Sankhyā : The Indian Journal of Statistics
1988, Volume **50**, Series A, Pt. 3, pp. 381-393.

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 θ_k and similar parameters appear in Schweder (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 $n^{\gamma}(\hat{\theta}_k - \theta_k(F))$ 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, 62G20,

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 θ_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 $\theta_k(F) = \int \{f^{(k)}(x)\}^2\, dx$ 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 $F \in \boldsymbol{F}_{m,\alpha,g}$ where $\boldsymbol{F}_{m,\alpha,g} = \{F : |f^{(m)}(x) - f^{(m)}(x+\xi)| \leqslant g(x)|\xi|^\alpha$ for all x 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 :

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

and h has $2k+1$ derivatives.

Divide the sample into two subsamples $X_1, \ldots, X_{n_1}$ 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, $\hat{f}_i(x) = \int h_\sigma(x-y)d\hat{F}_i(y)$, $i = 1, 2$. The dependence of $\hat{f}_i$ on σ is left implicit. Consider the following estimator of θ_0.

$$\hat{\theta}_0^*(X_1, \ldots, X_n\,; \sigma) = \frac{n_1}{n}\hat{\theta}_{01}^* + \frac{n_2}{n}\hat{\theta}_{02}^* \qquad \ldots \quad (2.1)$$

where $n = n_1 + n_2$

$$\begin{aligned}
&\hat{\theta}_{01}^*(X_1, \ldots, X_n; \sigma)\\
&= \int \hat{f}_2^2(x)dx + 2n_1^{-1}\sum_{i=1}^{n_1}(\hat{f}_2(X_i) - \int \hat{f}_2^2(x)dx) + \frac{1}{n_2}\int h_\sigma^2(x)dx\\
&= 2\int h_\sigma(x-t)d\hat{F}_1(t)d\hat{F}_2(x) - n_2^{-2}\sum_{n_1+1 \leqslant i \neq j \leqslant n}\int h_\sigma(x-X_i)h_\sigma(x-X_j)dx
\end{aligned} \qquad \ldots \quad (2.2)$$

and $\hat{\theta}^*_{02}$ is obtained by interchanging the roles of the two subsamples in $\hat{\theta}^*_{01}$. The first two terms of $\hat{\theta}^*_{01}$ can be recgonized as Hasminskii and Ibragimov's estimate of this parameter which they show is efficient in $\mathrm{F}_{0,\alpha,M}$ if $\alpha > 1/2$. This is the, by now, familiar one step estimate (see Bickel, 1982 ; Schick, 1986) using the estimated influence function $2(\hat{f}_2 - \int \hat{f}_2^2(x)dx)$. The last term in (2.2) removes the pure known bias component, $n_2^{-2} \sum_{i=n_1+1}^{n} \int h_\sigma^2(x-X_i)dx$ from

$$\int \hat{f}_2^2(x)dx = n_2^{-2} \sum_{i,j} \int h_\sigma(x-X_i)h_\sigma(x-X_j)dx. \qquad \ldots \quad (2.3)$$

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

$$\begin{aligned}\theta_k(F) &= (-1)^k \int f^{(2k)}(x)f(x)dx \\ &= (-1)^k E_F(f^{(2k)}(X)).\end{aligned}$$

This suggests, by the same process as above, estimates $\hat{\theta}^*_{k1}$, $\hat{\theta}^*_{k2}$ and $\hat{\boldsymbol{\theta}}^*_k$. For convenience we replace $\hat{\theta}^*_{01}$ by $\hat{\theta}_{01}$ where n_2^{-2} in (2.2) is replaced by $[n_2(n_2-1)]^{-1}$ and similar replacements are made in $\hat{\theta}^*_{02}$ and more generally $\hat{\theta}^*_k$. So the estimate we study is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_k(X_1, \ldots, X_n; \sigma) &= 2(-1)^k \int h_\sigma^{(2k)}(x-t)d\hat{F}_1(t)d\hat{F}_2(x) \\ &-n_2[n\, n_1(n_1-1)]^{-1} \sum_{1 \leqslant i < j \leqslant n_1} \int h_\sigma^{(k)}(x)-X_i)h_\sigma^{(k)}(x-X_j)dx \\ &-n_1[n\, n_2(n_2-1)]^{-1} \sum_{n_1+1 \leqslant i < j \leqslant n} \int h_\sigma^{(k)}(x-X_i)h_\sigma^{(k)}(x-X_j)dx. \qquad \ldots \quad (2.4)\end{aligned}$$

Our main results are summarized in the following two theorems. In the first we describe the performance of $\hat{\boldsymbol{\theta}}_k$ in terms of the assumed family $\boldsymbol{F}_{m,\alpha,g}$. The rate of convergence of $\hat{\boldsymbol{\theta}}_k$ to $\theta_k(F)$ is a function of $m+\alpha$ and $\hat{\boldsymbol{\theta}}_k$ 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 $\{F_1, F_2, \ldots\} \subseteq \boldsymbol{F}_{m,\alpha,g}$ where $0 \leqslant \alpha < 1$, $m+\alpha > k$ and $g \in L_2 \cap L_\infty$. Let $X_{n1}, \ldots, X_{nn}$ be i.i.d., $X_{n1} \sim F_n$ and let $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_k(X_{n1}, \ldots, X_{nn}; \sigma_n)$ where $\sigma_n = n^{-2/(1+4m+4\alpha)}$.*

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

$$\sqrt{n}\Big[\, \hat{\boldsymbol{\theta}}_k - \theta_k(F_n) - \frac{2}{n} \sum_{i=1}^{n} \{(-1)^k f_n^{(2k)}\ (X_{ni}) - \theta_k(F_n)\} \Big] \longrightarrow 0. \qquad \ldots \quad (2.5)$$

Let $I_k(F_n) = [Var\{f_n^{(2k)}(X_{n1})\}]^{-1}$. Then, $n\, I_k(F_n)E\, \{\hat{\boldsymbol{\theta}}_k - \theta_k(F_n)\}^2 \rightarrow 1$ and $L\{\sqrt{n}\, I_k^{1/2}(F_n)\, (\hat{\boldsymbol{\theta}}_k - \theta_k(F_n))\} \rightarrow N(0, 1)$ provided $\limsup_n I_k(F_n) < \infty$.

(ii) *If* $k < m+\alpha \leqslant 2k+1/4$ *then* $n^{2\gamma}E\{\hat{\theta}_k-\theta_k(F_n)\}^2$ *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 σ by cross validation to obtain an estimate $\hat{\theta}_k^* = \hat{\theta}_k(X_{n1}, \ldots, X_{nn} ; \hat{\sigma}_n)$ which does not depend on m and α but is equivalent to $\hat{\theta}_k$ which does so depend through σ_n given in the statement of Theorem 1.

Theorem 2 : (i) *The information bound (in the sense of Khoshevnik and Levit* (1976)) *for non parametric estimation of* $\theta_k(F)$, $F \in \boldsymbol{F}_{2k,\alpha,g}$ *is given by* $I_k(F)$ *as defined in Theorem* 1.

(ii) *Suppose* $k < m+\alpha \leqslant 2k+1/4$. *Then there is a small compact set* $\boldsymbol{F}^* \subsetneqq \boldsymbol{F}_{m,\alpha,g}$ *such that for any* $c_n \to \infty$ *and any sequence of estimators* $T_1, T_2, \ldots, T_n = T_n(X_1, \ldots, X_n)$, $X_1, X_2, \ldots, X_n$ *iid*, $X_1 \sim F$:

$$\liminf_{n} \sup_{F \in \boldsymbol{F}^*} P_F\{c_n n^\gamma | T_n-\theta_k(F) | \geqslant 1\} = 1 \qquad \ldots \quad (2.6)$$

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

The proof of the first part of Theorem 2 is quite standard and follows essentially the discussion in Hasminskii 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, θ_0 can be estimated at the $n^{-1/2}$ rate in any one dimensional sub model of $\boldsymbol{F}_{m,\alpha,g}$ and the information bound of Theorem 2i) 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* α, m *and* g *be such that* $\alpha > 0$ $m \geqslant 0$ *and* $g \in L_\infty$. *Then* $sup\{|f^{(i)}(x)| : x, F \in \boldsymbol{F}_{m,\alpha,g}\} < \infty$, $i = 0, 1, \ldots, m$.

Proof of Theorem 1 : Evidently to establish Theorem 1 it is enough to consider the asymmetric estimate

$$\hat{\theta}_{k2} = 2(-1)^k \int\int h_\sigma^{(2k)}(x-t)\, d\hat{F}_1(t)\, d\hat{F}_2(x).$$

$$-2\{n_1(n_1-1)\}^{-1} \sum_{1 \leq i < j \leq n_1} \int h_\sigma^{(k)}(x-X_{ni})\, h_\sigma^{(k)}(x-X_{nj})\, dx.$$

We begin by estimating the conditional bias

$$E(\hat{\theta}_{k2} \mid \hat{F}_1) - \theta_k(F_n) = 2(-1)^k \int \hat{f}_1^{(2k)}(x) f_n(x)\, dx$$

$$-2\{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{i-1} \int h_\sigma^{(k)}(x-X_{ni})\, h_\sigma^{(k)}(x-X_{nj})\, dx - \int \{f_n^{(k)}(x)\}^2\, dx.$$

But

$$(-1)^k \int \hat{f}_1^{(2k)}(x) f_n(x) dx = \int \hat{f}_1^{(k)}(x) f_n^{(k)}(x) dx$$

$$= n_1^{-1} \sum_{i=1}^{n_1} \int h_\sigma^{(k)}(x-X_{ni}) f_n^{(k)}(x) dx$$

$$= \{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{1 \leq j \neq i \leq n_1} \int h_\sigma^{(k)}(x-X_{ni}) f_n^{(k)}(x) dx.$$

Hence

$$E(\hat{\theta}_{k2} \mid \hat{F}_1) - \theta_k(F_n) = -2\{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{i-1} \int \{h_\sigma^{(k)}(x-X_{ni}) - f_n^{(k)}(x)\}$$

$$\{h_\sigma^{(k)}(x-X_{nj}) - f_n^{(k)}(x)\} dx. \quad \ldots \quad (3.1)$$

We obtain from (3.1) that

$$E\,\hat{\theta}_{k2} - \theta_k(F_n) = \int \{f_{n\sigma}^{(k)}(x) - f_n^{(k)}(x)\}^2 dx \quad \ldots \quad (3.2)$$

where $f_{n\sigma} = f_n * h_\sigma$.

But

$$f_{n\sigma}^{(k)}(x) - f_n^{(k)}(x) = \int h(t) \{f_n^{(k)}(x+\sigma t) - f_n^{(k)}(x)\} dt$$

$$= \int h(t) \left\{ \sum_{i=1}^{m-k-1} \frac{f_n^{(k+i)}(x)}{i\,!} \sigma^i t^i \right\} dt \quad \ldots \quad (3.3)$$

$$+ \int h(t) \frac{1}{(m-k)!} \{f_n^{(m)}(x+\sigma^* t) - f_n^{(m)}(x)\} \sigma^{m-k} t^{m-k} dt,$$

where $0 \leq \sigma^* \leq \sigma$. The first term in the RHS of (3.3) is null by the construction of h. Since $F_n \in \boldsymbol{F}_{m,\alpha,g}$ we can bound the integrand in the second term and obtain :

$$|f_{n\sigma}^{(k)}(x) - f_n^{(k)}(x)| \leq g(x) \sigma^{m+\alpha-k} \int |t|^{m+\alpha-k} |h(t)|\, dt. \quad \ldots \quad (3.4)$$

Combine (3.2) and (3.4) to conclude that

$$|E\,\hat{\theta}_{k2}-\theta_k(F_n)| \leqslant \|g\|_2^2\, n^{-4(m+\alpha-k)/(1+4m+4\alpha)}\, (\int |t|^{m+\alpha-k}\, |h(t)|\, dt)^2. \quad \ldots \quad (3.5)$$

Next we estimate var $(E(\hat{\theta}_{k2}\,|\,\hat{F}_1))$. Note that $E(\hat{\theta}_{k2}\,|\,\hat{F}_1)$ was written in (3.1) as a U-statistic, $E(\hat{\theta}_{k2}\,|\,F_1)-\theta_k(F_n) = 2\{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{i-1} U(X_{ni}, X_{nj})$ say.

By standard U-statistic theory,

$$\text{var}\,\{E(\hat{\theta}_{k2}\,|\,\hat{F}_1)\} = n^{-1}\,\{O(\text{var}[E\,(U\,(X_{n1}, X_{n2})\,|\,X_{n1}])$$
$$+O\,(n^{-1}\;\text{var}\;\;U(X_{n1}, X_{n2}))\}. \quad \ldots \quad (3.6)$$

Now

$$E\;U\,(x, X_{n2}) = \int \{h_\sigma^{(k)}(t-x)-f_n^{(k)}(t)\}\;\{f_{n\sigma}^{(k)}(t)-f_n^{(k)}(t)\}\;dt$$
$$= \int \delta(t)\;\{h_\sigma^{(k)}(t-x)-f_n^{(k)}(t)\}dt,$$

say. Hence,

$$\text{var}\;[E\{U(X_{n1}, X_{n2})\,|\,X_{n1}\}] = E[\int\delta(x)\;\{h_\sigma^{(k)}(x-X_{n1})-f_{n\sigma}^{(k)}(x)\}\;dx]^2$$
$$= E\int\int \delta(y)\;\delta(x)\;\{h_\sigma^{(k)}(y-X_{n1})-f_{n\sigma}^{(k)}(y)\}\;\{h_\sigma^{(k)}(x-X_{n1})-f_{n\sigma}^{(k)}(x)\}\;dx\;dy$$
$$\leqslant \int\int \delta(y)\;\delta(x)\;\int h_\sigma^{(k)}(y-t)h_\sigma^{(k)}(x-t)f_n\,(t)\;dt\;dx\;dy$$
$$= \int \{\int\delta(x)h_\sigma^{(k)}(x-t)dx\}^2 f_n\,(t)\;dt$$
$$\leqslant \|\delta\|_\infty^2\;\sigma^{-2k}\;\{\int |\,h^{(k)}\,(x)\,|\;dx\}^2 = O(\sigma^{2(m+\alpha-2k)}) \quad \ldots \quad (3.7)$$

by (3.4). At the same time, the random variable $\int h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})dx$ is bounded by $\sigma^{-2k-1}\,\|h^{(k)}\|_2^2$ and is equal to zero unless $|X_{n1}-X_{n2}| \leqslant 2\sigma$. Since f_n is bounded this last event has probability of the same order as σ.

Hence

$$\text{var}\quad \{\int h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})dx\} = O\;(\sigma.\,\sigma^{-4k-2}).$$

Since $|\int f_n^{(k)}(x)\,.\,h_\sigma^{(k)}(x-X_{n1})dx| \leqslant \|f_n^{(k)}\|_\infty\,\sigma^{-k}\int\;|h^{(k)}(x)|\;dx$ we conclude that

$$\text{var}\,\{U\,(X_{n1}, X_{n2})\}$$
$$= \text{var}[\int \{h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})-f_n^{(k)}(x)h_\sigma^{(k)}(x-X_{n1})-\,f_n^{(k)}(x)h_\sigma^{(k)}(x-X_{n2})\}\;dx)]$$
$$= O(\sigma^{-4k-1}). \quad \ldots \quad (3.8)$$

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

$$\text{var}\,\{E\,(\hat{\theta}_{k2}\,|\,\hat{F}_1)\} = O(n^{-1}\sigma^{2(m+\alpha-2k)}+n^{-2}\sigma^{-4k-1})$$
$$= O(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}).$$

for σ given in the statement of Theorem 1. Hence (3.5) implies that

$$E\{E(\hat{\theta}_{k2} \mid \hat{F}_1) - \theta_k(F_n)\}^2 = O(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}). \quad \ldots \quad (3.9)$$

We have proved that $E(\hat{\theta}_{k2} \mid \hat{F}_1) - \theta_k(F_n)$ is of the right order (in particular it is $o_P(n^{-1/2})$ if $m+\alpha > 2k+1/4$). We turn to the investigation of the behaviour of $\hat{\theta}_{k2} - E(\hat{\theta}_{k2} \mid \hat{F}_1)$. This will be carried on separately for the two cases : $2k+1/4 < m+\alpha$ and $k < m+\alpha \leqslant 2k+1/4$.

(i) Suppose $2k+1/4 < m+\alpha$. In the light of (3.9) we need only to consider the conditional variance of $\hat{\theta}_{k2}$ given the first sub sample. But, given $X_{n1}, \ldots, X_{nn_1}$, $\hat{\theta}_{k2}$ is just a sum of *i.i.d.* random variables, hence

$$\operatorname{var}\left\{\hat{\theta}_{k2} - \frac{2(-1)^k}{n-n_1} \sum_{i=n_1+1}^{n} f_n^{(2k)}(X_{ni}) + \theta_k(F_n) \mid \hat{F}_1\right\}$$

$$\leqslant \frac{4}{n-n_1} \int \{\hat{f}_1^{(2k)}(x) - f_n^{(2k)}(x)\}^2 f_n(x)\, dx.$$

So
$$E \operatorname{var}\{\hat{\theta}_{k2} - 2(-1)^k \int f_n^{(2k)}(x)\, d\hat{F}_2(x) + \theta_k(F_n) \mid \hat{F}_1\}$$

$$\leqslant \frac{4}{n-n_1} \int \{f_{n\sigma}^{(2k)}(x) - f_n^{(2k)}(x)\}^2 f_n(x) dx + \frac{4}{n-n_1} \int \{\operatorname{var} \hat{f}_1^{(2k)}(x)\} dx.$$

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

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 $k < m+\alpha \leqslant 2k+1/4$. We separate into two cases, $2k \leqslant m$, $2k > m$. If $2k \leqslant m$,

$$\begin{aligned}
|E\hat{f}_1^{(2k)}(x) - f_n^{(2k)}(x)| &= |\int h_\sigma^{(2k)}(x-t) f_n(t) dt - f_n^{(2k)}(x)| \\
&= |\int f_n^{(2k)}(x-t) h_\sigma(t) dt - f_n^{(2k)}(x)| \\
&= |\int (f_n^{(2k)}(x-\sigma t) - f_n^{(2k)}(x)) h(t) dt| \\
&= O(1)
\end{aligned}$$

so that
$$E\hat{f}_1^{(2k)}(x) = O(1). \quad \ldots \quad (3.11)$$

Also,

$$\operatorname{var}\{\hat{f}_1^{(2k)}(x)\} \leqslant \frac{1}{n_1} \int \{h_\sigma^{(2k)}(x-t)\}^2 f_n(t) dt$$

$$\leqslant \frac{1}{n_1} \|f_n\|_\infty \, \sigma^{-4k-1} \|h^{(2k)}\|_2^2. \quad \ldots \quad (3.12)$$

Then,

$$E \text{ var } (\hat{\theta}_{k2} \mid \hat{F}_1) \leqslant \frac{1}{n-n_1} \int E[\{f_1^{(2k)}(x)\}^2] f_n(x)dx$$
$$= O(n^{-2}\sigma^{-4k-1}+n^{-1})$$
$$= O(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}). \quad \dots \quad (3.13)$$

If $2k > m$ we compute,

$$|E\hat{f}_1^{(2k)}(x)| = |\int h_\sigma^{(2k)}(x-t)f_n(t)dt|$$
$$= |\int h_\sigma^{(2k-m)}(x-t)f_n^{(m)}(t)dt|$$
$$= \sigma^{-2k+m} |\int h^{(2k-m)}(t)f_n^{(m)}(x-\sigma t)dt|$$
$$= \sigma^{-2k+m} |\int h^{(2k-m)}(t)\{f_n^{(m)}(x-\sigma t)-f_n^{(m)}(x)\}dt|$$
$$\leqslant g(x)\,\sigma^{m+\alpha-2k} \int |h^{(2k-m)}(t)|\,dt \quad \dots \quad (3.14)$$

Again, by (3.12) and (3.14)

$$E \text{ var } (\hat{\theta}_{k2} \mid \hat{F}_1) = O(n^{-2}\sigma^{-(4k+1)}+n^{-1}\sigma^{m+\alpha-2k})$$
$$= O(n^{-8(m+\alpha-k)(1+4m+4\alpha)}) \quad \dots \quad (3.15)$$

The result follows by (3.13), (3.15) and (3.9). □

Proof of Theorem 2 : (i) Let $\{F_\nu\}$ be a sequence of distributions with densities f_ν and square root of densities s_ν. Suppose $\|s_\nu - s_0\|_2^2 \to 0$ and $\int \{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}^2 f_0(x)dx \to 0$.

Write, with some abuse of notation, $\theta_k(s_\nu) = \theta_k(F_\nu)$. Then,

$$\theta_k(s_\nu) = \int \{f_0^{(k)}(x)\}^2dx + 2\int f_0^{(k)}(x)\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}dx + \int\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}^2\,dx. \quad \dots \quad (3.16)$$

Now

$$\int f_0^{(k)}(x)\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}dx = (-1)^k \int f_0^{(2k)}(x)f_\nu(x)dx - \theta_k(s_0)$$
$$= \int\{(-1)^k f_0^{(2k)}(x)-\theta_k(s_0)\}f_\nu(x)dx, \quad \dots \quad (3.17)$$

and

$$\int\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}^2dx$$
$$= (-1)^k\int\{f_\nu(x)-f_0(x)\}\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$= (-1)^k\int\{s_\nu(x)-s_0(x)\}^2\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$+2(-1)^k\int s_0(x)\{s_\nu(x)-s_0(x)\}\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$\leqslant \|f_0^{(2k)}+f_\nu^{(2k)}\|_\infty\|s_\nu-s_0\|_2^2 + 2\|s_\nu-s_0\|_2\,[\int\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}^2f_0(x)dx]^{1/2}$$
$$= o(\|s_\nu-s_0\|_2). \quad \dots \quad (3.18)$$

(3.16), (3.17) and (3.18) imply that

$$\theta_k(s_\nu) = \theta_k(s_0) + 2\int\{(-1)^k f_0^{(2k)}(x) - \theta_k(F_0)\}f_\nu(x)dx + O(\|s_\nu - s_0\|_2).$$

This means that $\theta_k(s)$ is Fréchet differentiable along such paths with derivative $4\{(-1)^k f_0^{(2k)} - \theta_k(F_0)\}s_0$ 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 nth problem we observe $X_1,\ldots,X_n$ iid, $X_1 \sim F \epsilon \boldsymbol{F}_{m,\alpha,g}$. The loss function is $\boldsymbol{L}_n(\theta, d) = 1_{\{|\theta - d| > c_n^{-1} n^{-\gamma}\}}$. F is picked according to a measure Π_ν to be described next. Note that the sequence $\Pi_1, \Pi_2,\ldots$ is constructed such that the union of their supports $\boldsymbol{F}^*$ is compact with F_0 its only accumulation point. Let $F_0 \epsilon \boldsymbol{F}_{m,\alpha,g}$ be arbitrary. Clearly, f_0 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 $\sup_{x\epsilon[0,1]} g(x) \leqslant 1$.

We now describe Π_ν. Let h_i, $i = 0, 1,\ldots, \nu-1$ be a sequence of functions such that $\int_0^1 h_i(x)dx = 0$, $h_i^{(j)}(0) = h_i^{(j)}(1) = 0$, $j = 0,\ldots, m+1$, $\int\{h_i^{(k)}(x)\}^2\, dx = 1$ and $\int_{i/\nu}^{(i+1)/\nu} h_i^{(k)}(\nu x - i) f_0^{(k)}(x)dx = 0$. Let β equal $0, 1, \ldots, r-1$ with probability $1/r$ and let $\Delta_0,\ldots, \Delta_{\nu-1}$ be iid, independent of β and each equal to ± 1 with probability $1/2$. Let F be the random measure with density

$$f(x) = f_0(x) + \beta\nu^{-(m+\alpha)}\Delta_i h_i(\nu x - i) \text{ on } [i/\nu, (i+1)/\nu).$$

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

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

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

We show that if

$$n^2\nu^{-(4m+4\alpha+1)} \to 0 \qquad \ldots \quad (3.19)$$

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 Π_ν, Π_ν satisfies (3.19) and

$$\nu^{-2(m+\alpha-k)} c_n n^\gamma \to \infty \qquad \ldots \quad (3.20)$$

where

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

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

$$A_{nj} = \{|T_n - \theta_k(F_j)| < [c_n n^\gamma]^{-1}\}$$

then by construction for n sufficiently large the A_{nj} are disjoint. The Bayes risk for estimating $\theta_k(F)$ using our loss function is

$$R_n = \frac{1}{r}\sum_{j=1}^{r} P_j^{(n)}(A_{nj}^C)$$

$$= 1-\frac{1}{r}\sum_{j=1}^{r} P_j^{(n)}(A_{nj}).$$

But, by the equivalence of $P[\,\cdot\,|\beta=i]$ and $P[\,\cdot\,|\beta=j]$ we have observed

$$P_j^{(n)}(A_{nj})-P_0^{(n)}(A_{nj})\to 0 \text{ for each } j.$$

So,

$$\underline{\lim}_n R_n \geqslant 1-\frac{1}{r}\,\overline{\lim}\sum_{j=1}^{r} P_0^{(n)}(A_{nj})$$

$$= 1-\frac{1}{r}\,\overline{\lim}\, P_0^{(n)}\Big(\bigcup_{j=1}^{r} A_{nj}\Big) \geqslant 1-\frac{1}{r}.$$

Finally

$$\inf_{T_n}\sup_{F\epsilon F^*} P_F[c_n n^{\gamma}|T_n-\theta_k(F)| \geqslant 1] \geqslant R_n.$$

Hence, since r is arbitrary,

$$\underline{\lim}\,\inf_{T_n}\sup_{F\epsilon F^*} P_F[c_n n^{\gamma}|T_n-\theta_k(F)| \geqslant 1] = 1$$

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, $i=0,\ldots,\nu-1$ be the number of X's in $[i/\nu, (i+1)/\nu)$ and let $X_{i1},\ldots,X_{iN_i}$ be the set of observations in that interval. Note that the random vector $(N_0,\ldots,N_{\nu-1})$ is independent of β and $(\Delta_0,\ldots,\Delta_{\nu-1})$, and that the blocks $(X_{i1},\ldots,X_{iN_i})$ and $(X_{j1},\ldots,X_{jN_j})$, $i\neq j$ 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 $L=\prod_{i=0}^{\nu-1} L_i$ where

$$L_i = 1/2\prod_{j=1}^{N_i}\{1+\nu^{-(m+\alpha)}h_i(U_{ij})/f_0(U_{ij})\}+1/2\prod_{j=1}^{N_i}\{1-\nu^{-(m+\alpha)}h_i(U_{ij})/f_0(U_{ij})\}$$

$$= 1+\sum_{l=1}^{[N_i/2]}\nu^{-2(m+\alpha)l}\sum_{\substack{j_1,\ldots,j_{2l}\\ \text{all different}}}\!\!\cdots\sum \frac{h_i(U_{ij_1})}{f_0(U_{ij_1})}\cdots\frac{h_i(U_{ij_{2l}})}{f_0(U_{ij_{2l}})}$$

where $U_{ij}=\nu X_{ij}-i$ and $[x]$ is the greatest integer not larger than x.

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

Since $\int_0^1 h_i(x)dx = 0$,

$$E(L_i-1 \mid N_i) = 0. \qquad \ldots \quad (3.21)$$

Since $\|f_0\| < \infty$ by the lemma and the infimum of f_0 on $[0, 1]$ is > 0 by construction we obtain

$$\int_0^1 \frac{h_i^2(u)}{f_0^2(u)} f_i(u)du = \int_0^1 \frac{h_i^2(u)}{f_0(u)} \left[\nu\left\{F_0\left(\frac{i+1}{\nu}\right)-F_0\left(\frac{i}{\nu}\right)\right\}\right]^{-1} \leqslant \frac{1}{[\inf_{x\in[0,1]} f_0(x)]^2} < \infty.$$

Let $A = \sup_i \int_0^1 f_0^{-2}(u)\, f_i(u)\, h_i^2(u)\, du$. Then

$$\operatorname{var}(L_i-1 \mid N_i) \leqslant \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l},$$

and

$$\operatorname{var}\left\{\sum_{i=0}^{\nu-1} (L_i-1)\right\} = E\left\{\sum_{i=0}^{\nu-1} (L_i-1)^2\right\} \leqslant E \sum_{i=0}^{\nu-1} \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l}. \quad \ldots \quad (3.22)$$

Let $p_i = F_0((i+1)/\nu)-F_0(i/\nu)$. Straightforward calculations give

$$E \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l}$$

$$= \sum_{j=2}^{n} \binom{n}{j} p_i^j (1-p_i)^{n-j} \sum_{l=1}^{[j/2]} (A\nu^{-2(m+\alpha)})^{2l} \binom{j}{2l}$$

$$= \sum_{l=1}^{[n/2]} (A\nu^{-2(m+\alpha)})^{2l} \frac{n!}{(2l)!} \sum_{j=2l}^{n} \frac{1}{(j-2l)!\,(n-j)!} p_i^j (1-p_i)^{n-j}$$

$$= \sum_{l=1}^{[n/2]} (A\nu^{-2(m+\alpha)})^{2l} \binom{n}{2l} \sum_{j=0}^{n-2l} \frac{(n-2l)!}{j!\,(n-2l-j)!} p_i^{j+2l} (1-p_i)^{n-2l-j}$$

$$= \sum_{l=1}^{[n/2]} (A\, p_i\, \nu^{-2(m+\alpha)})^{2l} \binom{n}{2l}$$

$$\leqslant \sum_{l=1}^{[n/2]} \frac{1}{(2l)!} (nA\, p_i\, \nu^{-2(m+\alpha)})^{2l} \leqslant \exp\{nAp_i\, \nu^{-2(m+\alpha)}\}^2 - 1 \qquad \ldots \quad (3.23)$$

$$= (1+o(1))A^2 n^2 p_i^2\, \nu^{-4(m+\alpha)} = O(n\, \nu^{-(2m+2\alpha+1)})^2$$

since $\nu\, p_i < \|f_0\|_\infty$.

We obtain from (3.22) and (3.23) that

$$\operatorname{var}\left\{\sum_{i=0}^{\nu-1}(L_i-1)\right\} = O(n^{-2}\nu^{-(4m+4\alpha+1)})$$

Therefore, from (3.19) and (3.21) we obtain :

$$\sum_{i=0}^{\nu-1}(L_i-1) = o_P(1) \text{ and } \sum_{i=0}^{\nu-1}(L_i-1)^2 = o_P(1)$$

both under F_0. Hence

$$\log L = \sum_{i=0}^{\nu-1}(L_i-1)+O\left(\sum_{i=0}^{\nu-1}(L_i-1)^2\right) \xrightarrow{P} 0$$

under F_0 proving the assertion. □

Proof of Lemma 1: It is enough to prove that for any $\alpha_i > 0$ and $d_i < \infty$,

$$\sup_{\substack{x,y \\ 0<|x-y|\leqslant 1}} \{|f^{(i)}(x)-f^{(i)}(y)|/|x-y|^{\alpha_i}\} \leqslant d_i \qquad \dots \quad (3.24)$$

implies that

$$\|f^{(i)}\|_\infty \leqslant c_i \qquad \dots \quad (3.25)$$

where $c_i < \infty$ is a function of α_i and d_i only. Suppose (3.24) implies (3.25) then

$$|f^{(i-1)}(x)-f^{(i-1)}(y)| = |f^{(i)}(x^*)|\,|x-y| \leqslant c_i|x-y| \text{ for } 0<|x-y|\leqslant 1$$

and the lemma follows by backward induction from m.

Suppose (3.1) holds. Let b_i be an arbitrary number lying in (0, 1] and assume that $f^{(i)}(x) \geqslant d_i(b_i/2)^{\alpha_i}$ for a point $x \in R$. Then

$$f^{(i)}(y) \geqslant a_i = f^{(i)}(x)-d_i(b_i/2)^{\alpha_i} \geqslant 0 \qquad \dots \quad (3.26)$$

for all $y \in [x-b_i/2, x+b_i/2] \equiv J_i$.

Then $f^{(i-1)}(u)$ is monotone on J_i and $|f^{(i-1)}(y)|$, $y \in J_i$, can be smaller than $a_{i-1} \equiv 1/4a_ib_i$ only on an interval of length smaller than $1/2b_i$. This leaves an interval J_{i-1} of length $b_{i-1} \geqslant 1/4b_i$ on which either $\inf_{y\in J_{i-1}} \{f^{(i-1)}(y)\} \geqslant a_{i-1}$ or $\sup_{y\in J_{i-1}} \{f^{(i-1)}(y)\} \leqslant -a_{i-1}$. Continue this line of argument inductively and obtain that (3.26) entails that $f(y) \geqslant a_0 \geqslant a_ib_i^i/2^{i(i+1)}$ on the interval J_0 whose length is $b_0 \geqslant 4^{-i}\, b_i$. But $f(\cdot)$ is a probability density function and hence

$$1 \geqslant a_0b_0 \geqslant 2^{-i(i+3)}\, a_i\, b_i^{i+1}.$$

Therefore,

$$f^{(i)}(x) = a_i + d_i(b_i/2)^{\alpha_i}$$
$$\leqslant 2^{i(i+3)} b_i^{-(i+1)} + d_i(b_i/2)^{\alpha_i}.$$

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

BICKEL, P. J. (1982): On adaptive estimation. *Ann. Statist.*, **10**, 647-671.

DONOHO, D. L. and LIU, R. C. (1987): Geometrizing rates of convergence I–III (manuscript).

FARREL, R. H. (1972): On the best obtainable asymptotic rates of convergence of a density function at a point. *Ann. Math. Stat.*, **43**, 170-180.

HALL, P. and MARRON, J. S. (1987): Estimation of integrated squared density derivatives. To appear in *Statistical and Probability Letters.*

HASMINSKII, R. Z. (1979): Lower bound for the risks of nonparametric estimates of the mode. *Contributions to Statistics*, (J. Hájek Memorial Volume). *Academia, Prague*, 91-97.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1978): On the non parametric estimation of functionals. *Symposium in Asymptotic Statistics, Prague,* 41-52.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1981): *Statistical Estimation,* Springer Verlag, New York 237-240.

KHOSHEVNIK, YU, A. and LEVIT, B.YA, (1976): On a non-parametric analogue of the information matrix. *Theor. Probab. Appl.*, **21**, 738-753.

PFANZAGL, J. (1982): Contributions to a general asymptotic statistical theory. *Lecture Notes in Statistics*, **13**, Springer-Verlag, New York.

PRAKASA RAO, B. L. S. (1983): *Nonparametric Functional Estimation,* Academic Press, New York.

RITOV, Y. and BICKEL, P. J. (1987): Achieving information bounds in non and semi-parametric models. To appear in *Ann. Statist.*

SCHICK, A. S. (1986): On asymptotically efficient estimators in semiparametric models. *Ann. Statist.*, **14**, 1139-1151.

SCHWEDER, T. (1975): Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian J. of Statist.*, **2**, 113-126.

STONE, C. J. (1983): Optimal uniform rate of convergence for nonparametric estimators of a density function and its derivatives. *Recent Advances in Statistics*, Paper in Honor of H. Chernoff. Academic Press, New York.

Paper received : September, 1988.