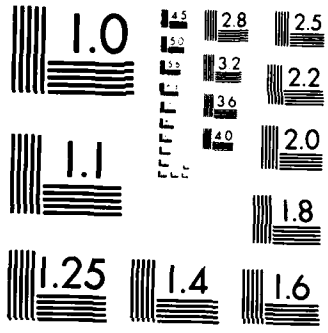
AD-A149 600 AN ASYMPTOTICALLY EFFICIENT SOLUTION TO THE BANDWIDTH PROBLEM OF KERNEL D.. (U) NORTH CAROLINA UNIV AT CHAPEL HILL DEPT OF STATISTICS J S MARRON APR 84 1/1

UNCLASSIFIED MIMEO-SER-1545 N00014-81-K-0373 F/G 12/1 NL

END

FILMED

DTIC



MIMEO SERIES #1545

AN ASYMPTOTICALLY EFFICIENT SOLUTION TO THE BANDWIDTH PROBLEM OF KERNEL DENSITY ESTIMATION (revised)

James Stephen Marron
University of North Carolina at Chapel Hill

Abstract

A data-driven method of choosing the bandwidth, h, of a kernel density estimator is proposed. It is seen that this means of selecting h is asymptotically equivalent to taken the h that minimizes a certain weighted version of the mean integrated square error. Thus, for a given kernel function, the bandwidth can be chosen optimally without making precise smoothness assumptions on the underlying density. The proposed technique is a modification of cross-validation.

AMS 1980 Subject Classification: Primary 62G05, Secondary 62G20

Keywords: Nonparametric density estimation, kernel estimator, bandwidth, smoothing parameter, cross-validation.

Research partially supported by ONR contract N00014-81-K-0373.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for Public Release: Distribution Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: Department of Statistics; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code): University of North Carolina, Chapel Hill, North Carolina 27514	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Office of Naval Research; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-81-K-0373
8c. ADDRESS (City, State and ZIP Code): Statistics & Probability Program, Arlington, VA 22217	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
NR	042	269	SRO 105

11. TITLE (Include Security Classification): An Asymptotically Efficient Solution . .(cont.)

12. PERSONAL AUTHOR(S): J.S. Marron

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
TECHNICAL	FROM _____ TO _____	April 1984	29

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Nonparametric density estimation, kernel estimator, bandwidth, smoothing parameter, cross-validation.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

A data-driven method of choosing the bandwidth, h, of a kernel density estimator is proposed. It is seen that this means of selecting h is asymptotically equivalent to taking the h that minimizes a certain weighted version of the mean integrated square error. Thus, for a given kernel function, the bandwidth can be chosen optimally without making precise smoothness assumptions on the underlying density. The proposed technique is a modification of cross-validation.

FULL TITLE (#11): An Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation (revised)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☒ DTIC USERS ☐	

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
C.R. Baker	(919) 962-2189	

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

1. Introduction

Consider the problem of estimating a univariate probability density function, f, using a sample $X_1, \ldots, X_n$ from f. Let $\hat{f} = \hat{f}(x, X_1, \ldots, X_n)$ denote an estimator. A common error norm is Mean Integrated Square Error, which is defined as follows. Let w(x) be some nonnegative "weight function." Define

$$\text{(1.1)} \qquad \text{MISE} = E\int [\hat{f}(x)-f(x)]^2 w(x)\, dx.$$

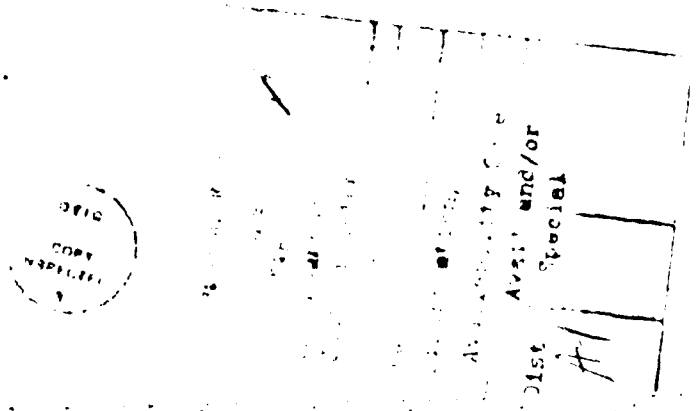
An estimator which has been studied extensively (see, for example, the survey by Wertz (1978)) is the kernel estimator which is defined as follows. Given a "kernel function," K (with $\int K(x)\, dx = 1$), and a "bandwidth," h>0, let

$$\text{(1.2)} \qquad \hat{f}(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) .$$

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 MISE or some other 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,..., n form the "leave one out" kernel estimator,

$$\text{(1.3)} \qquad \hat{f}_j(x,h) = \frac{1}{(n-1)h} \sum_{\substack{i=1 \\ i\neq j}}^{n} K\left(\frac{x-X_i}{h}\right).$$



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 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 (2.11).

Section 5 contains some remarks. The last section contains the proof of the optimality theorem.

2. Modification of cross-validation.

To see how $\hat{f}(x,\hat{h}_1)$ can be poorly behaved, consider the following example. Suppose the density f has cumulative distribution function F so that for some $\varepsilon>0$,

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

Such an F could easily be constructed to be infinitely differentiable. Let $X_{(1)}$ and $X_{(2)}$ denote the first two order statistics of $X_1,\ldots, X_n$. It follows from example 1.7.5 and Theorem 2.3.2 of Leadbetter, Lindgren and Rootzén (1983) that,

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

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.

By the familiar variance and bias^2 decomposition (see Rosenblatt (1971)) the mean square error may be written:

$$E[\hat{f}(x,h)-f(x)]^2 = O(\frac{1}{nh}) + O(h^{2s}) ,$$

where s represents the amount of smoothness that is assumed on f. Hence, it is apparent that the estimator $\hat{f}(x,\hat{h}_1)$ can behave very poorly in the mean square sense.

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

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

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. The assumption of the existence of such an interval seems easy for the practitioner to accept. Next 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 subopti-

mally 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, Ruppert has proposed eliminating this effect in the following way. First for $j=1,\ldots,n$ define

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

Next redefine the estimated likelihood

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

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

$$p = \int_a^b f(x)\,dx \,, \tag{2.2}$$

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

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

$$\frac{1}{n}\log\hat{L}_3(h) \approx \frac{1}{n}\sum_{j=1}^{n} 1_{[a,b]}(X_j)[\log\hat{f}(X_j,h)-\log\hat{p}]$$

(2.3)

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

But now by Jensen's Inequality,

$$(2.4)\qquad \int_a^b \frac{f(x)}{p}\log\left(\frac{\hat{p}\hat{f}(x,h)}{\hat{p}f(x)}\right)dx \le \log\left(\int_a^b \frac{\hat{f}(x,h)}{\hat{p}}\,dx\right) = 0 ,$$

with equality if and only if,

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

Hence

$$(2.5)\qquad \int_a^b f(x)\log\hat{f}(x,h)dx - p\log\hat{p} \le \int_a^b f(x)\log f(x)dx - p\log p.$$

Thus, $\hat{L}_3$ is essentially using the conditional Kullback-Leibler information (the left hand side of (2.4)) 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 (2.5) using the following device. Note that for $x,y>0$,

$$(2.6)\qquad y\log(x/y) \le x - y ,$$

with equality only when $x = y$. Hence

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

It now follows from (2.5) that

$$(2.7)\qquad \int_a^b f(x)\log\hat{f}(x,h)dx - \hat{p} \le \int_a^b f(x)\log f(x)dx - p ,$$

with equality if and only if $f(x) = \hat{f}(x,h)$ for almost all $x \in [a,b]$. Now rever-

sing the heuristic argument (2.3) 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 $\hat{L}_4(h)$.

Peter Bloomfield has pointed out that $\hat{L}_4(h)$ may be somewhat simplified, from the computational viewpoint, in the following way. Note that

$$\hat{p}_j = (n-1)^{-1} \sum_{i \neq j} \rho(X_i) ,$$

where

$$\rho(x) = \int_a^b \frac{1}{h} K(\frac{v-x}{h})dv \tag{2.8}$$

Hence, by a Strong Law of Large Numbers,

$$\prod_{j=1}^{n} \exp(-1_{[a,b]}(X_j)\hat{p}_j/p) = \exp(-\sum_{j=1}^{n} 1_{[a,b]}(X_j)(n-1)^{-1} \sum_{i \neq j} \rho(X_i)/p) =$$

$$= \exp(-\sum_{i=1}^{n} \rho(X_i)(n-1)^{-1} \sum_{j \neq i} 1_{[a,b]}(X_j)/p) \approx$$

$$\approx \prod_{i=1}^{n} e^{-\rho(X_i)} .$$

Thus redefine the estimated likelihood

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

Note this also avoids difficulties about the fact that p in $\hat{L}_4(h)$ is unknown.

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 non-negative kernels which guarantee that $\hat{f}$ is "range-preserving." However the proofs in this paper involve taking logarithms, so it is necessary to do some truncation. Define, for $x \in \mathbf{R}$,

$$\hat{f}^{+}(x,h) = \max(\hat{f}(x,h),0) , \tag{2.9}$$

and for j=1,..., n,

$$\hat{f}_j^{+}(x,h) = \max(\hat{f}_j(x,h),0) . \tag{2.10}$$

Now redefine the estimated likelihood

$$\hat{L}(h) = \prod_{j=1}^{n} \hat{f}_j^{+}(X_j,h)^{1_{[a,b]}(X_j)} e^{-\rho(X_j)} \tag{2.11}$$

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 the following. If for some h there is an $X_j \in [a,b]$ for which $\hat{f}_j(X_j,h)<0$, then $\hat{L}(h) = 0$. Hence, such an h can not be chosen to be $\hat{h}$. Thus, since

$$\hat{f}(X_j,h) = \frac{n-1}{n} \hat{f}_j(X_j,h) + \frac{1}{nh} K(0) ,$$

if $K(0)>0$, then for $j \in A$, $\hat{f}(X_j,\hat{h})>0$. Hence, the estimator $\hat{f}(x,\hat{h})$ has the property

that it is range-preserving (i.e.: >0) at each data point in [a,b]. Of course, the experimenter who requires that f be range-preserving outside the interval [a,b] can guarantee this by taking K nonnegative.

3. Asymptotic Optimality Theorems

It is well known (see, for example, Rosenblatt (1971)) that MISE admits the variance-bias^2 expansion

$$\text{(3.1)} \qquad \text{MISE}(h) = n^{-1}h^{-1}(\int f(y)w(y)dy)(\int K(u)^2du) + o(n^{-1}h^{-1}) + s_f(h),$$

where the bias^2 part is:

$$\text{(3.2)} \qquad s_f(h) = \int[\int K(u)f(y-hu)du - f(y)]^2\, w(y)dy.$$

Since the papers of Rosenblatt (1956) and Parzen (1962), expansions similar to the above have been handled as follows.

Assume K satisfies:

$$\text{(3.3)} \qquad \begin{aligned} &\int K(x)dx = 1, \\ &\int x^j K(x)dx = 0, \qquad j=1,\ldots,k-1, \\ &\int x^k K(x)dx > 0. \end{aligned}$$

Also assume f has a bounded k-th derivative. By Taylor's Theorem,

$$\text{(3.4)} \qquad s_f(h) = h^{2k}\int[f^{(k)}(y)]^2 w(y)dy[\int u^k K(u)du]/k! + o(h^{2k}) .$$

Now to find the "optimal bandwidth", ignore the terms $o(n^{-1}h^{-1})$ and $o(h^{2k})$ (which are of lower order, uniformly over h) in (3.1) and (3.4), and choose h to minimize

$$\text{(3.5)} \qquad An^{-1}h^{-1} + Bh^{2k},$$

where A and B are the obvious coefficients in (3.1) and (3.4).

While this solution to the bandwidth problem is theoretically pleasing, it is useless in practice because the quantities A and B are unknown. The

main theorem of this paper provides a means of overcoming this difficulty. In particular it is seen that (up to an additive constant), the function

$$-2n^{-1}\log L(h)$$

approximates MISE(h) in the same way as does (3.5) and so the h that maximizes L(h) is optimal in the same sense as the traditional "optimal bandwidth".

The main theorem of this paper also holds in a setting more general than that just discussed. In particular, it is well known that if f has only a bounded p-th derivative where p≤k then

$$s_f(h) = O(h^{2p}),$$

and the optimal (at least in the sense of exponent of algebraic convergence) h may be found by minimizing

$$O(n^{-1}h^{-1}) + O(h^{2p}) .$$

This is perhaps most clearly seen in the results of Stone (1980). It is also well known that p need not be an integer by either using Sobolev space methods or using Lipschitz conditions on derivatives. This setting is more difficult to handle than the above because there one knows the optimal h is of the form

$$cn^{-(2k-1)^{-1}}$$

and only c need be optimized, while here the exponent is also unknown.

In the closely related setting of nonparametric regression estimation, Stone (1982) has posed the problem (see his Question 3) of finding an optimal bandwidth when p is unknown. The theorem of this paper provides a solution to this problem, in the above sense, by showing that,

$$(3.6) \qquad -2n^{-1}\log L(h) = 2R + \mathrm{MISE}(h) + o_p(\mathrm{MISE}(h)),$$

uniformly over h, where the constant R is independent of h and is given by:

$$(3.7) \qquad R = p - n^{-1}\sum_{j=1}^{n} 1_{[a,b]}(X_j)\log f(X_j) .$$

The reason that the nonstandard notation $s_f(h)$ (see (3.4)) has been introduced is that it provides a powerful analytic tool. In the setting of p.k, the usual Taylor expansion techniques are useless for showing results like (3.6) because they only provide an upper bound on $s_f(h)$. Thus the quantity $s_f(h)$ itself is used everywhere in the proof. Another interesting role of $s_f(h)$ is that its tail behavior (as $h \to 0$) provides a measure of what is usually called "smoothness" of f which is more precise than the traditional Lipschitz conditions on derivatives or indices of Sobolev Spaces.

The main theorem of this paper will now be stated formally. First a very mild restriction will be placed on the bandwidth h. For some small $\delta > 0$, define the sequences $\{\underline{h}_n\}$ and $\{\overline{h}_n\}$ by

(3.8) $\quad \underline{h} = n^{-1+\delta}$ and $\overline{h} = n^{-\delta}$,

where here and below the dependence on n is suppressed. It will also be assumed that the density f satisfies:

(3.9) f is bounded above 0 on [a,b]

(3.10) there are constants $M, \gamma > 0$ so that for all x,y

$$|f(x)-f(y)| \leq M|x-y|^{\gamma}$$

Another assumption is that the kernel function K satisfies:

(3.11) $\int K(x)dx = 1$,

(3.12) K is compactly supported,

(3.13) There are constants $M, \gamma > 0$ so that for all x,y

$$|K(x)-K(y)| \leq M|x-y|^{\gamma}.$$

Finally it will be assumed that the weight function in (1.1) is given by

(3.14) $\quad w(x) = 1_{[a,b]}(x)/f(x)$.

Theorem 1: Under the assumptions (3.8) - (3.14), given $\varepsilon > 0$

$$\lim_{n\to\infty} P\left[\sup_{h\in[\underline{h},\bar{h}]} \left| \frac{-2n^{-1}\log\hat{L}(h)-R-MISE(h)}{MISE(h)} \right| > \varepsilon \right] = 0 .$$

A disturbing feature of this theorem is that it only applies to h in the vanishingly small interval $[\underline{h},\bar{h}]$. The above computations show that (by (3.10))

(3.15) $\quad s_f(h) = O(h^{2\gamma})$,

and thus the optimal bandwidth is easily inside the interval for δ sufficiently small. Also Monte Carlo experience with $\hat{L}(h)$ (see Bloomfield and Marron (1984)) indicates this assumption is not a problem in practice. Further reassurance along these lines is provided by

Theorem 2: Under (3.8) - (3.14), if $\hat{h} = \hat{h}(n)$ denotes any sequence of maxima of $\hat{L}(h)$, then

i) $\hat{h} \to 0 \quad$ a.s.

ii) $\lim_{c\to 0} \overline{\lim_{n}} P[\hat{h} < cn^{-(2\gamma+1)^{-1}}] = 0$.

It should be noted that while Theorem 2 does show $\hat{h} > \underline{h}$ (for δ 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 Marron (1983). 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 $\hat{f}(x,h)$. 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 the "vanishing moment" assumption (3.3) is not used in Theorem 1. That theorem says $\hat{f}(x,h)$ will have the best MISE that is possible for the given K, but how good that is is irrelevant

to the theorem. Of course one should choose a reasonably good K.

Remark 4.2. The fact that optimality is achieved only for a particular weight function should not be too disappointing. The one used here is quite natural because MISE is proportional to the expected relative square error:

$$E[(\frac{\hat{f}(X)-f(X)}{f(X)})^2 | X\epsilon[a,b]] \ .$$

It is seen in Marron (1982) that this error norm is precisely the one required for the application of density estimation to the classification problem. It may be seen without too much effort that the indicator function in (2.11) may be replaced by any bounded, measurable nonnegative function q(x), which is supported inside [a,b], and the theorem will still be true with

$$w(x) = q(x)/f(x) .$$

Remark 4.3. At first glance one might be disturbed by the fact that the MISE that is minimized here is limited to the interval [a,b]. In somewhat similar settings, in the case of estimating a regression function, Gasser and Müller (1979) and Rice and Rosenblatt (1983) have observed that such a MISE is strongly affected by the behavior of the unknown function at the endpoints and hence the bandwidth which minimizes MISE 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 MISE of this paper provides a very reasonable error criterion.

Remark 4.4 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. Preliminary computations (for the paper Bloomfield

and Marron (1984)) seem to validate theorem 1 and the heuristics of section 2.

5. Proof of Theorem 1.

This proof uses techniques developed in Hall (1982). It will be useful to define, for j=1,...,n

$$(5.1) \qquad \Delta_j = \frac{\hat{f}_j(X_j,h) - f(X_j)}{f(X_j)} \ , \ \Delta_j^+ = \frac{\hat{f}_j^+(X_j,h) - f(X_j)}{f(X_j)}$$

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

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

in probability. But by (1.2), (1.3), (3.11) and (3.13) letting $\sup_j$ denote supremum over j=1,...,n,

$$\sup_j \sup_x \sup_h nh \left| \hat{f}_j(x,h) - \hat{f}(x,h) \right| =$$

$$= \sup_j \sup_x \sup_h \left| (n-1)^{-1} \sum_{i\neq j} K\left(\frac{x-X_i}{h}\right) - K\left(\frac{x-X_j}{h}\right) \right| \le$$

$$\le 2 \sup_{u\in\mathbb{R}} |K(u)| \ .$$

Hence, by (3.9), using the notation #(A) to mean cardinality of the set

$$A = \{j=1,\ldots,n : X_j\in[a,b]\},$$

note that

$$\sup_h \sup_{j\in A} |\Delta_j^+| \le \sup_h \sup_{j\in A} |\Delta_j| \to 0 \ ,$$

in probability. Now for n=1,2,... define the event

$$U_n = \{\Delta_j^+ = \Delta_j \text{ for each } h\in[\underline{h},\overline{h}] \text{ and } j\in A\} \ .$$

It follows from the above that

$$\lim_{n\to\infty} P[U_n] = 1 \ .$$

From (2.11), (3.7) and the above it follows that, for $h\in[\underline{h},\overline{h}]$, on the event U_n ,

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

$$= n^{-1}\sum_{j=1}^{n}[1_{[a,b]}(X_j)(\Delta_j - \tfrac{1}{2}\Delta_j^2 + r_j) - \rho(X_j) + p]$$

$$= n^{-1}\sum_{j=1}^{n}[1_{[a,b]}(X_j)\Delta_j - \rho(X_j) + p] - \tfrac{1}{2}n^{-1}\sum_{j\in A}\Delta_j^2 + n^{-1}\sum_{j\in A} r_j ,$$

where r_j denotes the error term of the Taylor expansion of $\log(1+x)$.

The remainder of this proof will be split into two lemmas:

Lemma A: Given $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left[\sup_h \left| \frac{n^{-1}\sum_{j=1}^{n}[1_{[a,b]}(X_j)\Delta_j - \rho(X_j) + p]}{MISE(h)} \right| > \varepsilon \right] = 0 .$$

Lemma B: Given $\varepsilon > 0$

$$\lim_{n\to\infty} P\left[\sup_h \left| \frac{n^{-1}\sum_{j\in A}\Delta_j^2 - MISE(h)}{MISE(h)} \right| > \varepsilon \right] = 0 .$$

It is enough to establish these because from Lemma B it follows that, for $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left[\sup_h \left| \frac{n^{-1}\sum_{j\in A} r_j}{MISE(h)} \right| > \varepsilon \right] = 0 .$$

The proof of Lemma A is quite similar in spirit to that of Lemma B. Some details are different but these are very similar to the proof of Lemma 2a in Härdle and Marron (1984). Hence, this proof is omitted.

Proof of Lemma B:

First, for n=1,2,... partition the interval $[\underline{h},\bar{h}]$ by the following means. For $\ell=0,1,\ldots$ define

$$h_\ell = (n^{1-\ell} - n^{-3/4})^{-1} .$$

Then find L so that

$$h_{L-1} < \bar{h} \leq h_L ,$$

and redefine h_L to be $\bar{h}$. Note that the dependence of h_ℓ and L on n has been suppressed. Note also that

$$(5.2) \qquad |h_\ell^{-1} - h_{\ell+1}^{-1}| \leq n^{-3/\gamma}$$

and that, as $n \to \infty$,

$$(5.3) \qquad L = o(n^{1+3/\gamma}).$$

It will be convenient to adopt the shorthand:

$$(5.4) \qquad A(h) = \left[n^{-1} \sum_{j=1}^{n} 1_{[a,b]}(X_j) \Delta_j^2 - MISE(h) \right] MISE(h)^{-1} .$$

The idea of this proof is to show that A(h) converges uniformly over the "grid points", h_ℓ, and then to "fill in the gaps" with Lipschitz continuity. More formally, for $\varepsilon > 0$, note that

$$P[\sup_h |A(h)| > \varepsilon] \leq I + II ,$$

where the behavior at the grid points is controlled by

$$I = P[\sup_\ell |A(h_\ell)| > \varepsilon/2]$$

(where $\sup_\ell$ denotes supremum over $\ell=1,\ldots,L$), and the behavior between gridpoints is controlled by

$$II = P[\sup_{\ell,h} |A(h) - A(h_\ell)| > \varepsilon/2]$$

(where $\sup_{\ell,h}$ denotes supremum over $\ell=1,\ldots,L$ and $h \in [h_{\ell-1}, h_\ell]$). The proof of Lemma B will be complete when the following lemmas are established.

Lemma B1:

$$I \to 0.$$

Lemma B2:

$$II \to 0.$$

Proof of Lemma B1:

Note that by an obvious extension of the Chebychev Inequality, for M=2,4,6,...

$$P\left[\sup_{\ell}|A(h_\ell)| > \varepsilon/2\right] \le \sum_{\ell=1}^{L} P\left[|A(h_\ell)| > \varepsilon/2\right] \le$$

$$\le L \sup_{\ell} P[|A(h_\ell)| > \varepsilon/2] \le$$

$$\le L \sup_{\ell} E(2A(h_\ell)/\varepsilon)^M .$$

Thus, by (5.3), it is enough to show that

$$\sup_h n^{1+3/\gamma} E(A(h))^M \to 0,$$

for M sufficiently large.

Next observe that, by computations similiar to those leading to (3.1),

$$E(A(h)) = [E([\hat{f}_j(X_j,h) - f(X_j)]^2 f(X_j)^{-2} 1_{[a,b]}(X_j)) - MISE]\, MISE^{-1} =$$

$$= [E\int[\hat{f}_j(x,h) - f(x)]^2 w(x)dx - MISE]\, MISE^{-1} =$$

$$= [(n-1)^{-1}h^{-1}(\int f(y)w(y)dy)(\int K(u)^2 du) + o(n^{-1}h^{-1}) + s_f(h) - MISE]MISE^{-1},$$

and so

$$\sup_h E(A(h)) = O(n^{-1}) .$$

Now using a cumulant expansion (see, for example, (3.33) of Kendall and Stuart (1963)) of the M-th centered moment of A(h), to complete the proof of Lemma B1 it is enough to show that, for M sufficiently large, for m=2,...,M,

$$\sup_h cum_m(A(h),...,A(h)) = o(n^{-(1+3/\gamma)m/M})$$

where cum_m denotes the m-th order cumulant.

Observe that from (1.3), (5.1) and (5.4),

$$A(h) = \left[n^{-1}\sum_{j=1}^{n}\left[(n-1)^{-1}\sum_{i\neq j}\frac{1}{h}K(\frac{X_j - X_i}{h}) - f(X_j)\right]^2 f(X_j)^{-2} 1_{[a,b]}(X_j) - MISE\right] MISE^{-1} =$$

$$= n^{-1}\sum_{j=1}^{n}\left[\sum_{i\neq j} V_{ij}\right]^2 - 1 ,$$

where, by (3.14),

$$(5.5) \qquad V_{ij} = (n-1)^{-1}\left[\frac{1}{h}K\left(\frac{X_j - X_i}{h}\right) - f(X_j)\right] f(X_j)^{-\frac{1}{2}} w(X_j)^{\frac{1}{2}} MISE^{-\frac{1}{2}} .$$

S , by using linearity properties of cumulants (see, for example (iv) and (v) of theorem 2.3.1 in Brillinger (1979)) the proof of Lemma B1 will be complete when it is seen that, for M sufficiently large, m=2,...,M,

$$(5.6) \qquad \sup_h n^{-m}\sum_{j}\sum_{i}\sum_{i'} \mathrm{cum}_m(V_{i_1 j_1} V_{i'_1 j_1}, \ldots, V_{i_m j_m} V_{i'_m j_m}) = o(n^{-(1+3/\gamma)m/M}) ,$$

where $\sum_j$ denotes summation over $j_1,\ldots,j_m = 1,\ldots,n$, and where $\sum_i$ denotes summation over $i_1,\ldots,i_m = 1,\ldots,n$ subject to the restrictions $i_1 \neq j_1,\ldots,i_m \neq j_m$, and where $\sum_{i'}$ denotes summation over $i'_1,\ldots,i'_m = 1,\ldots,n$ subject to the restrictions $i'_1 \neq j_1,\ldots,i'_m \neq j_m$.

By another of the properties of cumulants, note that many of the terms in the summation (5.6) will be 0 because of the independence of $X_1,\ldots,X_n$. The nonzero terms will be handled by grouping them according to pattern of indices and proceeding casewise.

First note that by the usual moment expansion of cumulants (see, for example, (3.39) of Kendall and Stuart (1963)), each cum_m may be expanded into a linear combination, the first term of which is

$$(5.7) \qquad E[V_{i_1 j_1} V_{i'_1 j_1} \cdots V_{i_m j_m} V_{i'_m j_m}] ,$$

and the remaining terms of which are multiples of products of moments of all the various partitions of

$$\{V_{i_1 j_1} V_{i'_1 j_1}, \cdots, V_{i_m j_m} V_{i'_m j_m}\} .$$

Next a means of counting the nonzero terms in (5.6) will be developed. Since special attention must be paid to duplications among $i_1,\ldots,i_m$, $i_1',\ldots,i_m'$, and $j_1',\ldots,j_m'$, the following relabeling of the indices will be made:

(i) Suppose that r is the number of $j_1,\ldots,j_m$ that are distinct. Relabel these indices (each time they occur) by $j_1,\ldots,j_r$.

(ii) Suppose that the number of $i_1,\ldots,i_m$, $i_1',\ldots,i_m'$ that are the same as one of $j_1,\ldots,j_r$ is s. Each of these will now be denoted by the appropriate j.

(iii) Suppose that, out of what remain of $i_1,\ldots,i_m$, $i_1',\ldots,1_m'$ that of them are distinct. Denote these (each time they occur) by $i_1,\ldots,i_t$. Also let s_z denote the number of times the new i_1 appears, and similarly for $s_2,\ldots,s_t$.

Now by rearranging the V's, note that (5.7) may be rewritten as

$$E[V_{jj}\text{'s } V_{i_1 j}\text{'s}\cdots V_{i_t j}\text{'s}] ,$$

where "V_{jj}'s" denotes the product of all V's whose first index is one of $j_1,\ldots,j_r$, and where "$V_{i_1 j}$'s" denotes the product of all V's whose first index is i_1, etc. Since there seems to be no chance of confusion, both notations will be used in the following.

Now group the nonzero cumulant in (5.6) according to the pattern of duplication of indices (eg: $\text{cum}_2(V_{12}V_{12},V_{24}V_{14})$ is in the same group as $\text{cum}_2(V_{53}V_{53},V_{31}V_{51})$). Note that the number of cumulants falling into each group is (as $n\to\infty$) of the order $O(n^{r+t})$. So to verify (5.6), it is enough to show that, for each duplication group,

$$\sup_h n^{r+t-m}\,\text{cum}_m(\quad) = o(n^{-(1+3/\gamma)m/M}) ,$$

where M is sufficiently large, and m=2,3,...,M.

To this end, define the set

$$J = \{X_{j_1},\ldots,X_{j_t}\},$$

and let $E[\cdot|J]$ denote the usual conditional (on $X_{j_1},\ldots,X_{j_t}$) expectation operator. Now letting B denote a generic constant, by (3.9), (3.13), (5.5) and integration by substitution,

$$n^{r+t-m}|E(V_{jj}\text{'s } V_{i_1 j}\text{'s}\cdots V_{i_t j}\text{'s})| = n^{r+t-m}|E(V_{jj}\text{'s } E[V_{i_1 j}\text{'s}|J]\cdots E[V_{i_t j}\text{'s}|J])|$$

$$\leq Bn^{r+t-m}\frac{n^{-s_h-s}}{MISE^{s/2}}\cdot\frac{n^{-s_{l_h}-(s_l-1)}}{MISE^{s_l/2}}\cdots\frac{n^{-s_{t_h}-(s_t-1)}}{MISE^{s_t/2}} = \tag{5.8}$$

$$= Bn^{-(3m-r-t)}h^{-(2m-t)}MISE^{-m} = B[(nh)^{-m}MISE^{-m}](nh)^{-(2m-r-t)}h^{m-r}.$$

While this bound is sufficient to handle many of the patterns of duplication of indices, refined computations of several types are required for others.

To see what cases are necessary, note that in (5.6), the $cum_m(\ )$ are nonzero only when no subset of the arguments of cum_m is independent of the remaining arguments (see, for example, (iii) in Theorem 2.3.1 of Brillinger (1979)). In other words, there must be at least m-1 pairs of arguments of $cum_m(\ )$ which have an index in common. Thus, for each nonzero $cum_m(\ )$, the following counting argument is valid:

$$\begin{aligned} m-1 &\leq \#(\text{pairs with a common index}) \leq \\ &\leq \#(\text{pairs with an i in common}) + \#(\text{pairs with a j in common}) \leq \\ &\leq [\#(\text{i's available}) - \#(\text{distinct i's})] + \\ &\qquad + [\#(V_{jj}\text{'s}) + \#(\text{j's available}) - \#(\text{distinct j's})] \leq \\ &\leq [(2m-s)-t] + [s+m-r] = 3m-t-r. \end{aligned} \tag{5.9}$$

It follows from this that

$$2m-t-r \geq -1. \tag{5.10}$$

The bound (5.8) will now be either used or refined in a casewise manner.

Case 1: $m-r \geq m/12$ and $2m-r-t \geq 0$.

It follows from (3.1) that

$$(5.11) \qquad \sup_h (nh)^{-1}\mathrm{MISE}(h)^{-1} = O(1).$$

Thus from (3.8) and (5.8)

$$\sup_h n^{r+t-m}|E(V_{jj}'s\ V_{i_1j}'s\cdots V_{i_tj}'s)| = O(h^{m/12}) = O(n^{-m/12}) .$$

But similar computations show that the same bound may be obtained for the other products of moments appearing in $\mathrm{cum}_m(\ )$. Thus

$$\sup_h n^{r+t-m}|\mathrm{cum}_m(V_{i_1j_1}V_{i_1'j_1'},\ldots,V_{i_mj_m}V_{i_m'j_m'})| = O(n^{-\delta m/12}) = o(n^{-(1+3/\ell)m/M}),$$

by taking M sufficiently large.

Case 2: $m-r \geq m/12$ and $2m-r-t = -1$.

Here the basic bound (5.8) needs some modification. Since $r \leq m$, note that

$$m-t = r-m-1 < 0$$

and hence

$$t > m.$$

Thus at least two of $s_1,\ldots,s_t$ must be equal to 1. Now relabel $i_1,\ldots,i_t$ so that s_{t-1} and s_t are both 1. The bound (5.8) may now be modified to give

$$n^{r+t-m}|E(V_{jj}'s\ V_{i_1j}'s\cdots V_{i_tj}'s)| \leq$$

$$\leq Bn^{r+t-m}\frac{n^{-s_h-s}}{\mathrm{MISE}^{s/2}}\cdot\frac{n^{-s_1}h^{-(s_1-1)}}{\mathrm{MISE}^{s_1/2}}\cdots\frac{n^{-s_{t-2}}h^{-(s_{t-2}-1)}}{\mathrm{MISE}^{s_{t-2}/2}}$$

$$\cdot E_J E[V_{i_{t-1}j}|J]E[V_{i_tj}|J]_+ .$$

But from (5.5)

$$E[V_{ij}|J] = (n-1)^{-1}\left[\int\frac{1}{h}K(\frac{X_j-x}{h})f(x)dx - f(X_j)\right]f(X_j)^{-1/2}w(X_j)^{1/2}\mathrm{MISE}^{-1/2} .$$

Thus, by (3.2) and integration by substitution,

$$(5.12)\qquad E(E[V_{ij}|J]^2) = (n-1)^{-1}s_f(h)MISE(h)^{-1},$$

and so, by the Schwartz Inequality,

$$E|E[V_{i_{t-1}j}|J]E[V_{i_tj}|J]| \le (n-1)^{-1}s_f(h)MISE(h)^{-1} .$$

Hence, from (3.1)

$$\sup_h n^{r+t-m}|E(V_{jj}'s\ V_{i_1j}'s\cdots V_{i_tj}'s)| \le$$

$$(5.13)\qquad \le \sup_h B[(nh)^{-(m-1)}s_f(h)MISE(h)^{-m}](nh)^{-(2m-r-t+1)}h^{m-r} \le$$

$$\le O(h^{m/12}) = O(n^{-\delta m/12}) .$$

But similar computations show that the same bound may be obtained for the other products of moments appearing in $cum_m(\ )$. Thus

$$\sup_h n^{r+t-m}|cum_m(V_{i_1j_1}\ V_{i_1'j_1'},\ldots,V_{i_mj_m}\ V_{i_m'j_m'})| = O(n^{-\delta m/12}) = o(n^{-(1+3/\gamma)m/M}),$$

by taking M sufficiently large.

Case 3: $m-r < m/12$ and $2m-r-t \ge m/12$.

It follows from (3.8) that

$$(nh)^{-1} = n^{-\delta} .$$

Thus, since $r \le m$, by (5.8) and (5.11)

$$\sup_h n^{r+t-m}|E(V_{jj}'s\ V_{i_1j}'s\cdots V_{i_tj}'s)| = O(n^{-\delta m/12}) .$$

Hence, as above,

$$\sup_h n^{r+t-m}|cum_m(V_{i_1j_1}\ V_{i_1'j_1'},\ldots,V_{i_mj_m}\ V_{i_m'j_m'})| = O(n^{-\delta m/12}) = o(n^{-(1+3/\delta)m/M}),$$

for M sufficiently large.

Case 4: $m-r < m/12$, $0 \le 2m-r-t < m/12$, and $s \ge m/3$.

For this case, consider the factors $E[V_{ij}'s|J]$ appearing in the computation

(5.8). It will be convenient to apply the name "singleton" to those for which the corresponding $s_i=1$. Note that

$$t - \#(\text{singletons}) \le \#(\text{i's available}) - \#(\text{places to put i's}) = (2m-s)-t .$$

Thus,

$$\#(\text{singletons}) \ge 2(t-m)+s > 2(m-r-m/12) + s \ge -2m/12 + m/3 = m/6. \tag{5.14}$$

Now it is desired to use the computation (5.12) to generate extra factors of $s_f(h)$ from the above singletons. To do this, for each of $j_1,\dots,j_r$, at most 2 singletons having that particular j may be employed. Let u count the number of singletons that may be used. Since

$$r = \#(\text{distinct } j\text{'s}) \le 11m/12 ,$$

note that

$$u \ge m/6 - m/12 = m/12. \tag{5.15}$$

Now relabel $i_1,\dots,i_t$ so that the above singletons are indexed by $i_{t-u+1},\dots,i_t$. Note that the computation (5.8) may be refined to give

$$n^{r+t-m}|E(V_{jj}\text{'s } V_{i_1 j}\text{'s}\cdots V_{i_t j}\text{'s})| \le$$

$$\le Bn^{r+t-m}\frac{n^{-s_h-s}}{MISE^{s/2}}\frac{n^{-s_1 h-(s_1-1)}}{MISE^{s_1/2}}\cdots\frac{n^{-s_{t-u} h-(s_{t-u}-1)}}{MISE^{s_{t-u}/2}}\cdot$$

$$\cdot E(E[V_{i_{t-u+1} j}|J]\cdots E[V_{i_t j}|J])\le$$

$$\le Bn^{-(3m-r-t)}h^{-(2m-t)}s_f(h)^{u/2}MISE^{-m} \le$$

$$\le B[(nh)^{-m}MISE^{-m}](nh)^{-(2m-r-t)}h^{-(m-r)}s_f(h)^{m/24} .$$

But now, from (3.8) and (3.15), as above

$$\sup_h n^{r+t-m}|E(V_{jj}\text{'s } V_{i_1 j}\text{'s}\cdots V_{i_t j}\text{'s})| = O((\bar{h}^{2\gamma})^{m/24}) = O(n^{-\delta\gamma m/12}) .$$

Thus, as above,

$$\sup_h \left| \mathrm{cum}_m(V_{i_1 j_1} V_{i_1' j_1'}, \ldots, V_{i_m j_m} V_{i_m' j_m'}) \right| = O(n^{-3m/12}) = o(n^{-(1+3/\gamma)m/M}),$$

for M sufficiently large.

Case 5: $m-r \leq m/12$, $2m-r-t = -1$, and $s \geq m/3$.

This case is an extension of Case 4 in the same way that Case 2 extends Case 1. Note that in the present case, the computation (5.14) can be improved to

$$\#(\text{singletons}) \geq 2(m-r+1)+s \geq 2+m/3 .$$

Thus (5.15) can be improved to

$$u \geq 2+m/4 .$$

The extra two singletons are used to generate an extra $s_f(h)$ which is used as in (5.13). The result is:

$$\sup_h \left| \mathrm{cum}_m(V_{i_1 j_1} V_{i_1' j_1'}, \ldots, V_{i_m j_m} V_{i_m' j_m'}) \right| = O(n^{-\delta\gamma m/4}) = o(n^{-(1+3/\gamma)m/M}),$$

for M sufficiently large.

Case 6: $m-r \leq m/12$, $2m-r-t \geq 0$, and $s < m/3$.

First recall that $\mathrm{cum}_m(\ )$ is nonzero only if at least $(m-1)$ pairs of arguments of $\mathrm{cum}_m(\ )$ have an index in common. Let v denote the number of such pairs which have an i in common, but different j's. The counting argument (5.9) may be modified to give

$$m-1 \leq v + \#(\text{pairs where common index is a } j) \leq v + [s+m-r].$$

Thus,

$$v \geq m-1 - [s+m-r] \geq m-1 - [m/3 + m/12] = 7m/12 - 1. \tag{5.16}$$

Note that a "pair with an i in common, but different j's " arises from having factors V_{ij} and $V_{ij'}$ (for some $j \neq j'$). Now given $X_{j_1}, \ldots, X_{j_r}$, define the random variable Z by:

$$Z = \#(\text{such pairs with } |X_j - X_{j'}| \leq 2K^* h),$$

where K^* denotes the length of the compact support of the kernel function K.

Note that, for $z=0,\ldots,v$,

$$P[Z=z] = O(h^{v-z}) .$$

Also note that if one of the above pairs comes from $V_{i_1 j_1}$ and $V_{i_1 j_2}$ (for example) and if $|X_{j_1} - X_{j_2}| > 2K^*h$, then

$$E[V_{i_1 j}{}'s \mid J] = E[V_{i_1 j_1}^{q_1} V_{i_1 j_2}^{q_2} \cdots V_{i_1 j_r}^{q_r} \mid J] = \int E[V_{j_1}(x)^{q_1} \cdots V_{j_r}(x)^{q_r} \mid J] f(x)dx ,$$

where $q_1, q_2 > 0$, and where

$$V_j(x) = (n-1)^{-1} \left| \frac{1}{h} K\left(\frac{X_j - x}{h}\right) - f(X_j) \right| f(X_j)^{-1/2} w(X_j)^{1/2} MISE^{-1/2} .$$

From which it follows that

$$|E[V_{i_1 j}{}'s \mid J]| \leq \int |E[V_{j_1}(x)^{q_1} \cdots V_{j_r}(x)^{q_r} \mid J]| \, f(x)dx =$$

$$= \int_{\{x: |x-X_{j_1}| \leq K^*h\}} dx + \int_{\{x: |x-X_{j_2}| \leq K^*h\}} dx +$$

$$+ \int_{\{x: |x-X_{j_1}| > K^*h \text{ and } |x-X_{j_2}| > K^*h\}} dx \leq$$

$$\leq h^{q_1} \int dx + h^{q_2} \int dx + h^{q_1+q_2} \int dx \leq Bh \cdot \frac{n^{-s_1} h^{-(s_1-1)}}{MISE^{s_1/2}} .$$

By similar computations, for each of the above pairs V_{ij} and $V_{ij'}$, on the event $\{|X_j - X_{j'}| > 2K^*h\}$ at least one of the $E[V_{ij}{}'s \mid X_j{}'s]$ allows the factoring out of an additional power of h. Thus when $Z=z$, an additional h^{v-z} may be used in the basic bound (5.8). Letting $E(\ ; Z=z)$ denote expectation only over the event $\{Z=z\}$, (5.8) may be modified to

$$n^{r+t-m} |E(V_{jj}{}'s \, V_{i_1 j}{}'s \cdots V_{i_t j}{}'s)| \leq$$

$$\leq \sum_{z=0}^{v} n^{r+t-m} |E(V_{jj}{}'s \, V_{i_1 j}{}'s \cdots V_{i_t j}{}'s; \, Z=z)| \leq$$

$$\leq \sum_{z=0}^{v} n^{r+t-m} Bh^{(v-z)} n^{-(3m-r-t)} h^{-(2m-t)} MISE^{-m} h^{z} = O(h^{v}) .$$

Thus,

$$\sup_{h} n^{r+t-m} |E(V_{jj}'s\ V_{i_1 j}'s \cdots V_{i_t j}'s)| = O(n^{-\delta v}) .$$

At first glance, it looks like the above techniques may not be useful for handling the other "products of moments" which appear in the expansion of $cum_m(\ )$. This is because the above effect of "generating a factor of h" from $E[V_{ij}'s|J]$ will be lost when V_{ij} and $V_{ij'}$ fall into different subsets of the partition. But this is actually not a problem, because splitting into partitions already generates extra factors of h^d (from integration by substitution), so that the above bounds still apply. Thus, using (5.16) and the fact that $m \geq 2$,

$$\sup_{h} n^{r+t-m} |cum_m(V_{i_1 j_1} V_{i_1' j_1}, \ldots, V_{i_m j_m} V_{i_m' j_m})| =$$

$$= O(n^{-\delta v}) = O(n^{-\delta m/12}) = o(n^{-(1+3/\gamma)m/M}) ,$$

for M sufficiently large.

Case 7: $m-r < m/12$, $2m-r-t = -1$, and $s < m/3$.

Recall that Case 5 extends Case 4 in the same way that Case 2 extends Case 1. The present case extends Case 6 in the same way. The details are straight forward and hence are omitted.

Now by (5.10) all cases have been exhausted. This verifies (5.6) which completes the proof of Lemma B1.

Proof of Lemma B2:

To check that $II \to 0$, from (3.14), (5.1) and (5.4) write

$$A(h) = N(h)MISE(h)^{-1} - 1,$$

where

$$N(h) = n^{-1}\sum_{j=1}^{n}[\hat{f}_j(X_j,h) - f(X_j)]^2 f^{-1}(X_j)w(X_j).$$

Note that for $\ell=1,\ldots,L$ and $h\in[h_{\ell-1},h_\ell]$,

$$|A(h) - A(h_\ell)| \le \frac{|N(h)-N(h_\ell)|}{MISE(h)} + \frac{|N(h_\ell)|}{MISE(h_\ell)} \cdot \frac{|MISE(h_\ell)-MISE(h)|}{MISE(h)}.$$

Thus, by Lemma B1, the proof of Lemma B2 will be complete when it is seen that

$$(5.17)\qquad \sup_{\ell,h} |N(h)-N(h_\ell)|MISE(h)^{-1} \to 0$$

in probability, and that

$$(5.18)\qquad \sup_{\ell,h} |MISE(h_\ell)-MISE(h)|MISE(h)^{-1} \to 0.$$

To verify (5.17), note that by (1.3) and the algebraic identity $a^2-b^2=(a-b)(a+b)$,

$$|N(h)-N(h_\ell)| \le n^{-1}\sum_{j=1}^{n} |(\hat{f}_j(X_j,h)-\hat{f}_j(X_j,h_\ell)) \cdot$$

$$\cdot (\hat{f}_j(X_j,h)+\hat{f}_j(X_j,h_\ell)-2f(X_j))f^{-1}(X_j)w(X_j)| \le$$

$$\le n^{-1}\sum_{j=1}^{n}(n-1)^{-1}\sum_{i\neq j}\left|\frac{1}{h}K\left(\frac{X_j-X_i}{h}\right)-\frac{1}{h_\ell}K\left(\frac{X_j-X_i}{h_\ell}\right)\right| \cdot$$

$$\cdot \left(\sup_x \hat{f}_j(x,h)+\sup_x \hat{f}_j(x,h_\ell) + 2f(X_j)\right)f^{-1}(X_j)w(X_j).$$

But now by (1.3), (3.8) and (3.13), for B a generic constant

$$\sup_h \sup_x \hat{f}_j(x,h) \le \sup_h \sup_u \frac{1}{h}|K(u)| \le Bh^{-1} = Bn^{1-\delta},$$

and also by (5.2)

$$\left|\frac{1}{h}K\left(\frac{y-x}{h}\right) - \frac{1}{h_\ell}K\left(\frac{y-x}{h_\ell}\right)\right| \le |h^{-1}-h_\ell^{-1}| \sup_u |K(u)| + h_\ell^{-1} M|h^{-1}-h_\ell^{-1}| \le$$

(5.19)

$$\le B(n^{-3/\gamma} + n^{1-\delta}(n^{-3/\gamma})^\gamma).$$

It follows from the above and (3.1) that

$$\sup_{\ell,h} \frac{|N(h)-N(h_\ell)|}{MISE(h)} \le \frac{Bn^{-1-2\delta}}{n^{-1}} \to 0 .$$

To check (5.18) note that by the above method

$$|MISE(h_\ell)-MISE(h)| \le \int E|\hat{f}(x,h_\ell)-\hat{f}(x,h)| \cdot |\hat{f}(x,h_\ell)+\hat{f}(x,h)-2f(x)|w(x)dx.$$

But by (1.2) and (5.19),

$$|\hat{f}(x,h_\ell)-\hat{f}(x,h)| \le Bn^{-2-\delta} .$$

Thus, since $E|\hat{f}(x,h_\ell)|$ is bounded,

$$\sup_{\ell,h} \frac{|MISE(h_\ell)-MISE(h)|}{MISE(h)} \le \frac{Bn^{-2-\delta}}{n^{-1}} \to 0.$$

This completes the proof of Lemma B2 and hence also that of Lemma B.

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. The author is also indebted to Charles J. Stone for suggesting a key step in the proofs.

REFERENCES

Bloomfield, P. and Marron, J. S. (1984). 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, Inc.

Cheng, S.H. (1983). On a problem concerning spacings. Center for Stochastic Processes Tech. Rept. #27, Statistics Dept., UNC, Chapel Hill, NC.

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. Theor. Prob. 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. Compstat 1974: Proceedings in computational statistics. (G. Bruckman, ed.) 101-110. Vienna: Physica Verlag.

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 Institute of Statistics Mimeo Series #1546.

Kendall, M. G. and Stuart, A. (1963). The Advanced Theory of Statistics, Vol 1: Distribution Theory, Butler and Tanner Ltd.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). Extremes and related properties of random sequences and processes. Springer (New York).

Marron, J. S. (1982). Optimal rates of convergence to Bayes risk in nonparametric discrimination. Ann. Statist. 11, 1142-1155.

Marron, J. S. (1983). Uniform convergence properties of a cross-validated density estimator. North Carolina Institute of Statistics Mimeo Series #1519.

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 nonparanetric estimators. Ann. Statist. 8, 1348-1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10, 1040-1053.

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.

END

FILMED

2-85

DTIC