ESTIMATING INTEGRATED SQUARED DENSITY DERIVATIVES: SHARP BEST ORDER OF CONVERGENCE ESTIMATES

By P. J. BICKEL
University of California at Berkeley
and
Y. RITOV
The Hebrew University of Jerusalem

SUMMARY. Estimation of the integral of the square of a derivative of the probability density function is considered. The estimators we propose and their properties are a function of the amount of smoothness assumed. The rate of convergence of the appropriate estimator is shown to be optimal given the amount of smoothness assumed. In particular the appropriate estimator achieves the information bound when estimation at an \( n^{-1/2} \) rate is possible.

1. INTRODUCTION

Suppose \( X_1, X_2, \ldots, X_n \) are i.i.d., each with distribution function \( F \). Let \( f(.) \) be the probability density function of \( F \), \( f^{(k)} \) its \( k \)-th derivative and \( \theta_k(F) = \int [f^{(k)}(x)]^2 \, dx \). These functionals appear in the asymptotic variance of the Wilcoxon statistic and in the asymptotics of the integrated M.S.E. for kernel density estimates. Discussion of the estimation of \( \theta_k \) and similar parameters appear in Schwerder (1975), Hasminskii and Ibragimov (1978), Pfanzagl (1982), Prakasa Rao (1983), Donoho and Liu (1987) and Hall and Marron (1987).

Ritov and Bickel (1987) show that the standard semiparametric information bound for the estimation of \( \theta_0(F) \) fails to give an achievable rate of convergence. In fact, the information is strictly positive when \( f \) is bounded, promising that the \( n^{-1/2} \) rate is achievable. Nevertheless, there is no rate that can be achieved uniformly in small compact neighborhoods (in the total variation norm) of a given distribution. Moreover, even if the uniformity requirement is dropped then for any sequence of estimates \( \{\hat{\theta}_k\} \) there exists an (unknown) point \( F \) such that \( n^{\gamma}(\hat{\theta}_k(\theta) - \theta_k(F)) \) doesn't converge to 0 for any \( \gamma > 0 \).

In this paper we consider classes of \( F \) which satisfy Hölder conditions on \( f^{(m)} \) for suitable \( m \). We establish the rate achievable under these condi-

---

*Research supported by Office of Naval Research N00014-80-C-0163.
AMS (1980) subject classification: 62G05, 62G30,
Key words and phrases: estimation, density derivatives, semiparametric information bound, sharp rate convergence.

*An invited paper to Commemorate the 50-th volume of Sankhyā.
tions and exhibit estimators that achieve these rates. Our estimators converge uniformly and when improvement is possible faster than similar estimators suggested by Schweder (1975), Hasminskii and Ibragimov (1978), and Hall and Marron (1987). In particular we need to assume weaker Hölder conditions to obtain $n^{-1/2}$ rates and efficient estimators.

We believe that our proof of the best achievable rates is novel in that it cannot be reduced to considering a sequence of simple vs. simple testing problems and in effect requires the use of composite hypotheses of growing size. Note that $\theta_k$ can be estimated at the $n^{-1/2}$ rate in any fixed regular finite dimensional submodel.

2. MAIN RESULTS : THE ESTIMATORS AND THEIR PROPERTIES

Let $\theta_k(F) = \int \{ f^{(2k)}(x) \}^2 dx$ where $f$ is the (continuous) density of the distribution $F$. (In general we denote distribution functions by $F$ or $F_n$ and their densities by $f$ or $f_n$ respectively.) Let $\alpha > 0$, $m$ be a nonnegative integer and $g(\cdot) \in L_2 \cap L_\infty$. Suppose $X_1, \ldots, X_n$ is a random sample from $F$. How well can $\theta_k(F)$ be estimated if it is known a priori only that $F \in F_{m, \alpha, \theta}$ where $F_{m, \alpha, \theta} = \{ F : |f^{(2m)}(x) - f^{(2m)}(x+\xi)| \leq g(\xi)|\xi|^\alpha \text{ for all } x \text{ real } |\xi| < 1 \}$?

We begin by suggesting a family of estimators. Let $h_\sigma(x) = \sigma^{-1} h(x/\sigma)$ where $h$ is a kernel with the following properties:

- $h$ is symmetric about zero,
- $h(x) = 0$ for $|x| > 1$,
- $\int h(x) dx = 1$,
- $\int x^i h(x) dx = 0$, $i = 1, 2, \ldots, \max\{k, m-k\}$

and $h$ has $2k+1$ derivatives.

Divide the sample into two subsamples $X_{11}, \ldots, X_{n_1}$ and $X_{n_1+1}, \ldots, X_n$ with comparable sizes (i.e. $n_1/n$ is bounded away from 0 and 1). Let $\hat{F}_1$ and $\hat{F}_2$ be the empirical distribution functions of each subsample respectively.

Define, $\hat{f}_i(x) = \frac{1}{n_i} \sum_{l=1}^{n_i} h_{\sigma}(x-X_l)$, $i = 1, 2$. The dependence of $\hat{f}_i$ on $\sigma$ is left implicit. Consider the following estimator of $\theta_k$.

\[
\hat{\theta}_k(X_1, \ldots, X_n; \sigma) = \frac{n_1}{n} \hat{\theta}_{01}^* + \frac{n_2}{n} \hat{\theta}_{02}^*
\]  \hspace{1cm} \ldots (2.1)

where

\[
n = n_1 + n_2
\]
\[
\hat{\theta}_{0i}^*(X_1, \ldots, X_n; \sigma) = \int \hat{f}_i^*(x) dx + 2n_1 \sum_{l=1}^{n_1} \left( \int \hat{f}_2(x) dx - \int \hat{f}_1(x) dx \right) + \frac{1}{n_2} \int h_\sigma^2(x) dx
\]
\[
= 2 \int h_\sigma(x-t) d\hat{F}_1(t) d\hat{F}_2(x) - n_2^2 \sum_{n_1+1 \leq i \neq j \leq n} h_\sigma(x-X_i) h_\sigma(x-X_j) dx
\]  \hspace{1cm} \ldots (2.2)
and \( \delta_{02} \) is obtained by interchanging the roles of the two subsamples in \( \delta_{11} \). The first two terms of \( \delta_{02} \) can be re-recognized as Hasminskii and Ibragimov's estimate of this parameter which they show is efficient in \( F_{n, a, M} \) if \( \alpha > 1/2 \).

This is the, by now, familiar one step estimate (see Bickel, 1982; Schick, 1986) using the estimated influence function \( 2(\hat{f}_n - f(x))dx \). The last term in (2.2) removes the pure known function, \( n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{x}^{x+1} h_{n}(x-X_0)h_{n}(x-X_j)dx \).

Curiously enough this simple debiasing leads to efficient estimation in \( F_{n, a, M} \) for \( \alpha > 1/4 \) and (uniformly) \( \sqrt{n} \) consistent estimation on \( F_{0, 1/4, M} \). Moreover, \( \sqrt{n} \) consistent estimation is shown to be impossible for \( \alpha < 1/4 \). More generally, if \( f \) has 2k continuous derivatives,

\[
\delta_{k}(F) = (-1)^k \int f^{(2k)}(x) f(x) dx
\]

and

\[
\delta_{k}(F) = (-1)^k E f^{(2k)}(X). 
\]

This suggests, by the same process as above, estimates \( \delta_{k1}, \delta_{k2} \) and \( \delta_{k3} \). For convenience we replace \( \delta_{k1} \) by \( \delta_{k1} \) where \( n^{-\frac{1}{2}} \) in (2.2) is replaced by \( [n^{-\frac{1}{2}}(n-1)]^{-1} \) and similar replacements are made in \( \delta_{k2} \) and more generally \( \delta_{k3} \). So the estimate we study is

\[
\hat{\delta}_{k}(X_1, ..., X_n; \sigma) = 2(-1)^k \int f^{(2k)}(x-t) d\hat{F}_n(x-t) \hat{F}_n(x)
\]

\[
= n^{-\frac{1}{2}} \sum_{1 \leq i < j < n} h^{(k)}(x-X_i) h^{(k)}(x-X_j) dx
\]

\[
= n^{-\frac{1}{2}} \sum_{1 \leq i < j < n} h^{(k)}(x-X_i) h^{(k)}(x-X_j) dx. 
\]

(2.4)

Our main results are summarized in the following two theorems. In the first we describe the performance of \( \delta \) in terms of the assumed family \( F_{m, a, \sigma} \). The rate of convergence of \( \delta \) to \( \delta(F) \) is a function of \( m+\alpha \) and \( \delta \) is "efficient" when \( m+\alpha > 2k+1/4 \). In the second theorem we show that the rates given in the first theorem are, essentially, the best possible.

**Theorem 1**: Let \( \{F_1, F_2, ...\} \subset F_{m, a, \sigma} \) where \( 0 \leq a < 1 \), \( m+\alpha > k \) and \( g \in L_{a} \cap L_{a} \). Let \( X_{n1}, ..., X_{nn} \) be i.i.d., \( X_{n1} \sim F \) and let \( \delta = \delta(F, X_{n1}, ..., X_{nn}; \sigma) \) where \( \sigma = n^{-\frac{1}{2}}(1+4m+4\alpha) \).

(i) If \( m+\alpha > 2k+1/4 \) then

\[
\sqrt{n} \left| \hat{\delta} - \delta(F) - \frac{2}{n} \sum_{i=1}^{n} \left( (-1)^k f^{(2k)}(X_{ni}) - \delta(F) \right) \right| \rightarrow 0. 
\]

(2.5)

Let \( L_k(F_n) = \left[ \text{Var}[\hat{f}^{(2k)}(X_{ni})] \right]^{-1} \). Then, \( n L_k(F_n) \text{E} [\hat{\delta} - \delta(F)] \rightarrow 1 \) and \( L_k^{(2k)}(F_n) \rightarrow \mathcal{N}(0, 1) \) provided \( \text{lim sup} L_k(F_n) < \infty \).
(ii) If \( k < m + \alpha \leq 2k + 1/4 \) then \( n\gamma E[|\hat{\theta}_k - \theta_k(F_n)|^2] \) is bounded when \( \gamma = 4(m + \alpha - k)/(1 + 4m + 4\alpha) \).

We conjecture, but have not checked the details, that it is possible to estimate \( \sigma \) by cross validation to obtain an estimate \( \hat{\theta}_k^* = \theta_k(X_{n1}, \ldots, X_{nn}; \hat{\sigma}_n) \) which does not depend on \( m \) and \( \alpha \) but is equivalent to \( \hat{\theta}_k \) which does so depend through \( \sigma_n \) given in the statement of Theorem 1.

Theorem 2: (i) The information bound (in the sense of Khas'minskii and Levi (1976)) for nonparametric estimation of \( \theta_k(F) \), \( F \in F_{m,\alpha} \), is given by \( I_k(F) \) as defined in Theorem 1.

(ii) Suppose \( k < m + \alpha \leq 2k + 1/4 \). Then there is a small compact set \( \mathcal{F}^* \subseteq \mathcal{F}_{m,\alpha} \) such that for any \( c_n \to \infty \) and any sequence of estimators \( T_n = T_n(X_1, \ldots, X_n) \), \( X_1, X_2, \ldots, X_n \) iid, \( X \sim F \):

\[
\lim inf_{n} \sup_{F \in \mathcal{F}^*} P_F[c_n n^\gamma |T_n - \theta_k(F)| \geq 1] = 1 \quad (2.6)
\]

where \( \gamma = 4(m + \alpha - k)/(1 + 4m + 4\alpha) \). Moreover \( \mathcal{F}^* \) can be constructed so that its only accumulation point is any specified \( F_0 \in \mathcal{F}_{m,\alpha} \).

The proof of the first part of Theorem 2 is quite standard and follows essentially the discussion in Has'minskii and Ibragimov (1978). The proof of the second part of the Theorem is an extension of the ideas presented in Ritov and Bickel (1987). In our problem, \( \theta_0 \) can be estimated at the \( n^{-1/2} \) rate in any one-dimensional submodel of \( F_{m,\alpha} \) and the information bound of Theorem 2(i) is the best bound that can be achieved using these techniques. Yet for \( m + \alpha < 2k + 1/4 \) this bound is unachievable by uniformly \( n^{1/2} \) consistent estimates. In fact, for \( m + \alpha < 2k + 1/4 \) no uniformly \( n^{1/2} \) consistent estimate exists. Even uniformity can be dropped—see Ritov and Bickel (1987), Theorem 1. Our proof is based on the demonstration of a sequence of difficult multiparameter Bayesian problems.

3. PROOFS

We begin the proofs with the following technical lemma whose own proof is postponed to the end of the section.

Lemma 1: Let \( x, m \) and \( g \) be such that \( x > 0 \), \( m \geq 0 \) and \( g \in L_\infty \). Then

\[
\sup_i \{ |f_i^0(x)| : x, F_\in F_{m,\alpha} \} < \infty, \quad i = 0, 1, \ldots, m.
\]
Proof of Theorem 1: Evidently to establish Theorem 1 it is enough to consider the asymmetric estimate

\[ \theta_{2z} = 2(-1)^k \int \int h^{(2b)}(x-t) \ d\tilde{P}_1(t) \ d\tilde{P}_2(x). \]

\[ - 2\{n_1(n_1-1)\}^{-1} \sum_{i \neq j \neq n_1} \int h^{(b)}(x-X_{ni}) h^{(b)}(x-X_{nj}) \ dx. \]

We begin by estimating the conditional bias

\[ E(\theta_{2z} \mid \tilde{P}_1) - \theta_{2z}(F_n) = 2(-1)^k \int \int h^{(2b)}(x) f_n(x) \ dx \]

\[ - 2\{n_1(n_1-1)\}^{-1} \sum_{i \neq j} \sum_{n_1} h^{(b)}(x-X_{ni}) h^{(b)}(x-X_{nj}) \ dx - \int f_n^{(b)}(x)^2 \ dx. \]

But

\[ (-1)^k \int h^{(2b)}(x) f_n(x) \ dx = \int f_n^{(b)}(x)^2 \ dx \]

\[ = n_1^{-1} \sum_{i \neq j} \sum_{n_1} h^{(b)}(x-X_{ni}) f_n^{(b)}(x) \ dx \]

\[ = \{n_1(n_1-1)\}^{-1} \sum_{i \neq j} \sum_{n_1} h^{(b)}(x-X_{ni}) f_n^{(b)}(x) \ dx. \]

Hence

\[ E(\theta_{2z} \mid \tilde{P}_1) - \theta_{2z}(F_n) = - 2\{n_1(n_1-1)\}^{-1} \sum_{i \neq j} \sum_{n_1} \{h^{(b)}(x-X_{ni}) - f_n^{(b)}(x)\} \]

\[ h^{(b)}(x-X_{nj}) - f_n^{(b)}(x) \ dx. \]  \[ \cdots (3.1) \]

We obtain from (3.1) that

\[ E \theta_{2z} - \theta_{2z}(F_n) = \int \int f_n^{(b)}(x) - f_n^{(b)}(x) \ dx \]

\[ \cdots (3.2) \]

where \( f_n = f_n \circ h. \)

But

\[ f_n^{(b)}(x) - f_n^{(b)}(x) = \int h(t) \left\{ f_n^{(b)}(x+\sigma t) - f_n^{(b)}(x) \right\} dt \]

\[ = \int h(t) \left\{ \sum_{i=1}^{m-1} \frac{h^{(b+i)}(x)}{i} \right\} dt \]

\[ + \int h(t) \left\{ \frac{1}{(m-k)} \right\} \left\{ f_n^{(m)}(x+\sigma t) - f_n^{(m)}(x) \right\} dt. \]  \[ \cdots (3.3) \]

where 0 \( \leq \sigma^* \leq \sigma. \) The first term on the RHS of (3.3) is null by the construction of \( h. \) Since \( F_n \in F_{m,a,d} \) we can bound the integrand in the second term and obtain:

\[ |f_n^{(b)}(x) - f_n^{(b)}(x)| \leq g(x) \sigma^{m-a-k} \int |h(t)| \ dt. \]

\[ \cdots (3.4) \]
Combine (3.2) and (3.4) to conclude that

\[ |E \hat{\theta}_{g2} - \theta_0(F_n)| \leq \|h\|^2 n^{-d(n+a+b)(1+4n+4a)} \int |h(t)| \, dt. \quad \ldots \ (3.5) \]

Next we estimate \( \text{var} (E(\hat{\theta}_{g2} | \hat{F}_1)) \). Note that \( E(\hat{\theta}_{g2} | \hat{F}_1) \) was written in (3.1) as a U statistic, \( E(\hat{\theta}_{g2} | \hat{F}_1) - \theta_0(F_n) = 2(n(n_1-1))^{-1} \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_1} U(X_{ni}, X_{nj}) \) say.

By standard U-statistic theory,

\[ \text{var} (E(\hat{\theta}_{g2} | \hat{F}_1)) = n^{-1} \left( \text{var} (E(U(X_{n1}, X_{n2}) | X_{n1})) \right) + O(n^{-1} \text{var} U(X_{n1}, X_{n2})). \quad \ldots \ (3.6) \]

Now \( E U(x, X_{n2}) = \int \{ f_n^{(g)}(x-t) - f_n^{(g)}(t) \} \{ f_n^{(g)}(x-t) - f_n^{(g)}(t) \} \, dt \)

\[ = \int \delta(t) \{ f_n^{(g)}(x-t) - f_n^{(g)}(t) \} \, dt. \]

say. Hence,

\[ \text{var} [E(U(X_{n1}, X_{n2}) | X_{n1})] = E[\delta(x) \{ f_n^{(g)}(x-X_{n1}) - f_n^{(g)}(x) \} dx]^2 \]

\[ = E \int \int \delta(y) \delta(z) \{ f_n^{(g)}(y-X_{n1}) - f_n^{(g)}(y) \} \{ f_n^{(g)}(z-X_{n1}) - f_n^{(g)}(z) \} \, dx \, dy \]

\[ \leq \int \int \delta(y) \delta(z) \{ f_n^{(g)}(y-t) - f_n^{(g)}(t) \} \, dt \, dx \, dy \]

\[ = \int \{ f_n^{(g)}(x-t) - f_n^{(g)}(t) \} \, dt \]

\[ \leq \|h\|^2 \sigma^{-2k} \int |h(x)| \, dx \leq O(\sigma^{-2k+2}). \quad \ldots \ (3.7) \]

by (3.4). At the same time, the random variable \( \{ f_n^{(g)}(x-X_{n1}) - f_n^{(g)}(x-X_{n2}) \} \) is bounded by \( \sigma^{-2k} \|h\|^2 \) and is equal to zero unless \( |X_{n1} - X_{n2}| \leq 2\sigma \).

Since \( f_n \) is bounded this last event has probability of the same order as \( \sigma \).

Hence

\[ \text{var} \{ f_n^{(g)}(x-X_{n1}) \} = O(\sigma^{-2k+2}). \]

Since \( \|f_n^{(g)}(x-X_{n1})\|_\infty \leq \|f_n^{(g)}\|_{\infty} \sigma^{-k} \int |h^{(k)}(x)| \, dx \) we conclude that

\[ \text{var} (U(X_{n1}, X_{n2})) \]

\[ = \text{var} [\{ f_n^{(g)}(x-X_{n1}) - f_n^{(g)}(x-X_{n2}) \} \] \[ \leq O(\sigma^{-4k+1}). \quad \ldots \ (3.8) \]

We obtain from (3.1), (3.4), (3.6), (3.7), and (3.8) that

\[ \text{var} (E(\hat{\theta}_{g2} | \hat{F}_1)) = O(n^{-1} \sigma^{2(n+a+b)} \sigma^{-2k-4k+1}) \]

\[ = O(n^{-3(n+a+b)(1+4n+4a)}). \]
for $\sigma$ given in the statement of Theorem 1. Hence (3.5) implies that

$$E[E(\theta_{2n} | \hat{F}_1) - \theta_k(F_n)]^2 = O(n^{-8(m+\alpha-2)}/(1+4m+4\alpha^2)).$$  \quad (3.9)

We have proved that $E(\theta_{2n} | \hat{F}_1) - \theta_k(F_n)$ is of the right order (in particular it is $o_p(n^{-1})$ if $m+\alpha > 2k+1/4$). We turn to the investigation of the behaviour of $\theta_{2n} - E(\theta_{2n} | \hat{F}_1)$. This will be carried on separately for the two cases: $2k+1/4 < m+\alpha$ and $k < m+\alpha \leq 2k+1/4$.

(i) Suppose $2k+1/4 < m+\alpha$. In the light of (3.9) we need only to consider the conditional variance of $\theta_{2n}$ given the first sub sample. But, given $X_{n1}, \ldots, X_{nn1}$, $\theta_{2n}$ is just a sum of i.i.d. random variables, hence

$$\text{var}\{\theta_{2n} - \frac{2(-1)^k}{n-n_1} \sum_{n_1+1}^n f_n^{(2k)}(X_{n1} + \theta_k(F_n)) | \hat{F}_1\} \leq \frac{4}{n-n_1} \int \int f_n^{(2k)}(x)^2 f_n(x) dx.$$

So

$$E \text{var}\{\theta_{2n} - \frac{2(-1)^k}{n-n_1} \int f_n^{(2k)}(x) d\hat{F}_2(x) + \theta_k(F_n) | \hat{F}_1\} \leq \frac{4}{n-n_1} \int (f_n^{(2k)}(x) - f_n^{(2k)}(x))^2 f_n(x) dx + \frac{4}{n-n_1} \int \text{var}(f_n^{(2k)}(x)) dx,$$

and

$$= o_p(n^{-1}).$$ \quad (3.10)

Now (3.9) and (3.10) imply the validity of (2.5). Since by Lemma 1, $f_n$ is uniformly bounded, the first part of Theorem 1 follows.

(ii) Suppose $k < m+\alpha \leq 2k+1/4$. We separate into two cases, $2k \leq m$, $2k > m$. If $2k \leq m$,

$$|E(f_n^{(2k)}(x) - f_n^{(2k)}(x))| = |\int h_n^{(2k)}(x-t)f_n(t)dt - f_n^{(2k)}(x)|$$

$$= |\int h_n^{(2k)}(x-t)h_n(t)dt - f_n^{(2k)}(x)|$$

$$= |\int (f_n^{(2k)}(x-\sigma t) - f_n^{(2k)}(x))h(t)dt|$$

$$= O(1),$$

so that

$$E(f_n^{(2k)}(x)) = O(1).$$ \quad (3.11)

Also,

$$\text{var}(f_n^{(2k)}(x)) \leq \frac{1}{n_1} \int (h_n^{(2k)}(x-t))^2 f_n(t) dt$$

$$\leq \frac{1}{n_1} \|f_n\|_\infty \sigma^{4k-1} \|h_n^{(2k)}\|_2.$$ \quad (3.12)
Then,
\[
E \mathrm{var} (\hat{\theta}_n | \bar{F}_n) \leq \frac{1}{n-n_1} \int E[|f^{(2k)}_1(x)|^2]f_n(x)dx \\
= O \left( n^{-2} \sigma^{4k-1} + n^{-1} \right) \\
= O \left( n^{-3(m+2k-2)/(1+4m+4k)} \right).
\]
(3.13)

If \(2k > m\) we compute
\[
|E \hat{f}^{(2k)}_1(x)| = \left| \int h^{(2k)}_\sigma(x-t)f_n(t)dt \right|
\]
\[
= \left| \int h^{(2k-m)}_\sigma(x-t)f^{(m)}_n(t)dt \right|
\]
\[
= \sigma^{-2k+m} \int h^{(2k-m)}_\sigma(\cdot) f^{(m)}(x-\cdot t)dt
\]
\[
= \sigma^{-2k+m} \int h^{(2k-m)}_n(t) \left\{ f^{(m)}_n(x-\cdot t) - f^{(m)}_n(x) \right\} dt
\]
\[
\leq g(x) \sigma^{m+2k-2k} \int |h^{(2k-m)}_n(t)| dt
\]
(3.14)

Again, by (3.12) and (3.14)
\[
E \mathrm{var} (\hat{\theta}_n | \bar{F}_n) = O \left( n^{-2} \sigma^{(4k+1)-1} + n^{-1} \sigma^{m+2k} \right)
\]
\[
= O \left( n^{-3(m+2k-2)/(1+4m+4k)} \right).
\]
(3.15)
The result follows by (3.13), (3.15) and (3.9). \(\square\)

**Proof of Theorem 2**: (i) Let \(\{F_r\}\) be a sequence of distributions with densities \(f_r\) and square roots of densities \(s_r\). Suppose \(\|s_r - s_0\|_2 \to 0\) and
\[
\int \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}f_r(x)dx \to 0.
\]

Write, with some abuse of notation, \(\theta_2(s_r) = \theta_2(F_r)\). Then,
\[
\theta_2(s_r) = \int \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}f_r(x)dx + \int \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}f_0(x)dx.
\]
Now
\[
\int f_0^{(2k)}(x) \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}dx = (-1)^k \int f_0^{(2k)}(x)f_r(x)dx - \theta_2(s_0)
\]
\[
= \int \{(-1)^k f_0^{(2k)}(x) - \theta_2(s_0)\}f_r(x)dx,
\]\nand
\[
\int \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}^2dx
\]
\[
= (-1)^k \int \{f_r(x) - f_0(x)\} \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}dx
\]
\[
= (-1)^k \int \{s_r(x) - s_0(x)\}^2 \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}dx
\]
\[
+ 2 \int \{s_r(x) - s_0(x)\} \{f_r^{(2k)}(x) - f_0^{(2k)}(x)\}dx
\]
\[
\leq \|f_r^{(2k)} + f_0^{(2k)}\|_{\infty} \|s_r - s_0\|_2 + 2\|s_r - s_0\|_2 \left\{ \|f_r^{(2k)}(x) - f_0^{(2k)}(x)\|^2 f_0(x)dx \right\}^{1/2}
\]
\[
= o(\|s_r - s_0\|_2).
\]
(3.16), (3.17) and (3.18) imply that
\[
\theta_2(s_r) = \theta_2(s_0) + 2\int \{-(-1)^k f_0^{(2k)}(x) - \theta_2(F_0)\}f_r(x)dx + O(\|s_r - s_0\|_2).
\]
This means that $\theta_k(s)$ is Fréchet differentiable along such paths with
derivative $4((-1)^k \frac{d^{(2j-1)}}{d\theta^{2j-1}} \theta_k(F_0))s_0$, and the result follows by standard theory.

(ii) Here, as in Ritov and Bickel (1987) we prove the assertion by presenting a sequence of Bayes problems. In the $i$th problem we observe $X_1, \ldots, X_n$ iid, $X_i \sim F_{\epsilon(F_{m,n} \theta)}$. The loss function is $L_{\theta, d} = I(\theta - d > c_{\theta}^{-1} n^{-\gamma})$. $F$ is picked according to a measure $\Pi_\theta$ to be described next. Note that the sequence $\Pi_\theta$, $\Pi_{\beta,...}$ is constructed such that the union of their supports $F^\theta$ is compact with $F_0$ its only accumulation point. Let $F_0 \in F_{m,n, \theta}$ be arbitrary. Clearly, $f_\theta$ is bounded away from zero on some interval. For simplicity we take this interval to be $[0, 1]$. To simplify the notation we assume also that

$$\sup_{x \in [0, 1]} g(x) \leq 1.$$ 

We now describe $\Pi_\theta$. Let $h_i$, $i = 0, 1, \ldots, \nu - 1$ be a sequence of functions such that

$$\int_0^1 h_i(x)dx = 0, \quad h_i(0) = h_i(1) = 0, \quad j = 0, \ldots, m + 1, \quad \int_0^1 (h_i(x))^2 \, dx = 1 \quad \text{and} \quad \int_0^1 h_i(x)(x - i)f_\theta(x)dx = 0. \quad \text{Let } \beta \text{ equal } 0, 1, \ldots, \nu - 1 \text{ with probability } 1/\nu \text{ and let } \Delta_0, \Delta_1, \ldots, \Delta_{\nu - 1} \text{ be iid, independent of } \beta \text{ and each equal to } -1 \text{ with probability } 1/2. \text{ Let } F \text{ be the random measure with density}

$$f(x) = f_\theta(x) + \beta \nu^{-m+\alpha} \Delta h_i(x - i) \text{ on } [i/\nu, (i+1)/\nu].$$

The measure that governs the selection of $F$ is $\Pi_\theta$. Clearly, for any $F$ in the support of $\Pi_\theta$ by our assumptions of $h_i$,

$$\theta_k(F) = \theta_k(F_0) + \beta \nu^{-2(m+a) + \varepsilon \theta k}.$$

That is $\theta_k(F)$ equals $\theta_k(F_0) + \beta \nu^{-2(m+a-k)}$ if $\beta = j$.

We show that if

$$4^{(m+a-1)} \to 0 \quad \text{(3.19)}$$

then the variational distance between the probability measures of $X_1, \ldots, X_n$ under $\beta = i$ and $\beta = j$ tends to $0$. Assume that this is the case and $F$ is distributed according to $\Pi_\theta$. $\Pi_\theta$ satisfies (3.19) and

$$\nu^{-2(m+a-k)} \gamma n^\gamma \to \infty \quad \text{(3.20)}$$

where

$$\gamma = 4(m+a-k)/(1+4m+4x).$$

This is possible if $k < m+\alpha$. If

$$A_n \{ |T_n - \theta_k(F_0)| < [c_n n^\gamma]^{-1} \}$$
then by construction for \( n \) sufficiently large the \( A_{n \beta} \) are disjoint. The Bayes risk for estimating \( \theta_k (F) \) using our loss function is

\[
R_n = \frac{1}{r} \sum_{j=1}^{r} P^{(0)}_j (A_{nj})
\]

\[
= 1 - \frac{1}{r} \sum_{j=1}^{r} P^{(0)}_j (A_{nj}).
\]

But, by the equivalence of \( P \left[ . \right| \beta = i \) and \( P \left[ . \right| \beta = j \) we have observed

\[
P^{(0)}_j (A_{nj}) = P^{(0)}_i (A_{nj}) \to 0 \text{ for each } j.
\]

So,

\[
\lim_{n} R_n \geq 1 - \frac{1}{r} \lim_{n} \sum_{j=1}^{r} P^{(0)}_j (A_{nj})
\]

\[
\geq 1 - \frac{1}{r} \lim_{n} P^{(0)}_j \left( \bigcup_{j=1}^{r} A_{nj} \right) \to 1 - \frac{1}{r}.
\]

Finally

\[
\inf_{T_n} \sup_{F \in \mathcal{F}^*} P_{F} [c_n, n^2 | T_n = \theta_k (F) | \geq 1] \geq R_n.
\]

Hence, since \( r \) is arbitrary,

\[
\lim_{n} \inf_{T_n} \sup_{F \in \mathcal{F}^*} P_{F} [c_n, n^2 | T_n = \theta_k (F) | \geq 1] = 1
\]

as advertised. This combines ideas of Hasminskii (1979) and Stone (1983).

We turn to the proof that (3.22) implies convergence of the variational distance. Let \( N_i \), \( i = 0, \ldots, r - 1 \) be the number of \( X \)'s in \([i/n, (i+1)/n)\) and let \( X_{i_1}, \ldots, X_{i_{N_i}} \) be the set of observations in that interval. Note that the random vector \((N_0, \ldots, N_{r-1})\) is independent of \( \beta \) and \((A_0, \ldots, A_{r-1})\), and that the blocks \((X_{i_1}, \ldots, X_{i_{N_i}})\) and \((X_{j_1}, \ldots, X_{j_{N_j}}), i \neq j\) are independent given \( N_i \) and \( N_j \). Without loss of generality consider \( \beta = 0 \) and \( \beta = 1 \).

The likelihood ratio of \( \beta = 1 \) to \( \beta = 0 \) is \( L = \prod_{i=0}^{r-1} L_i \) where

\[
L_i = 1/2 \prod_{j=1}^{N_i} \left( 1 + \nu^{-2(m+\alpha)} h(U_{ij}) f_j(U_{ij}) \right) + 1/2 \prod_{j=1}^{N_i} \left( 1 - \nu^{-2(m+\alpha)} h(U_{ij}) f_j(U_{ij}) \right)
\]

\[
= 1 + \sum_{i=0}^{[r/2]} \nu^{-2(m+\alpha) \sum_{k \neq i} \delta_{\text{all different}}} \prod_{j=1}^{N_i} \left( h(U_{ij}) f_j(U_{ij}) \right)
\]

where \( U_{ij} = \nu X_{ij} - i \) and \([x]\) is the greatest integer not larger than \( x\).
INTEGRATED SQUARED DENSITY DERIVATIVES

Note that, \( f_0(x) := \left[ \nu \left\{ F_0 \left( \frac{i+1}{\nu} \right) - F_0 \left( \frac{i}{\nu} \right) \right\} \right]^{-1} f_0 \left( \frac{i+x}{\nu} \right) \) is the density of \( U_\nu \) under \( f_0 \). We show that \( L \to \nu \) under \( f_0 \), which implies that the variational distance between the two conditional distributions tends to 0.

Since \( \int_0^1 h_i(x)dx = 0 \),
\[
E(L_4 - 1|N_4) = 0. \tag{3.21}
\]

Since \( \|f_0\|_2 < \infty \) by the lemma and the infimum of \( f_0 \) on \([0, 1]\) is \( > 0 \) by construction we obtain
\[
\int h_i^2(u) f_0(u) du = \int h_i^2(u) \left[ \nu \left( F_0 \left( \frac{i+1}{\nu} \right) - F_0 \left( \frac{i}{\nu} \right) \right) \right]^{-1} \left[ \frac{1}{\inf_{x \in [0, 1]} f_0(x)} \right] < \infty.
\]

Let \( A = \sup_i \int f_0^{-\nu}(u) f_0(u) h_i^2(u) du \). Then
\[
\text{var} \left( L_i - 1 | N_i \right) \leq \sum_{i=1}^{[\nu t/2]} \nu^{-4(m+2)i} \left( \frac{N_i}{2l} \right) A^{2l},
\]
and
\[
\text{var} \left( \sum_{i=1}^{[\nu t/2]} (L_i - 1) \right) = E \left[ \sum_{i=1}^{[\nu t/2]} (L_i - 1)^2 \right] \leq E \sum_{i=1}^{[\nu t/2]} \nu^{-4(m+2)i} \left( \frac{N_i}{2l} \right) A^{2l}. \tag{3.22}
\]

Let \( p_t = F_0 \left( \frac{(i+1)/\nu}{F_0(i/\nu)} \right) \). Straightforward calculations give
\[
E \sum_{i=1}^{[\nu t/2]} \nu^{-4(m+2)i} \left( \frac{N_i}{2l} \right) A^{2l}
\]
\[
= \sum_{j=2}^{n} \binom{n}{j} p_t^j (1-p_t)^{n-j} \sum_{i=1}^{[\nu t/2]} \left( \frac{A \nu^{-2(m+2)i} j}{2l} \right)
\]
\[
= \sum_{i=1}^{[\nu t/2]} (A \nu^{-2(m+2)i}) \left( \frac{n!}{(2l)!} \right) \left( \frac{1}{(n-j)!} \right) \left( \frac{p_t(1-p_t)^{n-j}}{j} \right)
\]
\[
= \sum_{i=1}^{[\nu t/2]} (A \nu^{-2(m+2)i}) \left( \frac{n!}{(2l)!} \right) \left( \frac{1}{j} \right) \left( \frac{(n-2l)!(n-j)!}{(n-2l-j)!} \right) \left( p_t h_i^2 \right) \left( \frac{N_i}{2l} \right) A^{2l}
\]
\[
\leq \sum_{i=1}^{[\nu t/2]} \left( \frac{1}{(2l)!} \right) \left( n A p_t \nu^{-2(m+2)i} \right) \left( \frac{N_i}{2l} \right) \left( \frac{1}{j} \right) \left( \frac{(n-2l)!(n-j)!}{(n-2l-j)!} \right) \left( p_t h_i^2 \right) \left( \frac{N_i}{2l} \right) A^{2l}
\]
\[
= (1-o(1))A^2 n^2 p_t^2 \nu^{-4(m+2)} = O(\nu^{-2(m+2)})^2 = 1 \tag{3.23}
\]
since \( \nu p_t < \|f_0\|_\infty \).
We obtain from (3.22) and (3.23) that
\[ \text{var} \left\{ \sum_{t=0}^{r-1} (L_t - 1) \right\} = O(n^{-2} r^{-3(m+2r+3)}) \]

Therefore, from (3.10) and (3.21) we obtain:
\[ \sum_{t=0}^{r-1} (L_t - 1) = o_p(1) \text{ and } \sum_{t=0}^{r-1} (L_t - 1)^2 = o_p(1) \]
both under \( F_0 \). Hence
\[ \log L = \sum_{t=0}^{r-1} (L_t - 1) + O \left( \sum_{t=0}^{r-1} (L_t - 1)^2 \right) \xrightarrow{P} 0 \]
under \( F_0 \) proving the assertion. \( \square \)

**Proof of Lemma 1**: It is enough to prove that for any \( \alpha_t > 0 \) and \( d_t < \infty \),
\[ \sup_{0 < |x-y| < 1} \left| \frac{|f^{(0)}(x) - f^{(0)}(y)|}{|x-y|^t} \right| \leq d_t \]  \hspace{1cm} (3.24)
implies that
\[ |f^{(0)}|_\infty \leq c_t \]  \hspace{1cm} (3.25)
where \( c_t < \infty \) is a function of \( \alpha_t \) and \( d_t \) only. Suppose (3.24) implies (3.25) then
\[ |f^{(t-1)}(x) - f^{(t-1)}(y)| = |f^{(0)}(x^*)| \cdot |x-y| \leq c_t |x-y| \text{ for } 0 < |x-y| \leq 1 \]
and the lemma follows by backward induction from \( m \).

Suppose (3.1) holds. Let \( b_t \) be an arbitrary number lying in \((0, 1]\) and assume that \( f^{(0)}(x) \geq d_t(b_t/2)^t \) for a point \( x \in R \). Then
\[ f^{(0)}(y) \geq a_t = f^{(0)}(x) - d_t(b_t/2)^t \geq 0 \]  \hspace{1cm} (3.26)
for all \( y \in [x-b_t/2, x+b_t/2] = J_t \).

Then \( f^{(t-1)}(u) \) is monotone on \( J_t \) and \( f^{(t-1)}(y), y \in J_t \), can be smaller than \( a_{t-1} \equiv \frac{1}{4d_t b_t} \) only on an interval of length smaller than \( 1/2b_t \). This leaves an interval \( J_{t-1} \) of length \( b_{t-1} \geq 1/4b_t \) on which either \( \inf_{y \in J_{t-1}} f^{(t-1)}(y) \geq a_{t-1} \) or \( \sup_{y \in J_{t-1}} f^{(t-1)}(y) \leq -a_{t-1} \). Continue this line of argument inductively and obtain that (3.26) entails that \( f(y) \gg a_0 \gg a_{b_t}^{(g(t+1))} \) on the interval \( J_0 \) whose length is \( b_0 \gg 4^{-t} b_t \). But \( f(\cdot) \) is a probability density function and hence
\[ 1 \geq a_0 b_0 \geq 2^{-4(t+1)} a_t b_t^{t+1}. \]
Therefore,
\[ f^{(i)}(x) = a_i d_i (b_i/2)^{q_i} \]
\[ \leq 2^{(i+2)} b_i^{-1/2(i+1)} + d_i (b_i/2)^{q_i}. \]

Hence \( f^{(i)} \) is bounded and the lemma follows. \( \square \)

**Acknowledgment.** P. Hall and S. Marron pointed out a gap in our original proof of Theorem 2 which we have corrected.

**References**


*Paper received: September, 1988.*