The Annals of Statistics
1988, Vol. 16, No. 4, 1390–1420

ONE-SIDED INFERENCE ABOUT FUNCTIONALS OF A DENSITY[1]

BY DAVID L. DONOHO

University of California, Berkeley

This paper discusses the possibility of *truly* nonparametric inference about functionals of an unknown density. Examples considered include: discrete functionals, such as the number of modes of a density and the number of terms in the true model; and continuous functionals, such as the optimal bandwidth for kernel density estimates or the widths of confidence intervals for adaptive location estimators. For such functionals it is not generally possible to make two-sided nonparametric confidence statements. However, one-sided nonparametric confidence statements *are* possible: e.g., "I say with 95% confidence that the underlying distribution has *at least* three modes." Roughly, this is because the functionals of interest are semicontinuous with respect to the topology induced by a distribution-free metric. Then a neighborhood procedure can be used. The procedure is to find the minimum value of the functional over a neighborhood of the empirical distribution in function space. If this neighborhood is a nonparametric $1 - \alpha$ confidence region for the true distribution, the resulting minimum value lowerbounds the true value with a probability of at least $1 - \alpha$. This lower bound has good asymptotic properties in the high-confidence setting α close to 0.

1. Introduction. Let F denote a cumulative distribution function, f and $f^{(k)}$ its density and derivatives (when they exist).

Nonlinear functionals of f and $f^{(k)}$ are ubiquitous in the literature of theoretical statistics. A brief catalog would include:

$\mathbf{L}_2$-norms

$$L_2^k(F) = \left(\int (f^{(k)})^2\right)^{1/2}; \tag{1.1}$$

$\mathbf{L}_p$-norms

$$L_p^k(F) = \left(\int |f^{(k)}|^p\right)^{1/p}, \tag{1.2}$$

$1 \le p < \infty$; and $\mathbf{L}_\infty$-norms

$$L_\infty^k(F) = \sup_x |f^{(k)}(x)|; \tag{1.3}$$

Received January 1985; revised February 1988.

[1]Research partially supported by a postdoctoral fellowship at the Mathematical Sciences Research Institute in 1983 and by an NSF Postdoctoral Fellowship.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62G15.

Key words and phrases. One-sided inference, bandwidth selection, density estimation, multimodality, semicontinuity of statistical functionals, subdifferentiability of statistical functionals, convex functionals.

Fisher information

$$I(F) = \int (f')^2/f; \tag{1.4}$$

and the (Shannon) negentropy

$$\Lambda(F) = \int f \log f. \tag{1.5}$$

Two integer-valued nonlinear functionals worth mentioning are mixture complexity $K(F)$ (the number of mixture terms needed to represent a density, see Section 4 below) and $M(F)$, the number of modes of F.

The contexts in which these occur are varied, but many of them are of distinct interest from the point of view of actual applications. For example, knowledge of L_2^0 and I allows one to set confidence intervals for the Hodges–Lehmann and adaptive location estimators, respectively. Knowledge of L_2^1 and L_2^2 allows one to choose the smoothing parameter for histograms and kernel density estimates. The problem of selecting model size $K(F)$ is of real practical interest. L_p^0 can be related to the power of certain tests for uniformity. And I and Λ have useful applications in deconvolution and projection pursuit.

Of course, in any applied setting, the value of such functionals at the unknown true distribution is unknown. It is therefore of interest to obtain, based on empirical data, information about the value of such functionals. However, there is an immediate logical difficulty with this. The functionals all depend on the true distribution having a well-defined density with a certain amount of regularity. For example, the existence of Fisher information requires that the distribution have a density whose square root is differentiable in quadratic mean. *Such an hypothesis is not empirically verifiable.* Thus one is in the position of attempting to obtain an empirical estimate of a quantity whose very existence is not subject to empirical test.

The point of this paper is that sense *can* be made out of what might otherwise seem a problem of circular reasoning. It is possible to make inferences about the values of functionals such as (1.1)–(1.4) with a truly nonparametric validity. The inferences, however, are of a restricted, one-sided nature. Thus one can make statements of the form, "I have 95% confidence that the number of modes of the underlying distribution is *at least* 3." These statements, if constructed by the procedure given below, have at least the indicated coverage probability, whatever be the underlying distribution (smooth, singular or discrete); they depend on no hypothesis about the underlying distribution and are totally empirical. Moreover, as the paper will show, the one-sided bounds set by this procedure are good in the sense that they are consistent (converge to the true value of the functional as the sample size increases) and converge (under regularity) at rapid rates.

In short, there are nonparametric lower bounds for many functionals of interest and these have good sampling properties. In most applications it is precisely the lower bounds that are of interest. Knowing that the number of modes or the number of mixture components is at least so big rules out many

crude models. Knowing that the Fisher information is at least so big sets an upper bound on the size of an asymptotic confidence interval and allows for conservative inference.

The contents of this paper are as follows. Section 2 discusses the problem of placing nonparametric upper bounds on functionals such as those discussed above and adapts an idea of Bahadur and Savage to show that in nonparametric settings it is not generally possible to do so. Section 3 gives an explicit construction of lower bounds for lower semicontinuous functionals. Called the neighborhood procedure, it involves solving an optimization problem: Find the minimum value of the functional over a neighborhood of the empirical distribution and report that value as a lower bound on the value of the functional at the true distribution. The nonparametric validity of the lower bound is immediate if the neighborhood is a nonparametric confidence region for the true distribution. That the lower bound is good, i.e., is not much smaller than the true value, depends on asymptotic analysis. The high confidence case is considered, where the coverage probability is near 1 in large samples. Using elementary properties of the lower envelope of a semicontinuous functional, consistency holds (i.e., the lower bound tends to the true value) at every point of semicontinuity and a speed of convergence holds at every semi-Lipschitz point.

Applications are given to the functionals described above in Sections 4 and 5, where lower semicontinuity and lower semi-Lipschitz properties are derived. Section 6 gives some applications; the nonparametric validity of the lower bounds is useful for constructing consistent empirical procedures with conservative properties. Section 7 considers some generalizations of the present work and Section 8 discuses some broader issues it raises.

Some notation. J refers to an unspecified functional such as those mentioned above (L_2^k, I, K, M, etc.). F, G and H represent distribution functions and f, g and h ordinary functions (generally densities). $\mathbf{P}$ is the set of all probabilities on the real line $\mathbf{R}$. $\mathbf{L}_p$ denotes the traditional space of locally integrable functions with integrable pth powers. Occasionally, notation will be abused by saying that a distribution F belongs to $\mathbf{L}_p$ or some other function space. The intended meaning is that F has a density that belongs to that space. Φ denotes the Gaussian distribution and ϕ denotes the Gaussian density. $F * G$ denotes the convolution of distributions F and G.

2. Nonexistence of upper confidence bounds. Bahadur and Savage (1956) showed that in general there is no way of setting confidence intervals for the mean which will give nonparametric validity. This section will show that this is true of many other functionals as well—especially the ones of interest for this paper.

The key idea is that near any distribution of interest, there are empirically indistinguishable distributions (indistinguishable at a given sample size) where the functional takes on arbitrarily large values. It is not possible then in principle to place an upper bound on the value of the functional *solely* in terms of empirical data: untestable a priori assumptions would be necessary.

Formalizing this idea requires some topological notions. Let $\mathbf{F}$ be the set of distributions in which the true distribution is known to lie: examples being the set of all distributions and the set of all distributions in some neighborhood of the normal distribution. Equip $\mathbf{F}$ with the testing topology as follows. Let τ denote the testing metric

$$\tau(F,G) = \sup_{0\le\Psi\le 1}\left|\int\Psi\,dF - \int\Psi\,dG\right|; \tag{2.1}$$

the supremum is over all measurable functions on $\mathbf{R}$ with values in $[0,1]$. This distance is usually called the variation distance. We call it the testing distance to emphasize the following basic fact: This distance is small if and only if the best test between F and G has a poor chance of distinguishing the two [Le Cam (1973) and (1986), Chapter 4]. Indeed, if $\tau(F,G)=\varepsilon$, then any test based on a single observation has a sum of type I and type II errors of at least $1-\varepsilon/2$. Moreover, by Lemma A.1, if $\tau(F,G)\le 2(1-(1-\varepsilon^2/8)^{1/n})$, then the best test based on n i.i.d. observations has a sum of type I and type II errors of at least $1-\varepsilon/2$. Thus, at any given sample size n, if $\tau(F,G)$ is small enough, F and G are difficult to tell apart based on n observations.

Let J be the functional of interest and let $\mathbf{J}\subset\mathbf{R}$ be its range (e.g., $\mathbf{R}, \mathbf{R}^+, \mathbf{Z}, \mathbf{Z}^+$). Give $\mathbf{J}$ the topology it inherits from $\mathbf{R}$ and give the product space $\mathbf{F}\times\mathbf{J}$ the product topology. We suppose that J is well defined on a dense subset $\mathrm{dom}(J)$ of $\mathbf{F}$ (e.g., the set of densities with a finite $\mathbf{L}_2$-norm, etc.).

Now define the graph of J over $\mathbf{F}$ in the obvious way as a set of ordered pairs:

$$\mathrm{graph}(J,\mathbf{F}) = \{(F, J(F)) : F\in\mathrm{dom}(J)\cap\mathbf{F}\} \tag{2.2}$$

and let the epigraph be everything "above the graph":

$$\mathrm{epigraph}(J,\mathbf{F}) = \{(F, j) : F\in\mathrm{dom}(J)\cap\mathbf{F} \text{ and } j\ge J(F)\}. \tag{2.3}$$

DEFINITION. J is said to satisfy the dense graph condition (DGC) if

$$\mathrm{graph}(J,\mathbf{F}) \text{ is dense in } \mathrm{epigraph}(J,\mathbf{F}). \tag{2.4}$$

The DGC cannot generally be satisfied by a τ-continuous functional, since continuity implies a closed graph and DGC implies something at the other extreme. A functional satisfying DGC is badly discontinuous: In any neighborhood of a point, it takes on arbitrarily large values. As a result, there is no useful upper confidence bound valid uniformly over $\mathbf{F}$.

THEOREM 2.1. *Let $X_1, X_2, \ldots, X_n$ be a random sample from some distribution $F\in\mathbf{F}$. Let C_n be a confidence interval determined by the value of this sample. Let J satisfy DGC over $\mathbf{F}$. If somewhere in $\mathbf{F}$, C_n asserts a nontrivial upper bound, then somewhere in $\mathbf{F}$ the coverage probability of C_n is 0. Formally, if $B=\sup\{j: j\in\mathbf{J}\}$ (e.g., $B=+\infty$) and if*

$$\sup_{F\in\mathbf{F}} P_F\{B\notin C_n\} = 1,$$

then

$$\inf_{F \in \mathbf{F}} P_F\{J(F) \in C_n\} = 0.$$

PROOF. Let $\delta > 0$ and choose a distribution $F_\delta \in \mathbf{F}$ at which

$$P_{F_\delta}\{B \notin C_n\} \geq 1 - \delta.$$

The upper endpoint of C_n, $B_n = \sup\{x \in C_n\}$, is smaller than B with F_δ-probability at least $1 - \delta$. Let β be a $1 - 2\delta$ quantile of the distribution of B_n, i.e.,

$$P_{F_\delta}\{B_n \leq \beta\} \geq 1 - 2\delta. \tag{2.5}$$

β can be chosen less than B because of the last remark. Now let G_δ be a distribution so close to F_δ that

$$\tau(F_\delta, G_\delta) \leq 2\left(1 - \left(1 - \delta^2/8\right)^{1/n}\right) \tag{2.6}$$

and (by the DGC)

$$J(G_\delta) > \beta.$$

Consider the function Ψ_n on $\mathbf{R}^n$ which is the indicator of the event $\{B_n \leq \beta\}$. Since $\{B_n \leq \beta\}$ implies $\{J(G_\delta) \notin C_n\}$ we have

$$P_{G_\delta}\{J(G_\delta) \notin C_n\} \geq P_{G_\delta}\{B_n \leq \beta\} = E_{G_\delta^{(n)}}\Psi^{(n)}, \tag{2.7}$$

where $G_\delta^{(n)}$ is the product measure on $\mathbf{R}^n$ obtained from G_δ. We wish to show that the expectation on the right-hand side is large. We now invoke Lemma A.1, which shows that the bound (2.6) on the one-dimensional testing distance implies that

$$\tau^{(n)}\left(G_g^{(n)}, F_\delta^{(n)}\right) \leq \delta, \tag{2.8}$$

where $F_\delta^{(n)}$ is the product measure on $\mathbf{R}^n$ obtained from F_δ and $\tau^{(n)}$ denotes the n-dimensional testing distance. The lemma is based on inequalities between variation and Hellinger distance; see Le Cam [(1973), (1986)] or Pitman (1979). Now (2.8) yields

$$|E_{G_\delta^{(n)}}\Psi_n - E_{F_\delta^{(n)}}\Psi_n| \leq \delta,$$

which, plugged into (2.7), yields

$$\begin{aligned} P_{G_\delta}\{J(G_\delta) \notin C_n\} &\geq E_{F_\delta^{(n)}}\Psi_n - \delta \\ &= P_{F_\delta}\{B_n \leq \beta\} - \delta \geq 1 - 3\delta, \end{aligned}$$

the last inequality following from (2.5). As $\delta > 0$ was arbitrary, this completes the proof. □

While functionals satisfying the DGC may seem wildly pathological, they are common when $\mathbf{F}$ is truly nonparametric. Intuitively, a truly nonparametric family of distributions has the property that, when it contains a distribution F, it also contains all other distributions which cannot be reliably distinguished from F at a given sample size based on any empirical tests. No distribution

which produces samples very much like those actually seen should be ruled out a priori.

Following this line of reasoning, we arrive at the requirement that, for **F** to be nonparametric, it should contain at least a small τ neighborhood around essentially every point. Formally,

DEFINITION. The family **F** is *strongly nonparametric* if its τ-interior is dense in **F**.

Under this definition, the following are strongly nonparametric families:

1. the set of all distributions,
2. the set of all distributions satisfying a goodness-of-fit test, $\{F\colon \delta(F_n; F) \leq d\}$, where δ is a consistent goodness-of-fit statistic such as the Cramér–von Mises or Kolmogorov.

The following are not strongly nonparametric:

3. a finite-dimensional parametric model,
4. a set of distributions all satisfying a quantitative regularity condition, such as $\{F\colon L_2^k(F) < 128.364\}$,
5. a set of distributions with common compact support, e.g., $[0, 1]$.

While (3)–(5) may seem appealing models in many settings, it should be emphasized that if one claims to know that the unknown distribution lies in such a set, one is using information external to the sample to do so.

Under this definition, nonparametric upper bounds are unavailable for the functionals of Section 1.

THEOREM 2.2. *If* **F** *is strongly nonparametric, the following functionals satisfy DGC over* **F**:

1. *the number of modes,* $M(F)$,
2. *the mixture complexity,* $K(F)$,
3. *any* $\mathbf{L}_p$*-norm of any derivative of the density,*
4. *Fisher information,* $I(F)$,
5. *negentropy,* $\Lambda(F)$.

This proof is given in the Appendix. Intuitively, these functionals are all measures of the wiggliness of a density. But it is possible to make the density arbitrarily more wiggly via an arbitrarily small τ perturbation. Figure 1 shows how a density can be perturbed in this way.

The theorem says that data alone cannot rule out the possibility of wiggles in the density at a scale too narrow to have been observed at the given sample size. Consequently, measures of wiggliness cannot be upperbounded based on the data alone.

What about lower bounds? It is clear from the above that if both J and $-J$ satisfy DGC, lower bounds are not available either. Using this observation, one

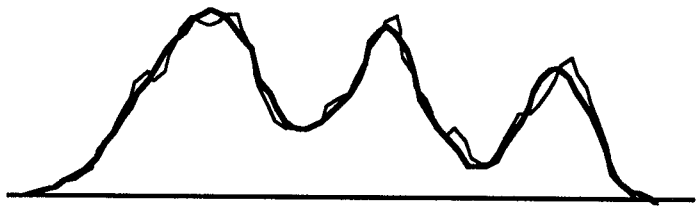

FIG. 1. *A density with three modes and a small perturbation of it with nine modes.*

can show that if **F** is large enough, there are no nonparametric lower or upper bounds for functionals such as the mean. This was the point of Bahadur and Savage's original work; their argument is at the core of Theorem 2.1. The present abstract approach is adapted to general functionals and can be used to show (surprisingly) that neither upper nor lower bounds are available for the negentropy Λ. One can also show that the only nonparametric bounds for functionals such as the density indicator,

$$D(F) = \begin{cases} 1, & \text{if } F \text{ has a density,} \\ 0, & \text{otherwise,} \end{cases}$$

are the trivial ones, i.e.,

$$D(F) \leq 1,$$
$$D(F) \geq 0.$$

However, the paper will now focus on positive results.

Actually, the functionals introduced in Section 1 (except for the negentropy Λ) *do* admit of lower bounds. Intuitively, this is because in a neighborhood of a smooth density one can find densities which are much wigglier, but none that are much smoother. Mathematically, this is because these functionals (despite their pathological graphs) have *closed* epigraphs. The significance of this for inference is apparent from Figure 2. Suppose that by empirical or other means we are able to conclude that the true distribution generating the data must lie in a small neighborhood N. This information gives a lower bound to $J(F)$—namely, the least value of J in that neighborhood; this is indicated in Figure 2.

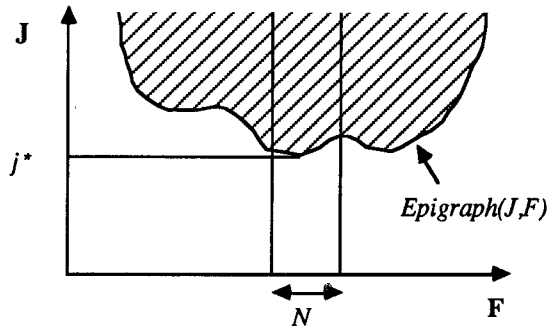

FIG. 2. *A functional with closed epigraph. The information that the true F lies in a neighborhood N provides the lower bound j* on the value of J(F).*

A functional with closed epigraph is called *lower semicontinuous*. The basic idea just described is that it should be possible to set lower confidence bounds for lower semicontinuous functionals. This idea is developed in detail in the next section.

3. Construction of lower confidence bounds. The construction of lower confidence bounds, as just indicated, depends on a topological property of J. It will be convenient to switch gears somewhat and focus on a weaker topology than the testing topology.

The Kolmogorov, or sup-norm, is defined by

$$|F - G| = \sup_t |F(t) - G(t)|. \tag{3.1}$$

This distance makes sense between any distributions, is bounded by 1, is strictly smaller than the testing distance (i.e., gives bigger balls for the same nominal radius) and, most importantly, the Kolmogorov distance between the empirical and the true distribution is distribution-free. Let F_n denote the empirical distribution of a sample from F and U_n denote the empirical of a sample from the uniform distribution U on $[0, 1]$. Then for each $\varepsilon > 0$,

$$\text{prob}\{|F_n - F| \leq \varepsilon\} \geq \text{prob}\{|U_n - U| \leq \varepsilon\} \tag{3.2}$$

with equality for every continuous F. This sort of property is fundamental to nonparametric constructions.

The functional J will be said to be norm lower semicontinuous ($|\cdot|$-l.s.c.) if, for every sequence F_n of distributions satisfying $|F_n - F| \to 0$, we have

$$\liminf_{n \to \infty} J(F_n) \geq J(F). \tag{3.3}$$

Among other things, this requires that at every point where the definition of J could be ambiguous, J takes the least value. As indicated above, (3.3) is fundamental to our construction of lower confidence bounds.

3.1. *The lower envelope of a functional.* Let J be an arbitrary functional and define the ε-lower envelope of J by

$$J(F; \varepsilon) = \inf\{J(G): |G - F| \leq \varepsilon\}. \tag{3.4}$$

That is, $J(F; \varepsilon)$ is a lower bound on the value of $J(G)$ given the information that $|F - G| \leq \varepsilon$. This lower bound has two parameters, F and ε, and regular behavior with respect to each of these.

LEMMA 3.1 (Behavior in ε). *$J(F; \cdot)$ is a monotone decreasing function of ε. If J is lower semicontinuous, $J(F; \cdot)$ takes the lower value at jumps, and its value at $\varepsilon = 0$ is just $J(F)$.*

PROOF.

$$J(F; \varepsilon) \text{ is monotone decreasing in } \varepsilon \tag{3.5}$$

because the neighborhoods $\{F: |F - G| \leq \varepsilon\}$ increase with increasing ε.

$$J(F; \varepsilon) \text{ is lower semicontinuous in } \varepsilon \tag{3.6}$$

if J is lower semicontinuous in F. This is immediate from the alternate definition of lower semicontinuity, namely that the sets $\{G\colon J(G) > j\}$ are open. If $J(F; \varepsilon)$ took the larger value at a jump, we could show that such a set was not open. Define

$$J(F; 0) = \lim_{\varepsilon \to 0} J(F; \varepsilon). \tag{3.7}$$

This makes sense by (3.5). Combining (3.7) and (3.3) we conclude that if J is $|\cdot|$-l.s.c.,

$$J(F; 0) = J(F). \tag{3.8}$$ □

LEMMA 3.2 (Behavior in F). *If $J(\cdot)$ has any of these properties, so does $J(\cdot; \varepsilon)$:*

(3.9) *semicontinuity*

(3.10a) *convexity*

(3.10b) *translation invariance*

(3.10c) *scale invariance*

(3.10d) *decrease under convolution.*

PROOF. (3.9) follows just as (3.6). (3.10b) and (3.10c) both follow from the scale and translation invariance of the Kolmogorov neighborhoods: $|G - F| \le \varepsilon$ if and only if $|G((\cdot - t)/s) - F((\cdot - t)/s)| \le \varepsilon$. (3.10a) and (3.10d) are due to the convexity of the Kolmogorov neighborhoods: $(1 - \delta)N_\varepsilon(F_0) + \delta N_\varepsilon(F_1) \subset N_\varepsilon((1 - \delta)F_0 + \delta F_1)$, where $N_\varepsilon(F) = \{G\colon |G - F| \le \varepsilon\}$. □

In short, $J(F; \varepsilon)$ is at least a regular of functional as F as J. Often it is considerably more regular continuous or even Lipschitz continuous.

LEMMA 3.3 [Bounds on $J(F; \varepsilon)$]. *For every G and F,*

$$J(G, \varepsilon) \ge J(F; \varepsilon + |F - G|) \tag{3.11}$$

and if $|F - G| \le \varepsilon$,

$$J(F) \ge J(G, \varepsilon) \ge J(F, 2\varepsilon). \tag{3.12}$$

PROOF. Since an $\varepsilon + |F - G|$ ball around F must contain an ε ball around G (use the triangle inequality), the infimum of J over the larger ball cannot be larger than the infimum over the smaller. This establishes (3.11). To get (3.12), combine (3.11) and (3.4). □

3.2. *Statistical applications of the lower envelope.* Let $X_1, \ldots, X_n$ be i.i.d. according to some unknown distribution function F. Let F_n be the empirical distribution function $F_n(x) = \#\{i\colon X_i \le x,\ 1 \le i \le n\}/n$. The basic idea for setting a lower confidence bound on J is to let ε_n be some fixed positive number and use $J(F_n; \varepsilon_n)$ as a lower bound for $J(F)$.

To see that this can work, note that

$$P_F\{J(F_n; \varepsilon_n) \le J(F)\} \ge P_F\{|F_n - F| \le \varepsilon_n\}$$

because, by (3.4), the event $\{|F_n - F| \le \varepsilon_n\}$ implies $\{J(F_n; \varepsilon_n) \le J(F)\}$. On the other hand the probability on the right-hand side is distribution free, by (3.2). Thus we have:

PROPOSITION. *If ε_n is the $1 - \alpha$ quantile of the distribution of $|U - U_n|$, then*

$$P_F\{J(F_n; \varepsilon_n) \le J(F)\} \ge 1 - \alpha \tag{3.13}$$

whatever F is. In other words, $J(F_n; \varepsilon_n)$ is a lower confidence bound for $J(F)$ with a nonparametric coverage probability of at least $1 - \alpha$.

It may help to think of this procedure in terms of neighborhoods. One is putting down a $1 - \alpha$ confidence region in function space for the true distribution. The value $J(F_n; \varepsilon_n)$ is the least value of the functional J over this region. With probability $1 - \alpha$, the true distribution lies in this region, and when this happens, $J(F_n; \varepsilon_n)$ is less than $J(F)$. Consequently the lower bound has the advertised coverage probability.

Of course, having a nonparametric coverage probability is only part of the story. The trivial bound $-\infty$ has an excellent nonparametric coverage probability. One also wants to have the slack $J(F) - J(F_n; \varepsilon_n)$ be small while still satisfying (3.13).

It is difficult to study the properties of this slack, for general J, except asymptotically. That is, instead of considering one fixed sample size, one considers a sequence of problems of size $n \to \infty$. Then, allowing ε_n to depend on n, one attempts to understand the behavior of $J(F) - J(F_n; \varepsilon_n)$ for large n.

There are two interesting ways that ε_n can depend on n. First, one can have

$$P_F\{|F_n - F| \le \varepsilon_n\} \to 1; \tag{3.14}$$

this is the high confidence setting where we want a very conservative lower bound on J. Alternatively, one might have

$$P_F\{|F_n - F| \le \varepsilon_n\} \to 1 - \alpha; \tag{3.15}$$

this is the moderate confidence setting. As it turns out, the first setting is much simpler to analyze than the second.

3.3. *Analysis in the high confidence setting.* A slight strengthening of (3.14) will be quite useful. Say that ε_n goes to 0 *slowly enough* if $\varepsilon_n \to 0$

$$P_F\{|F_n - F| \le \varepsilon_n \text{ for almost all } n\} = 1 \tag{3.16}$$

(below, "almost all n" means "for all but finitely many n" and will be abbreviated a.a.n.). By the Chung–Smirnov law of the iterated logarithm, a sufficient condition for (3.16) is

$$\liminf_{n\to\infty} \varepsilon_n \sqrt{\frac{n}{\log\log n}} > 2^{-1/2}; \tag{3.17}$$

see Csörgő and Révész [(1981), page 157]. In other words, if ε_n goes to 0 a little more slowly than $n^{-1/2}$, the event $\{|F_n - F| \le \varepsilon_n\}$ will happen for all but finitely many n, with probability 1. But then (3.12) can be used with F_n playing the role of G, and we conclude:

THEOREM 3.4. *If* $\varepsilon_n \to 0$ *slowly enough,*

$$J(F) \ge J(F_n; \varepsilon_n) \ge J(F; 2\varepsilon_n) \quad a.a.n. \tag{3.18}$$

with probability 1.

This bound is useful because it brackets the random quantity $J(F_n; \varepsilon_n)$ between two deterministic ones. It has two main corollaries.

COROLLARY (Consistency). *If* J *is norm-l.s.c. and if* $\varepsilon_n \to 0$ *slowly enough, then* $J(F_n; \varepsilon_n)$ *is universally consistent*:

$$J(F_n; \varepsilon_n) \to_{a.s.} J(F)$$

and converges from below:

$$J(F_n; \varepsilon_n) \le J(F) \quad a.a.n.,$$

with probability 1.

PROOF. By (3.7), $J(F; 2\varepsilon_n) \to J(F)$. Now use (3.18). □

DEFINITION. J is said to be lower semi-Lipschitz of order ρ at F if

$$J(F) - J(F; \varepsilon) = O(\varepsilon^\rho). \tag{3.19a}$$

COROLLARY (Rates). *If* (3.19*a*) *holds and if* $\varepsilon_n \to 0$ *slowly enough,*

$$J(F) - J(F_n; \varepsilon_n) = O_{a.s.}(\varepsilon_n^\rho). \tag{3.19b}$$

PROOF. If (3.19a) holds, then the same is true with 2ε in place of ε. Now use (3.18). □

These two corollaries relate statistical properties of $J(F_n; \varepsilon_n)$ (convergence and rates of convergence) to analytical properties of the functional J (semicontinuity and semi-Lipschitz bounds). In the cases where J is either integer-valued or convex, these analytical properties are not hard to establish.

4. Discrete functionals. This section applies the lower bounds technology of the last section to two functionals with discrete values: the number of modes of a distribution and the order of the true model.

4.1. *Lower semicontinuity.* Consider first the functional M counting the number of modes. There is some ambiguity in the definition of this functional: For example one might argue that the uniform distribution U has only 1 mode or that it has infinitely many. Of course, the lower semicontinuous version of the functional will have only 1 mode.

To define such a version, use a device introduced by Silverman (1981). Let F be an arbitrary (e.g., discrete) distribution, and Φ_h denote the Gaussian distribution with variance h. Consider the convolution $\Phi_h * F$; this is a distribution with smooth derivatives of any order (since $(d/dt)^k(\Phi_h * F) = [(d/dt^k)\Phi_h] * F$). In particular, it has a smooth second derivative. It therefore makes sense to define the number of modes of $\Phi_h * F$, $M(\Phi_h * F)$ as the number of downcrossings in the second derivative of this curve. Silverman's observation is that $M(\Phi_h * F)$ is monotone decreasing in h. This uses two special properties of the Gaussian distribution: the variation-diminishing property [Karlin (1968)] and the semigroup property $\Phi_{h_1} * \Phi_{h_2} = \Phi_{h_1+h_2}$. In any event, we can define

$$M(F) = \lim_{h \to 0} M(\Phi_h * F) \tag{4.1}$$

and this makes sense due to the indicated monotonicity in h.

LEMMA 4.1. *$M(F)$ is norm lower semicontinuous.*

PROOF. Fix $h > 0$. Let $F_n \to F$ in norm and let $f'_{h,n}$ and f'_h denote the second derivatives of $\Phi_h * F_n$ and $\Phi_h * F$. An integration by parts gives

$$|f'_{h,n} - f'_h| \le |F_n - F| \int |\Phi_h^{(3)}|,$$

so the derivative curves converge uniformly. It is then clear that eventually $f'_{h,n}$ must have as many downcrossings as f'_h:

$$\liminf_{n \to \infty} M(\Phi_h * F_n) \ge M(\Phi_h * F),$$

but by monotonicity $M(F_n) \ge M(\Phi_h * F_n)$ and so

$$\liminf_{n \to \infty} M(F_n) \ge M(\Phi_h * F);$$

so now the result follows from (4.1). □

Consider now a functional related to model selection. Let $\{G_\theta\colon \theta \in \Theta\}$ be a parameterized family of distributions and let $K(F)$, the mixture complexity of F, be the least number of G-components necessary to exactly represent F. Formally

$$K(F) = \inf\left\{k\colon F = \sum_{i=1}^{k} \beta_i G_{\theta_i}\right\}, \tag{4.2}$$

and $K = +\infty$ if F has no finite representation.

The family $\{G_\theta\}$ will be said to be *closed* if the set $\mathbf{G}_k = \{\sum_{i=1}^k \beta_i G_{\theta_i}\colon \Sigma\beta_i = 1, \beta_i \ge 0, \theta_i \in \Theta\}$ is closed under norm convergence for each k. The following analytic criterion can be used to establish closure.

LEMMA 4.2. *If norm convergence of $\int G_\theta \, d\mu_n(\theta) \to \int G_\theta \, d\mu(\theta)$ implies weak convergence $\mu_n \Rightarrow \mu$, then the family G_θ is closed.*

PROOF. If μ is a discrete measure with k points of support, weak convergence implies that eventually every member of the sequence μ_n will have at least k points of support. □

This lemma can be used to show that a translation family is closed if the Fourier transform of its kernel never vanishes. Thus the Gaussian translation family $G_\theta = \Phi(\cdot - \theta)$ is closed. Similar conclusions for the normal scale family follow from a condition on the Mellin transform. Any totally positive family will work as well.

Once closure is established, semicontinuity is easy:

LEMMA 4.3. *If the family $\{G_\theta\}$ is closed, the $K(F)$ is norm lower semicontinuous.*

PROOF. If the family is closed, then if $K(F) = k$, F is at a distance $\delta > 0$ from $\mathbf{G}_{k-1}$. Consequently, if $F_n \to F$, as soon as $|F_n - F| < \delta$, $K(F_n) > k - 1$. (3.3) follows. □

4.2. *Rate of convergence.* For all lower semicontinuous functionals the theory of Section 3 implies that if $\varepsilon_n \to 0$ slowly enough, the slack in the lower bound $J(F) - J(F_n; \varepsilon_n)$ tends almost surely to 0. For discrete functionals, such as M and K, something much stronger is true. By Lemma 3.1, $J(F; \varepsilon)$ is a nonincreasing integer valued function of ε, constant except for jumps, and at each jump takes the lower value. The picture is as in Figure 3.

Thus the graph of $J(F; \varepsilon)$ versus ε is a stairstep, and if $J(F) < \infty$, it has a last step extending from 0 out to some value $\varepsilon^*(F)$. For all ε in the range $[0, \varepsilon^*(F)]$, $J(F; \varepsilon) = J(F; 0) = J(F)$. Comparing with (3.18) one sees that as soon as $2\varepsilon_n < \varepsilon^*$, one has $J(F_n; \varepsilon_n) = J(F)$ whenever the high probability event $|F_n - F| \le \varepsilon_n$ occurs. There are standard bounds for the probability of this event.

THEOREM 4.4. *Every integer valued l.s.c. functional can be lower bounded with a slack tending to 0 at a rate that is ultimately exponential. Formally, at*

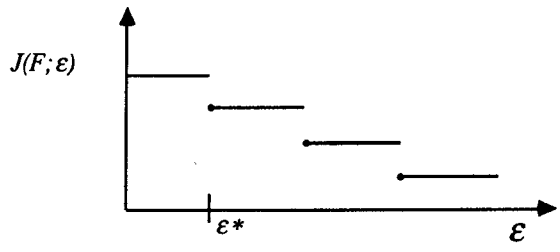

FIG. 3. *$J(F; \varepsilon)$ for an integer-valued functional.*

each F where $J(F) < \infty$, there is an $\varepsilon^*(F) > 0$ such that

$$P_F\{J(F_n; \varepsilon_n) < J(F)\} \leq 4\sqrt{2} \exp\{-2n(\varepsilon^* - \varepsilon_n)_+^2\}. \tag{4.3}$$

PROOF. Using (3.11), one sees that if $\varepsilon_n + |F_n - F| \leq \varepsilon^*$, then $J(F_n; \varepsilon_n) \geq J(F)$. Consequently, the probability on the left is no greater than the probability that $|F_n - F| > (\varepsilon^* - \varepsilon_n)_+$. The right side of (4.3) is simply an exponential bound on the probability of this event. The particular bound is an application of the Dvoretzky–Kiefer–Wolfowitz inequality with the constant $c = 4\sqrt{2}$ established by Hu (1985). Compare Shorack and Wellner [(1986), pages 354–356]. □

Of course, $\varepsilon^*(F)$ is unknown. If it were possible to bound ε^* away from 0 based on empirical observations, then one could construct an upper confidence bound on $J(F)$ using (4.3). But such upper bounds are not generally available, as Section 2 showed.

The theorem does not give information about the probability that $J(F_n; \varepsilon_n) > J(F)$; this depends on how the user has chosen the sequence ε_n. If ε_n goes to 0 very slowly, this probability goes to 0 rapidly; if ε_n goes to 0 rapidly [say as $(\log\log n/n)^{1/2}$], this goes to 0 slowly.

5. Convex functionals. This section applies the lower bound technique to the analysis of L_p^k and I.

5.1. *Lower semicontinuity.* The functionals of interest are not well defined on the space of all probability measures. For example, the Fisher information may be written as $I = 4\int((\sqrt{f}(x))')^2$. Now suppose that f is the density of the uniform $(0, 1)$ distribution. Then, one might argue that, as $d/dx\sqrt{f(x)} = 0$ a.e., we should put $I = 0$ in this case. On the other hand, as we know that one can estimate the location of the uniform $(\theta, 1 + \theta)$ family at a rate faster than $n^{-1/2}$, we might argue from the statistical interpretation of Fisher information that at the uniform distribution we must put $I = +\infty$.

Happily, there is no ambiguity about the value of $J(F)$ when F has a very smooth density, e.g., an infinitely differentiable density, for example a density containing a Gaussian convolution component. As the class of such smooth densities is weakly dense in the set of all densities, we can use the definition there to define an extension to all F which is lower semicontinuous and hence to which our inference methods apply. The Gaussian convolution technique we employ was also used by Port and Stone (1974).

LEMMA 5.1. *Let J be convex, bounded below and translation invariant. Suppose that $J_h(\cdot) = J(\Phi_h * \cdot)$ is continuous. Then the definition*

$$J_0(F) = \lim_{h \to 0} J(\Phi_h * F) \tag{5.1}$$

always makes sense and defines a convex lower semicontinuous functional.

PROOF. By the convexity of J and the semigroup property of Φ_h,

$$J(\Phi_{h_1+h_2} * F) = J(\Phi_{h_1} * \Phi_{h_2} * F) \leq \int J(\Phi_{h_2} * F(\cdot - t))\, d\Phi_{h_1}(t).$$

Using translation invariance of J this gives

$$J(\Phi_{h_1+h_2} * F) \leq J(\Phi_{h_2} * F)$$

so

(5.2) $J(\Phi_h * F)$ is monotone decreasing in h.

Consequently, the limit in (5.1) is well defined and exhibits J_0 as the pointwise supremum of a set of continuous functionals. Hence J_0 is l.s.c.; convexity follows by simple algebra. □

To apply this lemma to a particular J of interest one needs

(5.3) to show that the functionals $J_h(\cdot)$ are continuous;

(5.4) to show that $J_0 = J$ on the set where we agree that J is well defined.

Then one obtains a convex lower semicontinuous extension J_0 of J to all of P.

For the case $J = L_p^k$ where $1 \leq p < \infty$, the natural domain of this functional is $\mathbf{L}_p^k$, the $\mathbf{L}_p$ Sobolev space of order k [Adams (1975)]. Then we have

LEMMA 5.2. *Let $k \geq 0$, $1 \leq p < \infty$.* (a) *If $\{F_n\}$ converges in norm to F, then*

$$\phi_h * F_n \to \phi_h * F \quad \text{in } \mathbf{L}_p^k \text{ as } n \to \infty.$$

(b) *If $f \in \mathbf{L}_p^k$, then*

$$\phi_h * f \to f \quad \text{in } \mathbf{L}_p^k \text{ as } h \to 0.$$

The proof is based on standard ideas in Sobolev spaces and is in the Appendix. Since L_p^k is a continuous functional on $\mathbf{L}_p^k$, part (a) implies that J_h is norm continuous; part (b) implies that $J_0 = J$ on $\mathbf{L}_p^k$. For the case $p = \infty$ one obtains that $J_0 = J$ only if the supremum in (1.3) interpreted as an essential supremum.

Now let $J = I$. The domain of I is the set of distributions with root densities differentiable in quadratic mean. The relevant tool is

LEMMA 5.3. (a) *Let $\{F_n\}$ be a sequence of distributions converging in norm to F. Then*

$$\sqrt{\phi_h * F_n} \to \sqrt{\phi_h * F} \quad \text{in } \mathbf{L}_2^1 \text{ as } n \to \infty.$$

(b) *Let F have a root density $\sqrt{f} \in \mathbf{L}_2^1$. Then*

$$\sqrt{\Phi_h * f} \to \sqrt{f} \quad \text{in } \mathbf{L}_2^1 \text{ as } h \to 0.$$

There is quite a bit one can say about this method of defining these functionals. For example, the use of the Gaussian kernel Φ_h is not really necessary; defining $J_1(F) = \sup_{h>0} J(K_h * F)$, where $K_h(t) = K(t/h)$, one can show that $J_0 = J_1$ for the functionals J of interest, whenever K is reasonably smooth. Moreover, one can show (see Lemmas A.2 and A.3 below) that under this

approach, the condition $J_0 < \infty$ is equivalent, if $J = L_p^k$, to $F \in \mathbf{L}_p^k$ and if $J = I$, to F having a root density that is differentiable in quadratic mean.

For a concrete example, in the case of Fisher information, we get that $I_0 = +\infty$ at the uniform distribution, as it should on statistical grounds.

Let us discuss the statistical interpretation of these extensions. One can show that if $J = L_p^k$ or $J = I$, these extensions have the property that $J_0(F)$ is the smallest possible limit point of all sequences $\{J(F_n)\}$ with $\{F_n\}$ a sequence of distributions with C^∞ densities converging weakly to F. Thus J_0 is a kind of smallest extension of J; and statistically, this means it is a kind of most conservative version of J. Since these functionals are generally used in lower bounds, conservativeness means that the extended version may often be used in place of the original one and that this replacement is generally valid. For example, work of Donoho and Liu (1987) introduces a functional called the geometric information; this can be shown equivalent to the l.s.c. extension of Fisher information used here. They show that if one uses this functional in place of Fisher information, the information inequality (Cramér–Rao lower bound) still holds—without any smoothness hypothesis on the family to which it is applied. Devroye and Györfi [(1985), pages 77–79] use the lower semicontinuous extension of L_2^2 and show that if one uses this functional in place of L_2^2, bounds for the risk of kernel density estimation which were developed for the case where f has two nice derivatives remain valid for all f.

In short, the lower semicontinuous extension of these statistical functionals off the set of smooth densities makes good statistical sense. See also the two examples of Section 6. So below, we treat the extended version J_0, as the functional of interest.

5.2. *Rates of convergence.* The lower semicontinuity of J implies that the lower confidence bound $J(F_n; \varepsilon_n)$ converges to J as $n \to \infty$. The rate of this convergence is of interest, as one obviously wants the slack $J(F) - J(F_n; \varepsilon_n)$ to be small. Using the result (3.19) at the end of Section 3, it is obviously of interest to find conditions under which $\rho = 1$ in (3.19a). For convex functionals this is actually easy to do.

DEFINITION. Let J be convex. A *support functional* for J at a point F is a linear functional λ_F satisfying

$$J(G) - J(F) \geq \lambda_F(G - F).$$

If J has a linear Gateaux derivative J' at F, then J' is such a support functional. In general, a convex function has a set of subgradients ∂J at each F —each subgradient is a support functional.

DEFINITION. The dual norm $|\lambda|_*$ of a linear functional λ on the space of differences of distributions is

$$|\lambda|_* = \sup_{F,G} \frac{\lambda(G - F)}{|G - F|}. \tag{5.5}$$

THEOREM 5.4. *If the convex functional J has a support functional λ_F at F with finite dual norm, then* (3.19a) *holds with* $\rho = 1$. *In detail, we have*

$$J(F) - J(F, \varepsilon) \le \xi^*(F)\varepsilon, \tag{5.6a}$$

where

$$\xi^*(F) = \inf_{\lambda_F \in \partial J(F)} |\lambda_F|_*. \tag{5.6b}$$

PROOF. Let λ_F be a support functional for J at F. Then for any G,

$$J(G) - J(F) \ge \lambda_F(G - F). \tag{5.7a}$$

By definition of dual norm

$$\lambda_F(G - F) \ge -|\lambda_F|_* |G - F|, \tag{5.7b}$$

so

$$\inf_{|G-F| \le \varepsilon} J(G) \ge J(F) - |\lambda_F|_* \varepsilon. \tag{5.8}$$

□

To apply this proposition, we will compute support functionals for different J of interest. For (the square of) L_2^k we get, using integration by parts k times,

$$\lambda_F(\Delta) = (-1)^k \int 2 f^{(2k)}\, d\Delta, \tag{5.9}$$

whenever $L_2^{2k}(F) < \infty$. For I we get, after one integration by parts,

$$\lambda_F(\Delta) = \int f^{1/2}(\Psi' - \Psi^2/2)\, d\Delta, \tag{5.10}$$

where $\Psi = -f'/f$ and Ψ' is supposed to exist in measure. For L_∞^k,

$$\lambda_F(\Delta) = \operatorname{sgn}\big(f^{(k)}(x_0)\big)\, \Delta^{(k)}(x_0) \tag{5.11}$$

if $f^{(k)}$ has a unique extremum at x_0. For (the pth power of) L_p^0,

$$\lambda_F(\Delta) = \int p f^{p-1}\, d\Delta. \tag{5.12}$$

When do these functionals have a finite dual norm? A sufficient condition is supplied by looking at the Riesz representer.

LEMMA 5.5. *Every linear functional λ on the space of differences $\Delta = F - G$ with a representation $\lambda(\Delta) = \int \zeta\, d\Delta$ has $|\lambda|_* \le \text{variation}(\zeta)$.*

PROOF. Indeed, integrating by parts $\int \zeta\, d\Delta = -\int \Delta\, d\zeta$ so

$$|\lambda(\Delta)| \le \text{variation}(\zeta) |\Delta|.$$

□

COROLLARY. *Let* **BV** *denote the space of functions of bounded variation. These functionals have $\rho = 1$ in* (3.19) *under these conditions*:

L_2^k: $f^{(2k)} \in \mathbf{BV}$,

I: $f^{1/2}(\Psi' - \Psi^2/2) \in \mathbf{BV}$,

L_p^0: $f^{p-1} \in \mathbf{BV}$.

The interpretation of these results is as follows. Letting ε_n be of order $(\log\log n/n)^{1/2}$, then (3.19) implies that in these cases, $J(F_n; \varepsilon_n) \to J(F)$ at a rate very nearly $n^{-1/2}$.

On the other hand, for functionals such as L_∞^k and L_1^k, $k > 0$, expressions such as (5.11) involve evaluations of derivatives at various points and cannot be represented as integrals with kernels of bounded variation. It appears that in these cases $\rho = 1$ is not generally possible. The author's calculations indicate the rates

$$\rho\left(L_\infty^0, F\right) = \tfrac{2}{3} \tag{5.13}$$

if $f \in \mathbf{BV}$ and has a unique quadratic maximum and

$$\rho\left(L_\infty^1, F\right) = \tfrac{1}{2} \tag{5.14}$$

if $f' \in \mathbf{BV}$ and has a unique quadratic extremum. Using techniques of Donoho and Liu (1987) it may be shown that the rates for estimating these functionals certainly cannot be better than $\frac{4}{5}$ in the first case and $\frac{2}{3}$ in the second case. It should also be remembered that (3.19) only provides a bound on the rate of convergence; the actual rate may be faster.

The general conclusion of this section is that the convex functionals L_p^k and I are l.s.c. and so admit of nonparametric lower bounds which are universally consistent. At points where a semi-Lipschitz condition holds, lower bounds can be n^{-r} consistent for every r in the range $(0, \frac{1}{2})$. The semi-Lipschitz condition involves the boundedness of the subdifferential of J at a point. It is therefore essentially a smoothness condition on the functional.

It is important to note the constant $\xi^*(F)$ in (5.6b) cannot be upperbounded based on empirical observations. Otherwise there would exist nonparametric upper bounds for J, contradicting the results of Section 2.

6. Examples. The material of the last three sections can be used to present some existence results of practical interest. In general, the method is useful for generating consistent methods which are conservative in certain senses.

6.1. *Efficient bandwidth selection.* In order to estimate a density by the popular kernel or histogram methods, one has to choose a bandwidth or binwidth parameter. Several methods of automatically choosing such parameters have been proposed, ranging from the simple device of using the bandwidth appropriate for the normal density of the same scale, to various cross-validation schemes.

If one is interested in obtaining a density estimator with good integrated mean square error, the asymptotically optimal bandwidth for the kernel procedure has the form

$$b_n(F) = c_2\left(nL_2^2(F)^2\right)^{-1/5}, \tag{6.1}$$

where c_2 depends on the kernel employed; and the asymptotically optimal binwidth for the histogram is

$$h_n(F) = c_1\left(nL_2^1(F)^2\right)^{-1/3}. \tag{6.2}$$

(In each case it is assumed that the density has L_2^1 or L_2^2 finite.) Of course, in practice, one does not know b_n or h_n—after all, F is unknown.

The approach of this paper lends itself to placing *upper* bounds on b_n and h_n. That is, one places nonparametric lower bounds on $L_2^2(F)$ and $L_2^1(F)$ and plugs the lower bounds into (6.1) and (6.2). This method has good properties.

Consistency. Let $\varepsilon_n \to 0$ slowly enough. Then

$$\frac{b_n(F_n, \varepsilon_n)}{b_n(F)} \to_{\text{a.s.}} 1 \tag{6.3}$$

whenever $L_2^2(F) < \infty$ and

$$\frac{h_n(F_n; \varepsilon_n)}{h_n(F)} \to_{\text{a.s.}} 1$$

whenever $L_2^1(F) < \infty$.

Conservatism. With high probability, the resulting density estimate oversmooths, that is, $b_n(F_n; \varepsilon_n)$ and $h_n(F_n; \varepsilon_n)$ are upper bounds

$$P_F\{b_n(F_n; \varepsilon_n) \geq b_n(F)\} \geq P\{|U_n - U| \leq \varepsilon_n\}$$

(and similarly for h_n), where the right-hand side is tabulated as the distribution of the Kolmogorov–Smirnov statistic. Asymptotically, this probability tends to 1 if $\varepsilon_n \to 0$ slowly enough.

Efficiency. Under the assumption $L_2^2(F) < \infty$ and regularity conditions on the kernel and on F, the MISE of the kernel density estimate employing the empirical bandwidth $b_n(F_n; \varepsilon_n)$ is asymptotic to the MISE of the kernel estimate using b_n.

Thus one can make statements of the form: the right bandwidth may be smaller than this, but I have good confidence that it is not larger. This is an interesting result because it says that although one does not know how much to smooth, one knows it is not necessary to smooth more than a certain amount. Or, to put it another way, the best bandwidth, if we knew what it was, would reveal at least as much detail as our upper bound reveals.

At the same time, this conservatism does not cost anything in asymptotic efficiency. The upper bound and the quantity being bounded are asymptotically equivalent. The efficiency statement relies on a theorem of P. Hall; see the Appendix.

Terrell and Scott (1985) were first to propose a technique of bandwidth selection that always has the conservatism property mentioned above. However, their technique has neither the consistency nor efficiency properties, so the present result is something of an improvement.

6.2. *Construction of confidence intervals.* For several adaptive location estimators, construction of confidence intervals requires knowledge of a nonlin-

ear functional. The asymptotic variance of the Hodges–Lehmann estimator is $V = (12L_2^0(F)^4)^{-1}$; the asymptotic variance of the center of symmetry is $V = I(F)^{-1}$.

Under the approach of the present paper, the width of the confidence intervals would be upper bounded by first constructing a lower bound for L_2^0 or I. Plugging this into the formula for V, one would obtain an upper bound $\overline{V}$ on the asymptotic variance. Then one would say, "Although I do not know the width of the right interval, I know that with high probability it is not larger than this upper bound."

Let $C(t, v, n, \alpha)$ denote the interval

$$\left[t - \Phi^{-1}(1 - 2\alpha)(v/n)^{1/2}, t + \Phi^{-1}(1 - 2\alpha)(v/n)^{1/2}\right].$$

Whenever the $\overline{V}$ is greater than V, we have $C(\hat{\theta}, \overline{V}, n, \alpha)$ containing $C(\hat{\theta}, V, n, \alpha)$, so

$$P\{\theta \in C(\hat{\theta}, \overline{V}, n, \alpha)\} \geq P\{\theta \in C(\hat{\theta}, V, n, \alpha)\} - P\{V > \overline{V}\}.$$

But $P\{V > \overline{V}\} \leq P\{|F_n - F| > \varepsilon_n\}$, which tends to 0 in a known way if $\varepsilon_n \to 0$ slowly enough. It follows that the coverage probability of the procedure is not essentially worse than the coverage probability that would be possible if V were known. At the same time, because of the universal consistency $\overline{V} \to V$, the width of the interval based on $\overline{V}$ is not much larger than the width that of the V-known interval.

[What may be true, especially in small samples, is that the V-known interval is not very good. In small samples, one would want to work with nonasymptotic confidence intervals. In principle, the functional $C_{n,\alpha}(F)$ = "the shortest interval centered at $\hat{\theta}$ which contains θ with a P_F-probability of $1 - \alpha$" ought to be amenable to an upper bound approach.]

The insight that one could construct conservative confidence intervals for the adaptive location parameter is due to Bickel and Klaassen (1982).

7. Extensions. This section describes a few ways the approach of this paper can be used more generally—with other kinds of functionals and in other settings.

7.1. *Some convolution-decreasing functionals.* The author's interest in this topic arose in the study of some deconvolution problems [Donoho (1981)] and in projection pursuit [Huber (1985)]. In those problems it was of interest to study three particular functionals. Let $\mathrm{Var}(F)$ denote the variance of F. The standardized negentropy is

$$\tilde{\Lambda}(F) = \int f \log f + \log(\mathrm{Var}(F))/2. \tag{7.1}$$

The standardized Fisher information is

$$\tilde{I}(F) = \mathrm{Var}(F)I(F)$$

and the standardized fourth cumulant is

$$\tilde{C}_4(F) = \left| 3 - \frac{\int (x - \text{ave}(F))^4 \, dF}{\text{Var}^2(F)} \right|.$$

These three functionals are convolution decreasing:

$$J(F_0 * F_1) \leq \min(J(F_0), J(F_1)) \tag{7.2}$$

under appropriate conditions on F_0 and F_1.

This convolution-decreasing property is important in the formal solution of those problems. Roughly speaking if the functionals were norm continuous and if (7.2) held it would follow immediately that certain deconvolution problems could be solved, using the statistic $J(F_n)$. However, these functionals are not continuous—and $J(F_n)$ need not be well defined. They are semicontinuous, however. Because of (3.10d), (7.2) implies that

$$J(F_0 * F_1; \varepsilon) \leq \min(J(F_0; \varepsilon), J(F_1; \varepsilon)).$$

As $J(F; \varepsilon)$ is defined over a much broader class of distributions than $J(F)$, it offers the convolution-decreasing property without assuming the regularity conditions (e.g., existence of variance or of a regular density) involved in deriving relations such as (7.2). As a result, the formal approach to solving these problems can be made rigorous by substituting $J(F_n; \varepsilon_n)$ in place of $J(F)$. We hope to report on this work elsewhere.

7.2. *Use of other neighborhoods.* From an abstract point of view, many other neighborhood systems might be used to construct lower bounds of the type given here. If $\delta(F; G)$ denotes a measure of discrepancy between distributions, then defining

$$J_\delta(F_n; \varepsilon_n) = \inf\{J(G): \delta(F_n; G) \leq \varepsilon_n\}, \tag{7.3}$$

one has

$$P\{J_\delta \leq J\} \geq P\{\delta(F_n; F) \leq \varepsilon_n\}. \tag{7.4}$$

For discrepancies such as the Kuiper metric and the Cramér–von Mises goodness-of-fit measure, the right side of (7.4) is still distribution free; the approach still yields nonparametric lower confidence bounds. In this sense, the neighborhood approach is quite general. However, we have used the Kolmogorov neighborhood here for its combination of many useful properties:

1. triangle inequality,
2. translation invariance,
3. distribution freeness,
4. Glivenko–Cantelli property/law of iterated logarithm,
5. dual space with known properties.

To redo the analysis given in the sections above without some of these properties might force a real complication in the analysis.

7.3. *Applications to other kinds of data.* Although the paper focusses on functionals of a one-dimensional density, the approach can be used in other settings as well.

7.3.1. *Functionals of a spectral density.* One expects that the approach of the present paper could be applied to functionals of the spectral density of a stationary time series. Indeed, the normalized empirical spectral measure of a time series has many of the properties of the empirical measure of a random sample. First, the Kolmogorov distance between the empirical and true spectral measures is asymptotically distribution free. Second, a law of the (uniterated) logarithm for the Kolmogorov distance exists.

A number of spectral functionals seem to be of interest. The spectral entropy

$$e(f) = \int_0^{2\pi} \log f(\omega)\, d\omega \tag{7.5}$$

and spectral indeterminism

$$i(f) = \int_0^{2\pi} f^{-1}(\omega)\, d\omega \tag{7.6}$$

arise in prediction and interpolation theory; these are obviously semicontinuous. Other semicontinuous functionals include the number of spectral peaks $m(f)$ (defined much as the number of modes was defined in Section 4) and the order of the correct autoregressive model

$$k(f) = \text{least } k\text{: } f(\omega) = \left| \sum_0^k \beta_j e^{i\omega j} \right|^{-2} \quad \text{for some } (\beta_0, \ldots, \beta_k). \tag{7.7}$$

One-sided inference has obvious applications, "I don't know the exact order of the autoregression but I have good confidence it must be at least $k(F_n; \varepsilon_n) = 8$; and $k(F_n; \varepsilon_n)$ is universally consistent, so this lower bound can't be grossly conservative," etc.

7.3.2. *Linear inverse problems.* The following setup is encountered frequently in inverse problems in geophysics and astronomy. One observes data $\{y_i, x_i\}$ satisfying the relationship

$$y_i = F(x_i) + \varepsilon_i, \tag{7.8}$$

where the ε_i are supposed to be Gaussian with known variance $\sigma(x_i)$.

In this setting one might want to obtain information about a nonlinear functional of F. For definiteness, F might be related to some remote object f whose size is the functional of interest. For example, f may be a field in an inaccessible region (the Earth's core, say) while F is the field measured in an accessible region (the surface, say). Then F and f could be related by a Green function which propagates the remote field to the observer:

$$F(x) = \int k(x, \xi) f(\xi)\, d\xi. \tag{7.9}$$

The functional of interest might then be an energy of the remote field, a

quadratic functional such as

$$J = \int f(\xi)^2 \, d\xi.$$

The relation to the problem of inference about the L_2-norm of a density function, given information about the cumulative, should be evident. Indeed, (7.9) shows that in an explicit way f is a kind of derivative of F.

That upper bounds for J are generally not available (this, of course depends on the Green function k) is known and that one can obtain a lower bound via a neighborhood procedure is also known. For example, the paper by Shure, Parker and Backus (1982) discusses the problem of solving for the remote field with least J that fits to within a high P-value; in their case, J is a quadratic functional measuring ohmic dissipation in the (Earth's) core (in their notation, F_3). They mention on page 228 that their approach yields lower bounds on the ohmic dissipation and that physical arguments are necessary to place upper bounds.

It thus appears the geophysicists have observed the need for one-sided inference somewhat earlier than statisticians.

8. Discussion.

8.1. *One-sided inference.* Statisticians have generally focused their attention on confidence statements of a symmetric, two-sided nature, where data allows one to set confidence limits on a parameter from above and from below.

The present paper gives examples where, from a given type of data (in this case the empirical distribution of a random sample) one can only get one-sided information, unless prior or external information is available. The author believes that such situations are fairly common. In estimating quantities like the lifetime of the proton or the mass of a remote galaxy, the available data allow one to place lower bounds: The proton must have a long lifetime or else the universe would already have fallen apart, but we do not know if the lifetime is finite or not; the remote galaxy must have a large mass because it radiates a lot of energy, but we do not know how much dark mass there is.

The cases the author has examined have a common denominator: The quantity of interest is a measure of the complexity of a system—size, norm or number of components. This should make the phenomenon intuitively understandable. Empirical data can usually invalidate simple models, i.e., prove that a system possesses at least a certain degree of complexity. However, data can not usually rule out very complex models which differ from simpler ones in ways that are not detectable given the quantity and quality of data at hand. In short, measures of complexity usually admit of empirical lower but not upper bounds.

8.2. *Nonparametric versus asymptotic methods.* The term nonparametric, as originally used starting in the 1940s, meant essentially that no hypotheses about the form of the distribution function were necessary. E. J. G. Pitman, in his 1949 lecture notes on nonparametrics captured something of the spirit of nonparametrics when he said, "...often we have no knowledge of the nature of the distribution except what is supplied by the sample."

At present, the term is often applied to procedures such as density estimates. Although one is not making strict parametric assumptions when estimating a density, one ends up making hypotheses about the unknown distribution, involving the existence and regularity of various derivatives of the underlying density, in choosing the method of estimation (kernel, histogram, spline,...) and the tuning constants of the method.

Another distinctive feature of the original nonparametrics was that certain properties held exactly in small samples. The new nonparametric methods have generally only asymptotic properties, and it is simply not known how well they hold up in samples of any reasonable size. For example, about the bootstrap confidence intervals, the best one can generally say is that they are asymptotically nonparametric.

This paper has shown that some of the concerns of the new nonparametrics can be addressed while keeping the spirit of the old nonparametrics. One can discuss inference about certain quantities without making antiempirical, a priori assumptions, and there exist procedures with guaranteed properties (e.g., nonparametric coverage probability) in finite samples.

APPENDIX

Proofs.

LEMMA A.1 (Le Cam). *Let $F^{(n)}$ denote the product measure on $\mathbf{R}^n$ with one-dimensional marginal F. If $\tau(F, G) \le 2(1 - (1 - \delta^2/8)^{1/n})$, then $\tau^{(n)}(F^{(n)}, G^{(n)}) \le \delta$.*

PROOF. By Le Cam [(1986), pages 46–49], the Hellinger distance

$$H(F, G) = \left(2 - 2\int \sqrt{f}\sqrt{g}\, d\mu\right)^{1/2},$$

where f and g denote densities with respect to $\mu = F + G$, bounds the testing distance via the inequality

$$H^2\big(F^{(k)}, G^{(k)}\big) \le \tau^{(k)}\big(F^{(k)}, G^{(k)}\big) \le 2H\big(F^{(k)}, G^{(k)}\big) \tag{A.1}$$

for $k = 1, 2, \ldots$. Also, we can relate the distance between product measures $F^{(k)}, G^{(k)}$ to that between marginals via the identity

$$H^2\big(F^{(n)}, G^{(n)}\big) = 2\Big(1 - \big(1 - H^2(F, G)/2\big)^n\Big).$$

Combining the last two displays,

$$\begin{aligned}
\tau^{(n)}\big(F^{(n)}, G^{(n)}\big) &\le 2H\big(F^{(n)}, G^{(n)}\big) \\
&= 2\Big[2\Big(1 - \big(1 - H^2(F, G)/2\big)^n\Big)\Big]^{1/2} \\
&\le 2\big[2\big(1 - (1 - \tau(F, G)/2)^n\big)\big]^{1/2} \\
&\le \delta. \qquad \square
\end{aligned}$$

PROOF OF THEOREM 2.2. In this proof, F will be an interior point of $\mathbf{F}$ at which J is well defined, i.e., $F \in \mathrm{dom}(J)$. Since $\mathbf{F}$ is strongly nonparametric, it contains a testing neighborhood centered at F, $N(F)$, with radius $\varepsilon > 0$. The goal is to show that J satisfies DGC, i.e., that in any such neighborhood of F, J takes on values dense in $[J(F), \infty]$. Where $J(F) = \infty$, this is trivially true.

Note that since the radius of $N(F)$ is ε, every mixture of the form $(1 - \varepsilon)F + \varepsilon H$, where H is an arbitrary d.f., is in $N(F)$.

Consider M, the number of modes, defined as in Section 4. Suppose F has a finite number of modes. Pick a point x_1 outside of any modal interval and let H_1 be a point mass at x_1. Then $M((1 - \delta_1)F + \delta_1 H_1) = M(F) + 1$ for all $0 < \delta_1 \leq \varepsilon/2$. The process can be repeated, picking an x_2 and a $\delta_2 \leq \varepsilon/4$, yielding $M((1 - \delta_1 - \delta_2)F + \delta_1 H_1 + \delta_2 H_2) = M(F) + 2$. Continuing in this way, one can produce a distribution in $N(F)$ with $M(F) + j$ modes, for each $j > 0$.

Consider K, the mixture complexity. Suppose F is a mixture of exactly $K(F) < \infty$ components. Let t_1 be a θ-value distinct from those used in representing F. Then $K((1 - \delta_1)F + \delta_1 G_{t_1}) = K(F) + 1$, for each $0 < \delta_1 < \varepsilon/2$. Continue as with M.

Consider L_p^k as defined in Section 5 with $p > 1$. It is well defined on $\mathrm{dom}(L_p^k) = \mathbf{L}_p^k$, the $\mathbf{L}_p$-Sobolev space of order k. Put $N_0 = N(F) \cap \mathrm{dom}(L_p^k) = N(F) \cap \mathbf{L}_p^k$; this is a convex subset of $\mathbf{L}_p^k$. Let $\overline{N}_0$ denote the τ-closure of N_0. As may be checked by a smoothing argument, $\overline{N}_0$ consists of all absolutely continuous elements of $N(F)$. Now $\overline{N}_0 \setminus N_0$ is nonempty. Indeed we can construct $G \in \overline{N}_0 \setminus N_0$ as follows. Take an absolutely continuous $F \in N_0$ and modify its density on an interval (a, b) so that on that interval it is proportional to $(t - a)^{-(1+\eta)/p}$ (where η is a small positive number), but so that it still integrates to 1. Call the resulting distribution G. If $b - a$ is small enough, then the change on (a, b) affects things so little that $G \in N(F)$. G is absolutely continuous, but its density g is neither continuous nor in $\mathbf{L}_p$. Thus $G \in \overline{N}_0 \setminus N_0$.

Let G_n be a sequence of elements of N_0 converging in τ-distance to G (such a sequence exists as $G \in \overline{N}_0$). Then by Lemma A.3, $L_p^k(G_n) \to \infty$. It follows that L_p^k is unbounded on N_0.

Since N_0 is connected (it is convex) and since L_p^k is continuous on N_0 endowed with the Sobolev $\mathbf{L}_p^k$ topology, every value in the range $[L_p^k(F), +\infty)$ is taken on by L_p^k on N_0, even though the τ radius of $N(F)$ may be arbitrarily small.

The argument for Shannon entropy and Fisher information is similar. As the argument for I involves more technique, we give it here. Let F be a distribution with nice root density $\sqrt{f} \in \mathbf{L}_2^1$, so that $I(F) < \infty$. The testing ball $N(F)$ contains a Hellinger ball [by (A.1)]. Such a Hellinger ball contains all densities whose square roots are in an L_2 ball about $\sqrt{f}$. Call this ball of root densities $S(\sqrt{f})$ and put $S_0 = S(\sqrt{f}) \cap \mathbf{L}_2^1$. Let $\overline{S}_0$ be the $\mathbf{L}_2$-closure of S_0. Now $\overline{S}_0 \setminus S_0$ is nonempty. Indeed we can construct a root density r in $\overline{S}_0 \setminus S_0$. We simply take an arbitrary element of S_0 and modify it on a small interval (a, b) so that it becomes sufficiently discontinuous, but so that its L_2-norm remains 1. Making it proportional to $(t - a)^{-(1+\eta)/2p}$ with $p > 2$ and a small positive η will work. Then if $b - a$ is sufficiently small, we can guarantee that the resultant, r say, will be in $S(\sqrt{f})$. As r is a root density, it is in $\overline{S}_0$, but due to the discontinuity it

is not in $\mathbf{L}_2^1$ and hence not in S_0. Let r_n be a sequence of root densities in S_0 converging to r in $\mathbf{L}_2$-norm (this exists as $r \in \bar{S}_0$). As $r \notin \mathbf{L}_2^1$ but $\|r_n\|_2 = \|r\|_2 = 1$, Lemma A.2 implies that $\|r_n'\|_2 \to \infty$; thus $J(r) = \|r'\|$ is unbounded on S_0. As $J(r)$ is continuous in S_0 (endowed with the $\mathbf{L}_2^1$ topology) and as S_0 is connected, the image of S_0 under J contains every value in the range $[J(\sqrt{f}), +\infty)$. As $I(F)^{1/2} = 2J(\sqrt{f})$ we have established that the image of $N(F) \cap \text{dom}(I)$ under I contains all values in the range $[I(F), +\infty)$. □

We now give the lemmas referred to in the proof: first, a definition. A sequence $\{f_n\}$ of locally integrable functions converges distributionally to f if

$$\int \xi f_n \to \int \xi f$$

for all ξ that are C^∞ and of compact support.

LEMMA A.2. *If $f \notin \mathbf{L}_p^k$ and if $\{f_n\} \subset \mathbf{L}_p^k$ with $f_n \to f$ distributionally, then*

$$\liminf_n \|f_n\|_{\mathbf{L}_p^k} = +\infty.$$

PROOF. Suppose instead that a subsequence of $\{f_n\}$ exists, along which the $\mathbf{L}_p^k$ norm stays bounded. Without loss of generality let the subsequence be $\{f_n\}$ itself. Then $\{f_n\}$ is a bounded subset of the Banach space $\mathbf{L}_p^k$. Therefore by Alaoglu's theorem $\{f_n\}$ has a subsequence converging weakly in $\mathbf{L}_p^k$ to an element of $\mathbf{L}_p^k$. This weak convergence implies distributional convergence, as each linear functional $L(f) = \int \xi f$, where ξ is infinitely differentiable and of compact support, is a bounded linear functional on $\mathbf{L}_p^k$. Thus $\{f_n\}$ has a subsequence converging distributionally to an element of $\mathbf{L}_p^k$. By hypothesis the only distributional limit is f. As $f \notin \mathbf{L}_p^k$ we have reached a contradiction, which proves the lemma. □

LEMMA A.3. *If f is a density not in $\mathbf{L}_p^k$ and if $\{F_n\}$ is a sequence of distributions with densities in $\mathbf{L}_p^k$ with $\tau(F_n, F) \to 0$, then*

$$\liminf_n L_p^k(F_n) = +\infty.$$

PROOF. By equivalence of the $\mathbf{L}_p^k$ norm with $\|f^{(k)}\|_p + \|f\|_p$, there is $c > 0$ with

$$\|f^{(k)}\|_p + \|f\|_p \geq c\|f\|_{\mathbf{L}_p^k}.$$

Now convergence in testing distance implies distributional convergence (since the function ϕ in the definition of τ may be smooth and of compact support), so following Lemma A.2 we conclude

$$\|f_n^{(k)}\|_p + \|f_n\|_p \to +\infty. \tag{A.2}$$

Suppose $\|f_n^{(k)}\|_p$ is bounded in n. Then we must also have $\|f_n\|_p$ bounded. If $p = 1$ this is trivial since $\|f_n\|_1 = \int f_n = 1$ If $p > 1$, use the inequality in Lemma

A.4 to get

$$\| f_n \|_p^{p/(p-1)} \leq c \| f_n^{(k)} \|^{\alpha},$$

for some α, $0 < \alpha \leq 1$, so that $\| f_n \|_p$ stays bounded. But if both $\| f_n^{(k)} \|_p$ and $\| f_n \|_p$ stay bounded, we have reached a contradiction with (A.2). It follows that $\| f_n^{(k)} \|_p$ is not bounded in n nor is any subsequence. Lemma A.3 follows. □

LEMMA A.4. *Let f be a density which is $k-1$ times absolutely continuous, with $f^{(k)} \in \mathbf{L}_p$. Then*

$$\| f \|_p^{p/(p-1)} \leq \| f \|_{\infty} \leq C_{0,k,p} \| f^{(k)} \|_p^{\alpha}$$

for some constants $C_{0,k,p} < \infty$ and $\alpha = 1 - (k - 1/p)/(k + 1 - 1/p)$.

PROOF. As f is a density,

$$\int f^p = \int f^{p-1} f \leq \| f \|_{\infty}^{p-1}.$$

Magaril-Il'yaev (1984) gives the inequality

$$\| f \|_{\infty} \leq C_{0,k,p} \| f \|_1^{1-\alpha} \| f^{(k)} \|_p^{\alpha}$$

(and much more general results as well), with α as in the statement of the lemma. Combining these two displays, the lemma follows. □

PROOF OF LEMMA 5.2. Part (b) is standard. See Adams (1975).

For part (a), note that $\phi_h * F_n$ is a C^{∞} density. Gabushin [(1967), Theorem 2] gives the inequality

$$\| g^{(k)} \|_p \leq C_{p,k,\infty,1,l} \| g \|_{\infty}^{\alpha} \| g^{(l)} \|_1^{\beta},$$

valid for smooth functions, where $l > \max(k+1, 2)$ and

$$\alpha = \frac{l - k - 1 + 1/p}{l - 1}, \qquad \beta = \frac{k - 1/p}{l - 1}.$$

Putting $g = \phi_h * (F_n - F)$ and using

$$\| g^{(l)} \|_1^{\beta} \leq 2^{\beta} \| \phi_h^{(l)} \|_1^{\beta},$$

we have, from Gabushin's result,

$$\| g^{(k)} \|_p \leq C' \| g \|_{\infty}^{\alpha},$$

with $C' = C_{p,k,\infty,1,l} 2^{\beta} \| \phi_h^{(l)} \|_1^{\beta}$. Now as the sup-norm is convex and translation invariant,

$$\| g \|_{\infty} = | \phi_h * (F_n - F) | \leq \int \phi_h | F_n - F | = | F_n - F |,$$

and so

$$\| g^{(k)} \|_p \leq C' | F_n - F |^{\alpha}.$$

From equivalence of the $\mathbf{L}_p^k$ norm with $\|f\|_p + \|f^{(k)}\|_p$ we have

$$\|\phi_h * F_n - \phi_h * F\|_{\mathbf{L}_p^k} \leq C''\left(|F_n - F|^{\alpha_0} + |F_n - F|^{\alpha_k}\right)$$

with appropriate α_0, α_k and C''. This inequality shows that $|F_n - F| \to 0$ implies $\phi_h * F_n \to \phi_h * F$ in $\mathbf{L}_p^k$, as claimed. □

PROOF OF LEMMA 5.3. (a) By the equivalence of the norm of $\mathbf{L}_2^1$ with $\|f\|_2 + \|(f')\|_2$, it is enough to show that with $g_0 = \phi_h * F$ and $g_n = \phi_h * F_n$,

$$\int\left(\left(g_0^{1/2}\right)' - \left(g_n^{1/2}\right)'\right)^2 \to 0 \tag{A.3}$$

under the advertised conditions. Let $M > 0$ and partition the integral into the contributions for $|x| \geq M$ and $|x| < M$. Consider the $|x| \leq M$ part. It can be written as

$$\int_{-M}^{M}\left(\left(g_0^{1/2}\right)' - \left(g_n^{1/2}\right)'\right)^2 = \int_{-M}^{M}\left(g_n^{1/2}g_0' - g_n'g_0^{1/2}\right)^2/g_0g_n$$

$$\leq \left\{\inf_{[-M,M]} g_0g_n\right\}^{-1}\int_{-M}^{M}\left(g_n^{1/2}g_0' - g_n'g_0^{1/2}\right)^2.$$

The term in braces has a lim inf as $n \to 0$ of $\inf_{[-M,M]} g_0^2$ which can be bounded away from 0 because g_0 has a Gaussian convolution component. The integral can be bounded by

$$2\left\{\int_{-M}^{M} g_0\left(g_n' - g_0'\right)^2 + \int_{-M}^{M}\left(g_0'\right)^2\left(g_0^{1/2} - g_n^{1/2}\right)^2\right\}.$$

Both g_0 and $g_0'^2$ are in $\mathbf{L}_1[-M, M]$, and as in Lemma 4.1, one has the uniform convergence

$$|g_n' - g_0'| \to 0, \qquad |g_n - g_0| \to 0,$$

via an integration by parts. So the contribution to (A.3) of $[-M, M]$ tends to 0 as $n \to \infty$.

Consider now the part of the integral coming from outside of $[-M, M]$. Define $\rho_0 = g_0'g_0^{-1/2}$ and ρ_n similarly. As in Port and Stone (1974), fix $N \leq M$ and decompose ρ into the part from the middle of F and the part coming from the tails. Thus, write $\rho_0 = \rho_{0,1} + \rho_{0,2}$, where

$$\rho_{0,1}(x) = \left\{\int_{-N}^{N}\phi_h'(x - z)\,dF(z)\right\}/\sqrt{g_0(x)},$$

$$\rho_{0,2}(x) = \left\{\int_{|x|\geq N}\phi_n'(x - z)\,dF(z)\right\}/\sqrt{g_0(x)},$$

and similarly for ρ_n. Now by Lemma 2.6 of Port and Stone,

$$\int\rho_{0,2}^2 \leq h^{-2}(1 - F[-N, N])$$

and defining

$$B(M, N) = \sup_{F\in\mathbf{P}}\int_{|x|\geq M}\rho_{0,1}^2$$

we have, by Lemma 2.7 of Port and Stone, that for N fixed,

$$\limsup_{M\to\infty} B(M, N) = 0.$$

Putting these pieces together,

$$\begin{aligned}\int_{|x|\geq M}(\rho_n - \rho_0)^2 &\leq 2\left\{\int_{|x|\geq M}\rho_n^2 + \rho_0^2\right\}\\ &\leq 4\left\{\int_{|x|\geq M}\rho_{n,1}^2 + \rho_{0,1}^2 + \int\rho_{n,2}^2 + \rho_{0,2}^2\right\}\\ &\leq 4\{2B(M, N) + h^{-2}(2 - F[-N, N] + F_n[-N, N])\}.\end{aligned}$$

Letting first $n \to \infty$, then $M \to \infty$, then $N \to \infty$, we have the right-hand side tending to 0.

(b) Let r_h and r denote the root densities of $\Phi_h * f$ and f, respectively.

$$\int(r_h - r)^2 \leq \int|r_h^2 - r^2| = \int|\Phi_h * f - f|.$$

The rightmost term goes to 0 for every $f \in \mathbf{L}_1$ as $h \to 0$; so $r_h \to r$. Now $I(F)$ is proportional to the squared $\mathbf{L}_2$-norm of r', so from the convexity $I(\Phi_h * F) \leq I(F)$,

$$\int(r_h')^2 \leq \int(r')^2.$$

We can conclude that $\int(r_h')^2 \to \int(r')^2$ from the $\mathbf{L}_2$ convergence mentioned earlier. Also, $r_h \to r$ weakly in $\mathbf{L}_2^1$ because $\mathbf{L}_2$ is dense in the dual of $\mathbf{L}_2^1$. But these two facts—convergence of norms and weak convergence—imply strong convergence because $\mathbf{L}_2^1$ is a Hilbert space and so uniformly convex (again, see Adams (1975)]. □

PROOF OF THE EFFICIENCY RESULT. Theorem 2 of Hall (1983) establishes that, if f has compact support, and under regularity conditions on f and on the kernel k, the kernel density estimate $k_b * F_n$ [where $k_b(x) = k(x/b)/b$] has the integrated squared error

$$\text{(A.4)}\qquad \int(k_b * F_n - f)^2 = c_1(k)(nb)^{-1} + c_2(k)b^4L_2^2(F)^2 + o_p(n^{-4/5}),$$

where the o_p term is uniform in $a_1n^{-1/5} \leq b \leq a_2n^{-1/5}$, for $0 < a_1 < a_2 < \infty$. Here c_1 and c_2 depend on the kernel k:

$$c_1(k) = \int k^2, \qquad c_2(k) = \left(\int x^2k\,dx\right)^2/4.$$

Hall mentions in the last paragraph of his Section 2 that the result extends to f not of compact support under additional regularity. Assume this extension. Now by (3.18),

$$\text{(A.5)}\qquad b_n(F) \leq b_n(F_n; \varepsilon_n) \leq b_n(F; 2\varepsilon_n)$$

for almost all n, with probability 1. Now as $\varepsilon_n \to 0$, for all sufficiently large n,

$b_n(F; 2\varepsilon_n) \le b_n(F; 0.0001)$. Put

$$a_1 = \left[\frac{c_1(k)}{c_2(k)} \frac{4}{L_2^2(F)^2}\right]^{1/5}.$$

By Rosenblatt (1956), $a_1 n^{-1/5}$ is the asymptotically optimal bandwidth which was called b_n in Section 6. Put

$$a_2 = \left[\frac{c_1(k)}{c_2(k)} \frac{4}{L_2^2(F; 0.0001)^2}\right]^{1/5}$$

so that $a_2 > a_1$. Now (A.5) can be rewritten as

$$a_1 n^{-1/5} \le b_n(F_n; \varepsilon_n) \le a_2 n^{-1/5}. \tag{A.6}$$

This inequality holds whenever the event $\Omega_n = \{|F_n - F| < 0.0001\}$ is true. Now $P_F(\Omega_n) = P\{|U_n - U| < 0.0001\} \to 1$ exponentially fast [compare (4.3)].

Let $R_n(b)$ denote the remainder term that was written as $o_p(n^{-4/5})$ in (A.4). Define

$$R_n^*(a_1, a_2) = \sup_{a_1 n^{-1/5} \le b \le a_2 n^{-1/5}} |R_n(b)|;$$

this is a random variable, i.e., is measurable, by the continuity of $R_n(b)$ in b, which means the supremum actually can be taken over rational values of b. Hall's theorem (A.4) says that $R_n^*(a_1, a_2) = o_p(n^{-4/5})$. Now define R_n^{**} to give equality in

$$\mathrm{MISE}(b_n) = c_1(k)(nb_n)^{-1} + c_2(k) b_n^4 L_2^2(F)^2 + R_n^{**}.$$

Here $b_n = b_n(F_n; \varepsilon_n)$. When Ω_n is true, $|R_n^{**}| \le R_n^*(a_1, a_2)$. As $P_F(\Omega_n) \to 1$,

$$R_n^{**} = O_p(R_n^*(a_1, a_2)) = o_p(n^{-4/5}).$$

Now $a_1 n^{-1/5}$ is the asymptotically optimal bandwidth, and by Hall's theorem,

$$\mathrm{MISE}\left(a_1 n^{-1/5}\right) = M^* n^{-4/5} + R_n(a_1),$$

where

$$M^* = \tfrac{5}{4} c_1(k)^{4/5} c_2(k)^{4/5} L_2^2(F)^{2/5}.$$

Moreover, by consistency $L_2^2(F_n; \varepsilon_n) \to L_2^2(F)$, we have

$$b_n(F_n; \varepsilon_n) = a_1 n^{-1/5}(1 + \delta_n),$$

where $\delta_n = o_p(1)$. Some algebra gives that for a certain $q \in (0, 1)$,

$$\mathrm{MISE}(b_n(F_n; \varepsilon_n)) = M^* n^{-4/5}\left[q(1 + \delta_n)^{-1} + (1 - q)(1 + \delta_n)^4\right] + R_n^{**}.$$

Thus

$$\frac{\mathrm{MISE}(b_n(F_n; \varepsilon_n))}{\mathrm{MISE}\left(a_1 n^{-1/5}\right)} = \frac{\left[q(1 + \delta_n)^{-1} + (1 - q)(1 + \delta_n)^4\right] + R_n^{**}/(M^* n^{4/5})}{1 + R_n(a_1)/(M^* n^{4/5})}.$$

As $\delta_n = o_p(1)$, $R_n^{**} = o_p(n^{-4/5})$ and $R_n(a_1) = o_p(n^{-4/5})$, we have

$$\frac{\mathrm{MISE}(b_n(F_n; \varepsilon_n))}{\mathrm{MISE}(a_1 n^{-1/5})} = \frac{\left[q/(1 + o_p(1)) + (1 - q)(1 + o_p(1))^4\right] + o_p(1)\right]}{1 + o_p(1)}$$

$$= 1 + o_p(1),$$

as required. □

Acknowledgments. The author would like to thank Lucien Le Cam, Jim Pitman and Charles Stone for various discussions and Paul Robertson for preparing the manuscript.

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic, New York.

BAHADUR, R. R. and SAVAGE, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115–1122.

BICKEL, P. J. and KLAASSEN, C. A. J. (1982). Personal communication.

CSÖRGŐ, M. and RÉVÉSZ, P. (1981). *Strong Approximation in Probability and Statistics*. Academic, New York.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

DONOHO, D. L. (1981). On minimum entropy deconvolution. In *Applied Time Series Analysis II* (D. Findlay, ed.) 565–608. Academic, New York.

DONOHO, D. L. and LIU, R. C. (1987). Geometrizing rates of convergence, I, II, III. Unpublished.

GABUSHIN, V. N. (1967). Inequalities for the norms of a function and its derivatives in metric L_p. *Mat. Zametki* **1** 291–298. (Translated in *Math. Notes* **1** 194–198.)

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

HARTIGAN, J. A. and HARTIGAN, P. M. (1985). The dip test of unimodality. *Ann. Statist.* **13** 70–84.

HU, I. (1985). A uniform bound for the tail probability of Kolmogorov–Smirnov statistics. *Ann. Statist.* **15** 821–826.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–478.

KARLIN, S. (1968). *Total Positivity*. Stanford Univ. Press, Stanford, Calif.

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, Berlin.

MAGARIL-IL'YAEV, G. G. (1984). Inequalities for derivatives and duality. *Proc. Steklov Inst. Math.* **161** 199–212.

PITMAN, E. J. G. (1979). *Some Basic Theory for Statistical Inference*. Wiley, New York.

PORT, S. and STONE, C. J. (1974). Fisher information and the Pitman estimator of a location parameter. *Ann. Statist.* **2** 235–247.

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

SILVERMAN, B. R. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

SHURE, L., PARKER, R. L. and BACKUS, G. E. (1982). Harmonic splines for geomagnetic modeling. *Phys. Earth Planetary Interiors* **28** 215–229.

TERRELL, G. R. and SCOTT, D. W. (1985). Oversmoothed nonparametric density estimates. *J. Amer. Statist. Assoc.* **80** 209–214.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720