The Annals of Statistics
1982, Vol. 10, No. 4, 1217–1223

ON BANDWIDTH VARIATION IN KERNEL ESTIMATES—A SQUARE ROOT LAW[1]

BY IAN S. ABRAMSON

University of California, San Diego

We consider kernel estimation of a smooth density f at a point, but depart from the usual approach in admitting an adaptive dependence of the sharpness of the kernels on the underlying density. Proportionally varying the bandwidths like $f^{-1/2}$ at the contributing readings lowers the bias to a vanishing fraction of the usual value, and makes for performance seen in well-known estimators that force moment conditions on the kernel (and so sacrifice positivity of the curve estimate). Issues of equivariance and variance stabilization are treated.

Introduction. The idea of bandwidth variation across the sample is not new. Motivated possibly by the familiar temptation to combine class intervals in the tails of histograms, Breiman et al. (1977) pointed out the benefits of using diffuse kernel contributions where the data is thinly scattered, and spikier ones in the data-rich regions where there is less danger of raggedness in the resulting curve. But they proposed a dependence proportional to nearest neighbor distances (regardless of dimension), so that on the line, bandwidths would vary like f^{-1}, a dependence too strong in the light of our findings (which indicate $f^{-1/2}$). In two dimensions, fortuitously, our dependences agree. Indeed, in private communication, Breiman noted that the performance in their univariate study was considered disappointing, and in their bivariate study, excellent.

Let $X_1, \cdots, X_n$ be a sample from the p-variate Lebesgue density f assumed to have all second order partial derivatives $D_{jk}(f)$ continuous with $|D_{jk}f| \leq U_2$, a known bound imposed by the user. Automatically then, there are constants U_1 and U_0 say, bounding the $|D_jf|$ and f, and for technical reasons, we place a strict lower bound $L_0 > 0$ on f at $x = 0$, the target argument.

There is effectively nothing lost by choosing these bounds as generously as we like, and the requirements can in fact be localized to a neighborhood of $x = 0$ at some price in the elegance of the analysis. If the true density violates the bounds, the estimator we develop will perform like one with fixed bandwidths. With little loss we choose as kernel w, a smooth density supported in the cube $[-1, 1]^p$ and even in each argument. Our estimate of f at 0 will be given by

$$f_n(0) = n^{-1} \sum_{i=1}^{n} b_n^{-p} c(X_i)^p w(b_n^{-1} c(X_i) X_i), \tag{1}$$

where the bandwidths $b_n c(X_i)^{-1}$ show dependence not only on n, germane to any asymptotic study, but also on the contributing point through a yet-unassigned scalar function c on $\mathbb{R}^p$, anticipated to depend on the local behavior of f only. We disregard feasibility objections for the time being.

Our optimality criterion is mean squared error (MSE) at 0. (A practitioner might prefer a globalized measure such as integrated MSE, and under uniformity conditions the analyses are similar, though the question then arises of whether to admit bandwidth

Received October 1981; revised March 1982.

[1] This work was supported in part by National Science Foundation Grant MCS 80-02698 and by the U.S. Office of Naval Research Contract No. N00014-80-C-0163.

AMS 1980 *subject classification.* Primary 62G05; secondary 62F12.

Key words and phrases. Kernel estimate, bandwidth variation, inverse square root, bias reduction, equivariance, logogram.

dependence on the target argument, which renders the resulting curve estimate a dishonest density, and jars with the graphical spirit of kernel methods.)

We define a version of the density clipped below, viz. $\bar{f}(\cdot) = f(\cdot) \vee \frac{1}{10} f(0)$ (the small fraction $\frac{1}{10}$ is arbitrary), and are now equipped to state a theorem.

THEOREM. *Let $b_n \to 0$ be any sequence with $nb_n \to \infty$. In the present context, the estimator for $f(0)$ given by* (1) *with*

$$c(x) = \bar{f}(0)^{1/p-1/2} \cdot \bar{f}(x)^{1/2} \tag{2}$$

has

$$\text{MSE} = (nb_n^p)^{-1} f(0)^2 \int_{\mathbb{R}^p} w(v)^2 \, dv + o(b_n^4) + O(n^{-1} b_n^{1-p})$$

as $n \to \infty$.

NOTES. (i) Since $\bar{f}$ is bounded below, all bandwidths shrink uniformly to 0, freeing our estimate of dependence on the points falling outside a vanishing neighborhood of 0.

(ii) The usual bias term is missing from the MSE above; a fixed bandwidth estimator (taking $c(x) = c$, say) has an additional multiple of b_n^4 in its MSE, viz.

$$\tfrac{1}{4} b_n^4 \left(\sum \sum D_{jk} f(0) \int_{\mathbb{R}^p} y_j y_k w(y) \, dy \right)^2 c^{-4}$$

(see, e.g., Mack and Rosenblatt, 1979). We examine the implications presently.

(iii) The choice of multiplier $\bar{f}(0)^{1/p-1/2}$ in $c(x)$ (which might apparently be absorbed into b_n) stems from scale equivariance considerations. If the sequence b_n is to capture only the decay rate and be used for all permissible densities, then equivariance entails that for every $k > 0$, use of the scaled observations $\{kx_i\}$ to estimate the density from which *they* arose, viz. $k^{-p} f(k^{-1} \cdot)$, should yield simply $k^{-p} f_n(0)$ (where $f_n(0)$ is the estimate based on the unscaled observations using the same b_n).

(Short of requiring w to have spherical contours, we cannot obtain equivariance under more general matrix transformations.)

Noting that the scaling map $f \mapsto k^{-p} f(k^{-1} \cdot)$ and the clipping map $f \mapsto \bar{f}$ commute, we verify by simplification that indeed,

$$n^{-1} \textstyle\sum_{i=1}^n b_n^{-p} c_{k^{-p} f(k^{-1} \cdot)}(kX_i)^p w(b_n^{-1} c_{k^{-p} f(k^{-1} \cdot)}(kX_i) kX_i)$$

$$= k^{-p} n^{-1} \textstyle\sum_{i=1}^n b_n^{-p} c_f(X_i)^p w(b_n^{-1} c_f(X_i) X_i) \quad \text{for every} \quad k, f,$$

when (subscripting c by its corresponding density), $c_g(x)$ is given by $\bar{g}(0)^{1/p-1/2} \bar{g}(x)^{1/2}$ according to (2).

PROOF OF THEOREM. We treat the squared bias and variance contributions to the MSE separately, and for the purposes of showing that the bias is negligible to second order in b_n, our equivariance entitles us to make the artificial assumption that $f(0) = 1$.

The bias of (1) is given by

$$b_n^{-p} \int_{\mathbb{R}^p} \bar{f}(x)^{p/2} w(b_n^{-1} \bar{f}(x)^{1/2} x) f(x) \, dx - 1$$

$$= \int_{\mathbb{R}^p} [\bar{f}(b_n v)^{p/2} f(b_n v) w(\bar{f}(b_n v)^{1/2} v) - w(v)] \, dv, \tag{3}$$

and since w is compactly supported and $\bar{f}$ bounded below, the integration may be taken over a bounded range; call it I; then neglecting finitely many n if necessary, $\bar{f}$ may be

replaced throughout by f since they agree in a neighborhood of 0, giving

$$\int_I [w(f(b_n v)^{1/2}v)f(b_n v)^{1+p/2} - w(v)]\,dv = \int_I [J_v(f(b_n v)^{1/2}) - J_v(1)]\,dv,$$

where $J_v(y) = w(yv)y^{2+p}$, $y \in \mathbb{R}$, $v \in \mathbb{R}^p$.

Heading for a partial Taylor expansion under the integral and writing $E_n(v)$ for $f(b_n v)^{1/2} - 1$ (which is $o(1)$ as $n \to \infty$), the bias becomes

$$\int_I [E_n(v)J_v'(1) + \tfrac{1}{2} E_n(v)^2 J_v''(1 + tE_n(v))]\,dv,$$

with $-1 \le t \le 1$, a generic quantity not necessarily the same at each appearance. Calculating we obtain

$$J_v'(y) = y^{2+p} \sum D_r w(yv)v_r + (2+p)y^{1+p}w(yv)$$

$$J_v''(y) = y^{2+p} \sum\sum D_{rs} w(yv)v_r v_s + 2(2+p)y^{1+p} \sum D_r w(yv)v_r + (2+p)(1+p)y^p w(yv),$$

and

$$E_n(v) = \tfrac{1}{2} b_n \sum D_j f(0)v_j - \tfrac{1}{8} b_n^2 \{1 + E_n(tv)\}^{-3}\{\sum D_j f(tb_n v)v_j\}^2 + \tfrac{1}{4} b_n^2 \{1 + E_n(tv)\}^{-1} \sum\sum D_{jk} f(tb_n v)v_j v_k,$$

so that substituting and rearranging, bias$(f_n(0))$ becomes

$$\begin{aligned}
&\frac{1}{2} b_n \sum D_j f(0) \int_I v_j[\sum D_r w(v)v_r + (2+p)w(v)]\,dv \\
&+ \frac{1}{4} b_n^2 \int_I \{1 + E_n(tv)\}^{-1} \sum\sum D_{jk} f(tb_n v)v_j v_k[\sum D_r w(v)v_r + (2+p)w(v)]\,dv \\
&+ \frac{1}{8} b_n^2 \sum\sum D_j f(0)D_k f(0) \int_I v_j v_k[(2 + 3p + p^2)\{1 + tE_n(v)\}^p w(\{1 + tE_n(v)\}v) \\
&+ (4+2p)\{1 + tE_n(v)\}^{1+p} \sum D_r w(\{1 + tE_n(v)\}v)v_r \\
&+ \{1 + tE_n(v)\}^{2+p} \sum\sum D_{rs} w(\{1 + tE_n(v)\}v)v_r v_s]\,dv \\
&- \frac{1}{8} b_n^2 \sum\sum \int_I D_j f(tb_n v)D_k f(tb_n v)v_j v_k\{1 + E_n(tv)\}^{-3} \\
&\cdot [\sum D_r w(v)v_r + (2+p)w(v)]\,dv + O(b_n^3),
\end{aligned} \tag{4}$$

where $O(b_n^3)$ has arisen from the b_n^2 terms in $E_n(v)$ on squaring the latter. A uniform bound on the coefficient brings the integral under the $O(\cdot)$. Passing to the limit now, by bounded convergence,

$$\begin{aligned}
\text{bias} = &\frac{1}{2} b_n \sum D_j f(0) \int v_j[\sum D_r w(v)v_r + (2+p)w(v)]\,dv \\
&+ \frac{1}{4} b_n^2 \sum\sum D_{jk} f(0) \int v_j v_k[\sum D_r w(v)v_r + (2+p)w(v)]\,dv \\
&+ \frac{1}{8} b_n^2 \sum\sum D_j f(0)D_k f(0) \int v_j v_k \sum\sum D_{rs} w(v)v_r v_s\,dv \\
&+ \frac{1}{8} b_n^2 \sum\sum D_j f(0)D_k f(0) \int v_j v_k (2p + p^2)w(v)\,dv \\
&+ \frac{1}{8} b_n^2 \sum\sum D_j f(0)D_k f(0) \int v_j v_k (3+2p) \sum D_r w(v)v_r\,dv + o(b_n^2).
\end{aligned} \tag{5}$$

The first line vanishes by symmetry properties of w, then applying the multivariable identities,

$$\int v_j v_k \sum D_r w(v) v_r \, dv = -(p+2) \int v_j v_k w(v) \, dv$$

$$\int v_j v_k \sum \sum D_{rs} w(v) v_r v_s \, dv = (p+3)(p+2) \int v_j v_k w(v) \, dv,$$

all other leading terms fall away, leaving, as required,

$$\text{Bias} = o(b_n^2) \quad \text{as} \quad n \to \infty.$$

The variance requires no such expansion; simply

$$\text{var } f_n(0) = n^{-1}\left[\int_{\mathbb{R}^p} b_n^{-2p} \bar{f}(0)^{2-p} \bar{f}(x)^p w(b_n^{-1} \bar{f}(0)^{1/p-1/2} \bar{f}(x)^{1/2} x)^2 f(x) \, dx - \{E f_n(0)\}^2\right].$$

Changing variables, and arguing as before, this gives as required

$$\text{var } f_n(0) = (n b_n^p)^{-1} f(0)^2 \int_{\mathbb{R}^p} w(v)^2 \, dv + O(n^{-1} b_n^{1-p}),$$

and adding the contributions to the MSE, the proof is complete.

A feasible version. Evidently for implementation, a pilot estimate of f is needed. The asymptotic analysis is quite blind to hamhandedness in its construction, so long as crude consistency requirements are met. A user could presumably deviate with impunity from the following proposal.

Let $\hat{f}_n(0)$, $D_j \hat{f}_n(0)$, $D_{jk} \hat{f}_n(0)$ be any consistent estimators of f and its respective derivatives at 0, based on an independent auxiliary sample, a vanishingly small fraction of the observations, say. Such estimates do exist, e.g. as obtained by differentiating a kernel estimate with slowly decreasing bandwidths.

With no loss of consistency, we consider $|D_j \hat{f}_n(0)|$ and $|D_{jk} \hat{f}_n(0)|$ truncated to 0 in any region where they violate the bounds laid down on the corresponding true derivatives.

For a curve estimate, we construct the quadratic approximation

$$\hat{f}_n(x) = \hat{f}_n(0) + \sum D_j \hat{f}_n(0) x_j + \tfrac{1}{2} \sum \sum D_{jk} \hat{f}_n(0) x_j x_k$$

(accurate only near 0, but adequate for our analytic needs) and winsorize $\hat{f}_n(x)$ without notational change, above and below at the respective bounds U_0 and L_0. Again this does not harm consistency near 0 of $\hat{f}_n$ and its derivatives.

Denote by $\mathcal{F}_n$ the σ-field generated by the pilot sample, and suppose $\bar{f}$ in our infeasible estimator (1) is replaced by $\bar{\hat{f}} = \hat{f} \vee \frac{1}{10} \hat{f}(0)$. Then bias $(f_n(0)) = E E_{\mathcal{F}_n}[f_n(0) - f(0)]$, and the inner expectation is the same as at (3) with $\bar{\hat{f}}$ replacing $\bar{f}$.

The following random convergence lemma justifies invoking consistency under the integral.

LEMMA. *If* $\{Y_n(v)\colon n = 1, 2, \cdots, \infty;\ v \in K \subset \mathbb{R}^p$ *a compact*$\}$ *is a jointly measurable, uniformly bounded family of random variables, then* $Y_n(v) \to_{Pr} Y_\infty(v)$ *for (almost) every* v, *implies*

$$E \int Y_n(v) \, dv \to E \int Y_\infty(v) \, dv.$$

A proof (which could go through with weakened assumptions) follows standard lines in measure theory, and is left to the reader.

Examining (4), it is seen that we need uniform convergence to $f(0)$, $D_j f(0)$ and $D_{jk} f(0)$, of the respective quantities $\hat{f}(b_n v)$, $D_j \hat{f}(b_n v)$, $D_{jk} \hat{f}(b_n v)$ for v ranging over a fixed compact;

we argue for a first derivative; the others are no harder:

$$\sup_{|v|\le H}|D_j\hat{f}(b_n v) - D_j f(0)| = \sup_{|v|\le H}|D_j\hat{f}_n(0) + \textstyle\sum_\ell D_{\ell j}\hat{f}_n(0) b_n v - D_j f(0)|$$

$$\le p b_n H U_2 + |D_j\hat{f}_n(0) - D_j f(0)| = o_p(1)$$

by consistency of $\hat{f}_n(0)$, as required. Passing to the limit, and noting that our Y_∞ is actually not random, we have shown that our bias behaves just as before.

The variance is dealt with by noting its decomposition as

$$\operatorname{Var} f_n(0) = E \operatorname{Var}_{F_n} f_n(0) + \operatorname{Var} E_{F_n} f_n(0);$$

the expected conditional variance is handled just as the bias has been, and we observe that the conditional expectation in the other term has variance no larger than the mean squared conditional bias which can evidently be absorbed into $o(b_n^4)$.

The adaptation result is in the same spirit as one of Woodroofe's (1970), but is less critically dependent on a prescribed form of the pilot estimate, and more so on the independence between the pilot and the main phase. There is evidence that a tightness argument, which has been applied to the fixed bandwidth estimator (Abramson, 1982) would permit us to dispense with this independence as well, but the calculations are unwieldy.

Now by contrast with the fixed bandwidth estimator where (for fixed f) there is a "best" sequence $b_n = B_f n^{-1/5}$ (balancing squared bias against variance), and a corresponding "best" MSE $\sim k_f n^{-4/5}$ (Rosenblatt, 1971), the upshot of our findings here is that by taking B sufficiently large in $b_n = Bn^{-1/5}$, the asymptotic MSE can be shrunk to an arbitrarily small multiple of $n^{-4/5}$. In fact, there is even a rate at which B could grow with n to achieve MSE $= o(n^{-4/5})$, but this claim is marred by a real feasibility objection; the decay rate of the residual bias to 0 (which governs how fast B might grow) cannot be known more finely than our established $o(b_n^2)$, because there is critical dependence on the consistency rates available for f'', which in our Sobolev-like class, can be arbitrarily slow. As for uniform rates over this class, it is a matter for further investigation to what extent bandwidth variation copes simultaneously with the unfavorable densities (with second derivatives "only just" continuous near 0, hence capable of inflating the residual bias for finite n), and how it compares with its locally matched competitor, the estimate using fixed kernels with vanishing second moments. Of course the work of Farrell (1972) will preclude a uniform result such as

$$\inf \sup_f \limsup n^{4/5} E[\hat{f}(0) - f(0)]^2 = 0,$$

the infimum being taken over estimating procedures.

A Monte Carlo study (Abramson, 1981) indicates that the technique is no asymptotic curiosity; implementation is straightforward, and makes for sharp improvement over fixed bandwidths. By contrast, as Sacks and Ylvisaker (1981) comment, the constrained kernel estimators need prodigious sample sizes before their optimality detectably takes hold.

On the possibility of other solutions. Our proposal forcing the bias to vanish appears unmotivated, and the reader may question whether different dependences could achieve the same effect. Restricting ourselves to one dimension for simplicity, let us briefly rework the bias calculation from (3), carrying the unassigned function $c(\cdot)$ in place of $\bar{f}(\cdot)^{1/2}$.

Omitting the algebraic details (see Abramson, 1981) (5) becomes

$$\text{Bias} = b_n^2 \left[\frac{1}{2} c(0)^{-2} f''(0) \int v^2 w(v)\, dv + c(0)^{-3} c'(0) f'(0) \int v^2 w(v)\, dv \right.$$

$$\left. + c(0)^{-3} c'(0) f'(0) \int v^3 w'(v)\, dv + \frac{1}{2} c(0)^{-3} c''(0) f(0) \int v^2 w(v)\, dv \right.$$

$$+\frac{1}{2}c(0)^{-3}c''(0)f(0)\int v^3w'(v)\,dv + c(0)^{-4}c'(0)^2f(0)\int v^3w'(v)\,dv$$

$$+\frac{1}{2}c(0)^{-4}c'(0)^2f(0)\int v^4w''(v)\,dv\Big] + o(b_n^2)$$

which further reduces to

$$\text{Bias} = b_n^2\int v^2w(v)\,dv\,c(0)^{-2}\left[\frac{1}{2}f''(0) - 2\frac{c'(0)}{c(0)}f'(0) - \frac{c''(0)}{c(0)}f(0) + 3\left\{\frac{c'(0)}{c(0)}\right\}^2 f(0)\right].$$

Note how this generalizes Rosenblatt's (1971) formula for fixed bandwidths. The squared bias is asymptotically minimized when the expression in brackets vanishes. Furthermore, invoking a natural location equivariance argument viz. that $c(\cdot)$ corresponding to a shifted density should itself simply shift correspondingly, we replace all zero arguments in the bracket by a variable x, to get a differential equation

$$\frac{1}{2}f''(x) - 2\frac{c'(x)}{c(x)}f'(x) - \frac{c''(x)}{c(x)}f(x) + 3\left\{\frac{c'(x)}{c(x)}\right\}^2 f(x) = 0,$$

(an expression that the bias actually vanishes to order b_n^2 over a whole interval about 0). Solving this proved rather troublesome, but a substitution validates the general solution

$$c(x) = f(x)^{1/2}(Ax + B)^{-1/2};\quad A, B \text{ arbitrary constants.}$$

Again, location considerations render any nonzero assignment of A unappealing, and we are left essentially with our inverse square root proposal.

Variance stabilization—a logogram. It is well known that when using fixed bandwidth estimates to construct confidence intervals for univariate densities, a square root transformation (Tukey's Rootogram) approximately frees the variance of dependence on the underlying density. A benefit is that for small bandwidths, when the variance swamps the squared bias and an asymptotic normality law comes into force, an approximate confidence interval for $f(0)^{1/2}$ based on the rootogram has radius free of $f(0)$, and in fact confidence bands for the whole curve have constant width. Other powers are indicated for other dimensions. For our estimator the dependence of the variance on f is linear in $f(0)^2$ (regardless of dimension), and the asymptotic formula

$$\text{Var}\{h(\hat{f})\} \doteq \{h'(\hat{f})\}^2\text{Var}(\hat{f})$$

indicates the stabilizing transformation $h(f) = \log f$ in place of Tukey's square root.

Acknowledgment. This work formed part of the author's dissertation written under the supervision of Peter Bickel, whose guidance and suggestions are gratefully acknowledged.

REFERENCES

ABRAMSON, I. (1981). On kernel estimates of probability densities. Doctoral dissertation, unpublished.

ABRAMSON, I. (1982). Arbitrariness of the pilot estimator in adaptive kernel methods. *J. Multivariate Anal.* **12,** to appear.

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

FARRELL, R. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180.

MACK, Y. P. and ROSENBLATT, M. (1979). Multivariate k-nearest neighbor density estimates. *J. Multivariate Anal.* **9** 1–15.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

SACKS, F. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Math. Statist.* **9** 334–346.
WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.* **41** 1665–1671.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093