Ann. Inst. Statist. Math.
Vol. 46, No. 3, 521–535 (1994)

VARIABLE LOCATION AND SCALE KERNEL DENSITY ESTIMATION

M. C. JONES[1], I. J. MCKAY[2] AND T.-C. HU[3]

[1]*Department of Statistics, The Open University, Milton Keynes MK7 6AA, U.K.*
[2]*Department of Statistics, University of British Columbia, 2021 West Mall, Vancouver, Canada V6T 1W5*
[3]*Department of Mathematics, National Tsing Hua University, Hsinchu, Taiwan 30043, R.O.C.*

(Received April 2, 1993; revised October 29, 1993)

Abstract. Variable (bandwidth) kernel density estimation (Abramson (1982, *Ann. Statist.*, **10**, 1217–1223)) and a kernel estimator with varying locations (Samiuddin and El-Sayyad (1990, *Biometrika*, **77**, 865–874)) are complementary ideas which essentially both afford bias of order h^4 as the overall smoothing parameter $h \to 0$, sufficient differentiability of the density permitting. These ideas are put in a more general framework in this paper. This enables us to describe a variety of ways in which scale and location variation may be extended and/or combined to good theoretical effect. This particularly includes extending the basic ideas to provide new kernel estimators with bias of order h^6. Technical difficulties associated with potentially overly large variations are fully accounted for in our theory.

Key words and phrases: Bias reduction, smoothing, variable bandwidth.

1. Introduction

The (constant bandwidth) kernel estimator

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} h^{-1} K\{h^{-1}(x - X_i)\} \tag{1.1}$$

of a (univariate) density f based on an i.i.d. sample $X_1, X_2, \ldots, X_n$ has bias of order h^2 as $h = h(n) \to 0$, and variance of order $(nh)^{-1}$ as $n \to \infty$ and $nh \to \infty$. This holds provided f has at least two continuous derivatives. Here, h is the bandwidth, the parameter that controls the degree of smoothing applied to the data. K, the kernel function, will be taken to be a symmetric probability density throughout. For much good introductory material on kernel density estimation see Silverman (1986).

Variable (bandwidth) kernel density estimation extends this idea by replacing the constant h in (1.1) by $a(X_i)$, say, a different bandwidth for the kernel associated with each datapoint. The intuitive idea behind this is to allow for varying degrees of smoothing across the X-space, especially allowing for greater smoothing to be applied in areas of sparse data, and relatively less to be used in the "main body" of the sample. Abramson (1982) quantified this by making the important observation that if $a(z)$ were, essentially, taken to be $h/f^{1/2}(z)$—it remains convenient always to take out a scalar overall smoothing parameter and to continue to call it h—then (sufficient differentiability of f permitting) one obtains a bias of $o(h^2)$ as $h \to 0$ (while retaining a variance of $O(nh)^{-1}$). With care, $o(h^2)$ becomes $O(h^4)$. For practical application, as with all methods developed in this paper, "pilot" estimation of a is necessary, but even then, Silverman (1986), for example, has demonstrated real practical potential for the method.

Samiuddin and El-Sayyad ((1990), Section 4) have proposed varying the location of each kernel, as opposed to varying its scale. This amounts to adding a quantity $h^2A(X_i)$, say, to each X_i. It turns out that, if $A(z)$ is taken to be an appropriate constant times $f'(z)/f(z)$, then again $O(h^4)$ bias ensues (along with $O(nh)^{-1}$ variance). Intuitively, each datapoint is moved a little in the direction of increasing density and, in particular, this serves to accentuate features such as modes in the density estimate.

By allowing both scale and location variation at the same time, we observe an infinity of ways in which $O(h^4)$ bias (and $O(nh)^{-1}$ variance) can be achieved. The formula that drives this is made explicit in Section 3. All scale and location combinations mentioned in this paper are novel. As the two approaches alone work in apparently complementary ways, it is to be hoped that appropriate combination might afford good properties in practice too.

How might one achieve $O(h^6)$ bias? Samiuddin and El-Sayyad (1990) indicate how in the location-change case: add a further perturbation to each X_i of the form $h^4B(X_i)$, for appropriate B. We show that a similar idea—take $a(X_i) = h/\{\alpha(X_i)(1 + h^2\beta(X_i))\}$ for appropriate α and β—can be made to work in the scale-variation case. Indeed, our general framework affords another infinity of solutions to the $O(h^6)$ problem, as will be made clear in Section 4. We particularly exhibit two further tractable special cases which are combinations of scale and location variation; there are two because in one the scale variation is the principal factor and the location variation is secondary, while in the other, roles are reversed.

For practice, however, one can certainly hope that the step from $O(h^2)$ to $O(h^4)$ bias will correspond to a meaningful methodological development (if we are fortunate in our particular choices), but that the further step to $O(h^6)$ would be of much less importance. This is borne out by preliminary simulation experience (not given in detail). In this sense, the work of Section 4 in particular might well remain largely of academic interest only (especially considering the considerable pilot estimations required in practice and which are not considered at all here).

Manipulations, outlined in Section 2, are greatly simplified by use of a result of McKay (1993*a*) which greatly extends the bias formula of Hall (1990). This result also allows us to overcome a technical difficulty, due to possibly overly large bandwidths used in the tails in variable scale estimation, which was demonstrated

by an example of Terrell and Scott (1992). A closely related pathology can occur with variable location estimation; examples of both kinds are described in McKay (1993*b*). To avoid problems of this type we introduce the Condition A(II) in Section 2. This condition formalises the notion of *locality* implicit in comments of Terrell and Scott ((1992), p. 1239); it is related to the clipping method of Abramson (1982), the kernel truncation device of Hall *et al.* (1994) and the bandwidth-dependent clipping of McKay.

2. Basic formulae and technicalities

All estimators of interest in this paper are of the form

$$\tilde{f}(x) = n^{-1} \sum_{i=1}^{n} h^{-1}\gamma(X_i)K[h^{-1}\gamma(X_i)\{x - X_i - h^2 G(X_i)\}] \tag{2.1}$$

for appropriate functions γ and G, which may also depend on the smoothing parameter h. To obtain mean and variance properties of $\tilde{f}$, it is most convenient to employ McKay's (1993*a*) result which identifies, for a generic function J_h of two variables and dependent on h, the coefficients in an expansion of the form

$$h^{-1} \int J_h\{z, h^{-1}(x - z)\} f(z) dz = a_0(x) + \cdots + a_l(x)h^l + o(h^l),$$

subject to some mild conditions. If the moment functions $m_{k,h}(x) = \int z^k \cdot J_h(x, z)dz$, $k = 0, \ldots, l$ are sufficiently smooth these coefficients take the form

$$a_k(x) = (k!)^{-1}(-1)^k \{f(x)m_{k,h}(x)\}^{(k)},$$

which is immediately seen to generalize Hall's (1990) formula. We use this form to motivate our computations, but the most important version of our expansion only requires that the moment functions $m_{k,h}(v)$ have suitable Taylor expansions in powers of h.

Returning to the estimator defined by (2.1), for

$$J_h(v, u) = \gamma(v)K[\gamma(v)\{u - hG(v)\}],$$

we have the representation

$$\tilde{f}(x) = n^{-1} \sum_{i=1}^{n} h^{-1} J_h\{X_i, (x - X_i)/h\}.$$

Let $\tau_k = \int z^k K(z)dz$ for $k = 1, \ldots, 4$. The moment functions are given by

$$m_{k,h}(v) = \sum_{j=0}^{k} \binom{k}{j} \gamma^{-(k-j)}(v)\tau_{k-j} h^j G^j(v),$$

so, under suitable conditions, the leading terms in $E\{\tilde{f}(x)\}$ are

$$E\{\tilde{f}(x)\} = f(x) - h\{f(x)hG(x)\}' + \frac{1}{2}h^2[f(x)\{\tau_2\gamma^{-2}(x) + h^2G^2(x)\}]''$$
$$- \frac{1}{6}h^3[f(x)\{3\tau_2 h\gamma^{-2}(x)G(x) + h^3G^3(x)\}]'''$$
$$+ \frac{1}{24}h^4[f(x)\{\tau_4\gamma^{-4}(x) + 6\tau_2 h^2\gamma^{-2}(x)G^2(x) + h^4G^4(x)\}]^{iv} + \cdots.$$

Further progress can be made if we now, and for the rest of the paper, write γ and G in the following forms:

$$\gamma(z) = \alpha(z)(1 + h^2\beta(z)) \quad \text{and} \quad G(z) = A(z) + h^2B(z).$$

The functions α, β, A and B do not depend on h, and if they are sufficiently smooth, then the bias in using $\tilde{f}$ is

$$(2.2)\quad E\{\tilde{f}(x)\} - f(x) = h^2\left\{\frac{1}{2}\tau_2(\alpha^{-2}f)''(x) - (Af)'(x)\right\}$$
$$+ h^4\left\{\frac{1}{24}\tau_4(\alpha^{-4}f)^{iv}(x) - \tau_2(\alpha^{-2}\beta f)''(x)\right.$$
$$\left. -\frac{1}{2}\tau_2(\alpha^{-2}Af)'''(x) + \frac{1}{2}(A^2f)''(x) - (Bf)'(x)\right\}$$
$$+ R_x(h),$$

where the remainder $R_x(h)$ is no larger than $o(h^4)$. Much more will be made of this in the following sections.

The leading term in the variance of $\tilde{f}(x)$ is

$$(2.3)\quad (nh)^{-1}\kappa f(x)\alpha(x)$$

(compare 3.18 of Silverman (1986)). Notice that this variance remains of order $(nh)^{-1}$ whatever our choice of γ or G; here, $\kappa = \int K^2(z)dz$, assumed to be finite.

For a precise statement of the expansion (2.2) above, we introduce the following conditions. Let $l = 4$ or $l = 6$, and fix some point $x_0 \in \boldsymbol{R}$. We assume

A(I) f is bounded and integrable, and $f^{(l)}(v)$ exists and is continuous in a neighbourhood of x_0.

A(II) There is a non-negative integrable function $H(u)$ such that

$$\int(1 + |z|)^l H(z)dz < \infty \quad \text{and} \quad |J_h(v,u)| \leq H(u)$$

for all h in some neighbourhood of zero and all v. It is this condition that underlies the success of various clipping procedures (McKay (1993*b*)), and is therefore critically important.

A(III) The derivatives $\alpha^{(l)}(v)$, $A^{(l-1)}(v)$, $\beta^{(l-2)}(v)$ and $B^{(l-3)}(v)$ are continuous for all v in some neighbourhood of x_0.

Under these conditions the expansion in (2.2) is valid, with remainder $o(h^4)$ for $l = 4$, or $a_6(x)h^6 + o(h^6)$ when $l = 6$, uniformly for x in a neighbourhood of x_0. Furthermore, the expansion holds uniformly everywhere if the derivatives in A(I) and A(III) are assumed bounded and uniformly continuous for all $v \in \boldsymbol{R}$. The proof is given in the appendix; further details appear in McKay's thesis (1993*a*).

A few remarks regarding the expansion in (2.2) are in order. First, A(II) actually implies that $\gamma(v)$ is bounded above and away from zero, and that $hG(v)$ is bounded above. As well, A(II) imposes some mild conditions on K. In particular, K must possess at least l absolute moments, and if hG is not constant, then K must be bounded. Conversely, if these conditions are satisfied, and also $(1 + |z|)^{l+1+\varepsilon}K(z) \to 0$ as $|z| \to \infty$ for some $\varepsilon > 0$, then A(II) will be satisfied. These conditions are related to Abramson's (1982) "clipping" procedure and are necessary to avoid problems described by Terrell and Scott (1992) and McKay (1993*a*) (see also Hall *et al.* (1994)). Second, it is not necessary for our expansions to use a smooth kernel K, though this might be typical in practice, and third, the smoothness conditions A(I) and A(III) are essential if the expansion in (2.2) is to make sense. Lastly, the dominating function $H(u)$ in A(II) may be modified to allow for a dependence on h, if it has sufficiently many moments. Details for Abramson's estimator are discussed in McKay (1993*b*).

Before continuing further, we must briefly address a notational point related to assumption A(II). Abramson's (1982) square root law suggests the choice $\gamma(v) = f^{1/2}(v)$ and $G(v) = 0$ in (2.1). However this violates A(II), and the counterexample of Terrell and Scott (1992) shows that this can result in bias much larger than $O(h^4)$. McKay's thesis (1993*a*) presents additional examples, including a similar result for the variable location estimator of Samiuddin and El-Sayyad (1990), and shows that the condition A(II) effectively resolves the problem. To enforce the condition in A(II), McKay (1993*b*) used a smooth variation of Abramson's (1982) original "clipping" procedure. In effect, he imposed a mild condition on the tails of K and replaced $\gamma(v)$ by $\tilde{\gamma}(v) = r(\gamma(v))$ for some suitably smooth function r bounded away from zero, with $r(t) = t$ for all sufficiently large t. To see how such a function could be constructed see McKay (1993*b*). Up to a possible change of scale, one example considered there is given by

$$r(t) = \begin{cases} 1 + \dfrac{t^5}{32}\left\{1 - \dfrac{3}{2}(t-2) + \dfrac{5}{4}(t-2)^2 - \dfrac{5}{8}(t-2)^3\right\} & \text{for } 0 \le t \le 2 \\ t & \text{for } t > 2 \\ 1 & \text{for } t < 0. \end{cases}$$

For the variable location estimator of Samiuddin and El-Sayyad (1990), another example in McKay (1993*b*) demonstrates that we must replace $G(v)$ by $s(G(v))$ where s is bounded above, sufficiently smooth, and $s(t) = t$ for sufficiently small t. For the remainder of this paper, we shall assume that a similar device is used to force A(II) to hold. Thus the choices of γ and G recommended in Section 3 and Section 4 below must first be passed through some such clipping "filter" to guarantee the claimed bias. Though this is an important point, we consider our results to be more easily understood with this minor abuse of notation. In any actual implementation a function such as that described above could be used to construct the filter.

3. Achieving $O(h^4)$ bias

By setting the $O(h^2)$ bias term in (2.2) to zero, we immediately see that any choice of α, for scale variation, and A, for location variation, satisfying

$$\frac{\tau_2}{2}\left(\frac{f}{\alpha^2}\right)''(z) = (Af)'(z) \tag{3.1}$$

will achieve $O(h^4)$ bias. The two "familiar" special cases of this are:

(i) set $A(z) \equiv 0$ so that we need $(\alpha^{-2}f)''(z) = 0$. The latter is satisfied by $\alpha(z) = f^{1/2}(z)$, and we obtain precisely Abramson's (1982) variable kernel estimator, which we will call $\hat{f}_A$;

(ii) set $\alpha(z) \equiv 1$. Then, we require $\frac{1}{2}\tau_2 f''(z) = (Af)'(z)$. The choice $A(z) = \frac{1}{2}\tau_2(f'/f)(z)$ will do, and this yields precisely Samiuddin and El-Sayyad's (1990) variable location estimator, $\hat{f}_S$, say.

(Notice that there are also more general solutions to these differential equations involving extra constant or linear terms (see Abramson (1982), for (i)) which we are unable to exploit.)

Aside. The location variation in (ii) is quite different from an overall "shrinkage" of the data towards its mean (Jones (1991), and references therein). However, when $f(x) = \sigma^{-1}\phi(\sigma^{-1}(x-\mu))$, where ϕ is the standard normal density, and K is normal, each kernel is located at $X_i - (2\sigma^2)^{-1}h^2(X_i - \mu)$; this, with estimated μ and σ^2, is essentially the shrinkage formula (4) in Jones (1991). While such a device, which was motivated by the desire to correct for variance inflation, is certainly appropriate for normal f (Fryer (1976)), it is comparatively unsuccessful when used for other densities (Jones (1991)). One might now argue that it is $\hat{f}_S$ that is a more appropriate generalisation of variance correction to other situations than is application of the shrinkage formula itself.

We stress that the two existing variations of Abramson (1982) and Samiuddin and El-Sayyad (1990) are but (the most obvious) special cases satisfying (3.1). *Any* scale/location pair satisfying (3.1) will also work. Such combination might have especial appeal since singly scale and location variations tend to work in rather complementary ways. (The need to pilot estimate two functions in practice is less appealing, however.) For example, if we choose to take $\alpha(z) = f(z)$ alone, we get essentially the earliest proposals of Victor (1976) and Breiman *et al.* (1977). This has bias of order h^2 only, but if we combine it with the appropriate location variation wherein, from (3.1), $A(z) = \frac{1}{2}\tau_2(1/f)'(z)(1/f)(z) = -\frac{1}{2}\tau_2(f'/f^3)(z)$, we obtain another method with $O(h^4)$ bias. (Notice that an adjustment of a purely scale variation type, i.e. with nonzero β rather than A, cannot achieve this.) It turns out that this particular scale/location combination has links with an alternative viewpoint, and also has considerable promise; this may be developed elsewhere in joint work of the first author.

Comparisons between methods satisfying (3.1) can be made on the basis of their remaining bias (and variance). Write $g(z) = \alpha^{-2}(z)$. For any pairing such that $A(z) = \frac{1}{2}\tau_2(1/f)(z)(gf)'(z)$, the leading bias becomes

$$\frac{1}{24}h^4\left(\tau_4\{g(fg)\}^{iv}(x) - 6\tau_2^2\{g(fg)'\}'''(x) + 3\tau_2^2[f^{-1}\{(fg)'\}^2]''(x)\right). \tag{3.2}$$

The bias in using $\hat{f}_S(x)$ is, therefore,

$$\frac{1}{24}h^4[(\tau_4 - 6\tau_2^2)f^{iv}(x) + 3\tau_2^2\{f^{-1}(f')^2\}''(x)], \tag{3.3}$$

as given by Samiuddin and El-Sayyad (1990). Also, the bias in using $\hat{f}_A(x)$ is

$$\frac{1}{24}h^4\tau_4(f^{-1})^{iv}(x).$$

This simple formula was first noted by Hall (1990) and Jones (1990). (Earlier expressions include that of Hall and Marron (1988).) Non-integrability of squared bias for $\hat{f}_A$ is discussed in detail by Hall (1992).

That there seems to be little opportunity to use α and A to zero *both* h^2 and h^4 bias terms is indicated by the following argument. Specialise to use of a standard normal kernel so that $\tau_2 = 1$, $\tau_4 = 3$. Consider setting the second antiderivative of (3.2) to zero i.e. attempt to choose g such that $f(z)\{g(fg)\}''(z) - 2f(z)\{g(fg)'\}'(z) + \{(fg)'\}^2(z) = 0$ for all z. Manipulating this leads to $(g'/g)^2(z) = [\{ff'' - (f')^2\}/f^2](z)$. But this only has a solution if the right-hand side, equals $\{\log(f(z))\}''$, is positive. For log concave densities this is never the case, and for others, only at certain points z. We doubt whether reinstating greater generality could improve matters very much.

We will not pursue considerations of comparative $O(h^4)$ bias expressions further here.

4. Achieving $O(h^6)$ bias

We will now consider the vanishing of $O(h^4)$ bias in addition to that of $O(h^2)$ by introducing β and B in addition to α and A. With $\gamma(z) = \alpha(z)(1 + h^2\beta(z))$ and $G(z) = A(z) + h^2B(z)$ as in Section 2, choices of α, β, A and B satisfying (3.1) for $O(h^2)$ and, from (2.2),

$$\begin{aligned}\frac{1}{24}\tau_4(\alpha^{-4}f)^{iv}(x) - \tau_2(\alpha^{-2}\beta f)''(x) - \frac{1}{2}\tau_2(\alpha^{-2}Af)'''(x) \\ + \frac{1}{2}(A^2f)''(x) - (Bf)'(x) = 0,\end{aligned} \tag{4.1}$$

to zero the h^4 term will do. With four functions and only two differential equations for them to satisfy, there is again considerable scope for ways of achieving $O(h^6)$ bias. We will look briefly at only the four most tractable and appealing special cases.

4.1 *Scale variation only*

Set $A \equiv B \equiv 0$. Then $\alpha(z) = f^{1/2}(z)$ satisfies (3.1), of course, and we are left with an $O(h^4)$ term of

$$\frac{1}{24}\tau_4(1/f)^{iv}(x) - \tau_2\beta''(x)$$

which disappears if

$$\beta(z) = \frac{\tau_4}{24\tau_2}\left(\frac{1}{f}\right)''(z).$$

Thus, the extended pure scale variation

$$h/\left[f^{1/2}(X_i)\left\{1 + \frac{h^2\tau_4}{24\tau_2}\left(\frac{1}{f}\right)''(X_i)\right\}\right]$$

achieves bias of $O(h^6)$. This is an entirely novel extension of Abramson's (1982) result.

4.2 *Location variation only*

Set $\alpha \equiv 1$, $\beta \equiv 0$. Then, with $A(z) = \frac{1}{2}\tau_2(f'/f)(z)$ zeroing $O(h^2)$ bias, the $O(h^4)$ term becomes

$$\left(\frac{\tau_4}{24} - \frac{\tau_2^2}{4}\right) f^{iv}(x) + \frac{\tau_2^2}{8}\left\{\frac{(f')^2}{f}\right\}''(x) - (Bf)'(x)$$

and the extended pure location variation

$$X_i + \frac{1}{2}h^2\tau_2\left(\frac{f'}{f}\right)(X_i) + \frac{1}{24}h^4\frac{1}{f(X_i)}\left[(\tau_4 - 6\tau_2^2)f'''(X_i) + 3\tau_2^2\left\{\frac{(f')^2}{f}\right\}'(X_i)\right]$$

zeroes $O(h^4)$ bias too. A formula like this was suggested by Samiuddin and El-Sayyad (1990), but our expression, which is the simpler, appears to be a correction of theirs.

4.3 *Scale and location variation* I

Set $\beta \equiv 0 \equiv A$. In this, the first of two $O(h^6)$ bias joint scale and location variations, we let scale variation dominate in the sense that we remain with Abramson's square root law, $\alpha(z) = f^{1/2}(z)$, for the scale part. The choice

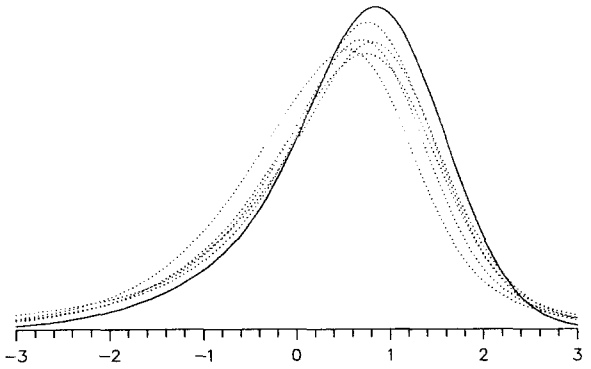
$$B(z) = \frac{\tau_4}{24}\frac{1}{f(z)}\left(\frac{1}{f}\right)'''(z)$$

turns out to be an appropriate ally to this in achieving $O(h^6)$ bias i.e. use bandwidth $h/f^{1/2}(X_i)$ and centre kernels at $X_i + \frac{1}{24}h^4\tau_4 f^{-1}(X_i)(1/f)'''(X_i)$.

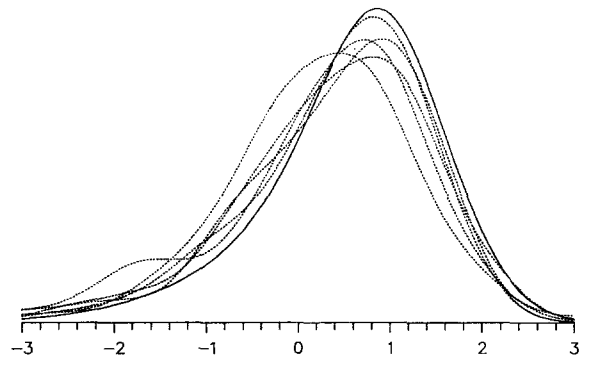
4.4 *Scale and location variation* II

Finally, set $\alpha \equiv 1$ and $B \equiv 0$. This time, location variation leads in the sense that we take $A(z) = \frac{1}{2}\tau_2(f'/f)(z)$ once again, and then find that the requisite formula for $\beta(z)$ is

$$\frac{1}{24\tau_2 f(z)}\left[(\tau_4 - 6\tau_2^2)f''(z) + 3\tau_2^2\left\{\frac{(f')^2}{f}\right\}(z)\right].$$



(a)

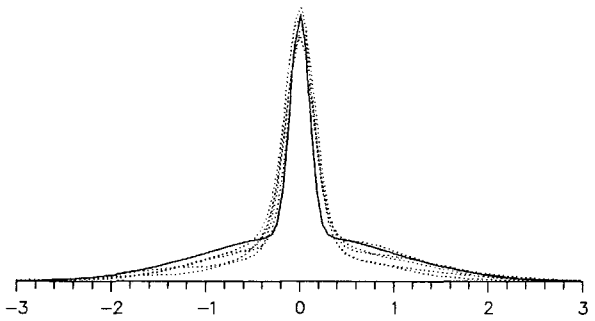


(b)

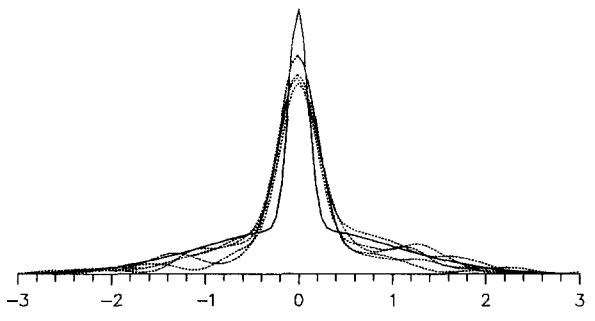
Fig. 1. Marron and Wand density 2 (solid line) together with estimates $\hat{f}_A$ with $h = 0.3$ (dotted lines in (a)) and $\hat{f}_S$ with $h = 0.4$ (dashed lines in (b)). The same 5 samples of $n = 100$ datapoints are used in each frame.

If we use a normal kernel again, we find that if the location variation is written $X_i + h^2 A(X_i)$, the associated scale variation is as $h/\{1 - \frac{1}{4}h^2 A'(X_i)\}$.

We tried out most of these ideas in some preliminary simulations (which fall far short of a full study). A fine testbed for nonparametric density estimation is the collection of densities appearing in Fig. 1 of Marron and Wand (1992). A very wide variety of density shapes is catered for although all the densities are normal mixtures. Only "ideal" forms of the estimators, i.e. ones using true density dependent quantities at the pilot stage, were implemented, and pictures of estimates formed using subjective choice of h were examined. No clipping was employed. The performance of Abramson's estimator $\hat{f}_A$ in its ideal form was most impressive, witness Figs. 1(a) and 2(a). Figures 1 and 2 display five realised estimates of Marron and Wand's densities 2, the "Skewed Unimodal", and 4, the "Kurtotic Unimodal", for each of $\hat{f}_A$ (dotted lines in (a)) and $\hat{f}_S$ (dashed lines in



(a)



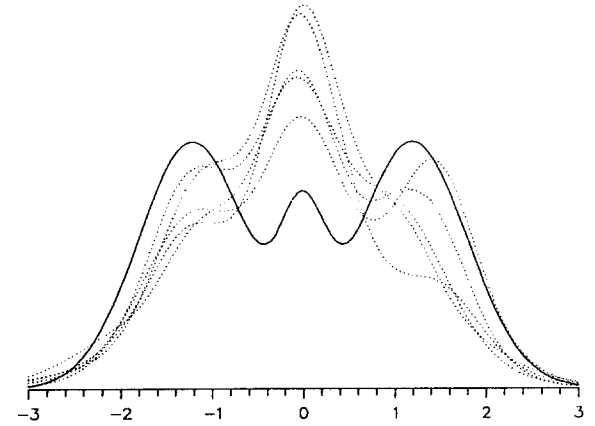
(b)

Fig. 2. Marron and Wand density 4 (solid line) together with estimates $\hat{f}_A$ with $h = 0.15$ (dotted lines in (a)) and $\hat{f}_S$ with $h = 0.2$ (dashed lines in (b)). The same 5 samples of $n = 100$ datapoints are used in each frame.

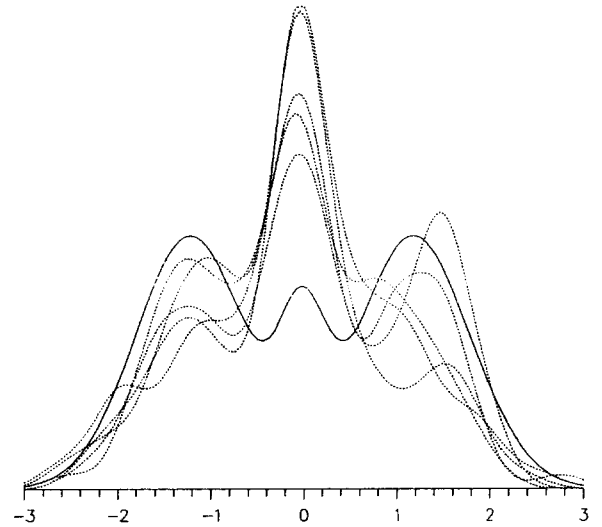
(b)); $n = 100$ throughout, h's of $0.3, 0.4, 0.15$ and 0.2 used in Figs. 1(a), 1(b), 2(a) and 2(b), respectively. Notice that $\hat{f}_S$ cannot begin to match $\hat{f}_A$'s performance at peaks without showing undesirable features elsewhere.

The location shift $\hat{f}_S$ is generally less good: its best performance is indeed reserved for cases with fairly "tight" peaks, although improved properties with respect to peak estimation tend to be at the expense of good behaviour elsewhere. This is further exemplified by Fig. 3 which repeats the above for Marron and Wand density 9, "Trimodal". Here $h = 0.175$ for $\hat{f}_A$ and $h = 0.25$ for $\hat{f}_S$. The point is that while $\hat{f}_A$ struggles to indicate three modes, $\hat{f}_S$ does so rather better, albeit with neither being very good at estimating heights of modes. Other than this kind of behaviour, $\hat{f}_A$ was usually the better of the two. Note that the very places where $\hat{f}_A$ is least good coincide with peak areas where $\hat{f}_S$ is at its best.

It must be said that preliminary simulations along these lines for methods with $O(h^6)$ bias were not especially encouraging. That is not to say that the $O(h^6)$ methods had inferior properties to the $O(h^4)$ ones, just that they did little or nothing to improve performance. This was especially true of the pure location



(a)



(b)

Fig. 3. Marron and Wand density 9 (solid line) together with estimates $\hat{f}_A$ with $h = 0.175$ (dotted lines in (a)) and $\hat{f}_S$ with $h = 0.25$ (dashed lines in (b)). The same 5 samples of $n = 100$ datapoints are used in each frame.

and location-leading location/scale combination methods (of Subsections 4.2 and 4.4); they usually followed $\hat{f}_S$ pretty closely. The other location/scale combination (Subsection 4.3) was much more closely related to $\hat{f}_A$. Sometimes the two were very similar, sometimes there were noticeable differences, but we did not get much impression of improvement when differences were apparent. If there was one of the $O(h^6)$ bias methods that might be worth further investigation, it seemed to be

the pure scale variation of Subsection 4.1. The amount of improvement in going from $\hat{f}_A$ ($O(h^4)$ bias) to this $O(h^6)$ method was very much smaller than that of moving from $\hat{f}$ ($O(h^2)$) to $\hat{f}_A$, as might be expected. But any small differences there were did appear to be in the right direction. (The extra pilot estimation required might, of course, nullify this in practice.)

We are confident that the best of the kind of method discussed above will prove to display greater practical benefit in small samples than the disappointing higher order kernels (Marron and Wand (1992)).

The work of Section 2 is straightforwardly extendible to obtain the $O(h^6)$ bias terms. We give only the general formula here for possible future reference, but do not bother to delve into any further theoretical comparison of the methods using it. It is

$$\begin{aligned}\frac{1}{720}\tau_6(\alpha^{-6}f)^{vi}(x) - \frac{1}{6}\tau_4(\alpha^{-4}\beta f)^{iv}(x) - \frac{1}{24}\tau_4(\alpha^{-4}Af)^{v}(x)\\ + \tau_2(\alpha^{-2}\beta Af)'''(x) - \frac{1}{2}\tau_2(\alpha^{-2}Bf)'''(x) + \frac{1}{4}\tau_2(\alpha^{-2}A^2f)^{iv}(x)\\ + \frac{3}{2}\tau_2(\alpha^{-2}\beta^2 f)''(x) + (ABf)''(x) - \frac{1}{6}(A^3f)'''(x).\end{aligned}$$

This is $a_6(x)$ in our earlier notation.

Acknowledgements

We are grateful to the referees for helpful comments.

Appendix

In this appendix we have included several key elements in the proof of the crucial bias expansion (2.2). Let $l = 4$ or 6, and fix the point of interest x sufficiently close to x_0. For any bounded integrable function g, denote by Tg the function

$$\begin{aligned}Tg(v) &= h^{-1}\int J_h\{z, h^{-1}(v-z)\}g(z)dz\\ &= \int J_h(v-hz, z)g(v-hz)dz,\end{aligned}$$

where J_h is as in Section 2 and satisfies A(II), A(III).

Our first proposition shows that only local behaviour of f matters.

PROPOSITION A.1. *If g is any bounded integrable function which agrees with f in some neighbourhood of x, then*

$$Tg(v) = Tf(v) + o(h^l)$$

uniformly for v sufficiently close to x.

PROOF. The function $f - g$ is bounded everywhere, and identically zero on some neighbourhood of x. Therefore, for some constant M and $\varepsilon > 0$, A(II) implies that for v sufficiently close to x,

$$|Tf(v) - Tg(v)| \leq \int_{|z|>\varepsilon/h} M|J_h(v - hz, z)|dz$$
$$\leq M(h/\varepsilon)^l \int_{|z|>\varepsilon/h} |z|^l H(z)dz$$

which is $o(h^l)$ by dominated convergence. □

This proposition reveals the true significance of A(II), namely that it disallows estimators which are not in some sense local. Furthermore, the proposition as stated does not require the smoothness conditions in A(I). A Taylor expansion of $f(v - hz)$ shows that if g also satisfies A(I) and has the same Taylor expansion at v, the result still holds.

If the expansion in (2.2) holds uniformly in a neighbourhood of x_0, then Taylor expansion of the coefficients shows that, for each x in a possibly smaller neighbourhood, there is a polynomial $P_x(v, h)$ of degree l such that if $|v| = O(h)$, the remainder $Tf(x + v) - P_x(v, h)$ is uniformly $o(h^l)$. Our next proposition shows that $Tf(x)$ has this property. The proof essentially follows suggestions of Hall and Marron (1988).

PROPOSITION A.2. *For h sufficiently close to 0 and x sufficiently close to x_0, there exists such a polynomial $P_x(v, h)$, with the remainder*

$$Tf(x + v) - P_x(v, h) = o(h^l)$$

uniformly in x, for $v = O(h)$.

PROOF. For any $\varepsilon > 0$, by Proposition A.1, we may assume that f has support within ε of x_0, and is l-times continuously differentiable everywhere. Without loss of generality, we may assume that $x = 0$. If we set $u = \gamma(v - hz)\{z - hG(v - hz)\}$, then by A(II) and A(III), for sufficiently small h, we can choose ε, ε' so that

$$\frac{du}{dz} = \gamma(v - hz) - hz\gamma'(v - hz) + h^2\{\gamma G\}'(v - hz)$$

exists, is continuous, and bounded away from zero, whenever $|hz| < \varepsilon$ and $|x - v| < \varepsilon'$. It follows by the implicit function theorem and the obvious change of variables that $Tf(v)$ is of the form

$$Tf(v) = \int f(v - hz)\gamma(v - hz)\{du/dz\}^{-1}K(u)\,du$$

where z is understood to be a function $z = z(v, h, u)$. For each u we can expand z in powers of v and h with a remainder that is $o(h^l)$ uniformly in x and u. To see

this, note that if we replace γ and G by appropriate Taylor series, we can define for each u an analytic function $\tilde{z}(v, h, u)$ which differs from z by $o(h^l)$. This yields the desired expansion. Further details are given in McKay (1993*a*). □

This argument was suggested by Hall and Marron (1988) to directly establish the formula in (2.2) for the special case of Abramson's (1982) estimator. However, it is tedious in the extreme to evaluate the coefficients in this manner. Instead, we employ a remarkably simple device which yields the coefficients quite easily.

THEOREM A.1. *Under* A(I)–A(III), *the expansion in* (2.2) *is valid, with remainder* $o(h^l)$ *uniformly in* x.

PROOF. By virtue of Proposition A.2, we have at each x in a neighbourhood Ω of x_0 a Taylor polynomial $P_x(v, h)$ of degree l. Setting $v = 0$ yields an expansion $Tf(x) = \sum_{k=0}^{l} b_k(x)h^k + o(h^l)$ uniformly for $x \in \Omega$. Note that each of the coefficients $b_k(x)$ is $l - k$ times continuously differentiable.

Let $\psi(v)$ be any infinitely differentiable function supported in Ω. Then we have

$$\int \psi(x)Tf(x)\,dx = \sum_{k=0}^{l} h^k \int \psi(x)b_k(x)\,dx + o(h^l).$$

On the other hand, by a Taylor series expansion of ψ, we have

$$\text{(A.1)} \qquad \begin{aligned} \int \psi(u)Tf(u)du &= \iint \psi(x + hz)f(x)J_h(x, z)dzdx \\ &= \sum_{k=0}^{l} h^k/k! \int \psi^{(k)}(x)f(x)\mu_k(x)dx + o(h^l) \end{aligned}$$

where $\mu_k(x) = \int z^k J_h(x, z)dz$. Note that the size of the remainder term depends only on the modulus of continuity of $\psi^{(l)}$. By A(III) we observe that each moment function μ_k is of the form

$$\mu_k(x) = c_{0k}(x) + hc_{1k}(x) + \cdots + h^l c_{lk}(x) + o(h^l)$$

for which the functions c_{jk} are $l - j$ times continuously differentiable. If we substitute these expressions into (A.1), and integrate each term by parts we obtain

$$\int \psi(x)Tf(x)\,dx = \sum_{k=1}^{l} h^k \int \psi(x)a_k(x)\,dx + o(h^l)$$

where the coefficient functions a_k are exactly those given in the expansion (2.2). Since ψ is arbitrary, and the coefficients are continuous functions, we can identify $a_k(x)$ with $b_k(x)$. □

REFERENCES

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

Breiman, L., Meisel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities, *Technometrics*, **19**, 135–144.

Fryer, M. J. (1976). Some errors associated with the non-parametric estimation of density functions, *Journal of the Institute of Mathematics and its Applications*, **18**, 371–380.

Hall, P. (1990). On the bias of variable bandwidth curve estimators, *Biometrika*, **77**, 529–535.

Hall, P. (1992). On global properties of variable bandwidth density estimators, *Ann. Statist.*, **20**, 762–778.

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

Hall, P., Hu, T.-C. and Marron, J. S. (1994). Improved variable window kernel estimates of probability densities, *Ann. Statist.* (to appear).

Jones, M. C. (1990). Variable kernel density estimates and variable kernel density estimates, *Austral. J. Statist.*, **32**, 361–371 (Correction: *ibid.* (1991). **33**, p. 119).

Jones, M. C. (1991). On correcting for variance inflation in kernel density estimation, *Comput. Statist. Data Anal.*, **11**, 3–15.

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

McKay, I. J. (1993*a*). Variable kernel methods in density estimation, Ph.D. Thesis, Queens University, Kingston, Ontario.

McKay, I. J. (1993*b*). A note on bias reduction in variable kernel density estimates, *Canad. J. Statist.*, **21**, 365–375.

Samiuddin, M. and El-Sayyad, G. M. (1990). On nonparametric kernel density estimates, *Biometrika*, **77**, 865–874.

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

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation, *Ann. Statist.*, **20**, 1236–1265.

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