The Annals of Statistics
1985, Vol. 13, No. 3, 1011–1023

AN ASYMPTOTICALLY EFFICIENT SOLUTION TO THE BANDWIDTH PROBLEM OF KERNEL DENSITY ESTIMATION[1]

BY JAMES STEPHEN MARRON

University of North Carolina at Chapel Hill

A data-driven method of choosing the bandwidth, h, of a kernel density estimator is heuristically motivated by considering modifications of the Kullback-Leibler or pseudo-likelihood cross-validation function. It is seen that this means of choosing h is asymptotically equivalent to taking the h that minimizes some compelling error criteria such as the average squared error and the integrated squared error. Thus, for a given kernel function, the bandwidth can be chosen optimally without making precise smoothness assumptions on the underlying density.

1. Introduction. Consider the problem of estimating a univariate probability density function, f, using a sample $X_1, \cdots, X_n$ from f. An estimator which has been studied extensively (see, for example, the survey by Wertz, 1978) is the kernel estimator which is defined as follows. Given a "kernel function," K (with $\int K(x)\,dx = 1$), and a "bandwidth," $h > 0$, let

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

The "bandwidth problem" consists of specifying $h = h(n)$ in some asymptotically (as $n \to \infty$) optimal fashion. Under very precise assumptions on the amount of smoothness of f, there are many results where $h(n)$ is given deterministically to asymptotically minimize some error norm. See, for example, Rosenblatt (1956), Parzen (1962), or Watson and Leadbetter (1963). Unfortunately, this type of result is virtually useless in practice because the optimal $h(n)$ is a function of the (unknown) smoothness of f. This may be seen especially clearly from the results of Stone (1980) who deals with a continuum of smoothness classes. Thus there has been a considerable search for techniques which use the data to specify h.

A popular technique of this type is the "cross-validated" or "pseudo-maximum-likelihood" method introduced by Habbema, Hermans and van den Broek (1974). This is defined as follows. For $j = 1, \cdots, n$, form the "leave one out" kernel estimator,

$$\hat{f}_j(x, h) = (1/(n - 1)h) \sum_{i=1, i\neq j}^{n} K((x - X_i)/h). \tag{1.2}$$

Then take $\hat{h}_1$ to maximize the "estimated likelihood,"

$$\hat{L}_1(h) = \prod_{j=1}^{n} \hat{f}_j(X_j, h).$$

A recent paper by Chow, Geman and Wu (1983) contains some interesting heuristics and a consistency theorem for the estimator $\hat{f}(x, \hat{h}_1)$, Despite these

Received April 1983; revised January 1985.

[1] Research partially supported by ONR Contract N00014-81-K-0373.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62G20.

Key words and phrases. Nonparametric density estimation, kernel estimator, bandwidth, smoothing parameter, cross-validation.

encouraging results, this estimator can be very poorly behaved. Section 2 contains examples which illustrate some of the pitfalls that may be encountered by this estimator. That section also contains a series of heuristically motivated modifications of $\hat{L}_1(h)$, leading to the version that is seen to be asymptotically optimal in the theorems of Section 3. The reader who is only interested in the form of the optimal estimator should skip all of Section 2 but (2.8).

Section 4 contains some remarks. Section 5 gives the results of some simulations. The last sections contain the proof of the main optimality theorem.

2. Modification of cross-validation. To see how $f(x, h_1)$ can be poorly behaved, consider the following example. Suppose the density f has cumulative distribution function F so that

$$F(x) = e^{-1/x} \quad \text{for} \quad x > 0.$$

Note that F is infinitely differentiable. Let $X_{(1)}$ and $X_{(2)}$ denote the first two order statistics of $X_1, \cdots, X_n$. It follows from Example 1.7.5 and Theorem 2.3.2 of Leadbetter, Lindgren and Rootzén (1983) that,

$$\lim_{\delta \to 0} \liminf_n P[X_{(2)} - X_{(1)} > \delta/(\log n)^2] = 1.$$

But for compactly supported K (such as, for example, the "optimal kernels" of Epanechnikov, 1969; or Sacks and Ylvisaker, 1981), $\hat{L}_1(h) = 0$ unless $h \geq c(X_{(2)} - X_{(1)})$ for some constant c. Thus, the cross-validated $\hat{h}_1$ must converge to 0 slower than any algebraic rate. Yet it is well known that an algebraic rate such as $n^{-1/5}$ (depending on the assumptions made) is necessary for reasonable performance of the estimator.

Analogous, though not so dramatic, examples can be constructed by taking, for k large,

$$F(x) = x^k \quad \text{for} \quad x \in (0, \tfrac{1}{2}),$$

or by taking K no longer compactly supported, but with suitably "light tails." These examples indicate that, even when f is very smooth and compactly supported, ordinary cross-validated estimators can be drastically affected by data points where f is close to 0.

A reasonable way to eliminate the above difficulty is the following. Find an interval $[a, b]$ on which f is known to be bounded above 0. Redefine the estimated likelihood

$$\hat{L}_2(h) = \prod_{j=1}^{n} \hat{f}_j(X_j, h)^{1_{[a,b]}(X_j)}$$

and take $\hat{h}_2$ to maximize $\hat{L}_2(h)$. Note that cross-validation is performed only over those observations which lie in $[a, b]$.

The estimator $\hat{f}(x, \hat{h}_2)$ has been studied by Hall (1982), although he seems to have arrived at it by considerations different from the above. The notation used here (different from that of Hall) is due to Peter Bloomfield and will facilitate the rest of this discussion. Hall's results show that, while the above pathologies cause no problems, this version of cross-validation still behaves suboptimally with respect to the rate of convergence of mean square error. It is interesting to

note that the dominant term in his expansions depends only on the behavior of f at the endpoints of $[a, b]$.

David Ruppert has suggested the following heuristic explanation of this endpoint effect. Note that if $f'(a) < 0$, there will be more X_j's "just to the left" of a than "just to the right." Hence if h is taken to be relatively large, more probability mass (of the density $\hat{f}(x, h)$) will be moved into the interval $[a, b]$ which will thus increase $\hat{L}_2(h)$. Hence, there will be a tendency for cross-validation to "oversmooth" (i.e., take h too large). On the other hand, if $f'(a) > 0$, then, by the same argument, cross-validation will tend to "undersmooth" in order to keep as much probability mass inside $[a, b]$ as possible. When this effect is taken into account at both endpoints simultaneously, it is not surprising that Hall reports oversmoothing when $f'(b) - f'(a) > 0$ and undersmoothing when $f'(b) - f'(a) < 0$.

With this insight, Rupport has proposed eliminating this effect in the following way. For $j = 1, \cdots, n$, define

$$\hat{p}_j = \int_a^b \hat{f}_j(x, h)\, dx,$$

redefine the estimated likelihood

$$\hat{L}_3(h) = \prod_{j=1}^n (\hat{f}_j(X_j, h)/\hat{p}_j)^{1_{[a,b]}(X_j)},$$

and take $\hat{h}_3$ to maximize $\hat{L}_3(h)$.

This estimator will now be investigated using heuristics developed by Chow, Geman and Wu (1983). First it will be convenient to define

$$p = \int_a^b f(x)\, dx, \quad \hat{p} = \int_a^b \hat{f}(x, h)\, dx. \tag{2.1}$$

For the heuristics, assume K is nonnegative and $f(x)\log f(x)$ is integrable. By a Law of Large Numbers,

$$\begin{aligned}(1/n)\log \hat{L}_3(h) &= (1/n) \textstyle\sum_{j=1}^n 1_{[a,b]}(X_j)[\log \hat{f}(X_j, h) - \log \hat{p}] \\ &\approx \int_a^b f(x)\log \hat{f}(x, h)\, dx - p \log \hat{p}.\end{aligned} \tag{2.2}$$

But now by Jensen's Inequality,

$$\int_a^b \frac{f(x)}{p} \log\left(\frac{p\hat{f}(x, h)}{\hat{p}f(x)}\right) dx \le \log\left(\int_a^b \frac{\hat{f}(x, h)}{\hat{p}}\, dx\right) = 0, \tag{2.3}$$

with equality if and only if

$$\hat{f}(x, h)/\hat{p} = f(x)/p, \quad \text{a.e. on} \quad [a, b].$$

Hence,

$$\int_a^b f(x)\log \hat{f}(x, h)\, dx - p \log \hat{p} \le \int_a^b f(x)\log f(x)\, dx - p \log p. \tag{2.4}$$

Thus, $\hat{L}_3$ is essentially using the conditional Kullback-Leibler information (the left-hand side of (2.3)) as a measure of how well $\hat{f}(x, h)$ approximates $f(x)$. But this measure has the disturbing property that it fails to distinguish between $\hat{f}$ and f when they are unequal but proportional to each other.

Peter Bloomfield has suggested overcoming this difficulty by sharpening the inequality (2.3) using the following device. Note that for $x, y > 0$,

$$y \log(x/y) \leq x - y, \tag{2.5}$$

with equality only when $x = y$. Hence

$$p \log \hat{p} - p \log p \leq \hat{p} - p.$$

It now follows from (2.4) that

$$\int_a^b f(x)\log \hat{f}(x, h)\, dx - \hat{p} \leq \int_a^b f(x)\log f(x)\, dx - p, \tag{2.6}$$

with equality if and only if $f(x) = \hat{f}(x, h)$ for almost all $x \in [a, b]$. Now reversing the heuristic argument (2.2) it is apparent that the estimated likelihood should be redefined as

$$\hat{L}_4(h) = \prod_{j=1}^n [\hat{f}_j(X_j, h)e^{-\hat{p}_j/p}]^{1_{[a,b]}(X_j)}$$

and $\hat{h}_4$ taken to maximize $\hat{L}_4(h)$.

Using a Law of Large Numbers in a manner similar to the above, it can be shown that $\hat{L}_4(h)$ is very similar to the computationally simpler

$$\hat{L}_5(h) = \prod_{j=1}^n \hat{f}_j(X_j, h)^{1_{[a,b]}(X_j)}e^{-\rho(X_j)},$$

where

$$\rho(x) = \int_a^b \frac{1}{h} K\left(\frac{y - x}{h}\right) dy. \tag{2.7}$$

One last refinement will now be made. Many authors, starting with Parzen (1962) and Watson and Leadbetter (1963), have noticed that the asymptotic properties of K can be greatly improved by allowing $K(x)$ to be negative for some x. The results of this paper apply to either this type of kernel or the nonnegative kernels which guarantee that $\hat{f}$ is "range-preserving" (i.e., ≥ 0). However the proofs in this paper involve taking logarithms, so it is necessary to do some truncation. Define, for $x \in \mathbb{R}$,

$$\hat{f}^+(x, h) = \max(\hat{f}(x, h), 0),$$

and for $j = 1, \cdots, n$,

$$\hat{f}_j^+(x, h) = \max(\hat{f}_j(x, h), 0).$$

Now redefine the estimated likelihood

$$\hat{L}(h) = \prod_{j=1}^n \hat{f}_j^+(X_j, h)^{1_{[a,b]}(X_j)}e^{-\rho(X_j)} \tag{2.8}$$

and take $\hat{h}$ to maximize $\hat{L}(h)$. It will be seen in Section 3 that the estimator $\hat{f}(x, \hat{h})$ has excellent asymptotic properties.

An interesting side effect of the above truncation is that if $K(0) \geq 0$, $\hat{f}(x, \hat{h})$ is range-preserving, and in fact is positive, at each data point in $[a, b]$.

3. Asymptotic optimality theorems. Three compelling means of assessing the performance of $\hat{f}(x,h)$ in estimating for $f(x)$ are the Average Square Error,

$$\text{ASE}(h) = n^{-1} \sum_{j=1}^{n} [\hat{f}(X_j, h) - f(X_j)]^2 w(X_j), \tag{3.1}$$

the Integral Square Error,

$$\text{ISE}(h) = \int [\hat{f}(x, h) - f(x)]^2 w(x)\, dx, \tag{3.2}$$

and the Mean Integral Square Error,

$$\text{MISE}(h) = E(\text{ISE}(h)). \tag{3.3}$$

The assumptions of the theorems which demonstrate good ASE, ISE and MISE performance of $\hat{f}(x, \hat{h})$ will now be given. For some small $\sigma > 0$, define sequences $\{\underline{h}_n\}$ and $\{\bar{h}_n\}$ by

$$\underline{h} = n^{-1+\sigma}, \quad \bar{h} = n^{-\sigma}, \tag{3.4}$$

where, here and below, the dependence on n is suppressed. Note that assuming $h \in [\underline{h}, \bar{h}]$ is slightly stronger than the bandwidth assumptions made for the best known uniform consistency results. Also assume that the density f satisfies

$$f \text{ is bounded above 0 on } [a, b] \tag{3.5}$$

$$\text{there are constants } M, \quad \gamma > 0 \quad \text{so that for all} \quad x, y$$
$$|f(x) - f(y)| \leq M |x - y|^{\gamma}. \tag{3.6}$$

Another assumption is that the kernel function K satisfies

$$\int K(x)\, dx = 1, \tag{3.7}$$

$$\text{there are constants } M, \quad \xi > 0 \quad \text{so that for all} \quad x, y$$
$$|K(x) - K(y)| \leq M |x - y|^{\xi}. \tag{3.8}$$

$$\text{Either } X \text{ has some moment, or } K \text{ is compactly supported.} \tag{3.9}$$

The main theorem of this paper is

THEOREM 1. *Under the assumptions* (3.4)–(3.9), *if* $\hat{h} = \hat{h}(n)$ *denotes any sequence of maxima of* $\hat{L}(h)$, *subject to the restriction* $\hat{h} \in [\underline{h}, \bar{h}]$, *then*

$$\text{ERR}(\hat{h}) / \inf_{h \in [\underline{h}, \bar{h}]} \text{ERR}(h) \to 1 \quad a.s.$$

where ERR(h) *denotes any one of*:

(a) ASE(h), *with* $w(x) = f(x)^{-2} 1_{[a,b]}(x)$
(b) ISE(h), *with* $w(x) = f(x)^{-1} 1_{[a,b]}(x)$
(c) MISE(h), *with* $w(x) = f(x)^{-1} 1_{[a,b]}(x)$.

A drawback to this theorem is that it applies only to h in the interval $[\underline{h}, \bar{h}]$. This is not a big problem from the theoretical point of view because it is well known that any "optimal bandwidth" is easily inside the interval. Also, Monte Carlo experience with $\hat{L}(h)$ (see Section 5) indicates that this assumption is not a problem in practice. Further reassurance along these lines is provided by:

THEOREM 2. *Assuming* (3.5)–(3.8) *and in addition assuming that* K *is compactly supported, if* $\hat{h} = \hat{h}(n)$ *denotes any sequence of maxima of* $\hat{L}(h)$, *then*

i) $\hat{h} \to 0$ *a.s.*
ii) $\lim_{c \to 0} \limsup_n P[\hat{h} < cn^{-(2\gamma+1)^{-1}}] = 0$.

It should be noted that while Theorem 2 does show $\hat{h} > \underline{h}$ (for σ sufficiently small) it does not show $\hat{h} < \bar{h}$ or even establish the consistency of $\hat{f}(x, \hat{h})$. It is intended only to give some backing to the above remarks. To save space, the proof of Theorem 2 will not be given here. The interested reader can find it in the technical report of Marron (1983a). The proof of (i) is based on techniques of Chow, Geman and Wu (1983) and it appears that these techniques may be further extended to establish the consistency of $\hat{f}(x, \hat{h})$. The proof of (ii) is based on an order statistics result of Cheng (1983).

4. Remarks

REMARK 4.1. The reader may be surprised that no vanishing moment assumption is made on K (see, for example, Parzen, 1962). Theorem 1 says $\hat{f}(x, \hat{h})$ will have the best asymptotic properties for the given K, but how good that is is irrelevant to the theorem.

REMARK 4.2. An inspection of the proofs shows that extending the theorems to the case of f multivariate requires little more than notational changes. Also, since the hardest part of the proof is a consequence of the results of Marron and Härdle (1984), extension of these results to the case of histogram and orthogonal series estimators seems straightforward.

REMARK 4.3. At first glance, one might be disturbed by the fact that the error criteria that are minimized here are limited to the interval $[a, b]$. In somewhat similar settings and in the case of estimating a regression function, Gasser and Müller (1979) and Rice and Rosenblatt (1983) have observed that such criteria are strongly affected by the behavior of the unknown function at the endpoints and hence the bandwidths which minimize them can provide relatively poor estimates in the interior of $[a, b]$. However, with very little effort, one may see that such an "endpoint effect" does not occur in the present setting. This is because the density f extends (and is smooth) outside the interval $[a, b]$ and observations outside $[a, b]$ are employed in the estimator of this paper. Hence, the error criteria of this paper seem very reasonable.

REMARK 4.4. For the Least Squares cross-validation function, which is quite different from $\hat{L}(h)$, results similar to Theorem 1 have been established by Stone (1984) and Burman (1984). A major difference between Theorem 1 and those results is that there the weight function $w(x)$ in ISE is identically one. The weight functions used in this paper are quite natural because all three error criteria are roughly (see Lemma B in Section 6) proportional to the expected relative square error:

$$E[((\hat{f}(x) - f(x))/f(x))^2 \mid X \in [a, b]].$$

It is seen in Marron (1983b) that this error norm is the most useful for the application of density estimation to the classification problem. Dennis Boos has remarked that this is also more useful for application to minimum Hellinger distance estimation.

5. Simulations. As with any asymptotic theory, it still remains to check that the properties described by the asymptotics "take effect" for sample sizes which are not prohibitively large. The results of some simulations, in the spirit of Hall (1982), are reported here. Following Hall, K and f were taken to be standard normal, with $a = -1$, $b = 1$, and $n = 2{,}000$. The maximizer of $\hat{L}$ was 0.240, while the minimizer of ASE was 0.247. The expansions in the next few sections indicate that these should be close to the minimizer of the familiar asymptotic representation of MISE (given in (8.1)) which is 0.235 in this case. The heuristic argument against the cross-validation curve $L_2(h)$ is supported by the fact that $\hat{h}_2$ was 0.078 (note drastic undersmoothing, as predicted).

6. Proof of Theorem 1. This proof uses ideas developed by Hall (1982). Note that choosing h to maximize $\hat{L}(h)$ is the same as finding a maximizer of

$$n^{-1}\log \hat{L}(h) + R, \tag{6.1}$$

where

$$R = p - n^{-1} \textstyle\sum_{j=1}^{n} \log f(X_j) 1_{[a,b]}(X_j),$$

and p was defined in (2.1). For analyzing (6.1), it will be useful to define, for $j = 1, \cdots, n$,

$$\Delta_j = \frac{\hat{f}_j(X_j, h) - f(X_j)}{f(X_j)}, \quad \Delta_j^+ = \frac{\hat{f}_j^+(X_j, h) - f(X_j)}{f(X_j)} \tag{6.2}$$

and for $n = 1, 2, \cdots$, the event

$$U_n = \{\Delta_j^+ 1_{[a,b]}(X_j) = \Delta_j 1_{[a,b]}(X_j) \text{ for each } h \in [\underline{h}, \bar{h}] \text{ and } j = 1, \cdots, n\}.$$

The cornerstone of this proof is the fact that, for any $h \in [\underline{h}, \bar{h}]$, and on the event U_n, (6.1) admits the expansion

$$\begin{aligned} & n^{-1}\log \hat{L}(h) + R \\ & = n^{-1} \textstyle\sum_{j=1}^{n} [1_{[a,b]}(X_j)\log(1 + \Delta_j) - \rho(X_j) + p] \\ & = n^{-1} \textstyle\sum_{j=1}^{n} [1_{[a,b]}(X_j)\Delta_j - \rho(X_j) + p] - \tfrac{1}{2}\widetilde{\mathrm{ASE}}(h) + n^{-1} \textstyle\sum_{j=1}^{n} r_j 1_{[a,b]}(X_j), \end{aligned} \tag{6.3}$$

where $\widetilde{\text{ASE}}$ is the leave-one-out version of (3.1) given by

$$\widetilde{\text{ASE}} = n^{-1} \sum_{j=1}^{n} [\hat{f}_j(X_j, h) - f(X_j)]^2 f(X_j)^{-2} 1_{[a,b]}(X_j), \tag{6.4}$$

and where r_j denotes the remainder term of the Taylor expansion of the logarithm.

The fact that it makes sense to consider only what happens on the event U_n is established by

LEMMA A. *Under the assumption of Theorem* 1, *letting* U_n^c *denote the complement of* U_n,

$$P[U_n^c \text{ i.o.}] = 0.$$

The proof of Lemma A is in Section 7.

The fact that (3.1), (3.2), (3.3) and (6.4) are asymptotically the same error criterion is established by

LEMMA B. *Under the assumptions of Theorem* 1, *letting* ERR(h) *denote any of* $\widetilde{ASE}$, *ASE or ISE*,

$$\sup_h |\text{ERR}(h) - \text{MISE}(h)|/\text{MISE}(h) \to 0, \quad a.s.$$

where $\sup_h$ *denotes supremum over* $h \in [\underline{h}, \bar{h}]$.

The proof of Lemma B is in Section 8.

The fact that the first term on the right side of (6.3) is negligible is established by

LEMMA C. *Under the assumptions of Theorem* 1,

$$\sup_h |n^{-1} \sum_{j=1}^{n} [1_{[a,b]}(X_j)\Delta_j - \rho(X_j) + p]|/\text{MISE} \to 0, \quad a.s.$$

The proof of Lemma C is in Section 9.

The fact that the final term of (6.3) is negligible is established by

LEMMA D. *Under the assumptions of Theorem* 1,

$$\sup_h |n^{-1} \sum_{j=1}^{n} r_j 1_{[a,b]}(X_j)|/\text{MISE} \to 0, \quad a.s.$$

The proof of Lemma D is in Section 10.

To arrive at the conclusion of Theorem 1, Stone (1984) has noticed that it is enough to check that

$$\inf_{h,h'} \frac{|\text{ERR}(h) - \text{ERR}(h') + n^{-1}(\log \hat{L}(h) - \log \hat{L}(h'))|}{\text{ERR}(h) + \text{ERR}(h')}$$

almost surely, where $\inf_{h,h'}$ denotes infimum over $h, h' \in [\underline{h}, \bar{h}]$. But in view of Lemma B, this can be done by showing

$$\inf_{h,h'} \frac{|\text{ERR}(h) - \text{ERR}(h') + n^{-1}(\log \hat{L}(h) - \log \hat{L}(h'))|}{\text{MISE}(h) + \text{MISE}(h')} \to 0,$$

almost surely. However this follows easily from the expansion (6.3) and Lemmas A, B, C, and D. This completes the proof of Theorem 1.

7. Proof of Lemma A. By Lemma 1 of Härdle and Marron (1984), letting $\sup_x$ and $\sup_h$ denote supremum over $x \in [a, b]$ and $h \in [\underline{h}, \bar{h}]$, respectively,

$$\sup_x \sup_h | \hat{f}^+(x, h) - f(x) | \leq \sup_x \sup_h | \hat{f}(x, h) - f(x) | \to 0,$$

almost surely. In a similar spirit, letting $\sup_j$ denote supremum over $j = 1, \cdots, n$, the computations leading to (5.5) of Härdle and Marron (1984) yield

$$\sup_j \sup_x \sup_h | \hat{f}_j^+(x, h) - f(x) | \leq \sup_j \sup_x \sup_h | \hat{f}_j(x, h) - f(x) | \to 0, \tag{7.1}$$

almost surely. It follows from this and (3.5) that

$$\sup_j \sup_h | \Delta_j^+ | \leq \sup_j \sup_h | \Delta_j | \to 0,$$

almost surely.

Lemma A is an easy consequence of this.

8. Proof of Lemma B. In the case of ASE and ISE, Theorems 1 and 2, respectively, of Marron and Härdle (1984) show that Lemma B is true if the supremum is taken over any finite set of h whose cardinality increases algebraically fast. It is straightforward to use the Lipschitz continuity assumptions (3.6) and (3.8) to extend the supremum to $[\underline{h}, \bar{h}]$.

To check the case of $\widetilde{\text{ASE}}$, first observe that

$$\hat{f}_j(x, h) - \hat{f}(x, h) = (n - 1)^{-1} \hat{f}(x, h) - (n - 1)^{-1} h^{-1} K(0).$$

Now write

$$\widetilde{\text{ASE}} = \text{ASE} + A + B,$$

where

$$A = 2n^{-1} \sum_{j=1}^n [(n - 1)^{-1} (\hat{f}(x, h) - h^{-1} K(0))][\hat{f}(X_j, h) - f(X_j)] f(X_j)^{-2} 1_{[a,b]}(X_j),$$

$$B = n^{-1} \sum_{j=1}^n (n - 1)^{-2} (\hat{f}(x, h) - h^{-1} K(0))^2 f(X_j)^{-2} 1_{[a,b]}(X_j).$$

Lemma B in the case of $\widetilde{\text{ASE}}$ now follows easily from the case of ASE, from Lemma A, and from the well-known (see, for example, Rosenblatt, 1971) variance-bias^2 expansion,

$$\text{MISE}(h) = n^{-1} h^{-1} \left(\int f(x) w(x)\, dx \right) \left(\int K(u)^2\, du \right) + o(n^{-1} h^{-1}) + b(h), \tag{8.1}$$

where the bias^2 part has been denoted

$$b(h) = \int \left[\int K(u) f(x - hu)\, du - f(x) \right]^2 w(x)\, dx. \tag{8.2}$$

This completes the proof of Lemma B.

9. Proof of Lemma C. Note that, by (1.2) and (6.2)

$$n^{-1}\sum_{j=1}^{n}[1_{[a,b]}(X_j)\Delta_j - \rho(X_j) + p]$$

$$= n^{-1}(n-1)^{-1}\sum_{i\neq j}[((1/h)K((X_j - X_i)/h)f(X_j)^{-1} - 1) \cdot 1_{[a,b]}(X_j) - \rho(X_j) + p]. \tag{9.1}$$

Thus, to finish the proof of Lemma C, it is enough to show that

$$\sup_h n^{-1}(n-1)^{-1}\left|\sum_{i\neq j} V_{i,j}\right| \mathrm{MISE}^{-1} \to 0 \quad \text{a.s.},$$

where

$$V_{i,j} = (1/h)K((X_j - X_i)/h)f(X_j)^{-1}1_{[a,b]}(X_j) - \rho(X_i) - 1_{[a,b]}(X_j) + p. \tag{9.2}$$

Observe that $\rho(X_j)$ in (9.1) has been replaced by $\rho(X_i)$ in (9.2), to allow

$$E[V_{i,j} \mid X_i] = 0.$$

To make the other conditional mean also vanish (in effect), write

$$V_{i,j} = V'_{i,j} + W_j,$$

where

$$W_j = E[V_{i,j} \mid X_j].$$

Note that

$$E[V'_{i,j} \mid X_i] = E[V'_{i,j} \mid X_j] = 0, \quad E[W_j] = 0. \tag{9.3}$$

The proof of Lemma C will be complete when it is seen that

$$\sup_h n^{-2}\left|\sum_{i\neq j} V'_{i,j}\right| \mathrm{MISE}^{-1} \to 0 \quad \text{a.s.} \tag{9.4}$$

and that

$$\sup_h n^{-1}\left|\sum_{j=1}^{n} W_j\right| \mathrm{MISE}^{-1} \to 0 \quad \text{a.s.} \tag{9.5}$$

As above, using the Lipschitz continuity assumptions (3.6) and (3.8), it is enough to verify (9.4) and (9.5), when $\sup_h$ denotes supremum over any sequence of finite sets, H_n, whose cardinality increases algebraically fast (in n). For (9.4), note that given $\varepsilon > 0$ and $k = 1, 2, \cdots$

$$\sum_{n=1}^{\infty} P[\sup_{h\in H_n} n^{-2}\left|\sum_{i\neq j} V'_{i,j}\right| \mathrm{MISE}^{-1} > \varepsilon]$$

$$\leq \sum_{n=1}^{\infty} \#(H_n)\sup_{h\in H_n} E[n^{-2}\sum_{i\neq j} V'_{i,j}\mathrm{MISE}^{-1}\varepsilon^{-1}]^{2k},$$

so (9.4) will be established when it is seen that there is a constant $\tau > 0$, so that for $k = 1, 2, \cdots$ there are constants $\mathscr{C}_k$ so that

$$E[n^{-2}\sum_{i\neq j} V'_{i,j}\mathrm{MISE}^{-1}]^{2k} \leq \mathscr{C}_k n^{-\tau k}.$$

To verify this, by the cumulant expansion of the $2k$th centered (since $V'_{i,j}$ has mean 0) moment (see, for example (3.33) of Kendall and Stuart, 1963), it is enough to show that, for $k = 2, 3, \cdots$, there is a constant $\mathscr{C}_k$ so that

$$\left|\mathrm{cum}_k(n^{-2}\sum_{i\neq j} V'_{i,j}\mathrm{MISE}^{-1})\right| \leq \mathscr{C}_k n^{-\tau k},$$

where $\text{cum}_k(\cdot)$ denotes the kth order cumulant with all k arguments the same. But to check this, using the linearity property of cumulants (see, for example, Theorem 2.3.1 of Brillinger, 1979), it is enough to show that, for $k = 2, 3, \cdots$, there is a constant $\mathscr{C}_k$ such that

$$|n^{-2k}\text{MISE}^{-k} \sum \text{cum}_k(V'_{i_1,j_1}, \cdots, V'_{i_k,j_k})| \leq \mathscr{C}_k n^{-\tau k}, \tag{9.6}$$

where $\sum$ denotes summation over $i_1, j_1, \cdots, i_k, j_k = 1, \cdots, n$ subject to $i_1 \neq j_1, \cdots, i_k \neq j_k$.

To check (9.6), note that by (9.3), most of the terms in the summation will be 0. In particular, cum_k can be nonzero only when each of $i_1, j_1, \cdots, i_k, j_k$ is the same as one of the others. For each such term, let m denote the number of unique elements of $\{1, \cdots, n\}$ appearing among $i_1, j_1, \cdots, i_k j_k$. It follows from integration by substitution and the assumptions of Theorem 1 that there is a constant $\mathscr{C}_k$ so that

$$|\text{cum}_k(V'_{i_1,j_1}, \cdots, V'_{i_k,j_k})| \leq \mathscr{C}_k h^{-k+m/2}.$$

Next observe that there is a constant $\mathscr{C}_k$ so that for $m = 2, \cdots, k$, the number of nonzero terms in the summation of (9.6) with exactly m distinct indices is bounded by $\mathscr{C}_k n^m$. It follows from the above, together with the expansion (8.1), that there is a constant $\mathscr{C}_k$ so that the left side of (9.6) is bounded by

$$\mathscr{C}_k n^{-2k}(n^{-1}h^{-1})^{-k} \sum_{m=2}^{k} n^m h^{-k+m/2} = \mathscr{C}_k \sum_{m=2}^{k} n^{-k+m} h^{m/2}.$$

The inequality (9.6) follows easily from this and (3.4). This verifies (9.4).

To verify (9.5), write

$$\begin{aligned} W_j &= \int \frac{1}{h} K\left(\frac{X_j - x}{h}\right) f(x)\, dx 1_{[a,b]}(X_j) f(X_j)^{-1} \\ &\quad - \int_a^b \int \frac{1}{h} K\left(\frac{y - x}{h}\right) f(x)\, dx\, dy - 1_{[a,b]}(X_j) + \int_a^b f(y)\, dy \\ &= \int K(u)[f(X_j - hu) - f(X_j)]\, du f(X_j)^{-1} 1_{[a,b]}(X_j) \\ &\quad - \int_a^b \int K(u)[f(y - hu) - f(y)]\, du\, dy. \end{aligned} \tag{9.7}$$

It follows from Taylor's theorem and the assumptions of Theorem 1 that there is a constant $\mathscr{C}$ so that

$$|W_j| \leq \mathscr{C} h^{\gamma} = \mathscr{C} n^{-\delta\gamma}.$$

Since the variance of W_j is bounded by the second moment of the first term on the right of (9.7), note that

$$\sigma^2 = \text{var } W_j \leq b(h),$$

where the notation (8.2) has been used. Since the W_j are i.i.d., mean 0 random variables, an application of Bernstein's Inequality (see (2.13) of Hoeffding, 1963

with (in Hoeffding's notation)

$$\lambda = bt/\sigma^2, \quad \tau = nt/b, \quad b = \mathscr{C}n^{-\delta\gamma}, \quad t = \text{MISE} \cdot \varepsilon,$$

yields

$$P[|\, n^{-1} \textstyle\sum_{j=1}^{n} W_j| > \text{MISE} \cdot \varepsilon]$$
$$\leq \exp(-\tau\lambda/2(1 + \lambda/3)) = \exp(-nt^2/2(\sigma^2 + bt/3))$$
$$\leq \exp(-n \text{ MISE}/2) \leq \exp(-n^{\delta}/2),$$

for n sufficiently large.

The inequality (9.5) follows easily when the supremum is taken over finite sets whose cardinality increases algebraically fast. This completes the proof of Lemma C.

10. Proof of Lemma D. Lemma D follows immediately from Lemma B and (7.1).

Acknowledgement. The author is grateful to David Ruppert, Raymond Carroll and especially to Peter Bloomfield for many interesting and stimulating conversations during the course of the research presented in this paper. Thanks are also due C. J. Stone for suggesting a key step in the proofs.

REFERENCES

BLOOMFIELD, P. AND MARRON, J. S. (1985). Cross-validation in density estimation from a likelihood point of view. Manuscript in preparation.

BRILLINGER, D. R. (1979). *Time Series Data Analysis and Theory*. Holt, Rinehart and Winston, New York.

BURMAN, P. (1984). A data dependent approach to density estimation. Unpublished manuscript.

CHENG, S. H. (1983). On a problem concerning spacings. *Z. Wahrsch. verw. Gebiete* **66** 245–258.

CHOW, Y. S., GEMAN, S. and WU, L. D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25–38.

EPANECHNIKOV, V. (1969). Nonparametric estimates of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68.

HABBEMA, J. D. F., HERMANS, J. and VAN DEN BROEK, K. (1974). A stepwise discrimination analysis program using density estimation. In *Compstat 1974: Proceedings in computational statistics.* (G. Bruckman, ed.) 101–110. Physica Verlag, Vienna.

HALL, P. (1982). Cross validation in density estimation. *Biometrika* **69** 383–390.

HÄRDLE, W. and MARRON, J. S. (1984). Optimal bandwidth selection in nonparametric regression function estimation (revised). North Carolina Inst. Statist. Mimeo Series #1546.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KENDALL, M. G. and STUART, A. (1963). *The Advanced Theory of Statistics, Vol. 1: Distribution Theory*, Butler and Tanner.

LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

MARRON, J. S. (1983a). Uniform convergence properties of a cross-validation density estimator. North Carolina Inst. Statist. Mimeo Series # 1519.

MARRON, J. S. (1983b). Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.* **11** 1142–1155.

MARRON, J. S. and HÄRDLE, W. (1984). Random approximations to an error criterion of nonparametric statistics (revised). North Carolina Inst. Statist. Mimeo Series #1538.

PARZEN, E. (1962). On the estimation of a probability density and mode. *Ann. Math. Statist.* **33** 1065–1076.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

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

SACKS, J. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9** 334–346.

STONE, C. J. (1980). Optimal convergence rates for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

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

WATSON, G. S. and LEADBETTER, M. R. (1963). On the estimation of a probability density, I. *Ann. Math. Statist.* **34** 480–491.

WERTZ, W. (1978). Statistical density estimation: a survey. *Angewandte Statistique und Okonometrie* **13** van den Broek and Ruprecht.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514