Austral. J. Statist., **32**(3), 1990, 361–371

VARIABLE KERNEL DENSITY ESTIMATES AND VARIABLE KERNEL DENSITY ESTIMATES

M.C. JONES[1]

The Open University, England

Summary

The term "variable kernel density estimate" is sometimes used to mean a kernel density estimate employing a different bandwidth for each data point, and sometimes to denote a kernel density estimate with bandwidth a function of estimation location. This expository article stresses the importance of the distinction between these two definitions, both via an introductory description of the ideas involved and in terms of their comparative theoretical performance.

Key words: Local estimation; mean squared error; smoothing; varying bandwidths.

1. Introduction

The title of this expository article refers to two distinct modifications of the kernel density estimate (KDE), each of which is sometimes referred to in the literature as the "variable" KDE. Our objective is to alleviate any confusion caused by this and to stress the importance of the distinction between the two.

The context is the familiar one of nonparametric density estimation: we have available a random sample $X_1, \ldots, X_n$, of size n, from some distribution and wish to obtain an estimate of its density f, say. Our starting

Received January 1990; revised March 1990.

[1] Statistics Dept., The Open University, Walton Hall, Milton Keynes, England MK7 6AA.

Acknowledgements. This work was started at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. The author thanks the referees for interesting comments, and one referee and Matt Wand for additional references.

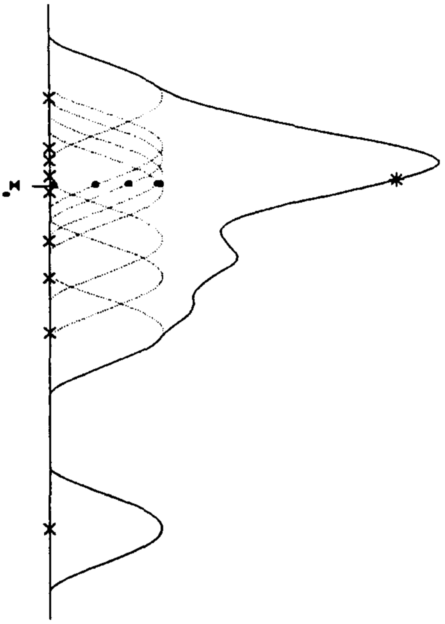

Fig. 1.—How $\hat{f}$ is made up.
The artificial data $X_1, \ldots, X_n$ are the crosses on the axis; here $n = 9$. The dotted curves are the individual kernels, $h^{-1}K\{(x - X_i)/h\}$, $i = 1, \ldots, n$. Non-zero values of these functions at the marked point x_0 are highlighted by bullets; the star marks their sum, $n\hat{f}(x_0)$. The solid line is the entire constant KDE $\hat{f}$.

point is the well-known KDE (see, for example, Silverman, 1986) given by

$$\hat{f}(x_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\Big(\frac{x_0 - X_i}{h}\Big). \tag{1}$$

Here, $\hat{f}(x_0)$ is an estimate of f at some point x_0. Two quantities in formula (1) are at the user's discretion. One is the symmetric "kernel function" K, the other the smoothing parameter, or "bandwidth", h. With these specified, Figure 1 demonstrates how $\hat{f}$ is made up, indeed for any point x (the solid line is a scaled-up version of $\hat{f}$), but with particular reference to x_0. Note that all three figures in this paper are purely illustrative; the X_is— which remain the same throughout — are chosen for convenience rather than being random, there are only $n = 9$ of them, and we have no intention of comparing our density estimators with a "true f" on the basis of these pictures! The KDE works by averaging out the dotted "bumps" in Figure 1, which are located at the data points; at x_0, the bullets emphasise the non-zero contributions to the total which is indicated by the star, and which, for clarity, is the value $n\hat{f}(x_0)$ rather than $\hat{f}(x_0)$ itself. (The solid $n\hat{f}$ coincides with the dotted kernels in places, especially about the largest X_i because of its distance from the remaining data). Here, K happens to be the biweight, or beta(3, 3), probability density function

$$K(x) = I_{[-1,1]}(x)15(1-x^2)^2/16,$$

where $I_A(x)$ is the indicator function (1 if $x \in A$, 0 otherwise). The bandwidth h controls the scale of the kernel K and hence the smoothness of the estimate $\hat{f}$ in a straightforward way, as illustrated by Silverman (1986, §2.4). The important point here is that h is a fixed constant which results in fixed width bumps at every data value and which is taken to be the same for all locations x at which $\hat{f}$ is evaluated. For this reason, it will be convenient to call $\hat{f}$ a "constant" KDE in what follows.

The two variable KDEs are described and contrasted in Section 2. The key idea is to allow h to be other than constant. The first type, which we call a "varying" KDE, allows h to vary with each data point X_i; the second, which we refer to as a "local" KDE is quite different, h being allowed to depend only on estimation location x_0. In Section 3, we address theoretical mean squared error (MSE) performance of the two. It turns out that the varying KDE behaves rather better than the local KDE in these terms, at least asymptotically. Various related topics are briefly addressed in Section 4; these include the practical status of these ideas, other modifications to the basic constant KDE, and a combined variable kernel idea.

The alternative name "adaptive" KDEs, proposed by Silverman (1986, §5.3) for the varying KDE and used in the regression context by various authors for the analogue of the local type, will be avoided because it is also used to mean constant KDEs employing a (single) data-based choice of bandwidth! This author does, however, have some sympathy with the terminology variable (or local) *bandwidth* density estimators.

2. The Two Methods

The general formula for the varying KDE is

$$\hat{f}_V(x_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(X_i)} K\left(\frac{x_0 - X_i}{h(X_i)}\right). \tag{2}$$

The only change from the constant KDE given by (1) is clear: instead of the single bandwidth h, $\hat{f}_V$ employs a different bandwidth $h(X_i)$ for each data

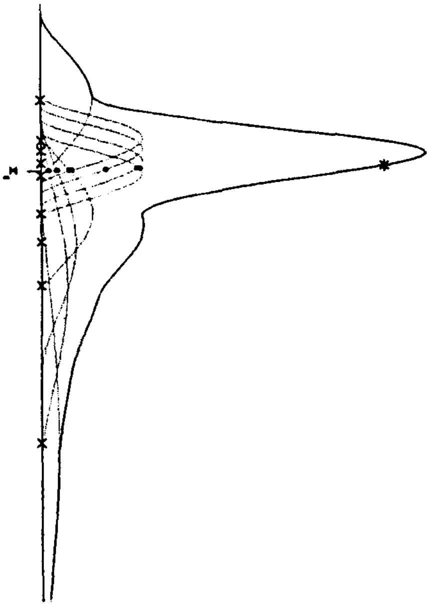

Fig. 2.—How $\hat{f}_V$ is made up.

The crosses on the axis are the same $X_1, \ldots, X_n$ as in Fig. 1. The dotted kernel curves are $h(X_i)^{-1}K\{(x - X_i)/h(X_i)\}$, $i = 1, \ldots, n$. Non-zero values of these functions at x_0 are highlighted by bullets; the star marks their sum, $n\hat{f}_V(x_0)$. The solid line is the entire varying KDE $\hat{f}_V$.

point. The motivation for this is "that a natural way to deal with long-tailed densities is to use a broader kernel in regions of low density. Thus an observation in the tail would have its mass smudged out over a wider range than one in the main part of the distribution" (Silverman, 1986, p.100). The idea is illustrated in Figure 2, which shows kernel functions and how they contribute to the total $n\hat{f}_V(x_0)$ in the same way as Figure 1 does for the constant KDE. Note that $\hat{f}_V(x_0)$ is still made up by averaging kernel bumps evaluated at x_0, but the bumps are of differing widths depending on the locations of their centres, $\{X_i, i = 1, \ldots, n\}$. The above intuitive notion of how $h(X_i)$ should be specified can be translated to saying that $h(X_i)$ should vary inversely with the (true) density f. In fact, theoretical work of Abramson (1982) shows that $h(X_i) \propto f^{-1/2}(X_i)$ is a good choice. Of course, $f^{-1/2}$ is unknown and an obvious practical strategy is to replace it with an appropriate "pilot estimate". Silverman (1986, §5.3) develops this approach for practical use. The $h(X_i)$ s in Figure 2 are, however, made up for illustrative convenience.

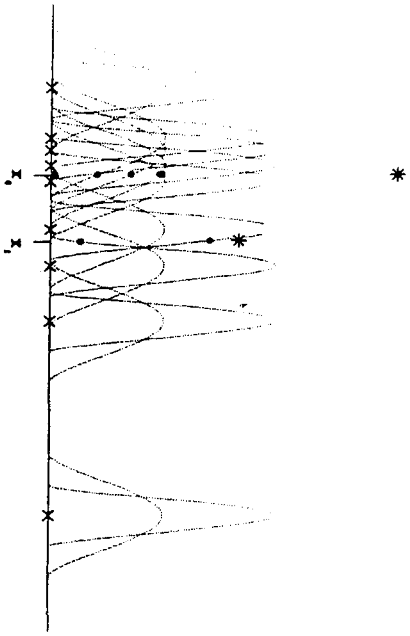

Fig. 3.—How $\hat{f}_L$ is made up.

The crosses on the axis are the same $X_1, \ldots, X_n$ as in Figs 1 and 2. The wider dotted kernel curves are the same as those in Fig. 1, now being thought of as $h(x_0)^{-1}K\{(x - X_i)/h(x_0)\}$, $i = 1, \ldots, n$. The narrower dotted kernel curves superimposed on these are plots of $h(x_1)^{-1}K\{(x - X_i)/h(x_1)\}$, $i = 1, \ldots, n$. At x_0, the same quantities are marked as in Fig. 1, with the total now denoted by $n\hat{f}_L(x_0)$; at the second point x_1, analogous quantities are marked for the estimate $\hat{f}_L(x_1)$ at x_1.

By way of contrast, the general formula for the local KDE is

$$\hat{f}_L(x_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_0)} K\left(\frac{x_0 - X_i}{h(x_0)}\right), \tag{3}$$

that is, h is thought of not as a single constant nor as a function of individual X_i s, but as a function of the point x_0 where f is being estimated. But at x_0, $h(x_0)$ is just a particular choice of single smoothing parameter — and the picture appropriate for $\hat{f}_L$ at x_0 is precisely the same as for $\hat{f}$ at x_0, namely Figure 1 without the solid line! The clear conceptual gulf between varying and local KDEs is therefore already apparent.

Globally, the formula (2) for $\hat{f}_V$ is equally applicable to points other than x_0, so the entire density estimate is given immediately by the solid line in Figure 2 (divided by n). Note that, if K is a probability density function, as here, then so is $\hat{f}_V$ as well as $\hat{f}$. The local KDE, on the other hand, differs from the constant KDE at points other than x_0, and is thus not given by the solid line in Figure 1. Rather, the overall local KDE is made up of a continuum of individual constant KDEs with different bandwidths, one for each x. This is illustrated in Figure 3, where $\hat{f}_L(x_0)$ and $\hat{f}_L(x_1)$, for $x_0 \neq x_1$, are calculated from two entirely different constant KDEs with bandwidths $h(x_0)$ and $h(x_1)$, respectively. In Figure 3, the wider dotted kernels and their contributions and sum at x_0 are as in Figure 1, while the narrower dotted kernels, still centred at the same locations of course, but with bandwidth $h(x_1)$ which we have taken to be less than $h(x_0)$, add up as shown to $n\hat{f}_L(x_1)$ at x_1. The theoretical work of Section 3 tells us that it would be best to take $h(x_0) \propto [f(x_0)/\{f''(x_0)\}^2]^{1/5}$ (when $f''(x_0) \neq 0$) i.e., that each individual bandwidth should be chosen to respond inversely to the (normalised) curvature, or "roughness", of the (true) density at the current point. The most well-known implementation of a local KDE is not usually thought of in this way directly, but takes instead the kth nearest neighbour distance, $r_k(x_0)$, say, from the data points to x_0 as $h(x_0)$, where k is now a smoothing parameter to be specified. A property of such nearest neighbour density estimates (see Silverman, 1986, §5.2) is that they do not integrate to one (but see the fix-up of Burman & Nolan, 1989) and neither does $\hat{f}_L$ in general. Incidentally, the original proposals of Victor (1976) and Breiman, Meisel & Purcell (1977) for a varying KDE took $h(X_i) \propto r_k(X_i)$.

3. Asymptotic Performance of the Methods

So, what implications does all this have for performance of the various density estimators? It is instructive to consider MSEs and, in particular, leading terms in the asymptotic expansions thereof (as $n \to \infty$, $h = h(n) \to 0$ such that $nh \to \infty$, and assuming f is smooth enough and K well-behaved enough for the formulae to hold). The expression for the MSE of the constant KDE $\hat{f}$ at x_0 is well-known (e.g., Silverman, 1986, §3.3.1):

$$\mathrm{MSE}(\hat{f}(x_0)) \simeq \tfrac{1}{4}\sigma_K^4 h^4 \{f''(x_0)\}^2 + (nh)^{-1} R(K) f(x_0). \qquad (4)$$

Here, $\sigma_K^2 = \int x^2 K(x)\,dx$ and $R(g)$ in general denotes $\int g^2(x)\,dx$. The $O(h^4)$ term is due to the squared bias of $\hat{f}$ and the $O((nh)^{-1})$ term to its variance. A consequence of the simplicity of (4) as a function of h is that we can easily work out the optimal value of h in the sense of minimising MSE at x_0; call it $h_0(x_0)$. In particular, $h_0(x_0) \propto n^{-1/5}$ and the corresponding minimal value of the MSE at x_0 is $O(n^{-4/5})$. Note, however, the dependence of formula (4), and hence of $h_0(x_0)$, on the true f through $\{f''(x_0)\}^2$, which we are assuming is non-zero for the moment, and through $f(x_0)$ itself.

For the varying KDE, $\hat{f}_V$, introduce the constant of proportionality, h_V, say, so that $h(X_i) = h_V\tilde{f}^{-1/2}(X_i)$. Here, $\tilde{f}$ refers to the pilot estimate of f and h_V plays the role of a single overall smoothing parameter analogous to h above. Then, Hall & Marron (1988) provide a careful development of the corresponding MSE expression for $\hat{f}_V$. Provided $\tilde{f}$ is a good enough estimate of f, it is

$$\mathrm{MSE}(\hat{f}_V(x_0)) \simeq \tfrac{1}{576}\delta_K^2 h_V^8 A_f^2(x_0) + (nh_V)^{-1}S(K)f^{3/2}(x_0)\,, \qquad (5)$$

where $A_f(x_0)$ is $(d^4/dx^4)\{1/f(x)\}$ evaluated at $x = x_0$, $\delta_K = \int x^4K(x)dx$ and $S(K) = \frac{3}{2}R(K) + \frac{1}{4}R(xK')$. Recognition that the f-dependent quantity A_f in the squared bias term of (5) has this simple form is new; it is not difficult to check that it does equal the complicated looking expressions given for it by Silverman (1986, p.105) and Hall & Marron (1988, p.41). See also the important paper of Hall (1990). Clearly, dependence on f and K in (5) is rather different from that in (4). Most important, however, is the fact that $\mathrm{MSE}(\hat{f}_V(x_0))$ can be made to converge to zero at a faster rate than can $\mathrm{MSE}(\hat{f}(x_0))$. This is due to the squared bias being $O(h_V^8)$ in (5) but $O(h^4)$ in (4), while the variance remains of order $(n\bar{h})^{-1}$ for both (where $\bar{h}$ denotes h_V or h as appropriate). It follows that if we choose h_V $(= h_V(x_0))$ optimally according to (5), it will be proportional to $n^{-1/9}$ and the optimal MSE of $\hat{f}_V(x_0)$ will be $O(n^{-8/9})$. This compares favourably with the $O(n^{-4/5})$ rate associated with $\hat{f}$.

This theoretical superiority of $\hat{f}_V$ at, recall, the single point x_0, translates immediately to the same superiority of $\hat{f}_V$ over $\hat{f}_L$ at x_0, of course, because of the correspondence between $\hat{f}_L$ and $\hat{f}$ described above. Therefore, at least for large enough sample sizes, the varying KDE outperforms the local KDE, even at a single point.

To compare the performance of $\hat{f}$, $\hat{f}_V$ and $\hat{f}_L$ as estimates of f at all points x simultaneously, it is natural to extend the above to consideration of the integrated MSE (IMSE). A little care needs to be taken with this, however. Simply integrating expression (5) for $\mathrm{MSE}(\hat{f}_V)$, we find that $\int A_f^2(x)\,dx$ will very often be infinite. To get around this, we might consider a weighted version of IMSE (strictly IWMSE, using an obvious acronym) where the weight is some suitable function of f, or we might restrict ourselves to some finite range on which f is of interest (a special case of IWMSE) and over which the integral is finite. For more on this, see Hall (1989). With this proviso, a globally optimal value of $h_V \propto n^{-1/9}$ can still be calculated and the corresponding version of IMSE will still be $O(n^{-8/9})$.

Integrating (4) gives the familiar expression for the IMSE of the constant KDE, $\hat{f}$, provided, as is usually the case, $R(f'') < \infty$. The globally optimal value for h follows easily; it is

$$h_0 \simeq \{\sigma_K^{-4} R(K)\}^{1/5} R(f'')^{-1/5} n^{-1/5}.$$

The resulting best possible IMSE is then given by

$$\mathrm{IMSE}_0(\hat{f}) \simeq \tfrac{5}{4}\{\sigma_K R(K)\}^{4/5} R(f'')^{1/5} n^{-4/5}. \tag{6}$$

(All this is as in Silverman, 1986, §3.3.2). Because h_0 depends on $R(f'')$, $\mathrm{IMSE}_0(\hat{f})$ is not achievable in practice, but it provides a useful lower bound on how well it is possible to do with a constant KDE. Note that $\mathrm{IMSE}_0(\hat{f})$ is, of course, also $O(n^{-4/5})$, so $\hat{f}_V$ remains asymptotically superior to $\hat{f}$.

The equivalent best possible version of $\hat{f}_L$ arises by using $h_L(x) \equiv h_0(x)$ for all x. Now,

$$h_0(x_0) \simeq \{\sigma_K^{-4} R(K)\}^{1/5} [f(x_0)/\{f''(x_0)\}^2]^{1/5} n^{-1/5}.$$

(Again, we skate over the issue of points x where $f''(x) = 0$ since it is not central to our arguments; in fact, as Schucany (1989) notes, at such x s, $\mathrm{MSE}(\hat{f}_L) = O(h^8 + (nh)^{-1})$ as is $\mathrm{MSE}(\hat{f}_V)$, unless $f^{iv}(x)$, too, is zero!). (Since $r_k(x)$ estimates $1/f(x)$, the above would only suggest use of a nearest neighbour density estimate when $f''(x) \propto f^3(x)$ for all x !). Pointwise, the corresponding $\mathrm{MSE}_0(\hat{f}_L(x_0))$ is

$$\mathrm{MSE}_0(\hat{f}_L(x_0)) \simeq \tfrac{5}{4}\{\sigma_K R(K)\}^{4/5}\{f(x_0)\}^{4/5}\{f''(x_0)\}^{2/5} n^{-4/5}.$$

Integrating this gives the best possible IMSE of $\hat{f}_L$; it is

$$\mathrm{IMSE}_0(\hat{f}_L) \simeq \tfrac{5}{4}\{\sigma_K R(K)\}^{4/5} R(\{f^2 f''\}^{1/5})\, n^{-4/5}. \tag{7}$$

Necessarily, this quantity is $O(n^{-4/5})$, so asymptotically $\hat{f}_L$, too, is inferior, globally, to $\hat{f}_V$. The local KDE does, however, improve on $\hat{f}$ in the sense that it has, by construction, a smaller constant multiplier of $n^{-4/5}$. It is certainly conceivable that the amount of improvement attained by $\hat{f}_L$ over $\hat{f}$ could be quite substantial at times.

How do $\hat{f}_V$ and $\hat{f}_L$ improve on $\hat{f}$? It appears that $\hat{f}_V$ provides a "first order" improvement (a better asymptotic rate of convergence) by responding to differences truly in the local density of the X_i s. On the other hand, $\hat{f}_L$ is concerned with variations in the local roughness of the density function, and this seems to be important only (by improving constants) in a "second order" sense.

4. Further Remarks

4.1. Practice

The practical status of these variable KDEs is not as yet entirely clear. The varying KDE, $\hat{f}_V$, is the nearer to being available for immediate application. Silverman (1986, §5.3) implements $\hat{f}_V$ and reports success in employing a pilot estimate to obtain $\tilde{f}^{-1/2}(X_i)$, $i = 1, \ldots, n$. Indeed, for prechosen h_V, Silverman displays examples which show considerable practical potential for the method. All that is missing — and in exploratory situations this is not always important — is a good technique for the data-based selection of the bandwidth h_V. Silverman (1986) and Hall (1989) adapt least squares cross-validation for this purpose, but there seems to be a case for providing better choices of h_V, possibly along the lines of recent work on selecting h in the constant KDE (e.g., Sheather & Jones, 1990).

Adaptive choice of $h(x_0)$, and hence of $h_L(x_0)$, for estimating f at the single point x_0 has also been addressed at times (e.g., Sheather, 1986). We are not aware of any work investigating extension of this to choice of the entire function $h_L(x)$. Indeed, it is not a foregone conclusion that use of the usual asymptotic MSE expansion for $\hat{f}_L$ will be successful. See, for instance, Schucany's (1989) exposition of how to avoid the $f''(x) = 0$ problem, and note that practical implementation of an approach based on this would need to involve estimation of f^{iv} and even f^{vi}.

4.2. Other Alternatives

There are other (potentially important) modified KDEs. Two are especially noteworthy. The first employs a kernel K which is not a probability density function, but takes negative values in places. Such "higher order" kernels can be chosen (sufficient smoothness of f permitting) to change the asymptotic bias order from the basic h^2 to any higher even power of h. KDEs involving "fourth order" kernels, which have $\sigma_K^2 = 0$ but $\delta_K \neq 0$, thus have the same order of magnitude of asymptotic MSE as does $\hat{f}_V$. This is all very well in theory, but work of Marron & Wand

(1990) indicates that very large samples are often needed before this theoretical advantage translates to practice. By the way, Gajek (1986) deals interestingly with the negative $\hat{f}$ problem arising here.

A second alternative is the use of transformations. Wand, Marron & Ruppert (1990) investigate this and exhibit considerable promise for the method in certain situations. They noted that their estimate, $\bar{f}$, say, can be written as

$$\bar{f}(x_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_0)} K\Big(\frac{x_0 - X_i}{h(\eta_i)}\Big), \tag{8}$$

for some $x_0 \le \eta_i \le X_i$. Comparing (8) with (2) and (3), we might loosely say that $\bar{f}$ is intermediate between $\hat{f}_V$ and $\hat{f}_L$. Since $\bar{f}$ has $h(x_0)$ rather than $h(X_i)$ "outside" K, we might, just as loosely, think $\bar{f}$ is "nearer" to $\hat{f}_L$, and it is interesting to note that it shares with $\hat{f}_L$ the property of improving on $\hat{f}$ only, but perhaps substantially, in terms of constants and not rates. If a suitable transformation of $X_1, \ldots, X_n$ can be chosen, this method has the advantage of becoming a constant KDE (which is well understood and for which, for example, good adaptive choices of h exist) applied to transformed data.

Of course, there are additionally yet more density estimators which are not immediately related to the KDE at all; see Silverman (1986) for some of these. Combinations of any of the above ideas are also a possibility.

4.3. Variable Variable KDEs

We have treated $\hat{f}_V$ and $\hat{f}_L$ as distinct ideas, but if both are successful there is no reason not to combine them too. Thus, we might think of employing $\hat{f}_V$ at each x_0, but choosing $h_V = h_V(x_0)$ differently for each x_0, as indicated in the text following (5). The resulting estimate might be written

$$\hat{f}_{VL}(x_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{f}^{1/2}(X_i)}{h_V(x_0)} K\Big(\frac{\tilde{f}^{1/2}(X_i)(x_0 - X_i)}{h_V(x_0)}\Big)$$

which is the natural member of the general family

$$\hat{f}_\bullet(x_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_0, X_i)} K\Big(\frac{x_0 - X_i}{h(x_0, X_i)}\Big)$$

for further consideration.

References

ABRAMSON, I.S. (1982). On bandwidth variation in kernel estimates — a square root law. *Ann. Statist.* **10**, 1217–1223.

BREIMAN, L., MEISEL, W. & PURCELL, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19**, 135–144.

BURMAN, P. & NOLAN, D. (1989). Hybrid density estimators. Abstract in *IMS Bulletin* **18**, 271.

GAJEK, L. (1986). On improving density estimators that are not bona fide functions. *Ann. Statist.* **14**, 1612–1618.

HALL, P. (1989). On bandwidth selection for variable bandwidth density estimation. Manuscript.

HALL, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika* **77**, 529–536.

HALL, P. & MARRON, J.S. (1988). Variable window width kernel estimates of probability densities. *Probab. Theory Related Fields* **80**, 37–50.

MARRON, J.S. & WAND, M.P. (1990). Exact mean integrated square error. To appear.

SCHUCANY, W.R. (1989). Locally optimal window widths for kernel density estimation with large samples. *Statist. Probab. Lett.* **7**, 401–405.

SHEATHER, S.J. (1986). An improved data-based algorithm for choosing the window width when estimating the density at a point. *Comput. Statist. Data Anal.* **4**, 61–65.

SHEATHER, S.J. & JONES, M.C. (1990). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, to appear.

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

VICTOR, N. (1976). Non-parametric allocation rules (with discussion). In *Decision Making and Medical Care: Can Information Science Help?*, eds. F.T. de Dombal and F. Grémy, pp.515–529. Amsterdam: North-Holland.

WAND, M.P., MARRON, J.S. & RUPPERT, D. (1990). Transformations in density estimation. *J. Amer. Statist. Assoc.*, to appear.