#1761

Lower bounds for Bandwidth Selection in Density Estimation

by

Peter Hall and J. S. Marron

May 11, 1988

ABSTRACT

This paper establishes asymptotic lower bounds which provide limits, in various contexts, as to how well one may select the bandwidth of a kernel density estimator. Earlier results are of this type are extended to the important case of different smoothness classes (for the underlying density), and it is also seen that very useful bounds can be obtained even in the presence of parametric knowledge of the density. An important feature of the results is that while the lower bound is unacceptably large (i.e. of order $n^{-1/10}$)when the error criterion is Integrated Squared Error, it can be quite acceptable (often of order $n^{-1/2}$) when the error criterion is Mean Integrated Squared Error. We feel this indicates that the latter should become the benchmark for this problem.

key words: bandwidth selection, cross-validation, density estimation, kernel estimators, rates of convergence.

grants: Research partially supported by National Science Foundation Grant DMS-8701201.

Subject Classification: Primary 62G05; secondary 62E20, 62H99.

1. Introduction

The density estimation problem is that of estimating a probability density f using a random sample, $X_1,\ldots,X_n$, from f. Given a bandwidth h, and a kernel function K, the kernel density estimator is

$$\hat{f}_h(x) = n^{-1} \Sigma_{i=1}^{n} K_h(x-X_i),$$

where $K_h(\cdot) = K(\cdot/h)/h$. The bandwidth controls the smoothness of the resulting curve estimate, with the result that bandwidth choice is crucial to performance of the estimator (see for example Devroye and Gyorfi (1985) or Silverman (1986)).

Widely considered means of assessing the performance of $\hat{f}$ include the Integrated Squared Error,

$$\Delta(h,f) = \int (\hat{f}_h - f),$$

and the Mean Integrated Squared Error,

$$M(h,f) = E\,\Delta(h,f).$$

(See Devroye and Gyorfi (1985) for another viewpoint.) The minimizers of these criteria, denoted $\hat{h}_f$ and h_f, are both reasonable choices of "optimal bandwidth". There is some controversy concerning which one is preferred. In particular, h_f seems appropriate for the same reasons that risk (as opposed to loss) is the main focus of decision theory. On the other hand, in the specific context of curve estimation, a case can be made for $\hat{h}_f$ being the most suitable target, because it is the bandwidth choice which makes the resulting estimator as close as possible to f for the set of data at hand, as opposed to the average over all possible data sets. See Haerdle, Hall and Marron (1988) and

Marron(1988) for further discussion concerning which should be called the "optimal" bandwidth.

The fact that there is a very substantial difference between $\hat{h}_f$ and h_f has been demonstrated by Hall and Marron (1987a), who have shown that the relative difference between these is (under common technical assumptions, such as those stated below) of the order $n^{-1/10}$. David Scott and Hans-Georg Mueller have expressed (in personal correspondence) the viewpoint that h_f may be a more reasonable goal, simply because it may be expected to be easier to estimate than $\hat{h}_f$. In this paper, it is seen that this intuition is substantially correct. In particular, it will be demonstrated in Section 2.2 that the relative rate of convergence to $\hat{h}_f$, of any data driven bandwidth, can never (in a minimax sense) be faster than $n^{-1/10}$ (the theorem in Section 2.2 is a generalization of the related results of Hall and Marron (1987b) which made use of much more stringent assumptions than those used there). On the other hand, Hall and Marron (1987c) (see their Remark 4.6), have shown that, under strong enough smoothness assumptions, much faster rates, even up to $n^{-1/2}$, are attainable when h_f is accepted as the target. We believe this demonstrates conclusively that $\hat{h}_f$, while intuitively attractive, is just too difficult to attain, and hence h_f should be the goal of data-based bandwidth selection methods.

In view of this, it makes sense to investigate how well one may choose the bandwidth $\hat{h}_f$. See Marron (1988) for a survey of proposed methods of using the data, $X_1,\ldots,X_n$, to objectively choose the bandwidth. To explore the best possible performance of not only these

bandwidth selectors, but also any that may subsequently be proposed, $\hat{h}$ will be thought of as a bandwidth selector, but it is allowed to be an arbitrary measurable function of the data. Under such an assumption on the data-driven bandwidth, in order find lower bounds on how close $\hat{h}$ may be to h_f (or in Section 2.2 to $\hat{h}_h$), it is necessary to consider more than one underlying density. A convenient means of doing this is through a minimax structure, where one considers suprema over a class of alternative densities (see the Theorems in Sections 2 and 3 for a precise formulation).

The results in Section 2 are connected to each other by the fact that, for each n, only two alternative densities need be considered. In addition to the bound obtained on the rate of convergence to $\hat{h}_f$ discussed above, it will be shown in Section 2.3 that a two-alternative class is sufficient to show that the relative rate of convergence to h_f can be no faster than $n^{-1/2}$, regardless of the smoothness of the densities under consideration. The surprising fact that these bounds require only two-alternative classes is explored further in Section 2.4, through considering the interesting special cases of scale and location change alternatives. In particular it will be shown that the same bounds hold (i.e. $n^{-1/10}$ for $\hat{h}_f$ and $n^{-1/2}$ for h_f), even when one has parametric knowledge about the underlying density.

An interesting question is when this bound of $n^{-1/2}$ can be achieved. In Remark 4.6 of Hall and Marron (1987c) it is seen that this bound can be achieved when one makes strong enough smoothness assumptions on the underlying density. However, this relies strongly on

having enough smoothness of the underlying density avaliable, and so one might suspect that, when not enough smoothness is available, the lower bound could be sharpened. The fact that this is indeed the case will be demonstrated in Section 3.2, where we obtain a lower bound on the relative rate of convergence of any data driven bandwidth to h_f, that is better than $n^{-1/2}$ when densities which are not too smooth are considered. The price for this improved result is that the two-alternative class is replaced by a much larger class which grows rapidly as the sample size increases. Our class of alternatives is similar to that developed by Stone (1982) and used by Hall and Marron (1987b).

Another application of larger alternative classes will be given in Section 3.3, where this idea will be used for some technical improvement of the results of Section 2.

All proofs may be found in Section 4

2. Bounds Involving Two Alternatives

2.1 Introduction and Summary

To obtain the lower bounds in the current section, it is enough to consider (for each n) only two alternative densities. A means of constructing these (in a way which yields useful lower bounds) is to start with a fixed density $f_0(x)$, and a function $\alpha(x)$, and consider the alternative density

$$f_1(x) \equiv \{1 + n^{-1/2}\alpha(x)\} f_0(x).$$

The fact that f_0 and f_1 are distant only $n^{-1/2}$ apart (note this representation entails that most of the usual norms will be of the order $n^{-1/2}$ when α is reasonable, as assumed below) means that our bounds will apply even in a parametric setting, not solely to nonparametric classes of densities. This will emerge particularly clearly in Section 2.4, where the case of a normal $N(\mu,\sigma^2)$ target density f will be discussed. In a parametric context, where f_1 represents a version of f_0 with "nuisance parameters" replaced by their estimates, f_0 and f_1 are indeed distant $n^{-1/2}$ apart.

To ensure that f_1 is a proper density (for n large enough), assume

(2.1.1) $\int \alpha f_0 = 0$ and $f_1 \geq 0$.

Also assume

(2.1.2) f_0 and $|\alpha| f_0$ are bounded,

(2.1.3) f_0 and αf_0 have five bounded derivatives,

(2.1.4) $\sigma^2 \equiv \int \alpha^2 f_0$ and $\int \{(\alpha f_0'')^2 f_0$ are nonzero and finite.

Convenient technical assumptions concerning the estimator are:

(2.1.5) K is nonnegative and symmetric, with $\int K = 1$,

(2.1.6) K is compactly supported with a Hoelder-continuous second derivative.

Assumption (2.1.5) is important to the effective behavior of the kernel estimator. It implies that K is a "second order kernel". Versions of our results for higher order kernels will be presented in Remarks 2.2.4, 2.3.5, and 3.2.4. Assumption (2.1.6) is made more for convenience. It is straighforward to weaken this assumption through the use of various

truncation arguments, but this is not done explicitly because the increased complexity of proof would detract from the main points.

Further useful notation is,

$$p \equiv 1 - \phi(\sigma/2),$$

where ϕ denotes the standard normal cumulative distribution function.

Sections 2.2 and 2.3 will provide lower bounds to convergence rates of general estimators of $\hat{h}_f$ and h_f, respectively. Section 2.4 will illustrate the main features of these results by considering density estimation in pararmetric problems where either scale or location is unknown. In particular the fact that the bounds obtained in Sections 2.2 and 2.3 apply even in the prescence of parametric knowledge is underscored.

2.2 Bounds in the Case of ISE

In addition to the technical assumptions made in Section 2.1, also assume that the alternative densities, f_0 and f_1, are distinct in the sense that

$$\int\{(d/dx)^2\alpha(x)f_0(x)\}f_0(x)dx \neq 0. \tag{2.2.1}$$

The implications of this condition will be made clear in Section 2.4. The following theorem shows that it is impossible to find a data-based bandwidth which is closer to $\hat{h}_f$, the minimizer of the Integrated Squared Error $\Delta(h,f)$, than $n^{-1/10}$ in a relative error sense.

Theorem 2.2: Under the assumptions (2.1.1) - (2.1.6) and (2.2.1), for $\hat{h}$ any measurable function of the data,

$$(2.2.2)\qquad \lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f(|\hat{h} - \hat{h}_f|/\hat{h}_f > \epsilon n^{-1/10}) \geq p,$$

$$(2.2.3)\qquad \lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f\left(\left\{\frac{\Delta(\hat{h},f) - \Delta(\hat{h}_f,f)}{\Delta(h_f,f)}\right| > \epsilon n^{-1/5}\right\} \geq p.$$

The proof of Theorem 2.2 will be given in Section 5.1.

Remark 2.2.1: If $\hat{h}$ is taken to be the bandwidth chosen by cross-validation then the convergence rates in (2.2.2) and (2.2.3) are achieved; see Hall and Marron (1987a). Therefore the convergence rates described by Theorem 2.2 are best possible. Theorem 2.2 is a substantial strengthening of Theorems 2.1 and 4.1 of Hall and Marron (1987b). Although the bound is the same, the class of alternatives is much smaller and simpler here.

Remark 2.2.2: The probability p may be increased to 1 if more than just the two alternatives f_0 and f_1 are considered. A method of doing this will be described in Section 3.3.

Remark 2.2.3: If there were really only two densities f_0 and f_1 under consideration, then "$\geq p$" would become "$= p$" if one took $\hat{h}$ to be the "likelihood ratio bandwidth", which chooses between $\hat{h}_{f_0}$ and $\hat{h}_{f_1}$ depending on whether the likelihood ratio is bigger or smaller than one. In fact the proof of the theorem is based on the fact that no discrimination rule can distinguish between f_0 and f_1 more effectively than the likelihood ratio rule.

Remark 2.2.4: If the kernel function K is allowed to take on negative values, then the rate of convergence of $\hat{f}_h$ to f may be improved

(see, for example, Section 3.6 of Silverman 1986). In particular the kernel function K is said to be of order r when

$$\int x^j K(x)dx = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \leq j \leq r-1 \\ \kappa \neq 0 & \text{if } j = r \end{cases} .$$

Assumption (2.1.5) ensures that K is of order 2. The advantage of K being of order r is that, when f is assumed to have r continuous derivatives and $h \sim n^{-1/(2r+1)}$, both $\Delta(h,f)$ and $M(h,f)$ are of size $n^{-2r/(2r+1)}$. Theorem 2.2 continues to hold under this type of assumption, with the rates of $n^{-1/10}$ and $n^{-1/5}$ replaced by $n^{-1/2(2r+1)}$ and $n^{-1/(2r+1)}$, respectively. The differential operator $(d/dx)^2$ in (2.2.1) should be replaced by $(d/dx)^r$.

2.3 Bounds in the Case of MISE

In this case, the assumption (2.2.1) concerning the difference between the alternative densities, f_0 and f_1, should be replaced by

(2.3.1) $$\int\{(d/dx)^j\alpha(x)f_0(x)\}f_0(x)dx = 0, \quad j = 2,4.$$

See Section 2.4 for an investigation of the implications of this condition. Our next result shows that it is impossible to use a data-based bandwidth which is closer to h_f, the minimizer of the Mean Integrated Squared Error $M(h,f)$, than $n^{-1/2}$ in a relative error sense.

Theorem 2.3: Under the assumptions (2.1.1) - (2.1.6) and (2.3.1), for $\hat{h}$ any measurable function of the data,

$$(2.3.2)\qquad \lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f((|\hat{h} - h_f|/h_f > \epsilon n^{-1/2}) \geq p,$$

$$(2.3.3)\qquad \lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f\{\left|\frac{M(\hat{h},f) - M(h_f,f)}{M(h_f,f)}\right| > \epsilon n^{-1}\} \geq p.$$

The proof of Theorem 2.3 will be given in Section 5.2.

Remark 2.3.1: For sufficiently large, but finite, values of ν (Hall and Marron (1987c) need $\nu = 4.5$), this bound is known to be best possible, in the sense that there is a bandwidth selector whose relative rate of convergence to h_f is $n^{-1/2}$.

Remark 2.3.2: Note that the lower bounds obtained here , $n^{-1/2}$ and n^{-1}, go to zero much more rapidly than the corresponding bounds in Theorem 2.2, $n^{-1/10}$ and $n^{-1/5}$. This demonstrates a fundamental difference between the error criteria, Integrated Squared Error $\Delta(h,f)$ and Mean Integrated Squared Error $M(h,f)$. This is what provides motivation for acceptance, as discussed above, of $M(h,f)$ as the more reasonable measure of error, on the grounds that $\Delta(h,f)$ appears to be simply too difficult to optimize in a reasonable fashion.

Remark 2.3.3: As in Section 2.2, the probability p may be increased to 1 if more than just the two alternatives f_0 and f_1 are considered. A method of doing this will be given in Section 3.3.

Remark 2.3.4: Also as in Section 2.2, the likelihood ratio bandwidth (adapted for M instead of Δ) gives equality in (2.3.2) and (2.3.3).

Remark 2.3.5: There is a version of Theorem 2.3, with exactly the same rates $n^{-1/2}$ and n^{-1} in (2.3.2) and (2.3.3) respectively, for higher

order kernels.

2.4 Example: Scale and Location Changes

Additional insight into the structure of the minimax bounds of Theorems 2.2 and 2.3 can be gained by consideration of some specific choices of the alternative densities f_0 and f_1. Particularly interesting features are emphasized if α is chosen to make f_1 approximately a scale or location change of f_0. Of course in such a context, one should never consider estimating a density with a kernel estimator, but this is worth studying because of the interesting implications for the bandwidth selection problem.

In the scale-change case, $f_1(x)$ may be represented as

$$(2.4.1) \qquad (1 + n^{-1/2})f_0\{(1 + n^{-1/2})x\}$$
$$= f_0(x) + n^{-1/2}\{f_0(x) + xf_0'(x)\} + O(n^{-1}).$$

Thus define $\alpha(x) = 1 + x\{f_0'(x)/f_0(x)\}$. Note that, under reasonable assumptions on f_0, conditions (2.2.1) and (2.3.1) are satisfied for this f_1, and so this "scale alternative" may be used in Theorems 2.2 and 2.3.

In this context, Theorems 2.2 and 2.3 are perhaps most vividly illustrated by considering the problem of estimating a normal $N(\mu,\sigma^2)$ density using a nonparametric density estimator, as follows. Suppose μ is known, but σ^2 is unknown. Estimate σ^2 using the sample variance $\hat{\sigma}^2$, and take $\tilde{f}$ to be the $N(\mu,\hat{\sigma}^2)$ density. Note that $n^{-1/2}$ is the order of magnitude of the distance between $\hat{\sigma}^2$ and σ^2, so we are essentially in the context of the previous paragraph. Take $\hat{h}_{\tilde{f}}$ (the

bandwidth which minimizes $\Delta(h,\tilde{f})$) as our estimate of $\hat{h}_f$ (the bandwidth which minimizes $\Delta(h,f)$). Likewise, let $\tilde{h}_f$ be our estimate of h_f. Then $(\hat{h}_f^{\sim} - \hat{h}_f)/\hat{h}_f$ is of precise order $n^{-1/10}$, as indicated by Theorem 2.2, and $(\tilde{h}_f - h_f)/h_f$ is of precise order $n^{-1/2}$, as indicated by Theorem 2.3. This simple example brings home strikingly the fact that, even in the presence of parametric knowledge about f, we cannot hope to estimate $\hat{h}_f$ with a relative error of less than $n^{-1/10}$. The goal of estimating h_f is clearly much different, because in the presence of such parametric knowledge we can achieve the usual parametric rate of $n^{-1/2}$.

However, the situation changes markedly if the unknown parameter is one of location rather than scale. In the location-change case, $f_1(x)$ may be represented as

$$f_0(x + n^{-1/2}) = f_0(x) + n^{-1/2} f_0'(x) + O(n^{-1}).$$

Hence, define $\alpha(x) = f_0'(x)/f_0(x)$. Note that conditions (2.2.1) and (2.3.1) are not satisfied by this choice of α. Indeed, not only are these assumptions not valid, but the conclusions of Theorems 2.2 and 2.3 fail. In particular, $\hat{h}$ may be chosen so that for any $\epsilon > 0$,

$$\max_{f \in \{f_0, f_1\}} P_f(|\hat{h} - \hat{h}_f|/\hat{h}_f > \epsilon n^{-1/10}) \to 0$$

and

$$\max_{f \in \{f_0, f_1\}} P_f(|\hat{h} - h_f|/h_f > \epsilon n^{-1/2}) \to 0.$$

Again, these features are perhaps best brought out by considering the problem of estimating a normal $N(\mu,\sigma^2)$ density. On this occasion, suppose μ is unknown and σ^2 is known. Estimate μ using the sample

mean $\hat{\mu}$, and take $\tilde{f}$ to be the $N(\hat{\mu},\sigma^2)$ density. Then $n^{-1/2}$ is the order of magnitude of the distance between $\hat{\mu}$ and μ, and so we are in the context of the previous paragraph. Let $\hat{h}_{\tilde{f}}$, $h_{\tilde{f}}$ be our parametric estimates of $\hat{h}_f$, h_f, respectively. Then $h_{\tilde{f}} = h_f$, so that our estimate of h_f is error-free. However, it may be shown that $(\hat{h}_{\tilde{f}}-\hat{h}_f)/\hat{h}_f$ is of precise order $n^{-3/5}$, which is considerably better than the error of the order $n^{-1/10}$ encountered in the scale-change problem, and even better than the error $n^{-1/2}$ which might have been expected, but not quite error-free. It turns out that a relative error of $n^{-3/5}$ is intrinsic to bandwidth selection for the ISE problem in this setting, as the following result shows.

Theorem 2.4: Under the assumptions (2.1.1) - (2.1.6) and (2.3.1), for $\hat{h}$ any measurable function of the data,

$$\lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f(|\hat{h} - \hat{h}_f|/\hat{h}_f > \epsilon n^{-3/5}) \geq p, \tag{2.4.2}$$

$$\lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f\left\{\left|\frac{\Delta(\hat{h},f) - \Delta(\hat{h}_f,f)}{\Delta(h_f,f)}\right| > \epsilon n^{-6/5}\right\} \geq p. \tag{2.4.3}$$

The proof of Theorem 2.4 is so close to the proof of Theorem 2.2 that it will not be given explicitly.

Finally we should mention the scale-change and location-change versions of α discussed above may not be proper densities, since they may violate the nonegativity part of condition (2.1.1). This in no way invalidates our conclusions - a correction for positivity is of order n^{-1}, and could be included, if one wishes, by simply incorporating a

term of order n^{-1}, as in (2.4.1).

3 Bounds Involving Multiple Alternatives

3.1 Introduction and summary

There are two points at which deeper insight can be gained by replacing the above two-alternative minimax results, by results which make use of multiple alternatives. The first point is in establishing a better bound on how well a bandwidth selector $\hat{h}$ may approximate h_f, the minimizer of the Mean Integrated Squared Error $M(h,f)$, in situations where the underlying density is not too smooth. The second point is in strengthening Theorems 2.2, 2.3 and 2.4 by replacing p by 1 on the right hand sides of (2.2.2), (2.2.3), (2.3.2), (2.3.3), (2.4.2) and (2.4.3).

When the underlying density is not too smooth, the lower bounds of Theorem 2.3 may be sharpened. In such cases, the rates of convergence depend on the amount of smoothness of the underlying density. To quantify this in a form convenient for minimax lower bound results, consider smoothness classes indexed by a parameter $\nu \geq 0$. In particular, given $B > 0$, let ℓ be the largest integer strictly less than $2 + \nu$, and define $G_\nu(B)$ to be the set of all probability densities which vanish outside of $(-B,B)$, have ℓ derivatives, and satisfy

$$\sup_{x,y} \left| f^{(\ell)}(x) - f^{(\ell)}(y) \right| / \left| x - y \right|^{2+\nu-\ell} \leq B.$$

A minimax lower bound for the relative rate of convergence of $\hat{h}$

to h_f, in terms of the smoothness index ν, will be stated in Section 3.2. The issue of increasing p to 1 will be treated in Section 3.3.

3.2 Bounds in the Case of MISE

The minimax lower bound of Theorem 2.3 may be sharpened, when the underlying density is not too smooth, to:

Theorem 3.2: Under the assumptions (2.1.5) and (2.1.6), for $\nu \geq 0$, $B > 0$ and $\hat{h}$ any measurable function of the data,

$$(3.2.1) \qquad \lim_{\epsilon \to 0} \liminf_{n \to \infty} \sup_{f \in G_\nu(B)} P_f(|\hat{h} - h_f|/h_f > \epsilon n^{-\rho}) = 1,$$

$$(3.2.2) \qquad \lim_{\epsilon \to 0} \liminf_{n \to \infty} \sup_{f \in G_\nu(B)} P_f\left\{\left|\frac{M(\hat{h},f) - M(h_f,f)}{M(h_f,f)}\right| > \epsilon n^{-2\rho}\right\} = 1,$$

where

$$\rho = (1+4\nu) / 2(5+2\nu).$$

The proof of Theorem 3.2 will be given in Section 5.3.

Remark 3.2.1: It is important to note that Theorem 3.2 and Theorem 2.3 each provide useful information for different values of ν, with $\nu = 2$ being the boundary point. In particular, for $\nu > 2$ we have $\rho > 1/2$, and then the lower bound of Theorem 2.3 is more informative. On the other hand, for $\nu < 2$ we have $\rho < 1/2$, so the present bound is more useful.

Remark 3.2.2: In the case $\nu = 0$, the relative error bound of $n^{-1/10}$

for $\hat{h}$, provided by Theorem 3.2, is known to be best possible (i.e. both an upper and a lower bound). It is achieved by the window selected by cross-validation (Hall and Marron 1987a).

Remark 3.2.3: The class of densities $G_\nu(B)$ is actually far bigger than is required to obtain the bound stated in the theorem. In particular, in the proof a much smaller class (finite for each n) of alternatives is constructed, and this is all that is necessary. The more general result is not stated here because it involves the introduction of considerably more notation, which has a tendency to obscure the main point of this section.

Remark 3.2.4: If the kernel K is of order r (as discussed in Renark 2.2.4), then $G_\nu(B)$ should denote a class of densities with $r + \nu$ "derivatives", instead of $2 + \nu$ as above. Then the only change to Theorem 3.2 is that ρ becomes $(1+4\nu)/2(2r+1+2\nu)$.

3.3 Probability One Bounds

In Theorems 2.2, 2.3 and 2.4 the probabilities p may all be sharpened to 1 if a larger class of alternatives is used. A simple way of constructing such a larger class is to consider all convex combinations of the f_0 and f_1 described above. In particular define

$$C(f_0,f_1) \equiv \{\omega f_0 + (1-\omega)f_1 : \omega \in [0,1]\}.$$

Then, if the set of alternatives, $\{f_0, f_1\}$, is replaced by $C(f_0,f_1)$, the values of p in Theorems 2.2, 2.3 and 2.4 may all be taken to be 1.

This is intuitively clear, because the minimax bounds calculated in these theorems come from the difficulty in using $X_1,\ldots,X_n$ to choose among the various possible density functions. If the class $\{f_0, f_1\}$ is enlarged by including convex combinations, then p, which is the smallest probability of misclassifying the underlying density, becomes larger. The limit of this process is the class $C(f_0, f_1)$, and $p = 1$. We do not include a specific proof of this fact, because the idea is the same as that used to verify (1.2) in Stone (1980).

4. Proofs

4.1 Proof of Theorem 2.2

We prove only (2.2.2), since the extension to (2.2.3) may be accomplished as in Hall and Marron (1987b, p. 171). Let $\tilde{h} = \hat{h}_{\tilde{f}}$ be an element of $\{\hat{h}_{f_0}, \hat{h}_{f_1}\}$ which minimizes $|\hat{h} - \tilde{h}|$ over those elements. If f is either f_0 or f_1 then

$$|\tilde{h}-\hat{h}_f| \le |\tilde{h}-\hat{h}| + |\hat{h}-\hat{h}_f| \le 2|\hat{h}-\hat{h}_f|.$$

Therefore result (2.2.2) will follow if we prove that

$$(4.1.1) \qquad \lim_{\epsilon\to 0} \lim_{n\to\infty} \inf \max_{f\in\{f_0 f_1\}} P_f(|\tilde{h}-\hat{h}_f| > \epsilon\, n^{-3/10}) \ge p.$$

Define $L(z) \equiv -z\,K'(z)$ and

$$\hat{g}_h(x) \equiv (nh)^{-1} \sum_i L\{(x-X_i)/h\}.$$

Arguing as in Hall and Marron (1987b, p. 169) we may deduce that

$$(4.1.1') \qquad \hat{h}_{\tilde{f}} - \tilde{h} = 2\xi(\tilde{h},f)/\tilde{h}\Delta^{(2)}(h^*,f),$$

where $\xi(h,f) \equiv \int(\hat{f}_h-\hat{g}_h)(\tilde{f}-f)$, where $\Delta^{(2)}(h,f)$ denotes the second derivation of $\Delta(h,f)$ with respect to h, and where h^* lies between

$\tilde{h}$ and $\hat{h}_f$. It is relatively easy to prove, as in Lemmas 4.2, 6.1 and 6.2 of Hall and Marron (1987b), that

$$\lim_{a\to 0, b\to\infty} \lim\inf_{n\to\infty} \min_{f\in\{f_0,f_1\}} P_f(an^{-1/5} \le \hat{h}_{f_0}, \hat{h}_{f_1} \le bn^{-1/5}) = 1,$$

and for all $0 < a < b < \infty$,

$$\lim_{\lambda\to\infty} \lim\sup_{n\to\infty} \max_{f\in\{f_0,f_1\}} P_f\{\min_{h\in(an^{-1/5},bn^{-1/5})} |\Delta^{(2)}(h,f)| > \lambda n^{-2/5}\} = 0.$$

Therefore result (4.1.1) will follow if we show that

$$(4.1.2)\quad \lim_{\epsilon\to 0} \lim\inf_{n\to\infty} \max_{f\in\{f_0,f_1\}} P_f\{\min_{h\in(an^{-1/5},bn^{-1/5})} |\xi(h,f)| > \epsilon n^{-9/10}\} \ge p.$$

If $\tilde{f} \ne f$ then

$$|\xi(h,f)| = |\int(\hat{f}_h - \hat{g}_h)(\tilde{f}-f)| = n^{-1/2}|\int(\hat{f}_h - \hat{g}_h)\alpha f_0|.$$

And by the Neyman-Pearson lemma,

$$\max_{f\in\{f_0,f_1\}} P_f(\tilde{f} \ne f) \ge (1/2)\{P_{f_0}(\tilde{f}=f_1) + P_{f_1}(\tilde{f}=f_0)\}$$

$$\ge (1/2)\{P_{f_0}(\bar{f}=f_1) + P_{f_1}(\bar{f}=f_0)\},$$

where $\bar{f}$ is the likelihood ratio rule for deciding between f_0 and f_1. Now,

$$P_{f_0}(\bar{f}=f_1) = P_{f_0}[\Sigma_i \log\{1 + n^{-1/2}\alpha(X_i)\} > 0]$$

$$= P_{f_0}\{n^{-1/2}\Sigma_i \alpha(X_i) - \tfrac{1}{2}n^{-1}\Sigma_i \alpha(X_i)^2 + o_p(1) > 0\}$$

$$\to 1 - \Phi(\sigma/2) = p,$$

and similarly $P_{f_1}(\bar{f}-f_0) \to p$. Therefore

$$(4.1.2')\quad \lim\inf_{n\to\infty} \max_{f\in\{f_0,f_1\}} P_f(\tilde{f}\ne f) \ge p,$$

and so (4.1.2) will follow if we prove that

$$(4.1.3)\quad \lim_{\epsilon\to 0}\ \liminf_{n\to\infty}\ \max_{f\in\{f_0,f_1\}} P_f\Big\{\min_{h\in(an^{-1/5},bn^{-1/5})} \Big|\int(\hat f_h-\hat g_h)\alpha f_0\Big| > \epsilon\, n^{-2/5}\Big\} = 1.$$

Put $A \equiv K-L$. Then

$$(4.1.3')\quad S = S(h) \equiv \int(\hat f_h-\hat g_h)\alpha f_0 = (nh)^{-1}\sum_i \int A\{(x-X_i)/h\}\alpha(x)f_0(x)dx = n^{-1}\sum_i \int A(y)\alpha(X_i+hy)f_0(X_i+hy)dy.$$

Now

$$(4.1.3'')\quad E_f(S) = \int A(y)dy \int \alpha(x+hy)f_0(x+hy)f(x)dx = h^2\{\int y^2A(y)dy\}\,[\int\{(d/dx)^2(\alpha(x)f_0(x))\}f_0(x)dx] + O(h^3+h^2n^{-1/2}) = h^2c + O(h^3+h^2n^{-1/2}),$$

say, where $c \neq 0$. (Here we have used (2.2.1) and the fact that $\int y^jA(y)dy = 0$ for $j = 0,1$.) By an inequality for sums of independent random variables (Burkholder, 1973, p. 40),

$$\max_{h\in(an^{-1/5},bn^{-1/5})}\ \max_{f\in\{f_0,f_1\}} E_f\{|S(h)-ES(h)|^{2r}\} \le C(a,b,r)n^{-r}$$

for all $r \ge 1$. Therefore if $\mathcal{H}_n$ is any set of elements of $(an^{-1/5}, bn^{-1/5})$ containing no more than n^d elements for any fixed $d > 0$, we have for large n,

$$\begin{aligned}\min_{f\in\{f_0,f_1\}} P_f\{\min_{h\in\mathcal{H}_n}|S(h)| > (1/3)a^2|c|n^{-2/5}\} &\ge 1 - \sum_{h\in\mathcal{H}_n}\ \max_{f\in\{f_0,f_1\}} P_f\{|S(h)-ES(h)| > (1/3)a^2|c|n^{-2/5}\}\\ &\ge 1 - O\{n^d(n^{2/5}n^{-1/2})^{2r}\} \to 1\end{aligned}$$

as $n \to \infty$, provided we choose $r > 5d$. Result (4.1.3) now follows via the continuity argument of Hall and Marron (1987b, p. 175). This

completes the proof of Theorem 2.2

4.2 Proof of Theorem 2.3

Let S be as at (4.1.3'). If (2.3.1) holds then, noting that $\int y^j A(y)dy = 0$ for $j = 0,1,3$, we deduce from (4.1.3'') that $E_{f_0}(S) = O(h^5) = O(n^{-1})$ and

$$E_{f_1}(S) = h^2 n^{-1/2} \{\int y^2 A(y)dy\} \int \{(d/dx)^2 \alpha(x) f_0(x)\} \alpha(x) f_0(x)dx + O(h^5 + h^4 n^{-1/2})$$

$$= h^2 n^{-1/2} \{2\int y^2 K(y)dy\} \int \{(d/dx)\alpha(x) f_0(x)\}^2 dx + O(n^{-1}).$$

Also, since

$$\int A(y)\alpha(x+hy) f_0(x+hy)dy = h^2 \{\int y^2 K(y)dy\} (\alpha f_0)''(x) + O(h^3),$$

then by the central limit theorem for sums of independent random variables,

$$P_f(S - ES \le h^2 n^{-1/2} z) \to \Phi(z/\tau)$$

for $f = f_0$ or f_1, where

$$\tau^2 \equiv \{\int y^2 K(y)dy\}^2 \int \{\alpha f_0)''(x)\}^2 f_0(x)dx$$

We may deduce from these results and the continuity argument of Hall and Marron (1987b, p. 175) that

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} \min_{f \in \{f_0, f_1\}} P_f\{ \min_{h \in (an^{-1/5}, bn^{-1/5})} |S(h)| > \epsilon n^{-9/10} \} = 1.$$

This establishes (4.1.3), with $n^{-2/5}$ replaced now by $n^{-9/10}$. Tracing through the argument preceding that result we deduce (2.3.2), and we may obtain (2.3.3) by arguing as in Hall and Marron (1987b, p. 171). This completes the proof of Theorem 2.3

4.3 Proof of Theorem 3.2

As for Theorems 2.2 and 2.3, we prove only (3.2.1). The circumstance $\nu = 0$ is similar in all essential details to $\nu > 0$, so we assume $\nu > 0$.

The first step is to construct a class of densities which are "hard to distinguish", yet at the same time "far apart". Following the ideas of Stone (1982), let ψ be a symmetric, six times differentiable function on $(-\infty,\infty)$, vanishing outside $(-1/4,1/4)$. Put $m \equiv n^{-1/(5+2\nu)}$, let $\tau = \{\tau_v: v = 1,\ldots,m\}$ be a sequence of 0's and 1's, let g_0 be a density which is constant at a nonzero value on $(-1/2,3/2)$ and vanishes outside $(-1,2)$, and define

$$\gamma_v \equiv m^{-(2+\nu)}\psi\{m(x-v/m)\},$$

$$f(x) = f(x|\tau) \equiv g_0(x) + \sum_{v=1}^{m} \tau_v \gamma_v(x),$$

and

$$\mathcal{F} \equiv \{f(x|\tau): \tau \text{ is a sequence of 0's and 1's}\}.$$

Note that for large n, $\mathcal{F}$ is a set of densities vanishing outside $(-1,2)$ and having uniformly continuous bounded $(2+\nu)$'th derivatives. Note that many related constructions are possible here. We choose this one, because it contains the necessary features with as little overhead, in terms of notation and length of proof, as possible.

Let $\tilde{h} = h_{\tilde{f}}$ minimize $|\hat{h}-h_{\tilde{f}}|$ over all $\tilde{f} \in \mathcal{F}$. Then $|\tilde{h}-h_f| \le 2|\hat{h}-h_f|$ for all $f \in \mathcal{F}$, and so it suffices to prove that

$$(3.1.1)\quad \lim_{\epsilon\to 0}\lim_{n\to\infty}\inf\max_{f\in\mathcal{F}} P_f(|\tilde{h} - h_f| > \epsilon n^{-(1/5)-(1+4\nu)/2(5+2\nu)}) = 1.$$

Our next step in establishing (3.1.1) is to develop an analogue of (4.1.1').

Let f, g be densities, and observe that

$$\begin{aligned} M(h,g) &= \int E_g(\hat{f}_h - g)^2 \\ &= \int E_g(\hat{f}_h - f)^2 + 2\int (E_g\hat{f}_h - f)(f-g) + \int (f-g)^2 \\ &= M(h,f) + \int (E_g - E_f)(\hat{f}_h - f)^2 + 2\int (E_g\hat{f}_h - f)(f-g) + \int (f-g)^2. \end{aligned}$$

Differentiating with respect to h we obtain

$$\begin{aligned} M^{(1)}(h,g) = M^{(1)}(h,f) &+ 2h^{-1}\int (E_f - E_g)(\hat{f}_h - \hat{g}_h)(\hat{f}_h - f) \\ &+ 2h^{-1}\int E_g(\hat{f}_h - \hat{g}_h)(g-f), \end{aligned}$$

where $M^{(j)}(h,g)$ denotes the jth derivative of $M(h,g)$ with respect to h and where $\hat{g}_h$ was defined in Section 4.1. Therefore with

$$\eta(h,f,g) \equiv \int \{(E_f - E_g)(\hat{f}_h - \hat{g}_h)(\hat{f}_h - f) + E_g(\hat{f}_h - \hat{g}_h)(g-f)\}.$$

we have

$$M^{(1)}(h,g) = M^{(1)}(h,f) + 2h^{-1}\eta(h,f,g).$$

Taking $(h,f,g) = (h_{f_1}, f, f_1)$ for $f_1 \in \mathcal{F}$, we find that

$$\begin{aligned} 0 = M^{(1)}(h_{f_1}, f_1) &= M^{(1)}(h_{f_1}, f) + 2h_{f_1}^{-1}\eta(h_{f_1}, f, f_1) \\ &= M^{(1)}(h_f, f) = M^{(1)}(h_{f_1}, f) + (h_f - h_{f_1})\, M^{(2)}(h^+, f), \end{aligned}$$

where h^+ lies between h_f and h_{f_1}. In consequence,

$$h_f - h_{f_1} = 2\eta(h_{f_1}, f, f_1)/h_{f_1} M^{(2)}(h^+, f),$$

whence

$$(4.3.2)\qquad h_f - \tilde{h} = 2\eta(\tilde{h}, f, \tilde{f})/\tilde{h} M^{(2)}(\tilde{h}^+, f),$$

where $\tilde{h}^+$ lies between $\tilde{h}$ and h_f. This is the desired analogue of

(4.1.1').

Arguing as in the proofs of Lemmas 4.2, 6.1 and 6.2 of Hall and Marron (1987b) we may show that for sufficiently large n_0,

$$0 < \inf_{n \geq n_0, f \in \mathcal{F}} n^{1/5} h_f \leq \sup_{n \geq n_0, f \in \mathcal{F}} n^{1/5} h_f < \infty,$$

and for any $0 < a < b < \infty$ and some $\lambda = \lambda(a,b) > 0$,

$$\sup_{h \in (an^{-1/5}, bn^{-1/5}), f \in \mathcal{F}} |M^{(2)}(h,f)| \leq \lambda n^{-2/5}.$$

In view of these results and (4.3.2) we see that (4.3.1) will follow if we prove that for each $0 < a < b < \infty$,

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \inf_{f \in \mathcal{F}} \max P_f\{ \min_{h \in (an^{-1/5}, bn^{-1/5})} |\eta(h,f,\tilde{f})| > \epsilon n^{-(4/5)-(1+4\nu)/2(5+2\nu)} \} = 1. \tag{4.3.3}$$

The next step is to simplify $\eta(h,f,f_1)$. Write

$$f = \sum_v \tau_v \gamma_v \quad \text{and} \quad f_1 = \sum_v \tau_{1v} \gamma_v,$$

and let A be as in Section 4.1. Define

$$J \equiv \int \mathcal{P}(y)[(1-n^{-1}) \iint A(w)K(x)\, \mathcal{P}\{y + hm(w+x)\}dwdx - \int A(w)\mathcal{P}(y+hmw)dw]dy,$$

not depending on v. Then:

LEMMA. $$\eta(h,f,f_1) = n^{-1} J \sum_v (\tau_v - \tau_{1v}).$$

PROOF OF LEMMA. Observe that

$$E_g\{\hat{f}_h(\hat{f}_h - \hat{g}_h)\} = (nh^2)^{-1} E_g[K\{(x-X)/h\}A(x-X)/h\}] + (1-n^{-1})(E_g\hat{f}_h)E_g(\hat{f}_h - \hat{g}_h).$$

Therefore

$$(4.3.4)\quad \eta(h,f,f_1) = \int \{(E_f-E_{f_1})(\hat{f}_h-\hat{g}_h)(\hat{f}_h-f) + E_{f_1}(\hat{f}_h-\hat{g}_h)(f_1-f)\}$$

$$= \int [(nh^2)^{-1}(E_f-E_{f_1})\, K\{(x-X)/h\}A\{(x-X)/h\}$$

$$+ (1-n^{-1})\{(E_f\hat{f}_h)\, E_f(\hat{f}_h-\hat{g}_h) - (E_{f_1}\hat{f}_h)\, E_{f_1}(\hat{f}_h-\hat{g}_h)\}$$

$$- (E_f-E_{f_1})(\hat{f}_h-\hat{g}_h)f + (f_1-f)E_{f_1}(\hat{f}_h-\hat{g}_h)]$$

$$= \int [(1-n^{-1})\{(E_f-E_{f_1})\hat{f}_h\, E_f(\hat{f}_h\, E_f(\hat{f}_h-\hat{g}_h)$$

$$+ (E_f-E_{f_1})(\hat{f}_h-\hat{g}_h)E_{f_1}(\hat{f}_h)\} - (E_f-E_{f_1})(\hat{f}_h-\hat{g}_h)f$$

$$+ (f_1-f)E_{f_1}(\hat{f}_h-\hat{g}_h)]$$

$$= \sum_v (\tau_v-\tau_{1v})m^{-(3+v)}I_v \,,$$

where

$$(4.3.5)\quad I_v \equiv m^{3+v}\int_{C_v}\gamma_v(y)\{\int [(1-n^{-1})\{K((x-y)/h)E_f(\hat{f}_h-\hat{g}_h)(x)$$

$$+ A((x-y)/h)\, E_{f_1}(\hat{f}_h)(x)\} - h^{-1}\, A\{(x-y)/h\}f(x)]dx$$

$$- E_{f_1}(\hat{f}_h-\hat{g}_h(y))dy$$

$$= \int \mathcal{P}(y-v)\{\int [(1-n^{-1})\{K(x)E_f(\hat{f}_h-\hat{g}_h)(m^{-1}\, y + hx)$$

$$+ A(x)\, E_{f_1}(\hat{f}_h)(m^{-1}y + hx)\} - A(x)f(m^{-1}y + hx)]dx$$

$$- E_{f_1}(\hat{f}_h-\hat{g}_h)(m^{-1}y))dy.$$

If $y \in v + (-1/4,\ 1/4)$, if K vanishes outside $(-1/4,1/4)$, and if x is in the support of K, then for n so large that $hm < 1/4$,

$$E_f(\hat{f}_h-\hat{g}_h)(m^{-1}y + hx) = h^{-1}\int A\{(m^{-1}y + hx-w)/h\}f(w)dw$$

$$= h^{-1}\sum_u \tau_u \int_{C_u} A\{(m^{-1}y + hx-w)/h\}\gamma_n(w)dw$$

$$= h^{-1}\tau_v \int_{C_v} A\{(m^{-1}y + hx-w)/h\}\gamma_v(w)dw$$

$$= m^{-(2+v)}\tau_v \int A(w)\, \mathcal{P}\{y-v + hm(x-w)\}dw \,.$$

Similarly,

$$E_{f_1}(\hat{f}_h(m^{-1}y + hx) = m^{-(2+v)}\tau_{1v} \int K(w)\,\psi\{y-v+hm(x-w)\}dw,$$

$$= f(m^{-1}y + hx) = m^{-(2+v)}\tau_v\psi(y-v+hmx),$$

$$E_{f_1}(\hat{f}_h - \hat{g}_h)(m^{-1}y) = m^{-(2+v)}\tau_{1v} \int A(w)\psi(y-v-hmw)dw.$$

Substituting into (4.3.5) we obtain

$$E_v = m^{-(2+v)} \int J(y)([\int (1-n^{-1})\{\tau_v K(x) \int A(w)\psi(y+hm(x-w))dw$$

$$+ \tau_{1v} A(x) \int K(w)\psi(y+hm(x-w))dw\}$$

$$- \tau_v A(x)\psi(y+hmx)]dx - \tau_{1v} \int A(w)J(y-hmw)dw$$

$$= m^{-(2+v)}(\tau_v + \tau_{1v})J,$$

where J is as defined prior to the statement of the lemma. We may now deduce from (4.3.3) that

$$\eta(h,f,f_1) = n^{-1} J \sum_v (\tau_v^2 - \tau_{1v}^2) = n^{-1} J \sum_v (\tau_v - \tau_{1v}),$$

completing the proof of the lemma.

Arguments in Hall and Marron (1987b, pp. 172-176) now provide the following analogue of their Lemma 4.1: for each $\epsilon_1 > 0$ there exists $\epsilon_2 > 0$ and a sequence $\{f_{(n)}\}$ with $f_{(n)} \in \mathcal{F}$ such that, for large n,

$$(4.3.6)\quad P_{f_{(n)}}\{\min_{h\in(an^{-1/5},bn^{-1/5})} |\eta(h,f_{(n)},\tilde{f})| > \epsilon_2 n^{-1}|J|m^{1/2}\} > 1 - \epsilon_1.$$

Now,

$$(4.3.7)\quad J = \int \psi(y)[(1-n^{-1})$$

$$\iint A(w)K(x)\{\tfrac{1}{2}(hm)^2(w+x)^2\psi''(y) + (1/24)(hm)^4(w+x)^4\psi^{(4)}(y)\}dwdx$$

$$- \int A(w)\{\tfrac{1}{2}(hm)^2w^2\psi''(y) + (1/24)(hm)^4w^4\psi^{(4)}(y)\}dw]dy + O\{(hm)^5\}$$

$$= \tfrac{1}{2} t_2 (\int \psi\psi'')(hm)^2 + (1/24)t_4(\int \psi\psi^{(4)})(hm)^4 + O\{(hm)^5 + n^{-1}(hm)^2\}$$

where

$$t_2 \equiv \iint A(w)K(x)(w+x)^2 dwdx - \int A(w)w^2 dw = 0,$$

$$t_4 \equiv \iint A(w)K(x)(w+x)^4 dwdx - \int A(w)w^4 dw$$

$$= 6\{\int w^2 A(w)dw\}\{\int x^2 K(x)dx\} \neq 0.$$

In consequence,

$$J = (1/24)t_4 \int (\varphi'')^2 (hm)^4 + o\{(hm)^4\},$$

and so the desired result (4.3.3) follows from (4.3.6). This completes the proof of Theorem 3.2.

Remark 4.3.1: To extend this proof to the case of a kernel of order r, as discussed in Remark 3.2.4, the only changes required in the above proof are

$$m \equiv n^{-1/(2r + 1 + 2\nu)},$$

and the consequences of this.

References

Burkholder, D. L. (1973), "Distribution function inequalities for martingales," *Annals of Probability*, 1, 19-42.

Devroye, L. and Györfi, L. (1984). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Härdle, W., Hall, P. and Marron, J. S. (1988), "How far are automatically chosen regression smoothers from their optimum?," to appear with discussion, *Journal of the American Statistical Association*.

Hall, P. and Marron, J. S. (1987a), "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation," *Probability Theory and Related Fields*, 74, 567-581.

Hall, P. and Marron, J. S. (1987b), "On the amount of noise inherent in bandwidth selection for a kernel density estimator, " *Annals of Statistics*, 15, 163-181.

Hall, P. and Marron, J. S. (1987c), "Estimation of integrated squared density derivatives", *Statistics and Probability Letters*, 6, 109-115.

Marron, J. S. (1988), "Automatic smoothing parameter selection: A survey", North Carolina Intstitute of Statistics, Mimeo Series #1746.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Stone, C. J. (1980), "Optimal convergence rates for nonparametric estimators," *Annals of Statistics*, 8, 1348-1360.

Stone, C. J. (1982), "Optimal global rates of convergence of nonparametric regression," *Annals of Statistics*, 10, 1040-1053.