JOURNAL OF MULTIVARIATE ANALYSIS **56**, 165–184 (1996)
ARTICLE NO. 0009

On the Accuracy of Binned Kernel Density Estimators

Peter Hall

Australian National University, Canberra, Australia

AND

M. P. Wand

University of New South Wales, Sydney, Australia

The accuracy of the binned kernel density estimator is studied for general binning rules. We derive mean squared error results for the closeness of this estimator to both the true density and the unbinned kernel estimator. The binning rule and smoothness of the kernel function are shown to influence the accuracy of the binned kernel estimators. Our results are used to compare commonly used binning rules, and to determine the minimum grid size required to obtain a given level of accuracy. © 1996 Academic Press, Inc.

1. Introduction

An important recent contribution to the practical application of kernel-type estimators is the idea of prebinning the data on an equally spaced mesh and then applying a suitably modified kernel estimator to the binned data. This approach results in what is usually referred to as the *binned* or *WARPed* (an acronym for *weighted averaging using rounded points*) version of the kernel estimator and leads to substantial computational savings compared to the direct computation of kernel estimators; see Silverman (1982), Scott (1985), Härdle and Scott (1992), Fan and Marron (1994), and Wand (1994). Binned kernel estimators are also appropriate for the common situation where the data are only available in a discretised format.

A question of considerable practical relevance concerns the accuracy of kernel estimators based on binned data. For simplicity's sake we will treat

Received March 25, 1994; revised March 1995.

AMS 1990 subject classifications: Primary 62G07; secondary 65D99.

Key words and phrases: Binning rules, Euler–Maclaurin summation formulae.

0047-259X/96 $18.00

the problem of estimating a probability density function f, although the main ideas are directly applicable to other settings such as nonparametric regression. Let $\tilde{f}$ be a binned kernel density estimator for f (defined in Section 2) and $\hat{f}$ be the ordinary kernel density estimator. The accuracy of $\tilde{f}$ can be assessed in two different ways. The first arises from treating $\tilde{f}$ as an estimator of f in its own right, and studying its estimation properties—as has been traditionally done for ordinary kernel estimators. The second concerns the closeness of $\hat{f}$ and $\tilde{f}$. This is worthwhile because $\hat{f}$ is the more natural and mathematically tractable of the two density estimators while, for reasons of speed, $\tilde{f}$ is the more appropriate estimator to use in practice.

In this article we investigate both measures of accuracy. In either case, the accuracy is shown to depend quite heavily on the rule used to bin the data. Various binning rules are discussed in Section 2. The presentation is facilitated through the characterization of a binning rule through a "binning kernel", that has similarities with the usual kernel function.

The properties of $\tilde{f}$ as an estimator for f are given a comprehensive treatment in Section 3. This is in the spirit of earlier work by Hall (1982), Scott and Sheather (1985), and Jones (1989), but it goes beyond their work by deriving concise asymptotic approximations rather than order-of-magnitude upper bounds. An important component of our results, not fully recognized by previous authors, is the effect of the smoothness of the kernel on the asymptotic performance of $\tilde{f}$. This is in contrast to ordinary kernel density estimation where smoothness properties of the kernel do not affect the asymptotics.

In Section 4 we derive results for the distance between $\hat{f}$ and $\tilde{f}$, generalizing previous work by Jones and Lotwick (1983). A noteworthy difference between these results and those of Section 3 is that they do not require the usual large sample asymptotics. This is because the individual performances of $\tilde{f}$ and $\hat{f}$ as density estimators are not of interest when measuring their closeness. Rather we use asymptotics that allow the binning mesh to become finer while keeping the sample size fixed. The linear binning scheme proposed by these authors is seen to perform very well in this regard. Our results also indicate that the goals of the estimation of f and closeness to $\hat{f}$ can be quite different; a binning rule resulting in better estimation properties of $\tilde{f}$ does not necessarily lead to improvement in the closeness of $\tilde{f}$ to $\hat{f}$.

The massive amounts of computation required for direct kernel estimation of multivariate data provide even greater motivation for the use of faster binned kernel estimators. In Section 5 we extend our univariate results to the multivariate density estimation context.

In practice $\tilde{f}$ is usually computed over a finite number of grid points. The choice of the size of this grid is very important since it involves a trade-off between minimizing the binning error and minimizing the computational time. In Section 6 we use our approximation results to obtain minimum grid sizes to achieve a prescribed accuracy. Our results give theoretical support for commonly used "rules of thumb" for choosing the grid size. For example, grid sizes up to about 500 grid points are seen to be adequate for most density types and sample sizes that arise in practice.

2. Binned Kernel Density Estimators

2.1. *General Approaches to Binning*

A binning rule may be represented by a sequence of functions $\{w_j(x, \delta), j \in \mathbf{Z}\}$ and asks that an observed data value X be distributed among "grid points" $g_j = j\delta$ in such a way that weight $w_j(X, \delta)$ is contributed to g_j. If we ask that for each real x and $\delta > 0$, $\sum_j w_j(x, \delta) = 1$, then it becomes clear that a binning rule divides each single data value into a number of parts and assigns them to different grid points. However, this condition is not always necessary and is violated by the higher order polynomial binning rules described in Section 3.5. We could also insist that each $w_j(x, \delta) \geqslant 0$, although this constraint is rather restrictive. It is like demanding that a kernel function be nonnegative, and that does exclude methods that are of both practical and theoretical interest.

Examples of binning rules include simple binning, where

$$w_j(x, \delta) = \begin{cases} 1 & \text{if} \quad x \in ((j - \frac{1}{2})\,\delta, (j + \frac{1}{2})\,\delta], \\ 0 & \text{otherwise;} \end{cases}$$

and common linear binning, where

$$w_j(x, \delta) = \begin{cases} 1 - |\delta^{-1}x - j| & \text{if} \quad |\delta^{-1}x - j| \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Most binning rules $w_j(x, \delta)$ can be characterized through the function $W(x) \equiv w_0(x, 1)$. The rule can then be written in terms of W as $w_j(x, \delta) = W(\delta^{-1}x - j)$. Asymptotic unbiasedness of the binned kernel estimator requires that $\int W = 1$. Also the order of the bias depends on the number of zero moments possessed by W; see Section 3.1. Because of its close analogy with the kernel of ordinary kernel density estimation we will call W the *binning kernel* associated with the binning rule $w_j(x, \delta)$. Figure 1 shows binning kernels for the two common binning rules described above, as well as an alternative "fourth-order" linear binning rule that we describe in Section 3.5.

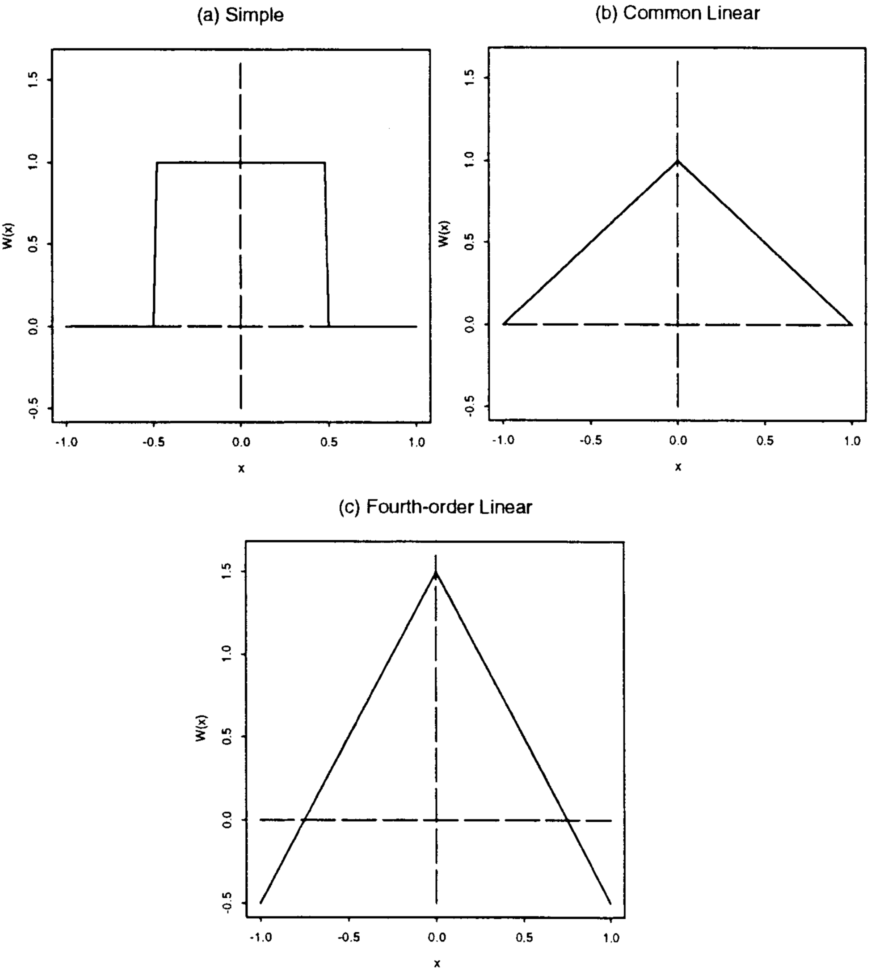

FIG. 1. Binning kernels for (a) simple binning. (b) common linear binning, and (c) fourth-order linear binning.

2.2. *Application of General Binning to Kernel Density Estimation*

Let $k \geqslant 2$ be an integer and let K be a bounded function, integrable against kth-degree polynomials and enjoying the property that

$$\int u^j K(u)\,du = \begin{cases} 1 & \text{if} \quad j = 0, \\ 0 & \text{if} \quad 1 \leqslant j \leqslant k-1, \\ k!\,\kappa \neq 0 & \text{if} \quad j = k. \end{cases}$$

In keeping with standard terminology we call K a kth-order kernel. Silverman (1986, p. 66ff) has discussed the use of such kernels in density estimation. An unbinned kernel estimator of a density f, based on a random sample $X_1, ..., X_n$ drawn from the population with density f, may be written as

$$\hat{f}(x; h) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \qquad -\infty < x < \infty,$$

where $K_h(u) = K(u/h)/h$. To reduce the computational labour in calculating $\hat{f}(\cdot; h)$ we may first bin the data, using a rule such as those discussed in Section 2.1, and then employ the resulting summary statistics to calculate an approximation to $\hat{f}$:

$$\tilde{f}(x; h) = n^{-1} \sum_{j \in \mathbf{Z}} N_j K_h(x - j\delta),$$

where $N_j = \sum_{i=1}^{n} w_j(X_i, \delta)$ denotes the "count" at grid point g_j. Note that N_j may be negative or nonintegral.

3. Accuracy of $\tilde{f}$ as a Density Estimator

3.1. *Effect of Binning on Bias*

A typical binning rule produces an expansion of the form

$$E\{w_j(X, \delta)\} = \delta f(j\delta) + c_1 \delta^2 f'(j\delta) + c_2 \delta^3 f''(j\delta) + \cdots, \tag{3.1}$$

where

$$c_i = (i!)^{-1} \int z^i W(z)\, dz.$$

Simple binning has $c_1 = 0$ and $c_2 = \frac{1}{24}$. Since even simple binning has $c_1 = 0$ then it is unlikely that one would ever employ a binning rule which had $c_1 \neq 0$. Common linear binning has $c_1 = 0$ and $c_2 = \frac{1}{12}$. A binning rule, based on fitting a polynomial of degree $m - 1$, may be constructed so that $c_1 = \cdots = c_{2m-1} = 0$; see Section 3.5 for examples.

Let c_s denote the first nonzero c_i. We will call s the *order* of the binning rule since its corresponding kernel is a sth-order kernel function. Assume that K is continuous, and $K^{(2t-1)}$ exists at all but a finite number of points and is piecewise continuous with only a finite number of discontinuities, at each of which both left-hand right-hand limits are well defined.

Assume that f has max $(s, 2t-1)$ continuous derivatives, and that $\delta = o(h)$; that is, bin width is of smaller order than bandwidth. We shall show in Appendix A that

$$E\{\tilde{f}(x; h)\} - E\{\hat{f}(x; h)\} \\ = c_s \delta^s f^{(s)}(x) + (\delta/h)^{2t}\, a(x, \delta, h) + o\{\delta^s + (\delta/h)^{2t}\}, \qquad (3.2)$$

where the function $a(x, \delta, h)$ is bounded and is determined by the jumps of $K^{(2t-1)}$. If there are no jumps then $a(\,\cdot\,, \delta, h) \equiv 0$. For example, if x is a grid point and if the jumps occur at points y such that hy is a grid point then

$$a(x, \delta, h) = -\{(2t)!\}^{2t} f(x)\, J(K^{(2t-1)}), \qquad (3.3)$$

where $J(K^{(2t-1)}) \equiv \sum_x \{K^{(2t-1)}(x+) - K^{(2t-1)}(x-)\}$ denotes the sum of the jumps of $K^{(2t-1)}$ and B_j is the jth Bernoulli number.

To appreciate the implications of this result, let us suppose that the kernel K is of order k, as defined in Section 2.2, and that f has k continuous derivatives. Then

$$E\{\hat{f}(x; h)\} - f(x) = \kappa h^k f^{(k)}(x) + o(h^k) \qquad (3.4)$$

as $h \to 0$, where $\kappa = (-1)^k (k!)^{-1} \int u^k K(u)\, du \neq 0$. See Silverman (1986, p. 67ff) for discussion of results such as this. Substituting (3.4) into (3.2) we see that if the bias of the binned estimator f is to be no larger than that of the unbinned estimator then, in addition to $\delta = o(h)$, we need

$$\delta^s + (\delta/h)^{2t} = o(h^k). \qquad (3.5)$$

(In theory we might hope that $O(h^k)$ on the right-hand side would suffice, but practical considerations suggest that the left-hand side should be of smaller size than h^k.) Second-order kernels, i.e., those with $k = 2$, are by far the most common in practice, and so, since we always have $s \geqslant 2$ (see the comments just below (3.1)) then the relation $\delta = o(h)$ implies $\delta^s = o(h^k)$. In this case, (3.5) is event to $(\delta/h)^{2t} = o(h^k)$, i.e., to $\delta = o(h^{1+\{k/(2t)\}})$. The commonly used Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)_+$ has $k = 2$ and $t = 1$, and for this we require that $\delta = o(h^2)$ if binning is not to have a significant effect on the bias of a kernel density estimator.

3.2. *Special Case of a Very Smooth Kernel*

For infinitely differentiable kernels, such as the standard normal kernel, t may be chosen arbitrarily large. Provided that $\delta = O(h^{1-\varepsilon})$ for some $\varepsilon > 0$ (no matter how small) and f has sufficiently many derivatives (that number depending on s and ε), the second and third terms on the right-hand side of (3.2) can be made of smaller order than δ^s by simply choosing t sufficiently large. (Note that if K is infinitely differentiable then $J(K^{(l)})=0$ for each l. This argument does require a sufficiently smooth density f.) In this case, (3.2) reduces to

$$E\{\tilde{f}(x;h)\}-E\{\hat{f}(x;h)\}=c_s\delta^s f^{(s)}(x)+o(\delta^s). \tag{3.6}$$

Therefore, there can be distinct advantages in using smooth kernel functions.

3.3. *Special Case of a Very Rough Kernel*

In Section 3.1 we assumed that K is continuous. The conclusions drawn there are not valid without this condition. Discontinuous kernels are rarely used in practice, except for the uniform kernel, $K(u)=\frac{1}{2}$ if $|u|\leqslant 1$ and $K(u)=0$ otherwise. In this case (3.2) should be replaced by

$$E\{\tilde{f}(x;h)\}-E\{\hat{f}(x;h)\}=\tfrac{1}{2}\,\delta h^{-1}(\Delta j+1-2h\delta^{-1})f(x)+o(\delta h^{-1}), \tag{3.7}$$

where $\Delta j=j_2-j_1$ and j_1 and j_2 are, respectively, the least and greatest integers j such that $|x-j\delta|\leqslant h$. Note particularly that $\Delta j+1-2h\delta^{-1}$ is of size 1, being bounded away from zero and infinity along a subsequence, but does not converge as $\delta h^{-1}\to 0$. Therefore, the difference between $E\{\tilde{f}(x;h)\}$ and $E\{\hat{f}(x;h)\}$ is genuinely of size δh^{-1}. A proof of (3.7) will be given in Appendix B.

The quantity δh^{-1} can be quite large unless δ is small, and it may be of larger order than the bias of the unbinned estimator. If K is the uniform kernel then, in order for the difference between $E\{\tilde{f}(x;h)\}$ and $E\{\hat{f}(x;h)\}$ to be of smaller size than $E\{\hat{f}(x;h)\}-f(x)$, it is necessary and sufficient that $\delta=o(h^3)$. In practical applications this condition can demand a relatively large number of bins. This phenomenon is illustrated through examples by Fan and Marron (1994).

3.4. *Effect of Binning on Variance and Mean Squared Error*

The variance of the unbinned estimator is well known to be given by

$$\operatorname{Var}\{\hat{f}(x;h)\}=(nh)^{-1}\left(\int K^2\right)f(x)-n^{-1}f(x)^2+o(n^{-1}) \tag{3.8}$$

(see Scott, 1992, p.131) while the difference between the variances of the binned and unbinned estimators has the property

$$\operatorname{Var}\{\tilde{f}(x;h)\}-\operatorname{Var}\{\hat{f}(x;h)\}=O[(nh)^{-1}\{\delta^s+(\delta/h)^{2t}\}] \tag{3.9}$$

under the conditions, and in the notation of Section 3.1. (The proof is similar to that in Appendix A). Since $\delta = o(h)$ then $(nh)^{-1}\,\delta^{s} = o(n^{-1})$, and so (3.8) and (3.9) together give

$$\operatorname{Var}\{\tilde{f}(x;h)\} = (nh)^{-1}\left(\int K^2\right) f(x) - n^{-1} f(x)^2 + O\{(nh)^{-1}\,(\delta/h)^{2t}\} + o(n^{-1}). \tag{3.10}$$

Whether or not the term $(nh)^{-1}\,(\delta/h)^{2t}$ is of smaller order than n^{-1}, and so may be dropped from (3.10), depends of course on the sizes of δ and h. It cannot generally be omitted, and so it is not always true that the variances of $\hat{f}(\,\cdot\,;h)$ and $\tilde{f}(\,\cdot\,;h)$ agree up to terms of order n^{-1}, i.e., to second order. However, if h is of a size which optimizes the performance of the unbinned estimator, and if δ is chosen so that the biases of the binned and unbinned estimators agree to first order, then the variances of the binned and unbinned estimators agree to second order. To appreciate why, observe that the first of these stipulations requires that bias of the unbinned estimator be of size $(nh)^{-1/2}$ and for the second that $(\delta/h)^{2t} = o\{(nh)^{-1/2}\}$. Therefore, (3.10) reduces to

$$\operatorname{Var}\{\tilde{f}(x;h)\} = (nh)^{-1}\left(\int K^2\right) f(x) - n^{-1} f(x)^2 + o(n^{-1});$$

compare (3.8).

In any event, the impact of binning is almost entirely through its effect on bias, as discussed in Sections 3.1–3.3. By combining (3.10) with appropriate bias formulae (see (3.2), (3.4), and (3.6), for example) we may deduce expansions for mean squared error; and by formally integrating those expressions we may obtain formulae for mean integrated squared error. (Formal integration is valid under a variety of regularity conditions, of which the simplest is that f have compact support, as well as satisfy the appropriate smoothness assumptions.) In particular, in the context of (3.2) we have the formula

$$\begin{aligned}\int E\{\tilde{f}(\,\cdot\,;h) - f\}^2 = (nh)^{-1}\int K^2 - n^{-1}\int f^2 \\ + \int \{\kappa h^k f^{(k)} + c_s \delta^s f^{(s)} + (\delta/h)^{2t}\, a(\,\cdot\,,\delta,h)\}^2 \\ + o[n^{-1} + \{h^k + \delta^s + (\delta/h)^{2t}\}^2] + O\{(nh)^{-1}(\delta/h)^{2t}\}\end{aligned} \tag{3.11}$$

for mean integrated squared error of the binned estimator, which compares with

$$\int E\{\hat{f}(\cdot\,;h)-f\}^2=(nh)^{-1}\int K^2-n^{-1}\int f^2+\kappa^2h^{2k}\int\{f^{(k)}\}^2+o(n^{-1}+h^{2k})$$

(see Scott, 1992, p. 133) for the unbinned estimator. In the context of a very smooth kernel, discussed in Section 3.2, the necessary modification of (3.11) is simply to drop the terms involving $(\delta/h)^{2t}$, obtaining

$$\int E\{\tilde{f}(\cdot\,;h)-f\}^2=(nh)^{-1}\int K^2-n^{-1}\int f^2+\int(\kappa h^k f^{(k)}+c_s\delta^s f^{(s)})^2 + o(n^{-1}+h^{2k}+\delta^{2s}). \tag{3.12}$$

3.5. *Polynomial Binning Rules*

One polynomial binning rule, quite different from more conventional binning rules (i.e., simple and common linear rules), is defined by

$$w_j(x,\delta)=\begin{cases}\sum_{i=0}^{m-1} b_i\,|\delta^{-1}x-j|^i & \text{if}\quad |\delta^{-1}x-j|\leqslant b,\\ 0 & \text{otherwise},\end{cases} \tag{3.13}$$

where $b>0$ is arbitrary and $b_0, ..., b_{m-1}$ are chosen to ensure that

$$\sum_{i=0}^{m-1}(i+2l+1)^{-1}\,b^{i+2l+1}b_i=\begin{cases}\frac{1}{2} & \text{if}\quad l=0,\\ 0 & \text{if}\quad 1\leqslant l\leqslant m-1.\end{cases}$$

The corresponding binning kernel is $W(x)=\sum_{i=0}^{m-1} b_i\,|x|^i$, $|x|\leqslant b$, and is zero otherwise. This binning rule has order $2m$. When $m=2$ this prescription produces a linear binning rule, with $b_0=3/(2b)$, $b_1=-2/b^2$, and $c_4=-b^4/15$. For the important special case of $b=1$ the fourth-order binning rule has binning kernel

$$W(x)=\tfrac{3}{2}-2|x|,\qquad |x|\leqslant 1 \tag{3.14}$$

(see Fig. 1). This linear binning rule has a weight function which takes negative values. That is certainly a drawback, but of course it is necessary if we are to achieve $s=4$. The common linear binning rule described in Subsection 2.2 has the property that, of all binning rules constructed according to (3.13) with $m=2$ and $b=1$, and satisfying (3.1) and $w_j\geqslant 0$, it has the smallest value of c_2.

3.6. *Comparison with Hall* (1982) *and Scott and Sheather* (1985)

Hall (1982) provided upper bounds to the effect of binning on bias and mean squared error. Later authors have interpreted his work as describing a "worst case" scenario which would not typically arise in practice. In particular, Scott and Sheather (1985) argued that, under the assumption that the kernel K is a continuous and compactly supported density, one may derive the following approximate expansions for mean integrated squared error of a kernel estimator computed by simple binning:

$$\int E\{\tilde{f}(\cdot\,;h)-f\}^2 \simeq (nh)^{-1}\int K^2 - n^{-1}\int f^2 + (\kappa h^2 + c_2\delta^2)^2 \int (f'')^2. \qquad (3.15)$$

(See Proposition 2 of Scott and Sheather, 1985, and note that in the context of that proposition, $k=s=2$ and $c_2=\frac{1}{24}$.)

Formula (3.15) does not include contributions which can result through lack of smoothness of K. For example, if K, $K^{(1)}$, $K^{(3)}$, ..., $K^{(2t-1)}$ are continuous but $K^{(2t-1)}$ has jump discontinuities then (3.15) should be replaced by (3.11) (with $k=s=2$); the latter formula correctly allows for contributions to bias arising from the lack of smoothness of $K^{(2t-1)}$.

As we noted in Section 3.1, the terms omitted from (3.15) can be significant in applications. For example, if one uses the Epanechnikov kernel then the correct version of (3.15) shows that $\delta=0(h^2)$ is necessary and sufficient for binning to have no asymptotic first-order effect on the performance of $\hat{f}(\cdot\,;h)$. However, (3.15) suggests that $\delta=0(h)$ is adequate.

Scott and Sheather (1985) conducted a simulation study which tends to confirm their conclusions. However, in this numerical work they used the standard normal kernel, not (for example) the Epanechnikov kernel. As we noted in Section 3.2, a normal kernel does not produce the same binning problems as other kernels which satisfy Scott and Sheather's regularity conditions. Indeed, the mean integrated squared error formula appropriate for a normal density is simply (3.12) in the case $k=s=2$; and that is essentially (3.15).

4. Accuray of One Estimator as an Approximation to the Other

4.1. *Motivation*

In practice $\tilde{f}$ is often thought of as an approximation to the idealized form, $\hat{f}$. In particular, the intuitive argument which leads one to consider kernel estimation—i.e., to place probability mass about each data value, and take the average of the masses—applies to $\hat{f}$ rather than $\tilde{f}$. Moreover, the vast majority of theory applies to $\hat{f}$ rather than $\tilde{f}$, and so it is of particular interest to know how close they are to one another. This problem may be solved without resorting to the usual large sample asymptotics, since the individual performances of $\tilde{f}$ and $\hat{f}$ are not of interest. More meaningful asymptotics result from simply letting the bin width δ

approach zero, with n and h fixed. It is difficult to give concise results of this type for general binning rules, so we will focus on three important special cases: simple binning (Subsection 4.2), common linear binning (Subsection 4.3) and the fourth-order linear binning rule having binning kernel (3.13) (Subsection 4.4).

4.2. *Approximation Accuracy of Simple Binning*

Suppose that $\tilde{f}$ is based on the simple binning rule. Assuming that K has a continuous derivative, Taylor expansion leads to

$$\tilde{f}(x; h) - \hat{f}(x; h) = \delta n^{-1} \sum_{i=1}^{n} K_h'(x - X_i)\, Q(\delta^{-1} X_i) + o_P(\delta)$$

as $\delta \to 0$, where $Q(x) = x -$ (closest integer to x), $-\infty < x < \infty$. Therefore, since $\int_{-1/2}^{1/2} Q(\delta^{-1} z)^2\, dz = \frac{1}{12} + o(1)$,

$$\begin{aligned} E\{\tilde{f}(x; h) - \hat{f}(x; h)\}^2 &= \delta^2 n^{-1} \int K_h'(x-y)^2\, Q(\delta^{-1} y)^2 f(y)\, dy + o(\delta^2) \\ &= \delta^2 n^{-1} \int K_h'(x-y)^2 \left\{ \int_{-1/2}^{1/2} Q(\delta^{-1} z)^2\, dz \right\} \\ &\quad \times f(y)\, dy + o(\delta^2) \\ &= \delta^2 (12n)^{-1}\, E\{K_h'(x - X)^2\} + o(\delta^2). \end{aligned}$$

The second line here may be explained intuitively by noting that as $\delta \to 0$, $Q(\delta^{-1} X)$ converges in distribution to a uniform random variable on $(-\frac{1}{2}, \frac{1}{2})$ that is independent of X (see, e.g., Hall, 1983, Lemma 3). The mean squared difference between $\tilde{f}(x)$ and $\hat{f}(x)$ is therefore of order δ^2 as $\delta \to 0$. Under appropriate integrability conditions we obtain

$$E \int (\tilde{f} - \hat{f})^2 = \delta^2 (12nh^3)^{-1} \int (K')^2 + o(\delta^2). \tag{4.1}$$

A closely related result was derived by Jones and Lotwick (1983).

4.3. *Approximation Accuracy of Common Linear Binning*

Suppose now that $\tilde{f}$ is based on common linear binning. Assuming that K has two continuous derivatives and noting that

$$\sum_{j \in \mathbf{Z}} (X - j\delta)(1 - |\delta^{-1} X - j|)\, I(-1 < \delta^{-1} X - j \leqslant 1) = 0$$

(here $I(E)$ is the indicator of the event E) we obtain for the common linear binned kernel density estimator:

$$\tilde{f}(x;h)-\hat{f}(x;h)=\tfrac{1}{2}\delta^2 n^{-1}\sum_{i=1}^{n} K_h''(x-X_i)\,R(\delta^{-1}X_i)\{1-R(\delta^{-1}X_i)\}+o_P(\delta^2),$$

where $R(x)=x$—(greatest integer not exceeding x), $-\infty<x<\infty$. Since $R(\delta^{-1}X_i)$ converges in distribution to a uniform random variable on (0, 1) as $\delta\to 0$, and independently of X (Hall, 1983, Lemma 3), we obtain

$$E\{\tilde{f}(x;h)-\hat{f}(x;h)\}^2=\delta^4[(120n)^{-1}E\{K_h''(x-X)^2\}+\tfrac{1}{144}(1-n^{-1})\{EK_h''(x-X)\}^2]+o(\delta^4).$$

Formal integration of this result leads to

$$E\int(\tilde{f}-\hat{f})^2=\delta^4\left\{(120nh^5)^{-1}\int(K'')^2+\tfrac{1}{144}(1-n^{-1})\int(K_h''*f)^2\right\}+o(\delta^4),\tag{4.2}$$

where $*$ denotes convolution. These results show that, in terms of how close $\hat{f}$ is to $\tilde{f}$, common linear binning is asymptotically superior to simple binning.

4.4. *Approximation Accuracy of Fourth-Order Linear Binning*

For $\tilde{f}$ based on the fourth-order linear binning rule we have

$$\tilde{f}(x;h)-\hat{f}(x;h)=\delta n^{-1}\sum_{i=1}^{n}K_h'(x-X_i)\{\tfrac{1}{2}-R(\delta^{-1}X_i)\}+o_P(\delta)$$

which leads to

$$E\{\tilde{f}(x;h)-\hat{f}(x;h)\}^2=\delta^2(12n)^{-1}E\{K_h'(x-X)^2\}+o(\delta^2).$$

Therefore the mean squared difference of $\tilde{f}$ based on fourth-order linear binning is asymptotically the same as that based on simple binning. Together with the results derived in Section 2 this result leads to the noteworthy conclusion that while the fourth-order linear binned estimate is a better estimate of f than the common linear binned estimate, its approximation by $\hat{f}$ is worse.

5. Extension to d-Dimensional Data

5.1. *Multivariate Binning Rules and Binned Multivariate Kernel Estimators*

Multivariate binning rules may be defined by taking the product of the univariate rules, as follows. Suppose $w_{ij}(x,\delta)$, denotes a univariate rule for each $1 \leqslant i \leqslant d$. In particular, w_{ij} could denote a polynomial binning rule such as those discussed in Section 4. Henceforth we let $j = (j_1, ..., j_d)$, $x = (x_1, ..., x_d)$, and $\delta = (\delta_1, ..., \delta_d)$ denote d-vectors and define the grid point g_j by

$$g_j = j\delta \equiv (j_1 \delta_1, ..., j_d \delta_d).$$

Put

$$w_j^d(x, \delta) = \prod_{i=1}^{d} w_{ij_i}(x_i, \delta_i),$$

the "product" binning rule based on the w_{ij}. A d-variate binning rule amounts to distributing a data-vector $X = (X_1, ..., X_d)$ among grid-points in such a way that the amount $w_j^d(X, \delta)$ is assigned to g_j. If W_i is the binning kernel associated with w_{ij} then $W^d(x) = \prod_i W_i(x_i)$ is the product binning kernel corresponding to w_j^d.

Likewise, d-variate kernel estimators may be defined multiplicatively. Let K be a k-th order univariate kernel, as defined in Section 2.2; let $h = (h_1, ..., h_d)$ denote a vector of bandwidths and

$$K_h^d(x) = K(x_1/h_1) \cdots K(x_d/h_d)/(h_1 \cdots h_d)$$

denote scalings of the d-variate product kernel $K^d(x) = \prod_i K(x_i)$ by h; and given a random sample $X_1, ..., X_n$ from a d-variate density f, define

$$\hat{f}(x; h) = n^{-1} \sum_{i=1}^{n} K_h^d(x - X_i).$$

The binned version of this estimator is given by

$$\tilde{f}(x; h) = n^{-1} \sum_{j \in \mathbf{Z}^d} N_j K_h^d(x - j\delta),$$

where $N_j = \sum_{i=1}^{n} w_j^d(X_i, \delta)$.

We shall show in the next subsection that formulae for bias, variance, and mean squared error in the multivariate case are straightforward analogues of their counterparts in a univariate setting. Therefore, much of the discussion in Section 3 of the effects of different binning rules and different kernel functions, is applicable without change in the multivariate case. For example, if the Epanechnikov kernel is employed then, to ensure that binning does not have a significant effect on bias or mean squared error, each bin dimension (i.e. each δ_i) should be of smaller order than the square of the corresponding bandwidth (i.e. h_i^2). Likewise, nonstandard polynomial binning rules can increase the value of s and so, depending on the order and smoothness of the kernel, reduce the effect of binning on bias or mean squared error.

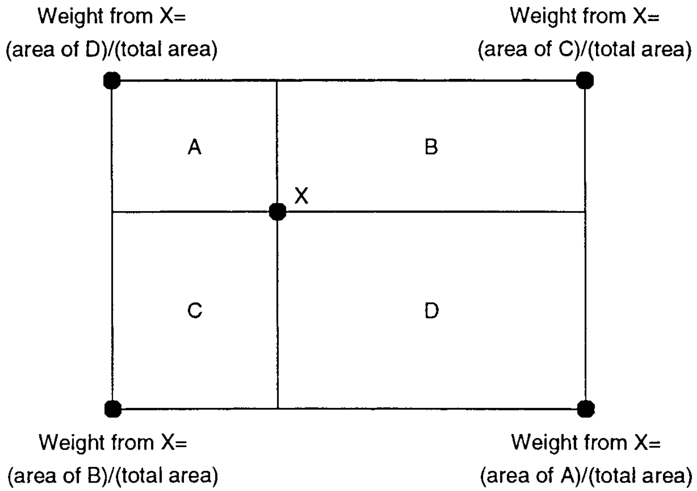

FIG. 2. Diagrammatic representation of the common linear binning rule when $d=2$. The data value at a point X is divided amongst neighbouring grid points according to the relative areas of opposite rectangles.

It is worth noting the bivariate version of the common linear binning rule. Given a data value X, construct the rectangle $\mathscr{R}$ with vertices corresponding to the grid points that neighbour X. Divide $\mathscr{R}$ into four subrectangles by inscribing lines through X, parallel to the sides of $\mathscr{R}$ to form four subrectangles. The weight assigned by X to a neighbouring grid point is the ratio of the area of the opposite subrectangle to the area of $\mathscr{R}$. This is illustrated in Fig. 2. The d-variate generalization of this rule, based on the relative contents of the 2^d rectangular prisms induced by X, is obvious.

5.2. *Effect of Binning on Bias, Variance and Mean Squared Error*

For the sake of simplicity we shall assume that the same binning rule, w_j, is used for each coordinate. That is, $w_{ij}=w_j$, and (with j now denoting a d-vector) $w_j^d(x, \delta)=\prod_i w_{j_i}(x_i, \delta_i)$. Let the univariate binning rule and univariate kernel K have the properties ascribed to them in Sections 2.2 and 3.1. In particular, the functions K, $K^{(1)}$, $K^{(3)}$, ..., $K^{(2t-3)}$ are continuous, and $K^{(2t-1)}$ has jump discontinuities. The multivariate analogue of formula (3.2), describing the difference between the expected values of binned and unbinned estimators, is

$$E\{\tilde{f}(x;h)\}-E\{\hat{f}(x;h)\}=c_s\sum_{i=1}^{d}\delta_i^s\,(\partial/\partial x_i)^s f(x)\sum_{i=1}^{d}(\delta_i/h_i)^{2t}\,a_i(x,\delta_i,h_i)$$
$$+o\left[\sum_{i=1}^{d}\{\delta_i^s+(\delta_i/h_i)^{2t}\}\right],$$

assuming that each $\delta_i = o(h_i)$. The multivariate analogue of (3.4), describing the bias of the unbiased estimator, is well known; it is

$$E\{\hat{f}(x; h)\} - f(x) = \kappa \sum_{i=1}^{d} h_i^k (\partial/\partial x_i)^k f(x) + o\left(\sum_{i=1}^{d} h_i^k\right)$$

(see Scott, 1992, p. 150). Formula (3.10), expressing the variance of $\tilde{f}(x; h)$, holds as before, provided $\int K^2$ is replaced by $(\int K^2)^d$. The d-variate analogue of (3.11), describing mean integrated squared error, is now obvious.

Analogues of the results described in Sections 3.2 and 3.3, for very smooth and very rough kernels respectively, are also straightforward. In particular, for very smooth kernels such as the standard normal or Student's t, (5.1) simplifies to

$$E\{\tilde{f}(x; h)\} - E\{\hat{f}(x; h)\} = c_s \sum_{i=1}^{d} \delta_i^s (\partial/\partial x_i)^s f(x) + o\left(\sum_{i=1}^{d} \delta_i^s\right).$$

5.3. *Approximation Accuracies of Multivariate Binning Rules*

The results of Section 4 can also be extended to the multivariate setting. For example, the d-variate extension of (4.1), for simple binning in each direction, is

$$E\int (\tilde{f} - \hat{f})^2 = (12n)^{-1} \int (K')^2 \left(\int K^2\right)^{d-1} \sum_{i=1}^{d} \delta_i^2 h_i^{-3} \prod_{j \neq i} h_j^{-1} + o\left(\sum_{i=1}^{d} \delta_i^2\right)$$

while the d-variate extension of (4.2), for common linear binning in each direction, is

$$E\int (\tilde{f} - \hat{f})^2 = \sum_{i=1}^{d} \delta_i^4 \left[(120n)^{-1} \int (K'')^2 \left(\int K^2\right)^{d-1} h_i^{-5} \prod_{j \neq i} h_j^{-1} \right.$$
$$\left. + \tfrac{1}{144}(1 - n^{-1}) \int \left\{\left(K''_{h_i} \prod_{j \neq i} K_{h_j}\right) * f\right\}^2 \right] + o\left(\sum_{i=1}^{d} \delta_i^4\right).$$

6. Minimum Grid Size Calculations

In practice $\tilde{f}$ is usually computed over a finite grid on an interval $[a, b]$ containing the data. Let

$$M = (b - a)/\delta + 1$$

be the number of grid points, a quantity that we shall refer to as the *grid size*. Since the amount of computing required for computation of $\tilde{f}$ varies directly with the grid size an issue of great practical relevance is that of how large a grid size is needed for the error due to binning to be "negligible" in some sense. It should be noted that there is no absolute answer to this question, since the amount of binning required to achieve a certain accuracy can be made arbitrarily large by choosing a density that is sufficiently wiggly and a sample size that is sufficiently large. Nevertheless, it is useful to determine minimum grid sizes for a selection of practical situations.

A convenient way of measuring the error due to binning is through the *relative*, *mean inteqrated*, *squared error*,

$$\text{RMISE} = E\int \{\tilde{f}(\cdot\,; h_0) - \hat{f}(\cdot\,; h_0)\}^2 / E\int \{\hat{f}(\cdot\,; h_0) - f\}^2,$$

where h_0 is the bandwidth that minimizes $E\int(\hat{f}-f)^2$. Observe that RMISE is the ratio of a distance measure between $\tilde{f}$ and $\hat{f}$ to that between $\hat{f}$ and f. Having RMISE equal to a small numbering α, such as $\alpha = 0.01$ or $\alpha = 0.001$, corresponds to the desirable situation where binning has a small effect on the overall error involved in the estimation process.

To determine minimum grid sizes we appeal to the "small δ" results derived in Section 4. In the case of simple binning, this involves the approximation

$$E\int \{\tilde{f}(\cdot\,; h_0) - \hat{f}(\cdot\,; h_0)\}^2 \simeq \delta^2 \int (K')^2/(12nh_0^3).$$

This leads to

$$M^*(\alpha) = \left\lceil (b-a)\left[\int (K')^2 \Big/ \left\{12\alpha n h_0^3 E\int \{\tilde{f}(\cdot\,; h_0) - f\}^2\right\}\right]^{1/2} + 1\right\rceil$$

(where $\lceil x \rceil$ is the smallest integer greater than or equal to x) being the smallest grid size required to approximately ensure that RMISE $\leqslant \alpha$. Similarly, the approximation (4.2) for common linear binning leads to

$$M^*(\alpha) = \Bigg\lceil (b-a)\Bigg[\Bigg\{\int (K'')^2/(120nh_0^5) + \tfrac{1}{144}(1-n^{-1}) \\ \times \int (K''_{h_0} * f)^2\Bigg\} \Big/ \left\{\alpha E\int \{\hat{f}(\cdot\,; h_0) - f\}^2\right\}\Bigg]^{1/4} + 1\Bigg\rceil.$$

Table I lists values of $M^*(0.01)$ for simple and common linear binning rules for each of the fifteen normal mixture densities in Marron and Wand (1992) if binned density estimates are to be computed over $[a, b] = [-3, 3]$ with K equal to the standard normal density. Sample

TABLE I

Minimum Grid Sizes to Achieve 1% Approximate Relative MISE for 15 Example Normal Mixture Densities

Sample size	$n=100$		$n=1000$		$n=10000$	
Density	Simple	Linear	Simple	Linear	Simple	Linear
1	31	16	47	25	71	39
2	32	20	49	32	76	50
3	133	71	268	145	462	253
4	145	77	257	139	419	230
5	276	139	425	224	659	355
6	33	18	55	30	86	47
7	47	25	75	41	118	65
8	38	20	69	38	113	62
9	34	18	63	34	107	59
10	115	62	217	118	359	198
11	30	17	39	25	93	63
12	49	28	148	82	376	206
13	29	17	94	53	214	127
14	67	39	229	127	595	325
15	53	33	255	139	469	257

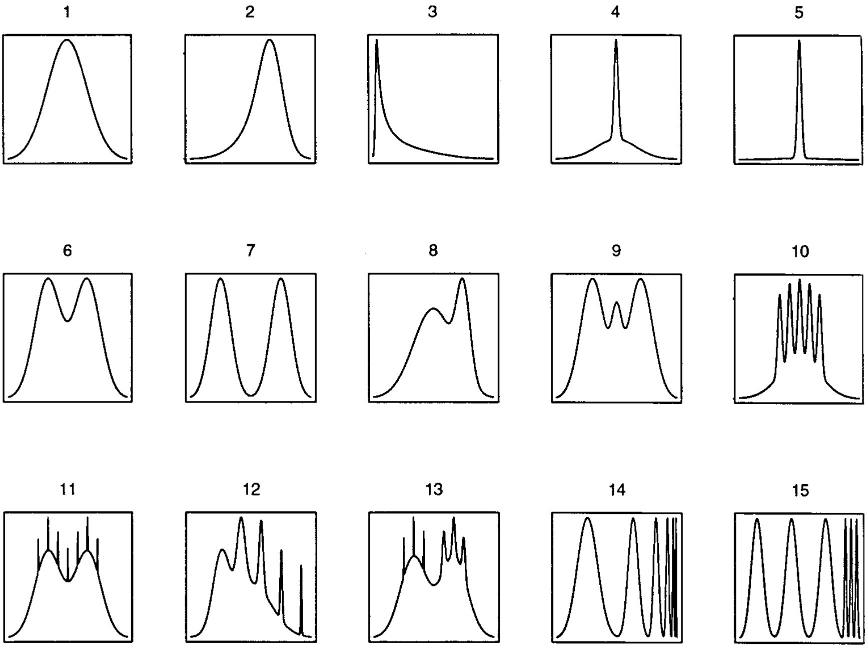

FIG 3. The fifteen example normal mixture densities.

sizes are $n=100$, $n=1000$ and $n=10000$. For convenience, the densities are plotted in Fig. 3. While the measure of accuracy RMISE $\simeq 1\%$ is somewhat arbitrary, it does represent a situation where binning has little effect on overall error and allows some important insight into the appropriate choice of the grid size.

The minimum grid sizes in Table I allow a direct comparison of simple and linear binning strategies. It is seen that simple binning requires between about 30% and 70% more grid points to achieve the same accuracy as linear binning. Therefore, linear binning has a clear-cut advantage over simple binning if economy of number of grid points is desirable. The results also give an indication of how many grid points one should use in practice. Surprisingly few grid points are required to achieve the prescribed level of accuracy for many of the "smoother" densities in Fig. 3; however this number increases considerably for densities having more "fine structure"—as well as for larger sample sizes since the smaller optimal bandwidths demand a finer mesh. For linear binning our results show that grid sizes of about $400-500$ are adequate for a wide range of practical situations.

APPENDIX A: Proof of (3.2)

In view of (3.1),

$$\begin{aligned} E\{\tilde{f}(x;h)\} &= I_0 + c_1\delta I_1 + c_2\delta^2 I_2 + \ldots \\ &= I_0 + c_s\delta^s I_s + o(\delta^s), \end{aligned} \tag{A.1}$$

where $I_i = \delta\sum_j g_i(j\delta)$ and $g_i(y) = f^{(i)}(y)\,K_h(x-y)$. Now, $g_0, g_0^{(1)}, ..., g_0^{(2t-3)}$ exist and are continuous on $(-\infty, \infty)$, and $g_0^{(2t-1)}$ is piecewise continuous, with total jump discontinuity given by $J(g_0^{(2t-1)})$. Noting Euler–Maclaurin summation formulae (see, for example, Abramowitz and Stegun, 1965, p. 886; Milne-Thomson, 1933, p. 187) we may deduce that if x is a grid point and if the jumps occur at points y such that hy is a grid point then

$$\begin{aligned} I_0 &= \int g_0 - \{1+o(1)\}\{(2t)!\}^{-1}\,B_{2t}\delta^{2t}J(g^{(2t-1)}) \\ &= E\{\hat{f}(x;h)\} + \{(2t)!\}^{-1}\,B_{2t}\delta^{2t}h^{-1}(-h^{-1})^{2t-1} \\ &\quad \times f(x)\,J(K^{(2t-1)}) + o(\delta^{2t}h^{-2t}). \end{aligned} \tag{A.2}$$

Furthermore, $I_s = f^{(s)}(x) + o(1)$, so by (A.1),

$$E\{\tilde{f}(x;h)\} = I_0 + c_s\delta^s f^{(s)}(x) + o(\delta^s). \tag{A.3}$$

Results (3.2) and (3.3) follow on combining (A.2) and (A.3).

We treat other cases by approximating g_0 as follows. Let g denote a function that is identical to g_0 on all bins, where g_0^{2t-1} has no discontinuities, and equal to g_0 at all bin endpoints. Within the closure of those bins, where g_0 has a discontinuity, let $g^{(2t-1)}$ denote a step function with jumps only at the bin ends. The sizes of those jumps are uniquely determined by the fact that g has to agree with g_0 at the ends. Since $g^{(2t-1)}$ is a step function then the total jump of $g^{(2t-1)}$ on any one of those bins, where $g_0^{(2t-1)}$ has a discontinuity equals that of $g_0^{(2t-1)}$ there, plus $o(1)$. Since g has its jumps at grid points then (3.2) applies to it; and since g is identical to g_0 at all grid points then the value of I_0 is not affected if g_0 in its definition is replaced by g. It follows that

$$I_0 = \int g - \{1 + o(1)\} \{(2t)!\}^{-1} B_{2t}(\delta/h)^{2t} f(x) \, J(K^{(2t-1)}).$$

Of course, we wish to replace $\int g$ on the right-hand side by $\int g_0$. The difference, $\int (g - g_0)$, reduces to the integral of $g - g_0$ over the union of those bins, where g_0 has a discontinuity. By Taylor expansion we may simplify this to the integral of $d_t = g^{(2t-1)} - g_0^{(2t-1)}$ against a polynomial of degree $2t-1$ which is of size δ^{2t-1} on each bin. Now, the function d_t is of size h^{-2t}, and so the required integral is of size $\delta\delta^{2t-1}h^{-2t} = (\delta/h)^{2t}$. Its exact value may be worked out in specific cases.

APPENDIX B: Proof of (3.7)

The argument in Appendix A remains valid, except that an alternative formula should replace (A.2). To derive that formula let j_1 and j_2 have the meanings ascribed to them in Section 3.3, and observe that

$$\begin{aligned}
I_0 &= \delta \sum_j g_0(j\delta) = \tfrac{1}{2}\,\delta h^{-1} \sum_{j=j_1}^{j_2} f(j\delta) \\
&= \tfrac{1}{2}\,\delta h^{-1} [\tfrac{1}{2} f(j_1\delta) + f\{(j_1+1)\,\delta\} + \cdots + f\{(j_2-1)\,\delta\} + \tfrac{1}{2} f(j_2\delta)] \\
&\quad + \tfrac{1}{4}\,\delta h^{-1} \{f(j_1\delta) + f(j_2\delta)\} \\
&= \tfrac{1}{2}\,\delta h^{-1} \int_{j_1}^{j_2} f(u\delta)\,du + \tfrac{1}{2}\,\delta h^{-1} f(x) + o(\delta h^{-1}) \\
&= (2h)^{-1} \left(\int_{x-h}^{x+h} + \int_{j_1\delta}^{x-h} + \int_{x+h}^{j_2\delta} \right) f(y)\,dy + \tfrac{1}{2}\,\delta^{-1} f(x) + o(\delta h^{-1}) \\
&= E\{\hat{f}(x; h)\} + (2h)^{-1} \{(j_2 - j_1 + 1)\,\delta - 2h\} f(x) + o(\delta h^{-1}).
\end{aligned}$$

Formula (3.7) follows from this result and the argument in Appendix A.

Acknowledgments

We are grateful to Professor J. S. Marron for helpful comments and to Mr. Eugene Dubossarsky for programming assistance. This article has also benefited from the helpful comments of an associate editor and two referees.

References

Abramowitz, M., and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Comput. Graph. Statist.* **3** 35–56.

Gasser, T., Müller, H.-G., and Mammitszch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B.* **47** 238–252.

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM J. Appl. Math.* **42** 390–399.

Hall, P. (1983). Edgeworth expansions of the distribution of Stein's statistic. *Math. Proc. Cambridge Philos. Soc.* **93** 163–175.

Härdle, W., and Scott D. W. (1992). Smoothing by weighted averaging of rounded points. *Comput. Statist.* **7** 97–128.

Jones, M. C. (1989). Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.* **84** 733–741.

Jones, M. C., and Lotwick, H. W. (1983). On the errors involved in computing the empirical characteristic function. *J. Statist. Comput. Simul.* **17** 133–149.

Jones, M. C., and Lotwick, H. W. (1984). Remark ASR50. A remark on algorithm AS176. Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **33** 120–122.

Marron J. S., and Wand, M. P. (1992). Exact mean integrated squared error. *Appl. Statist.* **20** 712–736.

Milne-Thomson, L. M. (1933). *The Calculus of Finite Differences*. Macmillan, London.

Scott, D. W. (1985). Average shifted histograms: Effective nonparametric density estimators in several dimensions. *Ann. Statist.* **13** 1024–1040.

Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley, New York.

Scott, D. W., and Sheather, S. J. (1985). Kernel density estimation with binned data. *Comm. Statist. Theory Methods* **14** 1353–1359.

Silverman, B. W. (1982). Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **31** 93–99.

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

Wand, M. P. (1994). Fast computation of multivariate kernel estimators. *J. Comput. Graph. Statist.* **3** 433–445.