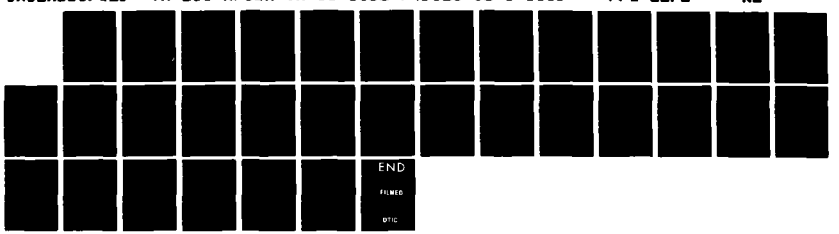
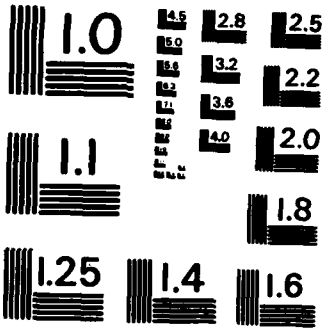
AD-A160 235 THE AMOUNT OF NOISE INHERENT IN BANDWIDTH SELECTION FOR A KERNEL DENSITY..(U) NORTH CAROLINA UNIV AT CHAPEL HILL CENTER FOR STOCHASTIC PROC.. P HALL ET AL. MAY 85 1/1

UNCLASSIFIED TR-100 AFOSR-TR-85-0631 F49620-82-C-0009 F/G 12/1 NL





CENTER FOR STOCHASTIC PROCESSES

Department of Statistics
University of North Carolina
Chapel Hill, North Carolina

THE AMOUNT OF NOISE INHERENT IN BANDWIDTH SELECTION FOR A KERNEL DENSITY ESTIMATOR

by

Peter Hall and James Stephen Marron

Technical Report No. 199

May 1985

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): Technical Report No. 100	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 85-0631
6a. NAME OF PERFORMING ORGANIZATION: Center for Stochastic Processes; 6b. OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: AFOSR
6c. ADDRESS (City, State and ZIP Code): Dept. of Statistics, UNC, Chapel Hill, NC 27514	7b. ADDRESS (City, State and ZIP Code): Bldg. 410, Bolling AFB, D.C. 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR; 8b. OFFICE SYMBOL (If applicable):	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: F49620-82-C-0009
8c. ADDRESS (City, State and ZIP Code): Bolling AFB, Washington, DC 20332	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2304	A5	

11. TITLE (Include Security Classification): THE AMOUNT OF NOISE INHERENT IN BANDWIDTH SELECTION FOR A KERNEL DENSITY ESTIMATOR

12. PERSONAL AUTHOR(S): Peter Hall and James Stephen Marron

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Technical	FROM 9/84 TO 8/85	May 1985	29

16. SUPPLEMENTARY NOTATION:

17. COSATI CODES		
FIELD	GROUP	SUB GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): bandwidth, cross-validation, data-driven estimate, density estimate, noise, second-order optimal, window width

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

Let $\hat{f}(\cdot|h)$ be a kernel estimator of a density f, using bandwidth h. The bandwidth $\hat{h}_f$ which minimizes the integrated square error of $\hat{f}$, depends on the unknown f. Therefore it is not a practical choice. Any data-driven attempt to minimize integrated square error must employ a bandwidth $\hat{h}$ which depends only on the sample. The integrated square error using $\hat{h}$ will exceed that using $\hat{h}_f$. In this paper we show that there is an unbridgeable gap between these two integrated square errors. In fact, we quantify the amount of noise inherent in any data-driven attempt to estimate $\hat{h}_f$. A bandwidth which minimizes this noise might be called "second-order optimal". We show that the cross-validatory bandwidth is second-order optimal.

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

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Brian W. Woodruff, Maj, USAF	(202)767-5027	NM

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

unclassified
SECURITY CLASSIFICATION OF THIS PAGE

THE AMOUNT OF NOISE INHERENT IN BANDWIDTH SELECTION FOR A KERNEL DENSITY ESTIMATOR

by

Peter Hall[1,2] and James Stephen Marron[3]

University of North Carolina, Chapel Hill

ABSTRACT. Let $\hat{f}(\cdot|h)$ be a kernel estimator of a density f, using bandwidth h. The bandwidth $\hat{h}_f$ which minimises the integrated square error of $\hat{f}$, depends on the unknown f. Therefore it is not a practical choice. Any data-driven attempt to minimise integrated square error must employ a bandwidth $\hat{h}$ which depends only on the sample. The integrated square error using $\hat{h}$ will exceed that using $\hat{h}_f$. In this paper we show that there is an unbridgeable gap between these two integrated square errors. In fact, we quantify the amount of noise inherent in any data-driven attempt to estimate $\hat{h}_f$. A bandwidth which minimises this noise might be called "second-order optimal". We show that the cross-validatory bandwidth is second-order optimal.

SHORT TITLE: Noise in bandwidth selection.

AMS (1980) SUBJECT CLASSIFICATION: Primary 62G05, Secondary 62E20, 62H99.

KEY WORDS AND PHRASES: bandwidth, cross-validation, data-driven estimate, density estimate, noise, second-order optimal, window width.

[1]On leave from Australian National University.

[2]Work of first author supported by AFOSR Grant No. F 49620 82 C 0009.

[3]Work of second author partially supported by NSF Grant DMS-8400602.

1. Introduction.

Let $\hat{f}(\cdot|h)$ be a nonparametric, kernel estimator of an unknown density f, with bandwidth (window size) h. A considerable amount has been written about "optimal" selection of bandwidth, usually in the context of minimising L^2 error (see e.g. Fryer [6], Wegman [23]). In particular, there are well-known asymptotic formulae for the window h_f which minimises mean integrated square error for a given f (see Parzen [14], Rosenblatt [17]). Of course, h_f depends intimately on the unknown density, and so is not a practical choice. Furthermore, a statistician who has been given a sample to analyse should really be interested in minimising integrated square error for that particular sample, not in minimising the average error over all possible samples. Unfortunately the window $\hat{h}_f$ which minimises integrated square error is also an intricate function of the unknown f.

Any practical method of constructing a bandwidth must depend only on ~~the~~ a practical sample, and should produce some sort of estimate of this bandwidth. The purpose of this paper is to show that there are well-defined limits to the accuracy of all data-driven bandwidth estimates. Put another way, there is an unbridgeable gap between the minimum integrated square error attained using ~~the~~ a optimal bandwidth $\hat{h}_f$, and the minimum achievable integrated square error using a data-driven bandwidth estimate. Additional keywords: Fourier series, cross validation, random variables.

We pause now to introduce notation. Let $X_1, \ldots, X_n$ be a random sample from an unknown density f, and let K be a kernel function. Here and during most of this paper we work in one dimension, although extensions to higher dimensions will be indicated at the end of Section 2. We assume at least

that K is continuous with compact support, and is constructed to suit a density f with $t(\geq 2)$ bounded derivatives:

$$(1.1) \qquad \int z^i K(z)dz = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{if } 1 \leq i \leq t-1 \\ d_K & \text{if } i = t, \end{cases}$$

where $d_K \neq 0$. (The most common case, where K is symmetric and positive, has $t = 2$ in these specifications, see Parzen [14] and Bartlett [1] for discussions of the general case. Our density estimate is

$$f(x|h) \equiv (nh)^{-1} \sum_{i=1}^{n} K\{(x-X_i)/h\}, \quad -\infty < x < \infty,$$

and has integrated square error

$$\Delta(h,f) \equiv \int\{\hat{f}(x|h) - f(x)\}^2 dx .$$

Mean integrated square error is given by

$$M(h,f) \equiv E\{\Delta(h,f)\} .$$

Define also:

$$D(h,f) \equiv \Delta(h,f) - M(h,f) .$$

Assume f has at least t continuous deviratives, and suppose for the sake of argument that f vanishes outside a compact interval. (This property permits us to avoid cumbersome regularity conditions, but is not essential.) Then the "optimal fixed bandwidth" h_f minimises M(h,f) and is asymptotic to a constant multiple of $n^{-1/(2t+1)}$, and the "optimal bandwidth" $\hat{h}_f$ minimises $\Delta(h,f)$ and satisfies $\hat{h}_f/h_f \to 1$ in probability as $n \to \infty$. Any practical procedure for constructing a bandwidth produces a random variable $\hat{h}$ which is

a function solely of the sample; it clearly must not depend on the unknown f. A statistician who claims that a certain procedure $\hat{h}$ is "best possible", is really saying: "In some sense, the closest you can come to minimizing $\Delta(h,f)$, is $\Delta(\hat{h},f)$". Of course, $\Delta(\hat{h},f)$ exceeds the true minimum $\Delta(\hat{h}_f,f)$, but we cannot realistically expect to close that gap. It is known [9] that if $\hat{h}_c$ is the cross-validatory window, then $n\{\Delta(\hat{h}_c,f) - \Delta(\hat{h}_f,f)\}$ has an asymptotic chi-square distribution with one degree of freedom. Therefore the distance between $\Delta(\hat{h},f)$ and $\Delta(\hat{h}_f,f)$ can be reduced to <u>at least</u> order n^{-1}. In Section 2 we shall show that in a minimax sense, order n^{-1} is a <u>lower bound</u> as well as an upper bound. From this point of view, least-squares cross-validation is <u>second-order optimal</u>; it is already known to be first-order optimal [7,8,21,4].

Throughout this discussion we have assessed optimality on the Δ-scale, not the h-scale. However, the two are interchangeable. To see this, observe that if the kernel K has two continuous derivatives then we may expand $\Delta(\hat{h},f)$ in a Taylor series about $\hat{h}_f$, obtaining:

$$\Delta(\hat{h},f) = \Delta(\hat{h}_f,f) + (\hat{h}-\hat{h}_f)\,\Delta^{(1)}(\hat{h}_f,f) + \tfrac{1}{2}(\hat{h}-\hat{h}_f)^2\,\Delta^{(2)}(h^*,f)\,,$$

where h^* lies inbetween $\hat{h}$ and $\hat{h}_f$, and

$$\Delta^{(i)}(h,f) \equiv (\delta/\delta h)^i\,\Delta(h,f)\,.$$

Since $\hat{h}_f$ minimises $\Delta(\cdot,f)$,

$$\Delta(\hat{h},f) - \Delta(\hat{h}_f,f) = \tfrac{1}{2}(\hat{h}-\hat{h}_f)^2\,\Delta^{(2)}(h^*,f)\,.$$

Suppose the data-driven bandwidth $\hat{h}$ has at least a chance of being "good", so that $\hat{h}/\hat{h}_f \to 1$ in probability. Then it may be shown, under the conditions stated earlier about f, that as $n \to \infty$,

$$n^{2(t-1)/(2t+1)}\,\Delta^{(2)}(h^*,f) \to c(f,K) > 0$$

in probability. In fact,

$$c(f,K) = \lim_{n\to\infty} n^{2(t-1)/(2t+1)} M(h_f,f)\,.$$

Therefore

$$\Delta(\hat{h},f) - \Delta(\hat{h}_f,f) = \tfrac{1}{2}\, c(f) n^{-2(t-1)/(2t+1)} (\hat{h}-\hat{h}_f)^2 \{1 + o_p(1)\}\,.$$

It follows from this expansion that whenever $\Delta(\hat{h},f) - \Delta(\hat{h}_f,f)$ is of order n^{-1}, we also have $\hat{h}-\hat{h}_f$ of order $n^{-3/2(2t+1)}$. Furthermore the fact that n^{-1} cannot be improved upon is equivalent to the statement that $\hat{h}$ and $\hat{h}_f$ must be at least $n^{-3/2(2t+1)}$ apart in some sense. Therefore a procedure $\hat{h}$ for which $|\hat{h}-\hat{h}_f| \approx n^{-3/2(2t+1)}$, is "best possible".

It is instructive to specialise these formulae to the important case t=2, where the kernel is usually taken to be positive. There, the bandwidths $\hat{h}$ and $\hat{h}_f$ are both asymptotic to a constant multiple of $n^{-1/5}$, and (we are claiming) their distance apart is at least $n^{-3/10}$, in a minimax sense. Therefore the fastest rate of convergence of $\hat{h}$ to $\hat{h}_f$ is excruciating slow: $(\hat{h}/\hat{h}_f) - 1$ can be no smaller than order $n^{-1/10}$, in a minimax sense.

For most of this paper we discuss our results on the h-scale, not the Δ-scale, since we feel statisticians are more familiar with bandwidth than they are with integrated square error. The statistician must make an explicit choice of bandwidth, but only chooses integrated square error indirectly. Our main results will be formulated in Section 2, and proved in the ensuing two sections. Section 3 will give introductory lemmas, while Section 4 will present main proofs.

It is worth pointing out that our results (as well as their proofs) are quite different in character from traditional works on "optimal rates

of convergence" for nonparametric density estimators [5,10,12,13,18,19,22,2]. The classical argument involves showing that a certain kernel estimator (for example) is asymptotically optimal in the class of all possible density estimators; that class includes orthogonal series estimators, spline estimators, etc. But in our case we confine attention not only to kernel estimators, but to kernel estimators constructed using a specific, fixed kernel K. The only variable is the bandwidth in that special estimator. We are, in effect, switching attention from the problem of "best estimates" of a density, to that of "best estimates" of the bandwidth $\hat{h}_f$. But $\hat{h}_f$ is a random varible, and our problem of "estimating a random variable" is quite different from that of estimating a density function. Works of Rice [16] is perhaps closest in spirit to this paper and [9], although Rice did not view the minimiser of integrated square error as the benchmark bandwidth. Rice's work is for the case of nonparametric regression, and a sequel to our paper will describe analogues of our results in that context.

2. Main results.

Minimax theory is usually developed by assessing performance over a specific "test class" Θ of distributions. It is clear that if Θ' is any class containing Θ, then the worst performance over Θ' is at least as bad as the worst performance over Θ. Therefore a basic result about distributions in Θ may be generalized in many ways.

To define Θ, we begin with any compactly supported density f_0 having $t+2$ derivatives on $(-\infty,\infty)$ and (for convenience) satisfying $f_0(x) \equiv c^{(0)} > 0$ for $x \in [0,1]$. Define $c^{(1)} \equiv \sup_{x;\, j \le t+2} \frac{1}{2}|f_0^{(j)}(x)|$. Let ψ be any function on $[0,\frac{1}{2}]$ which has $t+2$ derivatives and satisfies $\sup_{0 \le x \le \frac{1}{2}} |\psi^{(j)}(x)| < \frac{1}{2}c^{(0)}$, $|\psi^{(j)}(\frac{1}{4})| > 0$ and $\psi^{(j)}(0) = \psi^{(j)}(\frac{1}{2}) = 0$ for $0 \le j \le t+2$. Set $\psi(x) = -\psi(1-x)$ for $x \in [\frac{1}{2},1]$, and extend ψ from $[0,1]$ to $(-\infty,\infty)$ by periodicity. Let m equal the integer part of $n^{1/(2t+1)}$, and define

$$\gamma(x) = \gamma(x,n) \equiv \begin{cases} m^{-t}\,\psi(mx) & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

For $v = 0,\ldots,m-1$,

let $\gamma_v(x) = \gamma(x)$ on $C_v \equiv [vm^{-1}, (v+1)m^{-1}]$, and $\gamma_v(x) = 0$ off C_v.

Let $\{\tau_v,\ 0 \le v \le m-1\}$ be any sequence of length m all of whose elements are zeros and ones, and let

$$\theta(x) = \theta(\tau_0,\ldots,\tau_{m-1})(x) \equiv f_0(x)\ \{1 + \sum_v \tau_v\ \gamma_v(x)\},\ -\infty < x < \infty.$$

The set $\Theta = \Theta(n)$ is defined to be the class of all such functions θ.

The elements of Θ are all probability densities with support equal to the support of f_0 and satisfying

$$\sup_{x;\, j \le t} |\theta^{(j)}(x)| \le c^{(1)}.$$

In particular, the t'th derivatives of densities in Θ are all uniformly bounded. The kernel K specified by (1.1) is designed for just this type of density.

We are now in a position to state our main theorem. Let K be any compactly supported kernel satisfying (1.1), and having two Hölder-continuous derivatives on $(-\infty,\infty)$. Let $\hat{h}$ be a data-driven bandwidth estimate. Any positive function of the sample $X_1,\ldots,X_n$ is a candidate for $\hat{h}$. Recall that $\hat{h}_\theta$ is the bandwidth which minimizes $\Delta(h,\theta)$.

THEOREM 2.1. Under the above conditions on K,

$$(2.1)\quad \lim_{\varepsilon\to 0} \liminf_{n\to\infty} \sup_{\theta\in\Theta} P_\theta\left(|\hat{h}-\hat{h}_\theta| > \varepsilon n^{-3/2(2t+1)}\right) = 1.$$

In this sense, no data-driven bandwidth can get closer than order $n^{-3/2(2t+1)}$ to $\hat{h}_\theta$.

Next we introduce the cross-validatory bandwidth. Let $\hat{f}_i$ denote the kernel estimate obtained by leaving out the i'th sample value:

$$\hat{f}_i(x|h) \equiv \{(n-1)h\}^{-1} \sum_{j\neq i} K\{(x-X_j)/h\}.$$

Define

$$\delta(h,\theta) \equiv 2\int \hat{f}(x|h)\theta(x)dx - 2n^{-1}\sum_{i=1}^{n} \hat{f}_i(x_i|h),$$

$$(2.2)\quad CV(h) \equiv \Delta(h,\theta) + \delta(h,\theta) - \int\theta^2$$

$$= \int \hat{f}^2(x|h)dx - 2n^{-1}\sum_{i=1}^{n} \hat{f}_i(X_i|h).$$

The cross-validatory window, $\hat{h}_c$, is that value of h which minimizes CV(h).

Our next theorem is the natural complement of Theorem 2.1, in the case where $\hat{h} = \hat{h}_c$.

THEOREM 2.2 Under the same conditions on K,

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{\theta\in\Theta} P_\theta(|\hat{h}_c-\hat{h}_\theta|>\lambda\, n^{-3/2(2t+1)}) = 0 . \tag{2.3}$$

Thus, $\hat{h}_c$ is as close to $\hat{h}_\theta$ as it is possible to get, in a minimax sense.

Result (2.3) fails to hold if Θ is replaced by the class of densities f with t uniformly bounded derivatives, or even by the class $C(B_0,\ldots,B_t)$ of all compactly-supported densitives satisfying

$$\sup_{-\infty<x<\infty} |f^{(i)}(x)| \le B_i,\ 0 \le i \le t,$$

for given constants $B_0, \ldots, B_t$. To see this, suppose Z has density $f\in C(\underset{\sim}{B})$, and let $Z_\rho \equiv \rho Z$ for $\rho \ge 1$. The density f_ρ of Z_ρ is in $C(\underset{\sim}{B})$, and since scalar expansion of the data leads to an identical expansion of both $\hat{h}_c$ and $\hat{h}_f$, we have:

$$P_{f_\rho}(|\hat{h}_c-\hat{h}_{f_\rho}| > \lambda n^{-3/2(2t+1)}) = P_f(|\hat{h}_c-\hat{h}_f| > \rho^{-1}\,\lambda n^{-3/2(2t+1)}) .$$

Consequently,

$$\sup_{f\in C(\underset{\sim}{B})} P_f(|\hat{h}_c-\hat{h}_f| > \lambda n^{-3/2(2t+1)}) = 1$$

for each $\lambda > 0$ and each $n \ge 1$. (A similar property may be observed if we work on the Δ-scale instead of the h-scale.)

There are several ways of re-defining $C(\underset{\sim}{B})$ so as to avoid this type of behaviour. For example, we might insist that densities in $C(\underset{\sim}{B})$ be above a certain level over an interval of predetermined length. However, we prefer to avoid the obscuring technicalities involved in this specificaiton by using the same test class Θ to measure both upper and lower bounds to performance.

Theorems 2.1 and 2.2 have analogues on the Δ-scale. We state them together here, without proofs. Once again, $\hat{h}$ denotes an arbitrary data-driven window.

THEOREM 2.3. Under the same conditions on K,

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{\theta \in \Theta} P_\theta\{\Delta(\hat{h},\theta) - \Delta(\hat{h}_\theta,\theta) > \varepsilon n^{-1}\} = 1,$$

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{\theta \in \Theta} P_\theta\{\Delta(\hat{h}_c,\theta) - \Delta(\hat{h}_\theta,\theta) > \lambda n^{-1}\} = 0.$$

To obtain analogues of these results for p-dimensional density estimators, modify the class Θ along the lines of Stone [20]. Theorem 2.3 continues to hold without change.

3. Preparatory lemmas.

In this and the next section, the symbols C, C_1, C_2,... denote generic positive constants. $\tilde{E}$ denotes the complement of an event E. Superscript notation in $\Delta^{(j)}$, $\delta^{(j)}$, $M^{(j)}$ and $D^{(j)}$ indicates differentiation with respect to bandwidth, h. We keep our proofs very brief, leaving out all arguments whose development closely parallels work in [9]. There are no essential differences between arguments for different values of t, and so we work only with $t=2$, to simplify notation.

Several useful, intuitively obvious technical properties of densities from Θ are summarised in our first lemma. The proofs are tedious but straightforward, and so we give only an outline.

LEMMA 3.1 Take $t=2$ in all that follows. Then: for some $n_0 > 0$,

$$(3.1) \qquad 0 < \inf_{n \geq n_0, \theta \in \Theta} n^{1/5} h_\theta \leq \sup_{n \geq n_0, \theta \in \Theta} n^{1/5} h_\theta < \infty;$$

for any $\varepsilon > 0$ there exists $\eta = \eta(\varepsilon) > 0$ such that

$$(3.2) \qquad \inf_{|h-h_\theta| > \varepsilon n^{-1/5}} M(h,\theta) \geq (1+\eta) M(h_\theta, \theta)$$

for all $\theta \in \Theta$ and all large n; for some $n_0 > 0$,

$$(3.3) \qquad 0 < \inf_{n \geq n_0, \theta \in \Theta} n^{2/5} M^{(2)}(h_\theta, \theta) \leq \sup_{n \geq n_0, \theta \in \Theta} n^{2/5} M^{(2)}(h_\theta, \theta) < \infty ;$$

for any $\varepsilon > 0$ there exists $\eta = \eta(\varepsilon) > 0$ such that

$$(3.4) \qquad \sup_{|h-h_\theta| \leq \varepsilon n^{-1/5}} |M^{(2)}(h,\theta) - M^{(2)}(h_\theta, \theta)| \leq \eta(\varepsilon) n^{-2/5}$$

for all $\theta \in \Theta$ and all large n, and $\eta(\varepsilon) \to 0$ as $\varepsilon \to 0$.

OUTLINE OF PROOF: Write

$$M(h,\theta) = V(h,\theta) + B(h,\theta),$$

where

$$V(h,\theta) = n^{-1}h^{-1} \iint K(u)^2\, \theta(x-hu)\,du\,dx - n^{-1} \int [\int K(u)\theta(x-hu)\,du]^2 dx,$$

$$B(h,\theta) = \int [\int K(u)\{\theta(x-hu) - \theta(x)\}du]^2 dx .$$

The derivatives $M^{(1)}(h,\theta)$ and $M^{(2)}(h,\theta)$ may be studied by differentiating $V(h,\theta)$, then approximating as in Rosenblatt [17], and by differentiating $B(h,\theta)$ and using a Taylor expansion with integral form of the remainder. □

Proofs of Lemmas 3.2 and 3.3 below closely parallel those of Lemmas 3.1 and 3.2 in [9]. In establishing (3.10), note (3.1).

LEMMA 3.2. For each $0 < a < b < \infty$ and all positive integers ℓ,

$$(3.5) \qquad \sup_{n,\theta\in\Theta,a\le t\le b} E_\theta |n^{7/10} D^{(1)}(n^{-1/5}t,\theta)|^{2\ell} \le C_1(a,b,\ell),$$

$$(3.6) \qquad \sup_{n,\theta\in\Theta,a\le t\le b} E_\theta |n^{7/10} \delta^{(1)}(n^{-1/5}t,\theta)|^{2\ell} \le C_1(a,b,\ell) .$$

Furthermore, there exists $\varepsilon_1 > 0$, not depending on a,b or ℓ, such that

$$(3.7) \qquad E_\theta |n^{7/10}\{D^{(1)}(n^{-1/5}s,\theta) - D^{(1)}(n^{-1/5}t,\theta)\}|^{2\ell} \le C_2(a,b,\ell)|s-t|^{\varepsilon_1 \ell},$$

$$(3.8) \qquad E_\theta |n^{7/10}\{\delta^{(1)}(n^{-1/5}s,\theta) - \delta^{(1)}(n^{-1/5}t,\theta)\}|^{2\ell} \le C_2(a,b,\ell)|s-t|^{\varepsilon_1 \ell}$$

for all $\theta \in \Theta$ and $a \le s \le t \le b$.

LEMMA 3.3. For some $\varepsilon > 0$ and any $0 < a < b < \infty$,

$$(3.9) \qquad \sup_{\theta\in\Theta} P_\theta[\sup_{a\le t\le b} \{|D^{(1)}(n^{-1/5}t,\theta)| + |\delta^{(1)}(n^{-1/5}t,\theta)|\} > n^{-3/5-\varepsilon}] \to 0 .$$

Furthermore, for any $\varepsilon_2 > 0$ and $\eta > 0$,

$$(3.10)\quad \sup_{\theta\in\Theta} P_\theta\Big[\sup_{|t-n^{1/5}h_\theta|\le n^{-\varepsilon_2}} n^{7/10}\{|D^{(1)}(n^{-1/5}t,\theta) - D^{(1)}(h_\theta,\theta)| + |\delta^{(1)}(n^{-1/5}t,\theta) - \delta^{(1)}(h_\theta,\theta)|\} > \eta\Big] \to 0 .$$

LEMMA 3.4. *For any* $\varepsilon > 0$,

$$\sup_{\theta\in\Theta} P_\theta(|\hat{h}_\theta - h_\theta| > \varepsilon n^{-1/5}) \to 0 .$$

PROOF. It suffices to show that for any sequence of choices $\theta_1 = \theta_{1n} \in \Theta$, and for each $\varepsilon > 0$,

$$(3.11)\quad P_{\theta_1}(|\hat{h}_{\theta_1} - h_{\theta_1}| > \varepsilon n^{-1/5}) \to 0 .$$

We may easily prove that for some $b > 0$, $P_{\theta_1}(n^{-b} \le \hat{h}_{\theta_1} \le n^b) \to 1$. Let $H = H_n$ be a set of bandwidths in the range $[n^{-b}, n^b]$, and such that $\#(H) \le n^a$ for some $a > 0$. Arguing as in the proofs of Lemmas 2 and 4 of Stone [21] we may show that for each $\varepsilon > 0$,

$$(3.12)\quad P_{\theta_1}\{\sup_{h\in H} |\Delta(h,\theta_1) - M(h,\theta_1)|/M(h,\theta_1) > \varepsilon\} \to 0 .$$

Now use Hölder continuity of K to show that for any (random) bandwidth $\tilde{h}$ with $P_{\theta_1}(n^{-b} \le \tilde{h} \le n^b) \to 1$,

$$P_{\theta_1}\{|\Delta(\tilde{h},\theta_1) - M(\tilde{h},\theta_1)|/M(\tilde{h},\theta_1) > \varepsilon\} \to 0 .$$

Finally invoke (3.2), to obtain (3.11). ||

Recall that $\hat{h}_c$ is the cross-validatory window, chosen to minimise the function CV(h) at (2.2).

LEMMA 3.5. *For any* $\varepsilon > 0$,

$$\sup_{\theta\in\Theta} P_\theta(|\hat{h}_c - h_\theta| > \varepsilon n^{-1/5}) \to 0 .$$

PROOF. Again, it suffices to prove that for any $\varepsilon > 0$ and sequence $\theta_1 = \theta_{1n} \in \Theta$,

$$(3.13) \quad P_{\theta_1}(|\hat{h}_c - h_{\theta_1}| > \varepsilon n^{-1/5}) \to 0,$$

and it is easily shown that for some $b > 0$, $P_{\theta_1}(n^{-b} \le \hat{h}_c \le n^b) \to 1$. Define

$$CV(h,\theta) \equiv CV(h) + \int \theta^2 + 2n^{-1}(n-1)^{-1}(n+1) \sum_{i=1}^{n} \{\theta(X_i) - E_\theta \theta(X_i)\},$$

and let H be as in the proof of lemma 3.4. Minimising CV is equivalent to minimising $CV(\cdot,\theta)$, for any θ. Using the argument leading to Stone's [21] Lemmas 2,3 and 4, we may show that for any $\varepsilon > 0$,

$$P_{\theta_1}\{\sup_{h \in H} |CV(h,\theta_1) - M(h,\theta_1)|/M(h,\theta_1) > \varepsilon\} \to 0.$$

This formula serves as an analogue of (3.12) in the proof of Lemma 3.4. The proof of (3.13) may now be completed as was that proof. □

LEMMA 3.6 For some $\varepsilon > 0$,

$$\sup_{\theta \in \Theta} P_\theta(|\hat{h}_\theta - h_\theta| + |\hat{h}_c - h_\theta| > n^{-1/5-\varepsilon}) \to 0.$$

PROOF. Argue as in Lemma 3.3 of [9], but use Lemmas 3.4 and 3.5 above to replace the limit theorems $\hat{h}_0/h_0 \xrightarrow{p} 1$ and $\hat{h}_c/h_0 \xrightarrow{p} 1$ (in notation of [9]), and use our Lemma 3.3 in place of Lemma 3.2 of [9]. □

LEMMA 3.7

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{\theta \in \Theta} P_\theta(|\hat{h}_\theta - h_\theta| > \lambda n^{-3/10}) = 0.$$

PROOF. It suffices to show that for any sequences $\theta_1 = \theta_{1n} \in \Theta$ and $\lambda_n \uparrow \infty$,

$$(3.14) \quad P_{\theta_1}(|\hat{h}_{\theta_1} - h_{\theta_1}| > \lambda_n n^{-3/10}) \to 0.$$

Observe that

$$(3.15) \quad 0 = \Delta^{(1)}(\hat{h}_{\theta_1}, \theta_1) = M^{(1)}(\hat{h}_{\theta_1}, \theta_1) + D^{(1)}(\hat{h}_{\theta_1}, \theta_1) = (\hat{h}_{\theta_1} - h_{\theta_1}) M^{(2)}(h^*, \theta_1) + D^{(1)}(\hat{h}_{\theta_1}, \theta_1),$$

where h* lies inbetween $\hat{h}_{\theta_1}$ and h_{θ_1}. Define $c_1 = c_1(n)$ and $c_2 = c_2(n)$ by $h_{0_1} = c_1 n^{-1/5}$ and $M^{(2)}(h_{0_1}, \theta_1) = c_2 n^{-2/5}$. Then c_1 and c_2 are bounded away from zero and infinity as $n \to \infty$ (note (3.1) and (3.3) from Lemma 3.1). Given any $\xi > 0$, there exists $\eta(\xi) > 0$ such that $\eta(\xi) \to 0$ as $\xi \to 0$ and for large n,

$$\sup_{|h-h_{\theta_1}| \leq \xi n^{-1/5}} |M^{(2)}(h,\theta_1) - M^{(2)}(h_{\theta_1},\theta_1)| \leq \eta(\xi) n^{-2/5}.$$

(Note (3.4) of Lemma 3.1.) Let a_1, b_1 be fixed positive lower, upper bounds to c_1, respectively, and let a_2 be a fixed positive lower bound to c_2. Choose $\xi \in (0, \frac{1}{2}a_1)$ so small that $\eta(\xi) \leq \frac{1}{2}a_2$.

By (3.15),

$$|\hat{h}_{\theta_1} - h_{\theta_1}| \leq (\tfrac{1}{2}a_2 n^{-2/5})^{-1} |D^{(1)}(\hat{h}_{\theta_1}, \theta_1)|$$

$$\leq 2a_2^{-1} n^{2/5} \sup_{\frac{1}{2}a_1 \leq t \leq \frac{1}{2}a_1 + b_1} |D^{(1)}(n^{-1/5}t, \theta_1)|$$

whenever the event $E_1 \equiv \{|\hat{h}_{\theta_1} - h_{\theta_1}| \leq \xi n^{-1/5}\}$ holds. Let E_2 be the event $\{\sup_{a \leq t \leq b} |D^{(1)}(n^{-1/5}t, \theta_1)| \leq n^{-3/5-\varepsilon}\}$, where $a = \frac{1}{2}a_1$, $b = \frac{1}{2}a_1 + b_1$, and ε is as in (3.9) of Lemma 3.3. Whenever $E_1 \cap E_2$ holds, so does the event $E_3 \equiv \{|\hat{h}_{\theta_1} - h_{\theta_1}| \leq 2a_2^{-1} n^{-1/5-\varepsilon}\}$. Let E_4 be the event that

$$|D^{(1)}(\hat{h}_{\theta_1}, \theta_1) - D^{(1)}(h_{\theta_1}, \theta_1)| > n^{-7/10}.$$

Then:

$$(3.16) \quad P_{\theta_1}(|\hat{h}_{\theta_1} - h_{\theta_1}| > \lambda_n n^{-3/10}) \leq P_{\theta_1}(\tilde{E}_1) + P_{\theta_2}(\tilde{E}_2) + P_{\theta_1}(E_3 \cap E_4)$$

$$+ P_{\theta_1}\{|D^{(1)}(h_{\theta_1}, \theta_1)| > \lambda_n n^{-3/10}(2a_2^{-1}n^{2/5})^{-1} - n^{-7/10}\}.$$

Chebychev's inequality and (3.5) of Lemma 3.2 show that the last-written probability converges to zero as $n \to \infty$. Lemma 3.4 gives $P_{\theta_1}(\tilde{E}_1) \to 0$, (3.9) of Lemma 3.3 gives $P_{\theta_1}(\tilde{E}_2) \to 0$, and (3.10) of Lemma 3.3 gives $P_{\theta_1}(E_3 \cap E_4) \to 0$. Result (3.14) now follows from (3.16). □

LEMMA 3.8

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{\theta \in \Theta} P_\theta(|\hat{h}_c - h_\theta| > \lambda n^{-3/10}) = 0 .$$

PROOF. Use essentially the argument employed to prove Lemma 3.7, but replace (3.15) by

$$0 = CV^{(1)}(\hat{h}_c) = M^{(1)}(\hat{h}_c, \theta_1) + D^{(1)}(\hat{h}_c, \theta_1) + \delta^{(1)}(\hat{h}_c, \theta_1)$$

$$= (\hat{h}_c - h_{\theta_1}) M^{(2)}(h^*, \theta_1) + D^{(1)}(\hat{h}_c, \theta_1) + \delta^{(1)}(\hat{h}_c, \theta_1),$$

where h^* lies inbetween $\hat{h}_c$ and h_{θ_1}. □

We pause to introduce further notation. Let π be a kernel function, and let

$$\hat{p}(x|h) \equiv (nh)^{-1} \sum_{i=1}^{n} \pi\{(x - X_i)/h\}$$

be the corresponding density estimator. Set

$$s_v(h, \theta) \equiv \int_{C_v} \{\hat{p}(x|h) - \theta(x)\}\gamma(x)dx,$$

where γ is as in Section 2.

LEMMA 3.9. Assume π is Hölder continuous, vanishes outside a compact interval, and satisfies $\int \pi(x)dx = 1$, $\int x\,\pi(x)dx = 0$. Then for each $0 < a < b < \infty$ and each $\varepsilon > 0$,

$$\sup_{\theta} P_\theta\{\sup_{v;\, a \le t \le b} |s_v(n^{-1/5}t, \theta)| > n^{-1+\varepsilon}\} \to 0 .$$

PROOF. Using Hölder continuity of π and the fact that π vanishes outside a compact interval, we may choose $\lambda > 0$ so large that

$$|s_v(n^{-1/5}s,\theta) - s_v(n^{-1/5}t,\theta)| \le C_1 n^{-1}$$

uniformly in $n \ge 1$, $\theta\epsilon\Theta$, v, $a \le s \le t \le b$ with $|s-t| \le n^{-\lambda}$, and samples $X_1,\ldots,X_n$. Partition (a,b) in the manner $a = t_0 < t_1 < \ldots < t_\nu \le b < t_\nu$, where each $t_i - t_{i+1} = n^{-\lambda}$. It suffices to show that for each $\epsilon > 0$,

$$(3.17) \quad \sup_\theta P_\theta\{\sup_{v,i}|s_v(n^{-1/5}t_i,\theta)| > n^{-1+\epsilon}\} \to 0.$$

Let $\ell \ge 1$ be an integer, let $\|C_v\|$ denote the length of C_v, and notice that

$$|s_v(h,\theta)|^{2\ell} \le \|C_v\|^{2\ell-1} \int_{C_v} |\{\hat{p}(x|h) - \theta(x)\}\gamma(x)|^{2\ell}dx$$

$$\le C_2 n^{-(6\ell-1)/5} \int_{C_v} \{\hat{p}(x|h) - \theta(x)\}^{2\ell} dx$$

uniformly in θ, h and v. Therefore the left-hand side of (3.17) is dominated by

$$(3.18) \quad \sum_v \sum_i \sup_\theta P_\theta\{|s_v(n^{-1/5}t_i,\theta)| > n^{-1+\epsilon}\}$$

$$\le \sum_v \sum_i \sup_\theta E_\theta\{|n^{1-\epsilon}s_v(n^{-1/5}t_i,\theta)|^{2\ell}\}$$

$$\le C_2 n^{(4\ell+1)/5-2\epsilon\ell} \sum_i \sup_\theta \int_{-\infty}^{\infty} E_\theta\{\hat{p}(x|n^{-1/5}t_i) - \theta(x)\}^{2\ell}dx$$

$$\le C_3 n^{(4\ell+1)/5-2\epsilon\ell} \sum_i n^{-4\ell/5}$$

The last inequality uses the following facts:

$$\{\hat{p}(x|n^{-1/5}t_i) - \theta(x)\}^{2\ell} \le C_4\{\hat{p}(x|n^{-1/5}t_i) - E_\theta\hat{p}(x|n^{-1/5}t_i)\}^{2\ell}$$

$$+ C_4\{E_\theta\hat{p}(x|n^{-1/5}t_i) - \theta(x)\}^{2\ell},$$

$$\{E_\theta\hat{p}(x|n^{-1/5}t_i) - \theta(x)\}^{2\ell} \le C_5 n^{-4\ell/5},$$

$$E_\theta\{\hat{p}(x|n^{-1/5}t_i) - E_\theta\hat{p}(x|n^{-1/5}t_i)\}^{2\ell} \le C_6 n^{-4\ell/5},$$

all uniformly in x, i and θ, the latter by an inequality for centered sums of independent random variables (formula (21.4) of Burkholder [3]); and the integrand in (3.18) vanishes outside a compact set, independent of i and θ. The number of summands in the sum over i in (3.18) is of order n^{λ}, and so if we choose ℓ so large that $\lambda + (1/5) - 2\varepsilon\ell < 0$, the right-hand side of (3.18) converges to zero as $n \to \infty$. This proves (3.17). □

LEMMA 3.10. For each $0 < a < b < \infty$ and all positive integers ℓ,

$$\sup_{n,\theta\in\Theta,a\le t\le b} E_\theta |n^{1/2} D^{(2)}(n^{-1/5}t_i,\theta)|^{2\ell} \le C(a,b,\ell).$$

One consequence of this result and (3.1) of Lemma 3.1 is that for each $\varepsilon > 0$,

$$(3.19) \quad \sup_{\theta\in\Theta} P_\theta\{n^{2/5}|D^{(2)}(h_\theta,\theta)| > \varepsilon\} \to 0 .$$

PROOF OF LEMMA 3.10. Use the argument employed to derive (3.5) of Lemma 3.2. □

4. Main proofs.

Theorem 2.2 is immediate from Lemmas 3.7 and 3.8. (Remember that we are taking t=2 throughout, to simplify notation.) The remainder of this paper is devoted to proving Theorem 2.1. The classification argument used by Stone [20] and Marron [11] is an important element of our proof.

Given a data-driven bandwidth $\hat{h}$, define $\hat{\theta}$ to be any element of Θ such that $|\hat{h}_{\hat{\theta}} - \hat{h}| = \inf_{\theta\in\Theta} |\hat{h}_\theta - \hat{h}|$. Then $\hat{h}_{\hat{\theta}}$ is also a data-driven bandwidth - that is, it is a function of n and $X_1,\ldots,X_n$ alone; it does not employ any additional knowledge about the unknown density. For each $\theta_1\in\Theta$,

$$|\hat{h}_{\hat{\theta}} - \hat{h}_{\theta_1}| \le |\hat{h}_{\hat{\theta}} - \hat{h}| + |\hat{h} - \hat{h}_{\theta_1}| \le 2|\hat{h} - \hat{h}_{\theta_1}| .$$

Therefore result (2.1) will follow if we prove it for $\hat{h}_{\hat{\theta}}$ insead of $\hat{h}$:

$$(4.1) \qquad \lim_{\varepsilon \to 0} \lim_{n \to \infty} \inf_{\theta \in \Theta} \sup P_\theta(|\hat{h}_{\hat{\theta}} - \hat{h}_\theta| > \varepsilon n^{-3/10}) = 1 .$$

Choose $0 < a_1 < b_1 < \infty$ such that $2a_1 \le n^{1/5} h_\theta \le \frac{1}{2} b_1$ for all n and all $\theta \in \Theta$. (Note (3.1) of Lemma 3.1.) We keep a_1, b_1 fixed throughout this section. Define $\hat{h}'_{\hat{\theta}} = \hat{h}_{\hat{\theta}}$ if $a_1 n^{-1/5} < \hat{h}_{\hat{\theta}} < b_1 n^{-1/5}$, and $\hat{h}'_{\hat{\theta}} = \frac{1}{2}(a_1 + b_1)\, n^{-1/5}$ otherwise. Set $L(z) \equiv -zK'(z)$, and observe that $\int L(z)dz = 1$ and $\int zL(z)dz = 0$. (In the case of general t, if K satisfies (1.1) then so does L, although with d_K replaced by $(t+1)d_K$.) Let

$$\hat{g}(x|h) \equiv (nh)^{-1} \sum_{i=1}^{n} L\{(x - X_i)/h\}$$

be the density estimate constructed using kernel L instead of K. Define

$$\hat{\xi}(\theta) \equiv \int \{\hat{f}(x|\hat{h}'_{\hat{\theta}}) - \hat{g}(x|\hat{h}'_{\hat{\theta}})\}\{\hat{\theta}(x) - \theta(x)\}dx,$$

for $\theta \in \Theta$. The first step in establishing Theorem 2.1 is to prove:

PROPOSITION 4.1. Given $\eta_1 > 0$, we may choose $\eta_2 > 0$ and a sequence $\theta_1 = \theta_{1n} \in \Theta$ such that, for all large n,

$$P_{\theta_1}\{|\hat{\xi}(\theta_1)| > \eta_2\, n^{-9/10}\} > 1 - \eta_1 .$$

The proof is via a sequence of three lemmas. Let P_0 be the probability measure defined by

$$P_0(E) \equiv 2^{-m} \sum_{\theta \in \Theta} P_\theta(E) ,$$

and let E_0 denote expectation with respect to P_0. Under P_0, θ should be regarded as a random variable. There are precisely 2^m elements in Θ.

Writing $\theta = (1+ \sum_v \tau_v \gamma_v)f_0$ and $\hat{\theta} = (1 + \sum_v \hat{\tau}_v \gamma_v)f_0$ for sequences $\{\tau_v\}$ and $\{\hat{\tau}_v\}$ of 0's and 1's, we see that

$$\hat{\xi}(\theta) = - c_0 S ,$$

where c_0 is the constant value taken by f_0 on $\cup_v C_v$,

$$S \equiv \sum_v (\tau_v - \hat{\tau}_v)\hat{w}_v ,$$

and

$$\hat{w}_v \equiv \int_{C_v} \{\hat{f}(x|\hat{h}'_{\hat{\theta}}) - \hat{g}(x|\hat{h}'_{\hat{\theta}})\}\gamma(x)dx .$$

(The function γ was defined in Section 2.) Notice that S depends on θ only through the indicators τ_v; this observation is crucial to our argument.

Let X denote the sample $X_1,\ldots,X_n$. Under the probability measure P_0, and conditional on X, the τ_v's are independent Bernoulli random variables with

$$(4.2) \qquad \hat{q}_v \equiv P_0(\tau_v = 1|X) = [\Pi^{(v)}\{1 + \gamma(X_i)\}][1 + \Pi^{(v)}\{1 + \gamma(X_i)\}]^{-1} ,$$

where $\Pi^{(v)}$ denotes the product over indices i with $X_i \in C_v$. Thus,

$$\hat{\mu} \equiv E_0(S|X) = \sum_v (\hat{q}_v - \hat{\tau}_v)\hat{w}_v ,$$

$$\hat{\sigma}_v^2 \equiv \mathrm{var}_0(S|X) = \sum_v \hat{q}_v(1-\hat{q}_v)\hat{w}_v^2 ,$$

$$\hat{\beta} \equiv \sum_v E_0\{|(\tau_v - \hat{q}_v)\hat{w}_v|^3|X\} \leq \sum_v |\hat{w}_v|^3 .$$

The next two lemmas describe asymptotic properties of $\hat{\sigma}^2$ and $\hat{\beta}$.

LEMMA 4.1. <u>There exist fixed constants</u> $0 < d_1 < d_2 < \infty$ <u>such that</u>

$$P_0(d_1 n^{-9/5} < \hat{\sigma}^2 < d_2 n^{-9/5}) \to 1 .$$

PROOF. Let N_v denote the number of elements of X within C_v, and notice that the P_θ-distribution of the sequence $\{N_v\}$ does not depend on θ. Observe that for a constant $c > 0$, $E_\theta(N_v) = E_\theta(N_1) \sim cn^{4/5}$. Therefore for large n,

$$\begin{aligned} P_\theta(N_v > 3\, c_n^{4/5} \text{ for some } v) &\le C\, n^{1/5}\, P_\theta(N_1 > 3\, cn^{4/5}) \\ &\le C\, n^{1/5}\, P_\theta\{|N_1 - E_\theta(N_1)| > c\, n^{4/5}\} \\ &\le C\, n^{1/5}(c\, n^{4/5})^{-2}\, E\{|N_1 - E_\theta(N_1)|^2\} \\ &= O(n^{-3/5}) . \end{aligned}$$

Thus, if E is the event that no interval C_v contains more than $3\,c\,n^{4/5}$ elements of X, then $\inf_{\theta\in\Theta} P_\theta(E) \to 1$.

Let $\Sigma^{(v)}$ denote summation over indices i with $X_i \in C_v$, and observe that

$$\begin{aligned} \Pi^{(v)}\{1 + \gamma(X_i)\} &= \exp\Big[\sum_{j=1}^{\infty} (-1)^{j+1} j^{-1}\, \Sigma^{(v)}\{\gamma(X_i)\}^j\Big] \\ &= \exp(T_1^{(v)} + T_2^{(v)}) , \end{aligned}$$

where

$$T_1^{(v)} \equiv \Sigma^{(v)}\gamma(X_i), \quad |T_2^{(v)}| \le \sum_{j=2}^{\infty} j^{-1}\, \Sigma^{(v)}|\gamma(X_i)|^j \equiv T_3^{(v)} .$$

Bearing in mind that $\sup|\gamma| \le C_1\, n^{-2/5}$, we may easily prove that on the set E and for all large n, $T_3^{(v)} \le C_2$ uniformly in v.

Thus, for each $z > 0$ there exist numbers $0 < a_2(z) \le b_2(z) < \infty$ such that, on the set $\{|T_1^{(v)}| \le z\} \cap E$,

$$a_2(z) \le \Pi^{(v)}\{1 + \gamma(X_i)\} \le b_2(z)$$

for all v. Remembering the definition (4.2) of $\hat{q}_v$ in terms of $\Pi^{(v)}\{1+\gamma(X_i)\}$,

we now deduce the existence of a positive, decreasing function $a(z) \leq \frac{1}{2}$, such that on E, $|T_1^{(v)}| \leq z$ implies $|\hat{q}_v - \frac{1}{2}| \leq \frac{1}{2} - a(z)$. Therefore on E,

$$(4.3) \qquad \tfrac{1}{2}a(z) \sum_v \hat{w}_v^2 \, I\{|\Sigma^{(v)} \gamma(X_i)| \leq z\} \leq \hat{\sigma}_v^2 \leq \tfrac{1}{4} \sum_v \hat{w}_v^2 ,$$

for all $z > 0$.

Let

$$\hat{w}_v(h,z) \equiv I\{|\Sigma^{(v)} \gamma(X_i)| \leq z\} \int_{C_v} \{\hat{f}(x|h) - \hat{g}(x|h)\} \gamma(x)dx ,$$

$\mu_v(h,z,\theta) \equiv E_\theta\{\hat{w}_v^2(h,z)\}$ and $\mu(h,z,\theta) \equiv \Sigma_v \mu_v(h,z,\theta)$. We claim that the function $c(n,h,z,\theta)$ defined by $\mu(h,z,\theta) = c(n,h,z,\theta)n^{-9/5}$, is bounded away from zero and infinity uniformly in $n \geq 1$, $h \in n^{-1/5}(a_1,b_1)$, $z \geq z_0$ and $\theta \in \Theta$, for some $z_0 > 0$. This is relatively easy to verify if we take $z_0 = \infty$. To see that $z_0 < \infty$ is permissible, notice that

$$\mu_u(h,z,\theta) \geq$$

$$\geq \mu_u(h,\infty,\theta) - [P_\theta\{|\Sigma^{(u)} \gamma(X_i)| > z\}]^{\frac{1}{2}} [E_\theta\{\hat{w}_v^4(h,\infty)\}]^{\frac{1}{2}} ;$$

$E_\theta\{\hat{w}_v^4(h,\infty)\} \leq C_1 n^{-4}$ uniformly in $h \in n^{-1/5}(a_1,b_1)$ and $\theta \in \Theta$; and

$$P_\theta\{|\Sigma^{(v)} \gamma(X_i)| > z\}$$

$$\leq z^{-2} E_\theta\{|\Sigma^{(v)} \gamma(X_i)|^2\}$$

$$= z^{-2} E_\theta[N_v E\{\gamma^2(X_1)|X_1 \in C_v\} + (N_v^2 - N_v)E\{\gamma(X_1)\gamma(X_2)|X_1, X_2 \in C_v\}]$$

$$\leq z^{-2} C_2[E_\theta(N_v) \|C_v\|^{-1} \int_{C_v} \gamma^2(x)dx + E_\theta(N_v^2) \{\|C_v\|^{-1} \int_{C_v} \tau_v \gamma^2(x)dx\}^2]$$

$$\leq z^{-2} C_3 ,$$

uniformly in v and θ. (Remember $\|C_v\|$ denotes the length of C_v.)

Consequently,

$$\mu_\nu(h,z,\theta) \geq \mu_\nu(h,\infty,\theta) - (C_1C_3)^{\frac{1}{2}}\, n^{-2}\, z^{-1}$$

uniformly in ν,h,z and θ. Adding this inequality over ν, we see that the stated properties of the function c are available for some <u>finite</u> $z_0 > 0$.

Take $z = z_0$ in (4.3). In view of the properties of c established in the previous paragraph, Lemma 4.1 will follow via (4.3), if we prove that for each $\varepsilon > 0$, and for $z = z_0$ and $z = \infty$,

$$(4.4) \qquad \sup_{\theta\in\Theta} P_\theta\Big[\sup_{a_1\leq t\leq b_1} |\Sigma_\nu\{\hat{w}_\nu{}^2(n^{-1/5}t,z) - \mu_\nu(n^{-1/5}t,z,\theta)\}| > \varepsilon n^{-9/5}\Big] \to 0.$$

Using Hölder continuity of K and L, and the fact that these functions have compact support, we may choose $\lambda > 0$ so large that

$$(4.5) \qquad \Sigma_\nu\{|\hat{w}_\nu{}^2(n^{-1/5}s,z) - \hat{w}_\nu{}^2(n^{-1/5}t,z)| + |\mu_\nu(n^{-1/5}s,z,\theta) - \mu_\nu(n^{-1/5}t,z,\theta)|\}$$
$$\leq C\, n^{-2}$$

uniformly in n, $z = z_0$ and ∞, θ, ν, $a_1 \leq s \leq t \leq b_1$ with $|s-t| \leq n^{-\lambda}$, and samples $X_1,\ldots,X_n$. Let $a_1 = t_0 < t_1 < \ldots < t_{\nu-1} \leq b_1 < t_\nu$ be a partition of (a_1,b_1) with $t_i - t_{i-1} = n^{-\lambda}$ for each i. In view of (4.5), result (4.4) will follow if we show that for each $\varepsilon > 0$,

$$p_\theta \equiv \sum_i P_\theta[|\Sigma_\nu\{\hat{w}_\nu{}^2(n^{-1/5}t_i,z) - \mu_\nu(n^{-1/5}t_i,z,\theta)\}| > \varepsilon n^{-9/5}]$$

converges to zero uniformly in $\theta \in \Theta$ and $z = z_0$ and ∞.

Since K and L have compact support, and each $t_i \in (a_1,b_1)$, then for each i we may divide the subscripts ν among a fixed finite number k not depending on i or n) of sets $V_{i1},\ldots,V_{ik}$ such that for each i and j, and for $z = z_0$ and ∞, the variables $\hat{w}_\nu{}^2(n^{-1/5}t_i,z)$, $\nu \in V_{ij}$, are stochastically independent, and

for each i, each subscript v is contained in just one set V_{ij}. Consequently, for all integers $\ell \geq 1$,

$$p_\theta \leq \sum_i \sum_{j=1}^{k} P_\theta[\,|\sum_{v\in V_{ij}} \{\hat{w}_v{}^2(n^{-1/5}t_i,z) - \mu_v(n^{-1/5}t_i,z,\theta)\}| > \varepsilon\, k^{-1}n^{-9/5}]$$

$$\leq \sum_i \sum_{j=1}^{k} E_\theta[\,|(\varepsilon k^{-1}n^{-9/5})^{-1} \sum_{v\in V_{ij}} \{\hat{w}_v{}^2(n^{-1/5}t_i,z) - \mu_v(n^{-1/5}t_i,z,\theta)\}|^{2\ell}].$$

An inequality for moments of sums of independent random variables [3, formula (21.4)] now gives

$$p_\theta \leq C_1(\ell)\ (\varepsilon^{-1}k)^{2\ell}\ n^{18\ell/5} \sum_i \sum_{j=1}^{k} [\{\sum_{v\in V_{ij}} E_\theta(Y_{iv}^2)\}^\ell + \sum_{v\in V_{ij}} E_\theta(|Y_{iv}|^{2\ell})],$$

where $Y_{iv} \equiv \hat{w}_v{}^2(n^{-1/5}t_i,z) - \mu_v(n^{-1/5}t_i,z,\theta)$. The same moment inequality gives $E_\theta(|Y_{iv}|^{2\ell}) \leq C_2\, n^{-4\ell}$ uniformly in i, v and θ. Since the number of partition points t_i is of order n^λ, then

$$\sup_\theta p_\theta \leq C_3\, n^{18\ell/5} \sum_i \{(n^{1/5}n^{-4})^\ell + n^{1/5}\, n^{-4\ell}\} = O(n^{\lambda-\ell/5}) \to 0,$$

provided only that $\ell > 5\lambda$. □

LEMMA 4.2. For each $\varepsilon > 0$,

$$P_0(\hat{\beta} > n^{-14/5+\varepsilon}) \to 0.$$

PROOF. The argument used to prove Lemma 4.1 shows that for some $c_3 > 0$,

$$P_0(\sum_v \hat{w}_v{}^2 > c_3\, n^{-9/5}) \to 0.$$

Applying Lemma 3.9 twice, once with $\hat{p} \equiv \hat{f}$ and once with $\hat{p} \equiv \hat{g}$, we obtain:

$$P_0(\sup_v |\hat{w}_v| > n^{-1+\varepsilon}) \to 0.$$

Lemma 4.2 follows on combining these results. □

Let Φ be the standard normal distribution function.

LEMMA 4.3. For some fixed $c > 0$,

$$\liminf_{n \to \infty} \inf_{x>0} [P_0(|S| > n^{-9/10}x) - 2\{1 - \Phi(cx)\}] \geq 0 .$$

PROOF. Let E denote the event that $\hat{\beta}\,\hat{\sigma}^{-3} \leq n^{-1/20}$. According to Lemmas 4.1 and 4.2,

$$P_0(\tilde{E}) \leq P_0(\hat{\sigma}^2 \leq d_1 n^{-9/5}) + P_0\{\hat{\beta} > n^{-1/20}(d_1 n^{-9/5})^{3/2}\} \to 0 .$$

On the set E, the Berry-Esseen bound [15, page 111] gives

$$\sup_{-\infty<x<\infty} |P_0(S \leq x|X) - \Phi\{(x-\hat{\mu})/\hat{\sigma}\}| \leq A\, n^{-1/20} ,$$

where A is an absolute constant. Therefore on E, and for $x > 0$,

$$P_0(|S| > x|\, X) \geq 1 - \Phi\{(x-\hat{\mu})/\hat{\sigma}\} + \Phi\{(-x-\hat{\mu})/\hat{\sigma}\} - 2A\, n^{-1/20}$$

$$\geq 2\{1 - \Phi(x/\hat{\sigma})\} - 2A\, n^{-1/20} .$$

Taking expectations, and using Lemma 4.1 again, we obtain Lemma 4.3. □

To obtain Proposition 4.1 from Lemma 4.3, choose $x > 0$ so small that $2\{1 - \Phi(cx)\} > 1 - \frac{1}{2}\eta_1$, and let $\eta_2 = c_0^{-1}x$. Then for large n,

$$1 - \eta_1 \leq P_0(|c_0 S| > \eta_2 n^{-9/10})$$

$$= 2^{-m} \sum_{\theta \in \Theta} P_\theta\{|\hat{\xi}(\theta)| > \eta_2 n^{-9/10}\} .$$

Therefore there must exist some $\theta_1 \in \Theta$ such that

$$1 - \eta_1 \leq P_{\theta_1}\{|\hat{\xi}(\theta_1)| > \eta_2 n^{-9/10}\} .$$

Throughout the remainder of our proof of (4.1), we work with the "worst case" density $\theta_1 = \theta_{1n}$ specified by Proposition 4.1, for some fixed $\eta_1, \eta_2 > 0$.

Notice that

$$(\partial/\partial h)\ \hat{f}(x|h) = h^{-1}\{\hat{g}(x|h) - \hat{f}(x|h)\}$$

and

$$\Delta(h,\hat{\theta}) = \Delta(h,\theta_1) - 2\int \{\hat{f}(x|h) - \theta_1(x)\}\{\hat{\theta}(x) - \theta_1(x)\}dx$$

$$+ \int \{\hat{\theta}(x) - \theta_1(x)\}^2 dx\ .$$

Differentiate the latter formula with respect to h, and take $\hat{h} = \hat{h}_{\hat{\theta}}$, obtaining:

$$0 = \Delta^{(1)}(\hat{h}_{\hat{\theta}},\theta_1) + 2\ \hat{h}_{\hat{\theta}}^{-1} \int \{\hat{f}(x|\hat{h}_{\hat{\theta}}) - \hat{g}(x|\hat{h}_{\hat{\theta}})\}\quad \{\hat{\theta}(x) - \theta_1(x)\}dx.$$

That is,

$$\Delta^{(1)}(\hat{h}_{\hat{\theta}},\theta_1)\hat{h}_{\hat{\theta}} = -2\ \hat{\xi}(\theta_1)\ , \tag{4.6}$$

provided $\hat{h}_{\hat{\theta}} \in n^{-1/5}(a_1,b_1)$. Expand $\Delta^{(1)}(\hat{h}_{\hat{\theta}},\theta_1)$ in a Taylor series about $\hat{h}_{\theta_1}$:

$$\Delta^{(1)}(\hat{h}_{\hat{\theta}},\theta_1) = (\hat{h}_{\hat{\theta}}-\hat{h}_{\theta_1})\ \Delta^{(2)}(h^*,\theta_1), \tag{4.7}$$

where h* lies inbetween $\hat{h}_{\hat{\theta}}$ and $\hat{h}_{\theta_1}$. This definition of h* is used in all that follows. Define $\hat{h}^{\dagger} = h^*$ if $|\hat{h}_{\hat{\theta}}-\hat{h}_{\theta_1}| \le n^{-1/4}$, and $\hat{h}^{\dagger} = h_{\theta_1}$ otherwise.

LEMMA 4.4. For each $\varepsilon > 0$,

$$P_{\theta_1}\{n^{2/5}|\Delta^{(2)}(\hat{h}^{\dagger},\theta_1) - \Delta^{(2)}(h_{\theta_1},\theta_1)| > \varepsilon\} \to 0\ .$$

PROOF. From our definition of $\hat{h}^{\dagger}$,

$$|\hat{h}^{\dagger}-h_{\theta_1}| \le n^{-1/4} + |\hat{h}_{\theta_1}-h_{\theta_1}|\ .$$

Therefore by Lemma 3.7,

$$P_{\theta_1}(|\hat{h}^{\dagger}-h_{\theta_1}| > 2\ n^{-1/4}) \to 0. \tag{4.8}$$

It now follows from (3.4) of Lemma 3.1 that

$$P_{\theta_1}\{n^{2/5}|M^{(2)}(\hat{h}^{\dagger},\theta_1) - M^{(2)}(h_{\theta_1},\theta_1)| > \varepsilon\} \to 0,$$

and since (by (4.8))

$$(4.9) \qquad P_{\theta_1}\{\hat{h}^{\dagger},\ h_{\theta_1} \in n^{-1/5}(a_1,b_1)\} \to 1,$$

the proof of Lemma 4.4 will be completed if we show that

$$(4.10) \qquad P_{\theta_1}\{n^{2/5}|D^{(2)}(\hat{h}^{\dagger},\theta_1)| > \varepsilon\} \to 0 \quad \text{and} \quad P_{\theta_1}\{n^{2/5}|D^{(2)}(h_{\theta_1},\theta_1)| > \varepsilon\} \to 0.$$

Using Hölder continuity of K, K' and K'', and the fact that each of these functions has compact support, we may produce $\lambda > 0$ such that

$$|D^{(2)}(n^{-1/5}s,\theta_1) - D^{(2)}(n^{-1/5}t,\theta_1)| \le C\, n^{-1}$$

uniformly in $n \ge 1$, s and $t \in (a_1,b_1)$ with $|s-t| \le n^{-\lambda}$, and samples $X_1,\ldots,X_n$. Let $a_1 = t_0 < t_1 < \ldots < t_{\nu-1} \le b_1 < t_\nu$ be a partition of (a_1,b_1) such that $t_i - t_{i-1} = n^{-\lambda}$ for each i. In view of (4.9), to prove (4.10) it suffices to prove that

$$p \equiv \sum_i P_{\theta_1}\{n^{2/5}|D^{(2)}(n^{-1/5}t_i,\theta_1)\ | > \varepsilon\} \to 0.$$

But this is immediate from Lemma 3.10 and Markov's inequality:

$$p \le C \sum_i (\varepsilon^{-1}n^{-1/10})^{2\ell} \to 0,$$

provided ℓ is sufficiently large. □

In view of (3.19),

$$P_{\theta_1}\{n^{2/5}|\Delta^{(2)}(h_{\theta_1},\theta_1) - M^{(2)}(h_{\theta_1},\theta_1)| > \varepsilon\} \to 0$$

for each $\varepsilon > 0$, and so by Lemma 4.4,

$$(4.11) \qquad P_{\theta_1}\{n^{2/5}|\Delta^{(2)}(\hat{h}^{\dagger},\theta_1) - M^{(2)}(h_{\theta_1},\theta_1)| > \varepsilon\} \to 0.$$

Noting (3.3) of Lemma 3.1, let $0 < a_2 < \frac{1}{2} b_2 < \infty$ be constants such that $n^{2/5} M^{(2)}(h_{\theta_1}, \theta_1) \in (a_2, \frac{1}{2} b_2)$ for all n, and take $\epsilon = a_2$ in (4.11). Then

$$(4.12) \quad P_{\theta_1}\{0 < \Delta^{(2)}(\hat{h}^{\dagger}, \theta_1) < b_2 n^{-2/5}\} \to 1 .$$

Let η_1, η_2 be as in Proposition 4.1, and set $\eta_3 \equiv (b_1 b_2)^{-1} \eta_2$. Let E_1 be the event that $|\hat{h}_{\hat{\theta}} - \hat{h}_{\theta_1}| \leq n^{-1/4}$, E_2 the event that $|\Delta^{(2)}(\hat{h}^{\dagger}, \theta_1)| \leq b_2 n^{-2/5}$, and E_3 the event that $\hat{h}_{\hat{\theta}} \in n^{-1/5}(a_1, b_1)$. Remember that $\hat{h}^{\dagger} = h^*$ on E_1, and that (4.6) holds on E_3. By (4.6) and (4.7),

$$\begin{aligned}
(4.13) \quad & P_{\theta_1}(|\hat{h}_{\hat{\theta}} - \hat{h}_{\theta_1}| > \eta_3 n^{-3/10}; E_1) \\
& \geq P_{\theta_1}\{|\Delta^{(1)}(\hat{h}_{\hat{\theta}}, \theta_1)| > \eta_3 n^{-3/10} \cdot b_2 n^{-2/5}; E_1\} - P_{\theta_1}(\tilde{E}_2) \\
& \geq P_{\theta_1}\{|2\hat{\xi}(\theta_1)| > \eta_3 b_2 n^{-7/10} \cdot b_1 n^{-1/5}; E_1\} - P_{\theta_1}(\tilde{E}_2) - P_{\theta_1}(\tilde{E}_3) \\
& \geq P_{\theta_1}\{|\hat{\xi}(\theta_1)| > \eta_2 n^{-9/10}; E_1\} - P_{\theta_1}(\tilde{E}_2) - P_{\theta_1}(\tilde{E}_3) \\
& \geq P_{\theta_1}(E_1) - P_{\theta_1}(\tilde{E}_2) - P_{\theta_1}(\tilde{E}_3) - \eta_1 ,
\end{aligned}$$

the last line following from Proposition 4.1. Result (4.12) and Lemma 3.4 imply that $P_{\theta_1}(\tilde{E}_2) \to 0$ and $P_{\theta_1}(\tilde{E}_3) \to 0$, respectively. If n is so large that $\eta_3 n^{-3/10} < n^{-1/4}$, then by (4.13),

$$\begin{aligned}
& P_{\theta_1}(|\hat{h}_{\hat{\theta}} - \hat{h}_{\theta_1}| > \eta_3 n^{-3/10}) \\
& = P_{\theta_1}(|\hat{h}_{\hat{\theta}} - \hat{h}_{\theta_1}| > \eta_3 n^{-3/10}; E_1) + P_{\theta_1}(\tilde{E}_1) \\
& \geq \{P_{\theta_1}(E_1) - P_{\theta_1}(\tilde{E}_2) - P_{\theta_1}(\tilde{E}_3) - \eta_1\} + P_{\theta_1}(\tilde{E}_1) \\
& = 1 - P_{\theta_1}(\tilde{E}_2) - P_{\theta_1}(\tilde{E}_3) - \eta_1 \to 1 - \eta_1
\end{aligned}$$

as $n \to \infty$. This proves (4.1), and completes the proof of Theorem 2.1.

References

[1] BARTLETT, M.S. (1963). Statistical estimation of density functions. Sankhya Ser. A 25, 245-254.

[2] BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densities: risque minimax. Z. Wahrsch. Geb. 47, 119-137.

[3] BURKHOLDER, D.L. (1973). Distribution function inequalities for martingales. Ann. Prob. 1, 19-42.

[4] BURMAN, P. (1985). A data dependent approach to density estimation. to appear in Z. Wahrsch. Geb.

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

[6] FRYER, M.J. (1977). A review of some non-parametric methods of density estimation. J. Inst. Math. Appl. 20, 335-354.

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

[8] HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. Proc. Sixth Internat. Symp. Multivar. Anal. Pittsburgh, 25-29.

[9] HALL, P. and MARRON, J.S. (1985). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. Center for Stochastic Processes, Tech. Rep. No. 94.

[10] IBRAGIMOV, I.A. and HAS'MINSKII, R.Z. (1981). Statistical Estimation. Berlin: Springer.

[11] MARRON, J.S. (1983). Optimal rates of convergence to Bayes risk in nonparametric discrimination. Ann. Statist. 11, 1142-1155.

[12] MEYER, T.G. (1977). On fixed or scaled radii confidence sets: the fixed sample size case. Ann. Statist. 5, 65-78.

[13] MEYER. T.G. (1977). Bounds for estimation of density functions and their derivatives. Ann. Statist. 5, 136-142.

[14] PARZEN, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist. 33, 1065-1076.

[15] PETROV, V.V. (1978). Sums of independent random variables. Berlin: Springer.

[16] RICE, J. (1984). Bandwidth choice for nonparametric regression. Ann. Statist. 12, 1215-1230.

[17] ROSENBLATT, M. (1971). Curve estimates. Ann. Math. Statist. 42, 1815-1842.

[18] SACKS, J. and STRAWDERMAN, W. (1982). Improvements on linear minimax estimates, Statistical Decision Theory and Related Topics III, vol. 2 (S.S. Gupta and J.O. Berger, eds.) New York: Academic Press.

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

[20] STONE, C.J. (1982). Optimal global rates of convergence for non-parametric regression. Ann. Statist. 10, 1040-1053.

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

[22] WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. Ann. Statist. 3, 15-29.

[23] WEGMAN, E.J. (1972). Nonparametric probability density estimation: I. a summary of available methods. Technometrics 14, 533-546.

END

FILMED

12-85

DTIC