The Annals of Statistics
1992, Vol. 20, No. 2, 762–778

ON GLOBAL PROPERTIES OF VARIABLE BANDWIDTH DENSITY ESTIMATORS

BY PETER HALL

Australian National University

It is argued that mean integrated squared error is not a useful measure of the performance of a variable bandwidth density estimator based on Abramson's square root law. The reason is that when the unknown density f has even moderately light tails, properties of those tails drive the formula for optimal bandwidth, to the virtual exclusion of other properties of f. We suggest that weighted integrated squared error be employed as the performance criterion, using a weight function with compact support. It is shown that this criterion is driven by pointwise properties of f. Furthermore, weighted squared-error cross-validation selects a bandwidth which gives first-order asymptotic optimality of an adaptive, feasible version of Abramson's variable bandwidth estimator.

1. Introduction. A drawback to traditional methods of kernel density estimation is that fast rates of convergence can only be achieved using kernels which take negative values. This means that a classical kernel density estimator which enjoys a particularly fast rate of convergence must necessarily take negative values over part of its range, usually out in the tails. That feature can be disconcerting, to say the least, to a practical statistician.

The variable bandwidth method introduced by Victor [18], Breiman, Meisel and Purcell [3] and Abramson [1] overcomes these difficulties to a large extent. It provides many of the advantages of a fourth-order kernel estimator, without the disadvantage of negativity. One drawback is the difficulty of computing an appropriate bandwidth, and in this paper we address that important practical problem.

Our contributions are twofold. First, we show that unless the underlying density function has extremely heavy tails—so heavy that in one dimension the sampling distribution must have infinite third moment—then unweighted mean integrated squared error is not a useful measure of the performance of Abramson's variable bandwidth estimator. The reason is that if the unknown density f has even moderately light tails, then tail properties of f drive the bias contribution to mean integrated squared error. Indeed, tail properties dominate to such an extent that properties of f within any finite region are virtually overlooked if integrated squared error is used as the performance criterion. For example, these problems can arise if the tails of the univariate density f decrease like $|x|^{-\alpha}$ as $|x| \to \infty$, where $\alpha > 1$. We shall show in Section 2 that in this circumstance, the bandwidth which minimizes integrated

Received November 1989; revised May 1991.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62H12.
Key words and phrases. Cross-validation, density estimation, integrated squared error, square root law, variable bandwidth.

squared error of Abramson's estimator equals a constant multiple of $n^{-\beta(\alpha)}$, where $\beta(\alpha) = 1/9$ for $\alpha < 7/2$ and $\beta(\alpha) \uparrow 1/5$ as $\alpha \to \infty$. Now, $n^{-1/9}$ is the size of the optimal bandwidth for variable bandwidth density estimation at any fixed point, and $n^{-1/5}$ is the optimal bandwidth size for a traditional fixed bandwidth estimator. Therefore, the operation of minimizing integrated squared error will drive Abramson's variable bandwidth estimator away from its pointwise optimum in the direction of its traditional fixed bandwidth counterpart as the tails of f become lighter.

Second, we argue that in view of these results, one should consider *weighted* integrated squared error as a measure of performance, and one might use a weight function which vanishes outside a compact set. We show that in this circumstance, a version of squared-error cross-validation may be used to select the appropriate bandwidth for a practical, adaptive version of Abramson's variable bandwidth estimator and that this method produces a data-driven, asymptotically optimal estimator. This result will be discussed in Section 3.

Some discussion of the pros and cons of cross-validation is appropriate here. In the case of fixed-bandwidth density estimation, the disadvantages of cross-validation have been noted by several authors (e.g., [10, 13]). Those disadvantages include slow convergence rates and high sampling variability, and may be alleviated by using more recent plug-in methods (e.g., [10, 13, 15]). However, the latter approach requires explicit estimation of the integral of at least the dominant term in an asymptotic formula for squared bias. The analogue of that formula is highly complicated in the context of variable bandwidth density estimators. There, bias is proportional to the fourth derivative of the inverse of the density when data are univariate, and is even more complex in higher dimensions. This complexity is compounded by the need to estimate the first four derivatives of the density, and that is an awkward problem in itself. By way of comparison, cross-validation achieves this goal implicitly, without the complexity of an explicit approach, see also Jones [11]. Thus, it is not clear that plug-in rules offer a compelling alternative to cross-validation in variable bandwidth problems.

Construction of a practical variable bandwidth estimator entails a two-step procedure. The first stage produces a pilot estimator using a fixed bandwidth and the second stage yields the variable bandwidth estimator. To more clearly explain the way in which the two bandwidths operate, we must define both the pilot estimator and the variable bandwidth estimator. Let $X_1, \ldots, X_n$ denote a sample of p-vectors from the distribution having p-variate density f, which we wish to estimate. Let K, the kernel function, be a known p-variate density which is symmetric in each variable. The traditional form of kernel estimator based on kernel K and bandwidth h_1 is

$$\tilde{f}(x) = \frac{1}{nh_1^p} \sum_{i=1}^{n} K\left\{\frac{x - X_i}{h_1}\right\}; \tag{1.1}$$

see, for example, Silverman ([16], Chapters 3 and 4). Abramson [1] argued in favour of a variable kernel estimator in which the bandwidth for computations

involving X_i is taken inversely proportional to the square root of the density at X_i; thus, the estimator is

$$\hat{f}(x) = \frac{1}{nh^p} \sum_{i=1}^{n} f(X_i)^{p/2} K\left\{ \frac{(x - X_i) f(X_i)^{1/2}}{h} \right\}, \tag{1.2}$$

where h is the constant of proportionality. Of course, the ideal estimator $\hat{f}$ is not feasible because it depends on the unknown density f. A practical version would have f replaced by the pilot estimator $\tilde{f}$ defined at (1.1), giving

$$\check{f}(x) = \check{f}(x|h) = \frac{1}{nh^p} \sum_{i=1}^{n} \tilde{f}(X_i)^{p/2} K\left\{ \frac{(x - X_i) \tilde{f}(X_i)^{1/2}}{h} \right\}. \tag{1.3}$$

Construction of $\check{f}$ demands choice of two bandwidths—selection of h_1 for the pilot estimator $\tilde{f}$ and selection of h for the final estimator $\check{f}$.

We note that minor, but important, modifications should be made to the definitions at (1.2) and (1.3), to ensure that the estimators enjoy the excellent bias properties claimed for them by, for example, Abramson [1]. These modifications prevent very large X_i's, that is, X_i's a long way from x, from adversely influencing the bias formulae. Details are given in Hall, Hu and Marron [8]. It is sufficient to introduce an indicator function $I\{|x - X_i| \le (1 + |x|)\log h^{-1}\}$ into the definitions, obtaining

$$\hat{f}(x) = \frac{1}{nh^p} \sum_{i=1}^{n} f(X_i)^{p/2} K\left\{ \frac{(x - X_i) f(X_i)^{1/2}}{h} \right\} \times I\{|x - X_i| \le (1 + |x|)\log h^{-1}\}, \tag{1.4}$$

$$\check{f}(x) = \frac{1}{nh^p} \sum_{i=1}^{n} \tilde{f}(X_i)^{p/2} K\left\{ \frac{(x - X_i) \tilde{f}(X_i)^{1/2}}{h} \right\} \times I\{|x - X_i| \le (1 + |x|)\log h^{-1}\}. \tag{1.5}$$

Abramson [1], Silverman ([16], page 104f) and Hall and Marron [9] showed that the advantage of $\hat{f}$ over $\tilde{f}$ is that bias is reduced by an order of magnitude, from h_1^2 in the case of $\tilde{f}$ to h^4 in the case of $\hat{f}$, with relatively little impact on variance. Hall and Marron proved that these advantages extend to the feasible for $\check{f}$. Indeed, asymptotic bias formulae for $\hat{f}$ and $\check{f}$ are identical, provided h_1 is chosen reasonably close to the bandwidth which optimizes performance of $\tilde{f}$. While the variance of $\check{f}$ is a little larger than the variance of $\hat{f}$, both are of the same order of magnitude. General formulae exist ([6]) for the bias of one-dimensional variable bandwidth curve estimators.

2. Mean integrated squared error of the ideal estimator. In this section we show that in the univariate case, unless the underlying distribution has particularly heavy tails, asymptotic properties of mean integrated squared error (MISE) of the estimator $\hat{f}$, defined at (1.4), are driven by tail behaviour of f. Therefore it would be most inappropriate to use MISE as a global

criterion for measuring the performance of $\hat{f}$. Similar results may be proved for the p-variate case.

It turns out that for univariate distributions with densities whose tails decrease like $|x|^{-\alpha}$, where $\alpha > 1$, the condition $\alpha < 7/2$ is necessary and sufficient to ensure that MISE is not driven by tail behaviour of f. This condition is equivalent to $E(|X|^{(5/2)+\varepsilon}) = \infty$ for all $\varepsilon > 0$ and fails (for example) for the t distribution with $\nu \geq 3$ degrees of freedom. If $\alpha > 7/2$, then the bandwidth which minimizes MISE is of size $n^{-(\alpha-2)/(5\alpha-4)}$ and the minimum MISE is of size $n^{-2(2\alpha-1)/(5\alpha-4)}$. [The exponent $(\alpha - 2)/(5\alpha - 4)$ was formerly called $\beta(\alpha)$.] As $\alpha \to \infty$, $n^{-(\alpha-2)/(5\alpha-4)} \to n^{-1/5}$ and $n^{-2(2\alpha-1)/(5\alpha-4)} \to n^{-4/5}$, which is, of course, the size of MISE for an ordinary, nonvariable-bandwidth kernel estimator. Thus, as the tails of the distribution become lighter, the optimal global performance of $\hat{f}$ (as measured by MISE) converges to that of an ordinary kernel estimator.

A key step in verifying these and other results is to develop a formula for MISE. To that end, let us assume the usual bandwidth conditions $h \to 0$ and $nh \to \infty$. The variance contribution to MISE is easy to describe; it is

$$\int \operatorname{var} \hat{f} = \frac{1}{nh^2} \int \operatorname{var}\left[f(X)^{1/2} K\left\{ \frac{(x - X) f(X)^{1/2}}{h} \right\} \right.$$

$$\left. \times I\{|x - X| \leq (1 + |x|)\log h^{-1}\} \right] dx \tag{2.1}$$

$$\sim \frac{1}{nh}\left(\int K^2\right)\left(\int f^{3/2}\right).$$

A sufficient regularity condition is that K be bounded and compactly supported and f be bounded.

The bias contribution is more difficult to describe. Indeed, Silverman [16] used a computer algebraic manipulation package to calculate the bias of $\hat{f}(x)$ in the case of fixed x, and in the present instance we need to compute the integral of the square of this formula. It greatly simplifies our proof if we suppose that f is an analytic function, and so we shall make this assumption. We ask that there exist constants $C_1, C_2, C_3 > 0$ such that

$$f(x) \sim C_1|x|^{-\alpha}, \qquad \left|(d/dx)^4 f(x)^{-1}\right| \sim C_2|x|^{\alpha-4} \quad \text{as } |x| \to \infty, \tag{2.2}$$

and for all integers $s \geq 1$ and all x,

$$(s!)^{-1}\left|(d/dx)^{2s+2} f(x)^{-s}\right| \leq C_3^s(1 + |x|)^{(\alpha-2)s-2}. \tag{2.3}$$

Condition (2.3) is typically satisfied by an analytic function enjoying the property (2.2); consider for example the function $f(x) = C_4(1 + x^2)^{-\alpha/2}$. We assume in addition that K is a bounded symmetric, compactly supported probability density.

The bias contribution to MISE assumes three different forms, depending on whether $\alpha < 7/2$, $\alpha = 7/2$ or $\alpha > 7/2$. Define

$$a(x) = \left\{\int u^4 K(u)\,du\right\}(4!)^{-1}(d/dx)^4 f(x)^{-1}$$

if $\alpha < 7/2$,

$$a(x) = 2C_1\alpha^{-1}x^{(2/\alpha)-3}\int_0^\infty \big[K\{C_1u(1 + x^{(2/\alpha)-1}u^{-2/\alpha})\} + K\{C_1u(1 - x^{(2/\alpha)-1}u^{-2/\alpha})\}\big]u^{2-(2/\alpha)}\,du - C_1x^{-\alpha},$$

if $\alpha > 7/2$,

$$b = \begin{cases} \displaystyle\int_{-\infty}^{\infty} a(x)^2\,dx, & \text{if } \alpha < 7/2, \\ \displaystyle(8/3)\left\{\int u^4K(u)\,du\right\}^2(4!)^{-2}C_2^2, & \text{if } \alpha = 7/2, \\ \displaystyle 2\int_0^\infty a(x)^2\,dx, & \text{if } \alpha > 7/2, \end{cases}$$

$$H = \begin{cases} h^8, & \text{if } \alpha < 7/2, \\ h^8 \log h^{-1}, & \text{if } \alpha = 7/2, \\ h^{2(2\alpha-1)/(\alpha-2)}, & \text{if } \alpha > 7/2. \end{cases}$$

[The convergence of $\int a^2$ in the case $\alpha > 7/2$ follows from arguments in steps (i) and (ii) of Case (b) in the proof of Theorem 2.1.]

THEOREM 2.1. *Under the stated conditions,* $\int_{-\infty}^{\infty}(\text{bias } \hat f)^2 \sim bH$ *as* $h \to 0$.

REMARK 2.1. In the case $\alpha > 7/2$ we have $H = h^{2(2\alpha-1)/(\alpha-2)} = h^8h^{-2(2\alpha-7)/(\alpha-2)}$, implying that H decreases at a slower rate than h^8. Hence, it is only in the case $\alpha < 7/2$ that h decreases at rate h^8; the rate is slightly slower than h^8 when $\alpha \geq 7/2$.

REMARK 2.2. When $\alpha \geq 7/2$, the constant b appearing in the asymptotic formula $\int(\text{bias})^2 \sim bH$ depends on f only through C_1 and α, which describe only the extreme tails of f. Therefore, except for the case $\alpha < 7/2$, asymptotic properties of the bias contribution to MISE are driven by tail behaviour of f.

REMARK 2.3. Combining Theorem 2.1 and expansion (2.1) we see that MISE admits the formula

$$\text{MISE} = \int E(\hat f - f)^2 = \int \text{var } \hat f + \int(\text{bias } \hat f)^2$$

$$\sim (nh)^{-1}\left(\int K^2\right)\left(\int f^{3/2}\right) + bH.$$

This formula, and elementary calculus, may be used to deduce the bandwidth which asymptotically minimizes MISE to first order; it equals a constant multiple of η_1, where

$$\eta_1 = \begin{cases} n^{-1/9}, & \text{if } \alpha < 7/2, \\ (n \log n)^{-1/9}, & \text{if } \alpha = 7/2, \\ n^{-(\alpha-2)/(5\alpha-4)}, & \text{if } \alpha > 7/2. \end{cases}$$

The minimum MISE equals a constant multiple of η_2, where

$$\eta_2 = \begin{cases} n^{-8/9}, & \text{if } \alpha < 7/2, \\ (n^{-8} \log n)^{1/9}, & \text{if } \alpha = 7/2, \\ n^{-2(2\alpha-1)/(5\alpha-4)}, & \text{if } \alpha > 7/2. \end{cases}$$

REMARK 2.4. There exist analogues of all these results in the case of a density f whose upper tail decreases like $x^{-\alpha_1}$ and lower tail decreases like $|x|^{-\alpha_2}$, where $\alpha_1, \alpha_2 > 1$. Asymptotic properties of MISE are determined by the larger of α_1 and α_2. In particular, the minimum value of MISE is of order $n^{-8/9}$ if and only if $\max(\alpha_1, \alpha_2) < 7/2$, and in this case the bandwidth which minimizes MISE is of size $n^{-1/9}$.

PROOF OF THEOREM 2.1.

CASE (a). $\alpha < 7/2$. It is known (Silverman [16], pages 104–105 and Hall and Marron [9]; see also [6]) that $E\hat{f}(x) - f(x) = h^4 a(x) + O(h^6)$ for fixed x. Since $\alpha < 7/2$, then by (2.2), $\int a^2 < \infty$. Therefore the formula $\int(\text{bias})^2 \sim h^8 \int a^2$ is at least plausible. Rigorous verification is along lines outlined in step (i) of Case (b).

CASE (b). $\alpha > 7/2$. Let $0 < \varepsilon < \lambda < \infty$, where ε is small and λ large. Put $l = h^{-2/(\alpha-2)}$ and write

$$\begin{aligned} \int_{-\infty}^{\infty} (\text{bias})^2 &= \left(\int_{|x| \le \varepsilon l} + \int_{\varepsilon l < |x| \le \lambda l} + \int_{|x| > \lambda l} \right) (\text{bias})^2 \\ &= A_1 + A_2 + A_3, \end{aligned} \tag{2.4}$$

say. We estimate A_1, A_2 and A_3 separately, showing first that for any $\delta > 0$, we may choose ε so small and λ so large that for sufficiently small h, $A_1 \le \delta h^8 l^{2\alpha-7}$ and $A_3 \le \delta h^8 l^{2\alpha-7}$. Then we prove that for a constant $b_1 = b_1(\varepsilon, \lambda)$, whose limit as $\varepsilon \to 0$ and $\lambda \to \infty$ is b, we have $A_2 \sim b_1 h^8 l^{2\alpha-7}$. Note that $h^8 l^{2\alpha-7} = h^{2(2\alpha-1)/(\alpha-2)}$.

(i) Bound for A_1. In the range $|x| \le \varepsilon l$ we compute a bound for bias by using Taylor expansion. Techniques in [6] may be used to show that the coefficient of h^s in a Taylor expansion of $E\hat{f}(x) - f(x)$ is zero for odd s and is bounded in absolute value by $c_1^s(1 + |x|)^{-\alpha+(\alpha-2)s/2}$ for even s, where $c_1, c_2, \ldots$ denote

generic positive constants. Therefore,

$$\begin{aligned}\left|E\hat{f}(x) - f(x)\right| &\le \sum_{s=2}^{\infty} h^{2s} c_2^s (1 + |x|)^{(\alpha-2)s-\alpha} \\ &= h^4 c_2^2 (1 + |x|)^{\alpha-4} \left\{1 - h^2 c_2 (1 + |x|)^{\alpha-2}\right\}^{-1},\end{aligned} \tag{2.5}$$

assuming that $h^2 c_2 (1 + |x|)^{\alpha-2} < 1/2$. The latter inequality is satisfied if $|x| \le \varepsilon l$ and ε is sufficiently small. For such an ε,

$$A_1 = \int_{|x| \le \varepsilon l} \left\{E\hat{f}(x) - f(x)\right\}^2 dx \le c_3 h^8 \int_0^{\varepsilon l} (1 + x)^{2\alpha-8}\, dx \le c_4 h^8 (\varepsilon l)^{2\alpha-7}.$$

Since $\alpha > 7/2$, then given $\delta > 0$ we may choose $\varepsilon > 0$ so small that $A_1 \le \delta h^8 l^{2\alpha-7}$.

(ii) Bound for A_3. In the range $|x| > \varepsilon l$, we use the inequality

$$\begin{aligned}E\hat{f}(x) &\le \int_{-\infty}^{\infty} f(x - hy)^{3/2} K\left\{y f(x - hy)^{1/2}\right\} dy \\ &= 2\alpha^{-1} h^{-(2/\alpha)-1} |x|^{2/\alpha} \\ &\quad \times \int_0^{\infty} \Big[f\left(|x/hu|^{2/\alpha}\right)^{3/2} K\left\{h^{-1}\left(x - |x/hu|^{2/\alpha}\right) f\left(|x/hu|^{2/\alpha}\right)^{1/2}\right\} \\ &\qquad + f\left(-|x/hu|^{2/\alpha}\right)^{3/2} K\left\{h^{-1}\left(x + |x/hu|^{2/\alpha}\right) f\left(-|x/hu|^{2/\alpha}\right)^{1/2}\right\}\Big] \\ &\qquad \times u^{-(2/\alpha)-1}\, du.\end{aligned} \tag{2.6}$$

(To obtain the second identity, change variable in the first integral, from y to $u > 0$, where $y = h^{-1}x \pm h^{-1}|x/hu|^{2/\alpha}$.)

Let $L(x, u)u^{-(2/\alpha)-1}$ denote the integrand of the second integral on the right-hand side of (2.6). Assume that $|x| > \lambda l$. If in addition $|x/hu| \le 1$, then, since K is compactly supported, there exists a constant $c_1 > 0$ such that

$$L(x, u) \le 2(\sup f^{3/2})(\sup K) I(|x| \le c_1 h).$$

But $I(|x| \le c_1 h) = 0$ for $x > \lambda l$ and h sufficiently small, and so we may assume throughout our estimation of $L(x, u)$ that $|x/hu| > 1$.

Define $K_1(x, u) = K\{h^{-1}(x - |x/hu|^{2/\alpha}) f(|x/hu|^{2/\alpha})^{1/2}\}$, $K_2(x, u) = K\{h^{-1}(x + |x/hu|^{2/\alpha}) f(-|x/hu|^{2/\alpha})^{1/2}\}$. Since $|x| > \lambda l$, then $|x|^{-(\alpha-2)/2} h^{-1} \le \lambda^{-(\alpha-2)/2}$. If in addition $|x/hu|^{2/\alpha} > |x|/2$, then

$$u < 2^{\alpha/2} |x|^{-(\alpha-2)/2} h^{-1} \le 2^{\alpha/2} \lambda^{-(\alpha-2)/2}.$$

On the other hand, if $|x/hu|^{2/\alpha} \le |x|/2$, then, since $|x/hu| > 1$,

$$h^{-1}\left|x \pm |x/hu|^{2/\alpha}\right| f\left(\mp |x/hu|^{2/\alpha}\right)^{1/2} \ge c_2 h^{-1} |x|\, |hu/x| = c_2 u.$$

Therefore $K_j(x, u) \le (\sup K) I(u \le c_3)$; here we have used the fact that K is compactly supported. If λ is so large that $2^{\alpha/2} \lambda^{-(\alpha-2)/2} \le c_3$, then $K_j(x, u) \le$

(sup K)$I(u \le c_3)$ regardless of the value of $|x/hu|^{2/\alpha} - |x|/2$. It follows that $L(x, u) \le c_4|hu/x|^3 I(u \le c_3)$ and so by (2.6),

$$E\hat{f}(x) \le c_5 h^{-(2/\alpha)-1}|x|^{2/\alpha}\int_0^{c_3}|hu/x|^3 u^{-(2/\alpha)-1}\,du \le c_6 h^{2-(2/\alpha)}|x|^{(2/\alpha)-3}.$$

Trivially, $f(x) \le c_7|x|^{-\alpha}$ and so

$$\begin{aligned} A_3 &= \int_{|x|>\lambda l}\{E\hat{f}(x) - f(x)\}^2\,dx \\ &\le c_8\left(h^{4(\alpha-1)/\alpha}\int_{\lambda l}^{\infty}x^{2(2-3\alpha)/\alpha}\,dx + \int_{\lambda l}^{\infty}x^{-2\alpha}\,dx\right) \\ &\le c_9(\lambda^{(4/\alpha)-5} + \lambda^{1-2\alpha})(h^{4-(4/\alpha)}l^{(4/\alpha)-5} + l^{1-2\alpha}). \end{aligned}$$

But $h^{4-(4/\alpha)}l^{(4/\alpha)-5} = l^{1-2\alpha} = h^8 l^{2\alpha-7}$ and $c_9(\lambda^{(4/\alpha)-5} + \lambda^{1-2\alpha}) \le \delta$ if λ is sufficiently large. For such a λ, we have $A_3 \le \delta h^8 l^{2\alpha-7}$.

(iii) Formula for A_2. Take $x = lz$ in the analogue of (2.6), with the indicator function included, obtaining the following relation uniformly in $\varepsilon < z < \lambda$: $E\hat{f}(x) \sim l^{-\alpha}C_1\gamma(x)$, where

$$\begin{aligned} \gamma(x) = 2\alpha^{-1}|z|^{(2/\alpha)-3}\int_0^{\infty}\Big[&K\{C_1u(\operatorname{sgn} z - |z|^{(2/\alpha)-1}u^{-2/\alpha})\} \\ &+K\{C_1u(\operatorname{sgn} z + |z|^{(2/\alpha)-1}u^{-2/\alpha})\}\Big]u^{2-(2/\alpha)}\,du. \end{aligned}$$

Trivially, $f(x) \sim C_1 l^{-\alpha}|z|^{-\alpha}$. Therefore,

$$A_2 = l\int_{\varepsilon<|z|<\lambda}\{E\hat{f}(lz) - f(lz)\}^2\,dz \sim l^{1-2\alpha}b_1(\varepsilon,\lambda) = h^8 l^{2\alpha-7}b_1(\varepsilon,\lambda),$$

where

$$b_1(\varepsilon,\lambda) = C_1^2\int_{\varepsilon<|z|<\lambda}\{\gamma(z) - z^{-\alpha}\}^2\,dz.$$

CASE (c). $\alpha = 7/2$. The argument is virtually identical to that in Case (b). In particular, we divide $\int(\text{bias})^2$ into three parts as in (2.4). The estimation of A_2 and A_3 carried out in steps (iii) and (ii), respectively, proceeds as before. It yields $A_2 + A_3 \le \text{const.}\ h^8$. The contribution from A_1 dominates both A_2 and A_3. To determine a formula for A_1, observe that in place of (2.5),

$$\begin{aligned} |E\hat{f}(x) - f(x) - h^4 a(x)| &\le \sum_{s=3}^{\infty}h^{2s}c^s(1+|x|)^{(\alpha-2)s-\alpha} \\ &= h^6c^3(1+|x|)^{2\alpha-6}\{1 - h^2c(1+|x|)^{\alpha-2}\}^{-1}, \end{aligned}$$

where $a(x)$ has the definition from Case (a) ($\alpha < 7/2$). Thus it may be proved

that

$$A_1 = h^8 \int_{|x| \le \varepsilon l} a(x)^2 \, dx + O(h^8)$$

$$\sim 2c_0^2 h^8 \int_1^{\varepsilon l} C_2^2 x^{2\alpha - 8} \, dx$$

$$\sim 2c_0^2 C_2^2 \log l = (8/3) c_0^2 C_2^2 \log h^{-1},$$

where $c_0 = \{\int u^4 K(u) \, du\}(4!)^{-1}$. The desired result now follows from (2.4). □

3. Cross-validation for an adaptive estimator. In view of the results noted in Section 2, integrated squared error is not really a valid criterion for assessing the performance of variable bandwidth density estimators. A viable alternative is weighted integrated squared error,

$$\text{WISE} = \int (\check{f} - f)^2 w,$$

where w is an appropriate bounded, nonnegative function and $\check{f}$ denotes the p-variate adaptive estimator defined in Section 1. [Because we shall be treating only a compact set of values x, we may on this occasion replace the indicator function in (1.5) by $I(|x - X_i| \le A)$ for arbitrary $A > 0$. Thus,

$$\check{f}(x) = \frac{1}{nh^p} \sum_{i=1}^n \tilde{f}(X_i)^{p/2} K\left\{ \frac{(x - X_i)\tilde{f}(X_i)^{1/2}}{h} \right\} I(|x - X_i| \le A).$$

This definition is assumed throughout the present section.]

A thorough, detailed analysis of appropriate w's involves trade-offs between tail behaviour of f and tail behaviour of w. In the case $p = 1$, this study could be conducted along the lines of Section 2, but would be too specialized and intricate to have much practical bearing. Therefore, we shall instead assume that w has compact support, which allows us to work with quite general densities f and general $p \ge 1$.

One choice of w would be

$$w(\mathbf{x}) = \begin{cases} 1, & \text{for } \left\| \hat{\Sigma}^{-1/2}(\mathbf{x} - \hat{\mu}) \right\|^2 \le z_\gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{3.1}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ denote, respectively, the sample mean and variance, $\| \cdot \|$ is Euclidean distance and z_γ is the upper $(1 - \gamma)$-level critical point of the chi-squared distribution on p degrees of freedom. Appropriate γ's are $\gamma = 0.1, 0.2$. This particular w averages the performance of $\check{f}$ over that ellipsoid, centered at the sample mean, which contains a proportion $1 - \gamma$ of the distribution. Strictly speaking, the analysis which we shall give does not allow for a random w, but the random component in the definition at (3.1) is easily disposed of by a second, subsidiary argument.

In the prescription (1.1) for $\tilde{f}$ in Section 1, we shall take h_1 to be a bandwidth of size $n^{-1/(p+4)}$, the optimal size for estimating f using the usual

kernel estimator $\tilde{f}$. It is convenient in proofs to take h_1 to be nonrandom and satisfy

$$C_1 n^{-1/(p+4)} \le h_1(n) \le C_2 n^{-1/(p+4)} \tag{3.2}$$

for $n \ge 1$, where C_1, C_2 are fixed constants. However, a minor refinement of our proof permits h_1 to be data-dependent, satisfying (3.2) for all sufficiently large n with probability 1. In particular, h_1 could be taken to be the bandwidth chosen by the equivalent normal kernel method (e.g., Silverman [16], pages 45ff, 86f), or by other plug-in methods, or by ordinary cross-validation (according to the prescription suggested by Rudemo [14] and Bowman [2]), or by a weighted version of cross-validation (Marron [12]); see Hall [4, 5], Silverman [16], pages 48ff, page 87f and Stone [17] for accounts of ordinary cross-validation. If h_1 were chosen empirically, then the cross-validation procedure which we shall propose would be purely automatic, once the weight function w had been selected.

To develop our cross-validation procedure, consider expanding formula (3.1) for weighted integrated squared error:

$$\text{WISE} = \int \check{f}^2 w - 2\int \check{f} f w + \int f^2 w. \tag{3.3}$$

The last term in this expansion does not depend on h and so plays no role in any procedure for minimizing WISE; and the first term is known. Therefore it is only the integral

$$I = \int \check{f} f w,$$

which requires attention. Minimizing WISE is equivalent to minimizing

$$J = J(h) = \int \check{f}^2 w - 2I.$$

To effect an estimate of I, define

$$\tilde{f}_j(x) = \frac{1}{(n-1)h_1^p} \sum_{i \ne j} K\left\{\frac{x - X_i}{h_1}\right\},$$

$$\check{f}_j(x|h) = \frac{1}{(n-1)h^p} \sum_{i \ne j} \tilde{f}_j(X_i)^{p/2} K\left\{\frac{(x - X_i)\tilde{f}_j(X_i)^{1/2}}{h}\right\} I(|x - X_i| \le Ah),$$

$$\hat{I} = \hat{I}(h) = \frac{1}{n} \sum_{j=1}^{n} \check{f}_j(X_j|h) w(X_j).$$

We take the cross-validatory criterion to be

$$\hat{J} = \hat{J}(h) = \int \check{f}^2 w - 2\hat{I}. \tag{3.4}$$

Choose $\hat{h}$ to minimize $\hat{J}$ and take $\check{f}(x|\hat{h})$ to be the estimate of f.

Theoretical support for this rule is provided by Theorem 3.1. Before stating that result, we briefly discuss the basic properties of WISE. Hall and Marron [9] showed that there exist a function g and a constant c depending only on p and K, such that

$$\check{f}(x|h) - f(x) = (nh^p)^{-1/2} c f(x)^{(p+2)/4} N(x) + h^4 g(x) + o(h^4), \tag{3.5}$$

where $N(x)$ is asymptotically normal $N(0, 1)$. For distinct values $x_1, \ldots, x_m$, the variables $N(x_i)$ are asymptotically independent. In the case $p = 1$,

$$g(x) = (1/24)\left\{\int u^4 K(u)\, du\right\}(d/dx)^4 f(x)^{-1};$$

the formula for g is more complicated in higher dimensions. It follows from (3.5) that from the viewpoint of minimizing mean squared error, the optimal h is of size $n^{-1/(p+8)}$. The techniques developed by Hall and Marron [9] may be used to establish the following intuitively obvious consequence of (3.5) and of the fact that the $N(x)$'s are asymptotically independent:

$$\int (\check{f} - f)^2 w = (nh^p)^{-1} c^2 \int f^{(p+2)/2} w + h^8 \int g^2 w + o\{(nh)^{-1} + h^8\} \tag{3.6}$$

with probability 1. [Sufficient regularity conditions for (3.6) are stated below.] Our claim is that, except for terms which either do not depend on h or equal $o\{(nh)^{-1} + h^8\}$, $\hat{J}$ is identical to WISE. Indeed,

$$\hat{J} = \int (\check{f} - f)^2 w - \frac{2}{n} \sum_{i=1}^{n} f(X_i) w(X_i) + \int f^2 w + o\left\{\frac{1}{nh} + h^8\right\}. \tag{3.7}$$

This formula follows from (3.3), (3.4), (3.6) and (3.7). It is the key to the efficacy of $\hat{J}$ as a criterion for minimizing WISE: Minimizing $\hat{J}$ is asymptotically equivalent to minimizing WISE.

These results are valid under ther following regularity conditions. Let w be a measurable function vanishing outside a compact set $\mathscr{R} \subseteq \mathbb{R}^p$, let $\mathscr{R}_\varepsilon = \{x \in \mathbb{R}^p$: for some $y \in \mathscr{R}$, $\|x - y\| \le \varepsilon\}$, and assume that for some $\varepsilon > 0$, f has four bounded, continuous derivatives on $\mathscr{R}_\varepsilon$ and f is bounded away from zero on $\mathscr{R}_\varepsilon$. Let K be a compactly supported density function, symmetric in each variable and having at least $(p/2) + 2$ bounded derivatives. (The condition of compact support may be relaxed at the expense of a longer proof; for example, we may take K to be the p-variate standard normal density function.) For constants $B_2 > B_1 > 0$, put $\mathscr{H} = \{h$: $B_1 n^{-1/(p+8)} \le h \le B_2 n^{-1/(p+8)}\}$. [Note that $n^{-1/(p+8)}$ is the size of the asymptotically optimal h.] Assume that $h_1 = h_1(n)$ satisfies (3.2). Then we have:

THEOREM 3.1. *Under the above conditions and with probability* 1,

$$\hat{I}(h) - I(h) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) w(X_i) - \int f^2 w + o\left\{\frac{1}{nh} + h^8\right\} \tag{3.8}$$

uniformly in $h \in \mathscr{H}$.

REMARK 3.1. The analogue of (3.8) in the case of ordinary cross-validation is

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{f}_j(X_j) - \int \tilde{f}\tilde{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int f^2 + o\left\{\frac{1}{nh_1} + h_1^4\right\};$$

see [5].

REMARK 3.2. Let $h_0 = B_0 n^{-1/(p+8)}$ denote the bandwidth which minimizes the sum of the two dominant terms on the right-hand side of (3.6), and choose B_1, B_2 (in the definition of $\mathscr{H}$) such that $B_1 < B_0 < B_2$. Write $\hat{h}_0, \hat{h}$ for the bandwidths which minimize WISE, $\hat{J}$, respectively, over $h \in \mathscr{H}$. We may deduce from (3.6)–(3.8) that $\hat{h}_0/h_0 \to 1$ and $\hat{h}/h_0 \to 1$ with probability 1. This establishes first-order optimality of the cross-validatory bandwidth $\hat{h}$.

PROOF OF THEOREM 3.1. For the sake of brevity, we shall assume that the distribution of X is compactly supported. The sole purpose of this assumption is to allow us to ignore the indicator function in the definitions of $\tilde{f}$ and $\tilde{f}_j$, thereby simplifying notation.

Define $\delta_j(x) = \{\tilde{f}_j(x) - f(x)\}/f(x)$. It may be proved by standard arguments, using the Borel–Cantelli lemma, Bernstein's and Markov's inequalities and Hölder continuity of K, that with probability 1,

$$\sup_{x\in\mathscr{R}_\epsilon} \max_{1\le j\le n} |\delta_j(x)| = O\{n^{-2/(p+4)}(\log n)^{1/2}\}. \tag{3.9}$$

In consequence, there exist constants $c_1, \ldots, c_R$ [with $R \ge (p/2) + 1$] such that

$$\tilde{f}_j(x)^{p/2} = f(x)^{p/2}\{1 + c_1\delta_j(x) + \cdots + c_R\delta_j(x)^R\} + O\{n^{-2(R+1)/(p+4)}(\log n)^{(R+1)/2}\}.$$

This formula and Taylor expansion give

$$\tilde{f}_j(x)^{p/2}K\{u\tilde{f}_j(x)^{1/2}\} = f(x)^{p/2}\left\{\sum_{r=0}^{R}\delta_j(x)^r l_r\right\} + O\{n^{-2(R+1)/(p+4)}(\log n)^{(R+1)/2}\},$$

where $l_r = L_r\{uf(x)^{1/2}\}$ and each function L_r is even in each variable. In particular, $L_0 = K$. Thus,

$$\hat{I} = \frac{1}{n}\sum_{j=1}^{n} w(X_j)\sum_{r=0}^{R}\frac{\delta_j(X_j)^r}{(n-1)h^p}\sum_{i\ne j} f(X_i)^{p/2}L_r\left\{\frac{(X_j - X_i)f(X_i)^{1/2}}{h}\right\} + O\{n^{-2(R+1)/(p+4)}(\log n)^{(R+1)/2}h^{-p}\}.$$

A similar expansion may be developed for I. Indeed, if we define $\delta(x) = \{\tilde{f}(x) - f(x)\}/f(x)$, then we have

$$I = \sum_{r=0}^{R} \int w(x)\delta(x)^r f(x) \left[\frac{1}{nh^p} \sum_i f(X_i)^{p/2} L_r\left\{ \frac{(x - X_i) f(X_i)^{1/2}}{h} \right\} \right] dx$$

$$+ O\{n^{-2(R+1)/(p+4)} (\log n)^{(R+1)/2} h^{-p}\}.$$

To combine these formulae, put

$$\mu_r(x) = E\left[h^{-p} f(X)^{p/2} L_r\{(x - X) f(X)^{1/2}/h\} \right],$$

$$\Delta_{rj}(x) = (n-1)^{-1} \sum_{i \neq j} \left[h^{-p} f(X_i)^{p/2} L_r\{(x - X_i) f(X_i)^{1/2}/h\} - \mu_r(x) \right],$$

$$\Delta_r(x) = n^{-1} \sum_i \left[h^{-p} f(X_i)^{p/2} L_r\{(x - X_i) f(X_i)^{1/2}/h\} - \mu_r(x) \right].$$

Then

$$\hat{I} - I = \sum_{r=0}^{R} \left[\frac{1}{n} \sum_{j=1}^{n} w(X_j) \delta_j(X_j)^r \{\Delta_{rj}(X_j) + \mu_r(X_j)\} \right.$$

$$\left. - \int w(x)\delta(x)^r f(x)\{\Delta_r(x) + \mu_r(x)\}\, dx \right] \tag{3.10}$$

$$+ O\{n^{-2(R+1)/(p+4)} (\log n)^{(R+1)/2} h^{-p}\}.$$

The next step is to deal individually with the contributions to (3.10) from the cases $r = 0, \ldots, R$. Put $\lambda_r = \int L_r$. Since L_r is even in each variable, then

$$\int u^{(i)} L_r(u)\, du = \int u^{(i_1)} u^{(i_2)} u^{(i_3)} L_r(u)\, du = 0$$

for each i, i_1, i_2, i_3. This fact and the square root law imply that $\mu_r(x) = \lambda_r + O(h^4)$. It may be proved by standard methods that, analogously to (3.9),

$$\sup_{x \in \mathscr{R}_\varepsilon} \left\{ \max_{1 \le j \le n} |\Delta_{rj}(x)| + |\Delta_r(x)| \right\} = O\{(nh^p)^{-1/2} (\log n)^{1/2}\}.$$

Hence

$$\sup_{x \in \mathscr{R}_\varepsilon} \left\{ \max_{1 \le j \le n} |\Delta_{rj}(x) + \mu_r(x) - \lambda_r| + |\Delta_r(x) + \mu_r(x) - \lambda_r| \right\}$$

$$= O\{n^{-4/(p+8)} (\log n)^{1/2}\}$$

uniformly in $h \in \mathscr{H}$. From this formula, (3.9), the analogue of the latter for δ rather than δ_j and the fact that $\lambda_r = 0$ for $r \ge 1$, we see that terms in $r \ge 2$

make a contribution to (3.10) which equals $o(n^{-8/(p+8)})$. Therefore,

$$\hat{I} - I = \sum_{r=0}^{1} \left[\frac{1}{n} \sum_{j=1}^{n} w(X_j)\delta_j(X_j)^r \{\Delta_{rj}(X_j) + \mu_r(X_j)\} - \int w(x)\delta(x)^r f(x)\{\Delta_r(x) + \mu_r(x)\}\,dx \right] + o(n^{-8/(p+8)}) \tag{3.11}$$

with probability 1.

We claim that with probability 1,

$$\left| \frac{1}{n} \sum_{j=1}^{n} \left\{ w(X_j)\delta_j(X_j)\Delta_{1j}(X_j) - \int w(x)\delta(x) f(x)\Delta_1(x)\,dx \right\} \right| + \left| \frac{1}{n} \sum_{j=1}^{n} \left\{ w(X_j)\delta_j(X_j)\mu_1(X_j) - \int w(x)\delta(x) f(x)\mu_1(x)\,dx \right\} \right| = o\{(nh)^{-1}\} \tag{3.12}$$

uniformly in $h \in \mathscr{H}$. This claim may be verified by applying the Borel–Cantelli lemma, Markov's inequality and Hölder continuity of K, once it is shown that for some $\eta > 0$, the $2m$th moment of each of the quantities within absolute value signs above is dominated by $C(m)\{(nh)^{-1}n^{-\eta}\}^{2m}$ for all integers $m \geq 1$, where $C(m)$ is a constant. This may be accomplished by a rather lengthy argument which we now give in outline.

Let $\hat{\alpha}_j - \hat{\alpha}$ denote the jth summand in either of the sums over j on the left-hand side of (3.12). For example, in the case of the first series we have

$$\hat{\alpha}_j = w(X_j)\delta_j(X_j)\Delta_{1j}(X_j), \qquad \hat{\alpha} = \int w(x)\delta(x) f(x)\Delta_1(x)\,dx. \tag{3.13}$$

We wish to bound

$$T = E\left\{n^{-1}\Sigma\left(\hat{\alpha}_j - \hat{\alpha}\right)\right\}^{2m} = n^{-2m}\Sigma_{j_1} \cdots \Sigma_{j_{2m}} t(j_1, \ldots, j_{2m}), \tag{3.14}$$

where

$$t(j_1, \ldots, j_{2m}) = E\left\{\left(\hat{\alpha}_1 - \hat{\alpha}\right)^{r_1} \cdots \left(\hat{\alpha}_{2m} - \hat{\alpha}\right)^{r_{2m}}\right\} = u(r_1, \ldots, r_{2m}),$$

say, for integers $r_1, \ldots, r_{2m}$ depending on $j_1, \ldots, j_{2m}$ and satisfying $r_1 \geq \cdots \geq r_{2m} \geq 0$, $\Sigma r_i = 2m$. In the range $1 \leq j \leq 2m$, approximate to $\hat{\alpha}_j - \hat{\alpha}$ by an analogous term $\hat{\beta}_j - \hat{\beta}$ in which sums over $i \neq j$ are replaced by sums over $i \neq 1, 2, \ldots, 2m$. For example, if $\hat{\alpha}_j$ and $\hat{\alpha}$ are given by (3.13), then we define

$$\hat{\beta}_j = w(X_j)\delta_0(X_j)\Delta_0(X_j), \qquad \hat{\beta} = \int w(x)\delta_0(x) f(x)\Delta_0(x)\,dx,$$

where

$$\delta_0(x) = \{(n-2m)h_1^p\}^{-1} \sum_{i \neq 1,\ldots,2m} K\{(x-X_i)/h_1\} - f(x),$$

$$\Delta_0(x) = (n-2m)^{-1} \sum_{i \neq 1,\ldots,2m} \left[h^{-p} f(X_i)^{p/2} L_1\{(x-X_i)^{1/2}/h\} - \mu_1(x)\right].$$

Note that $v(r_1,\ldots,r_{2m}) = E\{(\hat\beta_1 - \hat\beta)^{r_1} \cdots (\hat\beta_{2m} - \hat\beta)^{r_{2m}}\} = 0$ if any one of the integers r_i equals unity, on account of the fact that $E(\hat\beta_j | \{X_i,\ i \neq j\}) = \hat\beta$. Arguing thus, it may be proved that if $t(j_1,\ldots,j_m) = u(r_1,\ldots,r_{2m})$ is replaced by $v(r_1,\ldots,r_{2m})$ in formula (3.14), then T changes to a quantity T' whose order of magnitude equals that of

$$n^{-m} E\left(\hat\beta_1 - \hat\beta\right)^{2m} = O\{n^{-m}(nh^p)^{-m}(nh_1^p)^{-m}\}$$

$$= O\{(nh^p)^{-1}(nh_1^p h^{-p})^{-1/2}\}^{2m} = O\{(nh^p)^{-1} n^{-\eta}\}^{2m},$$

where $\eta > 0$. The difference between T and T' may be handled by a subsidiary argument.

Results (3.10) and (3.12) together give us the simplified expansion

$$\hat I - I = n^{-1} \sum_{j=1}^{n} w(X_j)\{\Delta_{0j}(X_j) + \mu_0(X_j)\} - \int w(x) f(x)\{\Delta_0(x) + \mu_0(x)\}\,dx + o(n^{-8/(p+8)}), \tag{3.15}$$

with probability 1.

The next step in the proof involves rearrangement of the right-hand side of (3.15). Define

$$M(u,v) = h^{-p} f(u)^{p/2} w(v) K\{(v-u) f(u)^{1/2}/h\},$$

$$M_1(u) = E\{M(u,X)\}, \qquad M_2(u) = E\{M(X,u)\}, \qquad \mu = E\{M(X_1,X_2)\}$$

$$A(u,v) = M(u,v) + M(v,u) - M_1(u) - M_1(v) - M_2(u) - M_2(v) + 2\mu,$$

$$z_i = \sum_{j=1}^{i-1} A(X_i,X_j), \qquad S_1 = \frac{1}{n}\sum_{i=1}^{n} \{M_1(X_i) - \mu\},$$

$$S_2 = \frac{1}{n}\sum_{i=1}^{n} \{M_2(X_i) - \mu\},$$

$$S_3 = \frac{1}{n^2} \sum\sum_{i \neq j} \{M(X_i,X_j) - M_1(X_i) - M_2(X_j) + \mu\} = \frac{1}{n^2}\sum_{i=2}^{n} Z_i.$$

Then

$$\frac{1}{n}\sum_{i=1}^{n} w(X_j)\{\Delta_{0j}(X_j) + \mu_0(X_j)\} = \frac{1}{n(n-1)}\sum\sum_{i\neq j} M(X_i, X_j)$$

$$= S_1 + S_2 + (1 - n^{-1})^{-1} S_3 + \mu,$$

$$\int w(x) f(x)\{\Delta_0(x) + \mu_0(x)\}\, dx = \frac{1}{n}\sum_{i=1}^{n} M_1(X_i) = S_1 + \mu.$$

Hence by (3.15),

$$\hat{I} - I = S_2 + (1 - n^{-1})^{-1} S_3 + o(n^{-8/(p+8)}). \tag{3.16}$$

By dint of the square root law, $M_2(u) = w(u)f(u) + O(h^4)$, whence it may be proved that for each $\eta > 0$,

$$S_2 = n^{-1}\sum_{i=1}^{n}\{f(X_i)w(X_i) - Ef(X)w(X)\} + O(n^{-(1/2)+\eta}h^4) \tag{3.17}$$

uniformly in $h \in \mathscr{H}$. We claim that for each $\eta > 0$ and with probability 1,

$$S_3 = O(n^{-1+\eta}h^{-p/2}) \tag{3.18}$$

uniformly in $h \in \mathscr{H}$. The theorem follows from (3.16)–(3.18).

We conclude by outlining the proof of (3.18). This formula may be derived via the Borel–Cantelli lemma, Markov's inequality and Hölder continuity of K, once it has been shown that for each integer $m \geq 1$,

$$E(S_3^{2m}) \leq C_1(m)\left\{(n^{-1}h^{-p/2})^{2m} + (n^{-3/2}h^{-p})^{2m} nh^p\right\}, \tag{3.19}$$

where C_1, C_2, C_3 are constants not depending on n or h. To establish (3.19), observe that the Z_i's are martingale differences. Hence by Rosenthal's inequality (Hall and Heyde [7], page 23),

$$n^{4m}E(S_3^{2m}) \leq C_2(m)\left[E\left\{\sum_{i=2}^{n} E(Z_i^2|X_1, \ldots, X_{i-1})\right\}^m + \sum_{i=2}^{n} E(Z_i^{2m})\right]. \tag{3.20}$$

Put $U_i = E(Z_i^2|X_1, \ldots, X_{i-1})$ and note that

$$E\left(\sum_{i=2}^{n} U_i\right)^m = \sum_{i_1=2}^{n}\cdots\sum_{i_m=2}^{n} E(U_{i_1}\cdots U_{i_m})$$

$$\leq \sum_{i_1=2}^{n}\cdots\sum_{i_m=2}^{n}\{E(U_{i_1}^m)\cdots E(U_{i_m}^m)\}^{1/m} \leq \left\{\sum_{i=2}^{n}(EZ_i^{2m})^{1/m}\right\}^m.$$

Conditional on X_i, the variables $A(X_i, X_j)$, $1 \leq j \leq 1 - 1$, are independent with zero mean. Arguing thus we may prove that

$$E(Z_i^{2m}) \leq C_2(m)\left(\left[\sum_{j=1}^{i} E\{A(X_i, X_j)^2|X_i\}\right]^m + \sum_{j=1}^{i-1} E\{A(X_i, X_j)^{2m}\}\right)$$

$$\leq C_3(m)(i^m h^{-mp} + ih^{-(2m-1)p}).$$

Formula (3.19) follows on combining the results from (3.20) down. □

Acknowledgments. The comments of a referee and Associate Editor have led to significant improvements in the presentation of this paper, and are much appreciated.

REFERENCES

[1] ABRAMSON, I. S. (1982). On bandwidth estimation in kernel estimators—A square root law. *Ann. Statist.* **10** 1217–1223.

[2] BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

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

[4] HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

[5] HALL, P. (1985). Asymptotic theory of mean integrated squared error for multivariate density estimation. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 289–309. North-Holland, Amsterdam.

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

[7] HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic, New York.

[8] HALL, P., HU, T. C. and MARRON, J. S. (1991). Improved variable window width kernel estimates of probability densities. Unpublished manuscript.

[9] HALL, P. and MARRON, J. S. (1988). Variable window width kernel estimates of probability densities. *Probab. Theory Related Fields* **80** 37–49.

[10] HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 263–269.

[11] JONES, M. C. (1991). Prospects for automatic bandwidth selection in extensions to basic kernel density estimation. In *Nonparametric Functional Estimation and Related Topics* (G. G. Roussas, ed.). Kluwer, Dordrecht.

[12] MARRON, J. S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15** 152–162.

[13] PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

[14] RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

[15] SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

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

[17] STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

[18] VICTOR, N. (1976). Nonparametric allocation rules. In *Decision Making and Medical Care: Can Information Science Help?* (F. T. Dombal and F. Grémy, eds.) 515–529. North-Holland, Amsterdam.

DEPARTMENT OF STATISTICS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA ACT 2601
AUSTRALIA