackknifing kernel regression function estimators. *IEEE Trans. Inform. Theory* **32** 298–300.

HART, J. D. (1984). Efficiency of a kernel density estimator under an autoregressive dependence model. *J. Amer. Statist. Assoc.* **79** 110–117.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

TERRELL, G. R. and SCOTT, D. W. (1985). Oversmoothed density estimates. *J. Amer. Statist. Assoc.* **80** 209–214.

WAND, M. P. and SCHUCANY, W. R. (1990). Gaussian-based kernels. *Canad. J. Statist.* **18** 197–204.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514

DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TEXAS 77251-1892