Achieving Information Bounds in Non and Semiparametric Models

By

Y. Ritov[1]
The Hebrew University of Jerusalem

and

P.J. Bickel[1]
University of California, Berkeley

Technical Report No. 116
September 1987
(revised August 1988)

Department of Statistics
University of California
Berkeley, California

Abstract

We consider in this paper two widely studied examples of non and semiparametric models in which the standard information bounds are totally misleading. In fact no estimators converge at the $n^{-\alpha}$ rate for any $\alpha > 0$, although the information is strictly positive "promising" that $n^{-1/2}$ is achievable. The examples are the estimation of $\int p^2$ and the slope in the Engle et al. model. A class of models in which the parameter of interest can be estimated efficiently is discussed.

Running head: Achieving Information bounds

AMS 1980 Subject Classification G2G20 G2G05

Key Words and Phrases: Rate of convergence, Nonparametric estimations, Functionals of a density.

Achieving Information Bounds in Non and Semiparametric Models.

By

Y. Ritov[1]
The Hebrew University of Jerusalem

and

P.J. Bickel[1]
University of California, Berkeley

1. Introduction.

Consider the standard simple random sampling model on a sample space **X**:

$X_1, \ldots, X_n$ i.i.d. according to $P \in \mathbf{P}$, a set of probability measures on **X** dominated by μ. Let p denote the density of P and $\theta : \mathbf{P} \to R$ be a parameter.

Suppose **P** is a regular parametric model, that is

1) $\mathbf{P} = \{P_{(\theta,\eta)} : \theta \in R, \eta \in R^m\}$ where if $s(\theta,\eta) = \left[\frac{dP_{(\theta,\eta)}}{d\mu}\right]^{1/2}$ the map $(\theta,\eta) \to s(\theta,\eta)$ is continuously Fréchet differentiable from R^{m+1} to $L_2(\mu)$, with derivative $\dot{s}(\theta,\eta)$ an m + 1 vector of elements of $L_2(\mu)$.

2) The Fisher information matrix, $I(\theta,\eta) = 4\,[\int \dot{s}_i(\theta,\eta)\,\dot{s}_j(\theta,\eta)d\mu\,]_{(m+1)\times(m+1)}$ where the $\dot{s}_i$ are the components of $\dot{s}$, is nonsingular.

Then it is known, see for example Hájek (1972), that if θ is identifiable it can be estimated at rate $\frac{1}{\sqrt{n}}$. In fact there exist $\hat{\theta}_n$ of "maximum likelihood" type which have the property that, if I^{11} is the first element of I^{-1}, then

$$\mathbf{L}_\theta\, X(n^{1/2}(\hat{\theta} - \theta)) \to N(0, I^{11}(\theta,\eta))$$

uniformly on compact subsets of R^{m+1} and I^{11} is the smallest asymptotic variance achievable by uniformly converging estimates.

[1] Research supported by ONR Grant N00014-80-C-0163.

Levit (1978), Pfanzagl (1982), Begun et al. (1983) have used an idea of Stein (1956) to extend these lower bounds to **P** non or semiparametric, provided that θ is pathwise Hellinger differentiable on **P**.

In this paper we investigate the question: Under the conditions of the above authors, are the bounds necessarily sharp if we drop the restriction that **P** is a regular parametric model?

We begin, in Section 2, by showing in the context of two widely studied examples, estimation of $\int p^2$, and of the regression coefficient in the Engle et. al. (1986) model that the answer is, in general, no. In fact, the rate $n^{1/2}$ is not even achievable pointwise. Although the arguments are specific they can evidently be generalized to show similar results for much broader classes of parameters. A general view on these phenomena is given in Donoho and Liu (1988).

In Section 3 we show that the information bounds are valid for a general class of semiparametric models. This class includes the regular parametric models and is rich enough to contain models having essentially any tangent space structure.

2. The bounds are not sharp.

The first example we consider is:

$$\mathbf{P} \equiv \{P \text{ on } [0,1]\text{: } P \text{ absolutely continuous with density } p \leq M\}$$

where M is a finite constant and,

$$\theta(p) = \int p^2(x)\,dx.$$

Since the functional $\theta(p)$ is differentiable along every Hellinger path in **P**, the regularity conditions required for validity of the information bound are satisfied. This functional appears in the asymptotic variance of the Hodges-Lehmann estimator. Similar functions (the integral of the square of the derivative of the density) appear in the theory of optimal density estimation.

It is well known, Pfanzagl (1982), Donoho and Liu (1988), that the information bound in this case is

$$4\,\mathrm{Var}_p(X) = 4\int (p(x) - \theta(p))^2 p(x)\,dx. \tag{2.1}$$

Hasminskii and Ibragimov (1979) following work of Schweder (1975) exhibit an estimate $\hat{\theta}_n$ such that $\sqrt{n}(\hat{\theta}_n - \theta(p))/2[\mathrm{Var}_p(X)]^{1/2}$ converges in law to $\mathbf{N}(0,1)$ uniformly on $\{P$ with densities p such that $\|p\|_\infty + \|p'\|_\infty \leq L\}$. Yet we can establish

Theorem 1:

For any $\varepsilon > 0$, there exists a subset $\mathbf{P}_0 \subset \mathbf{P}$ (compact in the topology induced by the variational norm and having diameter less than ε) such that for every sequence of estimators $\hat{\theta}_n$ and every $\alpha > 0$ there exists $P \in \mathbf{P}_0$ such that

$$\underline{\lim}_n P[|\hat{\theta}_n - \theta| \geq n^{-\alpha}] > 0. \tag{2.2}$$

□

A consequence of this result is that the rate of convergence on $\mathbf{P}_0$, as defined for example by Stone (1980), is slower than $n^{-\alpha}$ for any $\alpha > 0$. In fact, no sequence of estimators which is n^{α} consistent at each point of $\mathbf{P}_0$ exists. So the information bound is totally misleading for **P**.

To see what goes wrong we consider the behaviour of a plausible type of estimator. It is proved in Pfanzagl (1982) — See also Bickel et al (1989) (which we refer to in the sequel as BKRW) — that if $\hat{\theta}$ is efficient, then

$$\hat{\theta}_{\mathrm{eff}} = \theta(p) + 2n^{-1}\Sigma_{i=1}^{n}(p(X_i) - \theta(p)) + o_p(n^{-1/2}).$$

The naive approach to estimating θ efficiently is to try $\tilde{\theta} = \theta(\hat{p}_n) + 2n^{-1}\Sigma_{i=1}^{n}[\hat{p}_n(X_i) - \theta(\hat{p}_n)]$ for $\hat{p}_n$ an estimator of the density. For simplicity suppose $\hat{p}_n(\cdot)$ is based on an auxiliary sample. If $\tilde{\theta} = \hat{\theta}_{\mathrm{eff}} + o_p(n^{-1/2})$ we would expect

$$E(\tilde{\theta} \mid \hat{p}_n) = \int p^2(x)\,dx + O_p(n^{-1/2}).$$

But,

$$E(\tilde{\theta} \mid \hat{p}_n) - \int p^2(x)\,dx = 2\int \hat{p}_n(x)\,p(x)\,dx - \int \hat{p}_n^2(x)\,dx - \int p^2(x)\,dx$$
$$= -\int (\hat{p}_n(x) - p(x))^2\,dx$$

By Bretagnolle and Huber (1979) to have this last term be of order $n^{-1/2}$ uniformly for $p \in \mathbf{P}$ we need a Hölder condition of order at least $\frac{1}{2}$ on p in $\mathbf{P}$, viz. $|p(x) - p(y)| \le c\,|x - y|^{1/2}$. A positive result when p is so restricted has been obtained by Ibragimov and Hasminskii (1979). This argument cannot be translated into a proof since we have considered only estimates of a particular type in the discussion of the rate at which p can be estimated. In fact, a cleverer construction — see Bickel and Ritov (1988) shows that a Hölder condition of order $\frac{1}{4}$ suffices. However, we hope the point is clear. The calculations leading to the information bound are local. They are irrelevant to actual performance if you can't even get to within $o_p(n^{-1/4})$ of $\theta(p)$.

We begin with a simpler construction which establishes

Theorem 2:

For any sequence of estimates $\hat{\theta}_n$ there exists a compact $\mathbf{P}_0$ for which the uniform rate of convergence is slower than a_n, for any sequence $a_n \to 0$, viz.

$$\underline{\lim}_n \sup_{\mathbf{P}_0} P[|\hat{\theta}_n - \theta| \ge a_n] > 0 \tag{2.3}$$

Note that (2.3) implies the existence of $\varepsilon > 0$ such that

$$\underline{\lim}_n \sup_{\mathbf{P}_0} P[|\hat{\theta}_n - \theta| \ge \varepsilon] > 0.$$

The main idea of the proof is a "Bayesian" construction. We exhibit a sequence of prior distributions π_n assigning mass $\frac{1}{2}$ each to finite subsets H_{0n} of $\{P: \theta(P) = 1 + \frac{4}{3}a_n\}$ and H_{1n} of $\{P: \theta(P) = 1 + \frac{16}{3}a_n\}$ whose size $k(n) \uparrow \infty$ such that the posterior probabilities of H_{1n}, H_{0n} given $X_1, \ldots, X_n$ are with probability tending to 1, still equal to $\frac{1}{2}$. More explicitly, the members p_{jln}, $l = 1, \ldots, k(n)$, of H_{jn}, $j = 0,1$ are equally likely a priori and are chosen so that with probability tending to 1

$$k^{-1}(n) \sum_{l=1}^{k(n)} \prod_{i=1}^{n} p_{0ln}(X_i) = k^{-1}(n) \sum_{l=1}^{k(n)} \prod_{i=1}^{n} p_{1ln}(X_i) = \prod_{i=1}^{n} p(X_i)$$

where p is the uniform distribution on (0, 1) (though this is inessential). Define $\mathbf{P}_0$ to

be this countable collection of P_{jlm}'s together with their limit, the uniform. An immediate consequence from which (2.3) follows is that,

$$\inf_{\hat{\theta}_n} \int P[|\hat{\theta}_n - \theta| \geq a_n] \pi_n(dP) \rightarrow \frac{1}{2},$$

and this establishes the theorem. This construction differs from similar constructions appearing in the density estimation literature where the corresponding H_{0n}, H_{1n} are simple (consist of one point).

Proof of Theorem 2:

Here is the sequence of priors the union of whose carriers is a set having the uniform distribution on (0, 1) as its limit. We prescribe π_n through some auxiliary variables

1. Let $\alpha_n = c_n$ with probability $\frac{1}{2}$

   $= 2c_n$ with probability $\frac{1}{2}$; the sequence $c_n \downarrow 0$ is to be chosen later.

2. Let $\Delta_0, \ldots, \Delta_m$, $m = n^3$, be independent identically distributed random variables independent of α_n and equal to ± 1 with probability $\frac{1}{2}$.

π_n is the distribution of the random density p given by

$$p((i + y)(m + 1)^{-1}) = 1 + \Delta_i \alpha_n h(y), \quad i = 0, \ldots, m \quad 0 \leq y \leq 1$$

where (say)

$$h(t) = t, \qquad 0 \leq t < \frac{1}{2}$$

$$= -(1 - t) \qquad \frac{1}{2} \leq t \leq 1.$$

The support of each π_n is finite and $\int |p - 1| \leq 2c_n$ with π_n probability 1, so the union of the supports of π_n is a sequence tending to the uniform. Now, if P corresponds to the random p

$$\theta(P) = \int p^2(x)\,dx = (m + 1)^{-1} \Sigma_{i=0}^{m} \int_0^1 (1 + \Delta_i \alpha_n h(y))^2\,dy$$

$$= 1 + \frac{\alpha_n^2}{12}.$$

This construction, since $m = n^3$, has the property that the π_n probability that at most one of the observed $X_1, \ldots, X_n$ will fall into any of the intervals $[\frac{i}{m+1}, \frac{i+1}{m+1})$

is $1 - O(n^{-1})$. But 1 observation in a cell gives no new information on whether $\alpha_n = c_n$ or $2c_n$ and so the posterior probability

$$(2.4) \qquad \pi_n\{\theta = 1 + \frac{c_n^2}{12} | X_1, \ldots, X_n\} = \pi_n\{\theta = 1 + \frac{c_n^2}{3} | X_1, \ldots, X_n\}$$

$$= \frac{1}{2} + o_{\pi_n}(1).$$

Let $c_n = 3a_n^{1/2}$. Then (2.4) implies that $\inf_{\hat{\theta}} P(|\hat{\theta}_n - \theta| > a_n | X_1, X_2, \ldots, X_n) \xrightarrow{\pi_n} 1/2$, or, for any $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$,

$$\int P[|\hat{\theta}_n - \theta| \geq a_n] \pi_n(dP) \rightarrow 1/2.$$

Then

$$\underline{\lim}_n \sup_{P_0} P[|\hat{\theta}_n - \theta| > a_n] \geq \underline{\lim}_n \int P[|\hat{\theta}_n - \theta| > a_n] \pi_n(dP)$$

$$= 1/2$$

and (2.3) follows. To check (2.4) note that if at most one X_i falls in each interval the posterior distribution of $(\alpha_n, \Delta_0, \ldots, \Delta_m)$ is

$$(2.5) \qquad \pi_n(\alpha, \Delta_0, \ldots, \Delta_m | X_1, \ldots, X_n) =$$

$$= 2^{-(m+2)} \Pi_{i=0}^m \{\frac{1 + \Delta_i}{2} f_\alpha^+(Y_i) + \frac{1 - \Delta_i}{2} f_\alpha^-(Y_i)\}^{\delta_i} c(X_1, \ldots, X_n)$$

$$= \Pi_{i=0}^m \{1 + \Delta_i \alpha h(Y_i)\}^{\delta_i}$$

where

$$f_\alpha^\pm(y) = 1 \pm \alpha h(y)$$

$$\delta_i = 1 \text{ if there exists } X_{j_i} \in [\frac{i}{m+1}, \frac{i+1}{m+1})$$

$$= 0 \text{ otherwise}$$

and Y_i is the fractional part of $(m + 1)X_{j_i}$. By symmetry, from (2.5),

$$\pi_n(\alpha_n = c_n | X_1, \ldots, X_n) = \frac{1}{2}$$

and (2.4) follows. □

Theorem 1 again uses a Bayesian construction. For the conclusion we can not reduce our problem from estimation to testing but have to construct a prior distribution with infinite support whose Bayes risk for the loss function $l_n(\theta, \hat{\theta}) = 1(|\hat{\theta} - \theta| \geq a_n)$ is

bounded away from 0.

Proof of Theorem 1:

We exhibit a $\mathbf{P}_0$ contained in the ε ball around $\mathbf{U}(0,1)$ and π_0 concentrating on $\mathbf{P}_0$ such that for all $\alpha > 0$

$$\underline{\lim}_n \inf_{\hat{\theta}_n} \int P[|\hat{\theta}_n - \theta| \geq n^{-\alpha}]\, \pi_0(dP) \geq \frac{1}{4}. \tag{2.6}$$

Then (2.2) follows. Otherwise, we could exhibit $\alpha > 0$, $\hat{\theta}_n$ such that for all P

$$P[|\hat{\theta}_n - \theta| \geq n^{-\alpha}] \to 0$$

which by dominated convergence would imply

$$\int P[|\hat{\theta}_n - \theta| \geq n^{-\alpha}]\, \pi_0(dP) \to 0$$

contradicting (2.6). Here is π_0. Let α_k, $\Delta_k(0), \ldots, \Delta_k(2^k - 1)$, $k = 1,2,\ldots$ be independent, $\alpha_k = 0$ or 1 with probability $\frac{1}{2}$, each $\Delta_k(i) = \pm 1$ with probability $\frac{1}{2}$ each. Define random functions

$$h_k(x) = \Delta_k(i), \qquad i2^{-k} \leq x < (i + \tfrac{1}{2})2^{-k} \tag{2.7}$$

$$= -\Delta_k(i), \quad (i + \tfrac{1}{2})2^{-k} \leq x < (i+1)2^{-k}.$$

Finally, the random density p is given by

$$p(x) = 1 + \sum_{k=1}^{\infty} c_k \alpha_k h_k(x)$$

where the c_k are positive $\Sigma_{k=1}^{\infty} c_k < \frac{\varepsilon}{2}$. Note that since $\int h_i(x)\,dx = 0$, $\int h_i h_j(x)\,dx = \delta_{ij}$

$$\theta(P) = 1 + \sum_{i=1}^{\infty} \alpha_i^2 c_i^2$$

$$= 1 + \Sigma_{i=1}^{m-1} \alpha_i^2 c_i^2 + \Sigma_{i=m}^{\infty} \alpha_i^2 c_i^2.$$

Let $\beta = (\alpha_1, \ldots, \alpha_{k-1})$ and $\pi_{0\beta}$ be the conditional distribution of all the α's and Δ's given β. For any bounded loss function $L(\theta, a)$

$$\inf_\delta E_{\pi_0} L(\theta, \delta) = \inf_\delta \int E_{\pi_{0\beta}} L(\theta, \delta)\, \nu(d\beta) \tag{2.8}$$

$$\geq \int \inf_\delta E_{\pi_{0\beta}} L(\theta, \delta)\, \nu(d\beta)$$

where δ ranges over all estimates of θ based on $X_1, \ldots, X_n$ and ν is the marginal distribution of β. Therefore, there exists a value β_0 of β such that the Bayes risk of π_0

is no smaller than the Bayes risk of $\pi_{0\beta_0} \equiv \pi_{00}$. Under π_{00}, if $m = [3\log_2 n]$ any interval of the form $[i2^{-m}, (i+1)2^{-m})$ contains at most one of $X_1, \ldots, X_n$ with probability $\geq 1 - (2n)^{-1}$. Arguing as before, under π_{00}, except on a set of probability $O(n^{-1})$, the conditional distribution of $\Delta \equiv \{\Delta_k(i): 1 \leq i \leq 2^k, k \geq m\}$ given $X_1, \ldots, X_n$ is the same as the marginal distribution. We claim that the same is true of the conditional distribution of $\alpha = \{\alpha_k, \ldots, k \geq m\}$. Write the joint density of $(\alpha, \Delta, X_1, \ldots, X_n)$ with respect to the measure μ where, under μ, the α_k's and $\Delta_k(i)$ have the distribution specified earlier and $X_1, \ldots, X_n$ are independent of α, Δ and uniform $(0,1)$ as

$$\Pi_{i=1}^{n}(1 + \Sigma_{k=1}^{m-1} c_k \alpha_{k0} h_k(X_i) + \sum_{k=m}^{\infty} c_k \alpha_k h_k(X_i)).$$

The posterior density, if at most one X_i is in each interval $[\frac{i}{2^k}, \frac{i+1}{2^k})$, $k \geq m$, is proportional to

$$\Pi_{i=1}^{n}(A_i(X_i) + \Sigma_{k=m}^{\infty} c_k \alpha_k \varepsilon_k(X_i)\Delta_{ki})$$

where $A_i(x) = 1 + \Sigma_{k=1}^{m-1} c_k \alpha_{k0} h_k(x)$, $\Delta_{ki} = \Delta_k(j)$ iff j is such that $X_i \in [j2^{-k}, (j+1)2^{-k})$ and

$$\varepsilon_k(X_i) = +1 \text{ if } X_i \in [j2^{-k}, (j + \tfrac{1}{2})2^{-k})$$

$$= -1 \text{ if } X_i \in [(j + \tfrac{1}{2})2^{-k}, (j+1)2^{-k}).$$

Then the posterior probability that $(\alpha_{m+1}, \ldots, \alpha_{m+t}) = (\alpha_{m+1}^0, \ldots, \alpha_{m+t}^0)$ given $X_1 = x_1, \ldots, X_n = x_n$ is proportional to

$$E_\mu\{\Pi_{i=1}^{n}(A_i(X_i) + \Sigma_{k=m+1}^{m+t} c_k \alpha_k^0 \varepsilon_k(X_i)\Delta_{ki} + \sum_{k=m+t+1}^{\infty} c_k \alpha_k \varepsilon_k(X_i)\Delta_{ki})$$

$$1(\alpha_{m+1} = \alpha_{m+1}^0, \ldots, \alpha_{m+t} = \alpha_{m+t}^0)\}.$$

But the α_k and the Δ_{ki} are independent under μ. Multiplying out the product and using the symmetry of the Δ_{ki} we obtain that the posterior probability is proportional to $\Pi_{i=1}^{n} A_i(X_i)$ and our claim follows. To complete the argument note that, under π_{00} if $B_m = \Sigma_{k=m}^{\infty} c_k^2(\alpha_k^2 - \frac{1}{2})$

$$P[B_m \geq \frac{1}{2}c_m^2] \geq P[\alpha_m = 1, \Sigma_{k=m+1}^{\infty} c_k^2(\alpha_k^2 - \tfrac{1}{2}) \geq 0]$$

$$\geq \frac{1}{4} \quad \text{by the symmetry and independence of } \alpha_m, \text{ and } \alpha_k^2 - \tfrac{1}{2}, k = m+1, \ldots$$

A similar argument shows

$$P[B_m \le -\frac{1}{2}c_m^2] \ge 1/4.$$

Hence, if at most one X_i falls in each interval,

$$\inf_a P[|\theta - a| \ge \frac{1}{2}c_m^2 | X_1, \ldots, X_n]$$

$$\ge \min\{P[B_m \ge \frac{1}{2}c_m^2 | X_1, \ldots, X_n],\ P[B_m \le -\frac{1}{2}c_m^2 | X_1, \ldots, X_n]\}$$

$$\ge \frac{1}{4} + O_p(n^{-1})$$

since except on a set of probability $O(n^{-1})$ the marginal and conditional distributions of B_m agree. So the Bayes risk of π_{00} for the loss function $L_m(\theta, a) = 1[|\theta - a| \ge \frac{1}{2}c_m^2]$ is $\ge \frac{1}{4} + O(n^{-1})$. If $c_m = 9\varepsilon^2[\log n]^{-1-\varepsilon}$, say, then (2.6) follows from (2.8). □

In the Engle et al. model (1986) we observe $X_i = (W_i, Z_i, Y_i)$, $i = 1, \ldots, n$ where

$$Y = \beta W + t(Z) + \varepsilon \tag{2.9}$$

and $\varepsilon \sim N(0, \sigma^2)$. The joint distribution of (W, Z) and t are unknown. In recent work, Hung Chen (1988) and Cuzick (1987) have exhibited, under various smoothness restrictions on t, estimates $\hat{\beta}$ which are asymptotically $N(0, \frac{I^{-1}}{n})$ where

$$I = \sigma^{-2}E(W - E(W|Z))^2 > 0 \tag{2.10}$$

unless W is a function of Z.

Local calculations yield this as the information bound whenever $W \in L_2$. Let

$$\mathbf{P} = \{\text{All distributions } (W, Z, Y) \text{ given by (2.9) such that } I > 0 \text{ and well defined.}\}$$

We show

Theorem 3: a) Even if $\sigma = 0$ (or equivalently I given by (2.10) equals ∞) there exists a subset $\mathbf{P}_0$ of $\mathbf{P}$ such that for all estimates $\hat{\beta}_n$

$$\sup_{\mathbf{P}_0} P[|\hat{\beta}_n - \beta| \ge \varepsilon] > 0 \quad \text{for any } \varepsilon > 0. \tag{2.11}$$

b) For $\sigma > 0$ there exists a compact subset $\mathbf{P}_0$ of $\mathbf{P}$ such that for all estimates $\hat{\beta}_n$ and all $\gamma > 0$,

$$\underline{\lim}_n \sup_{\mathbf{P}_0}[|\hat{\beta} - \beta| \ge n^{-\gamma}] > 0.$$

We argue as for Theorem 2.

Proof: a) We give the simpler construction for $\sigma = 0$ and $\mathbf{P}_0$ non compact and sketch it for $\sigma > 0$ and $\mathbf{P}_0$ compact. Here is the prior π_n. Take $W = \pm 1$ with probability $\frac{1}{2}$ and $0 \le Z \le 1$.

Let $\alpha, \Delta_0, \ldots, \Delta_m$,$m = n^3$ be i.i.d. and equal to ± 1 with probability $\frac{1}{2}$. If $\alpha = -1$ then $\beta = 0$, $Z \sim \mathbf{U}(0,1)$ independent of W and $t(z) \equiv 0$. If $\alpha = 1$, then $\beta = c$ and the conditional density of $Z|W$ and $t(\cdot)$ are given by,

$$\left.\begin{aligned} p(z|w) &= 1 - \Delta_i w \\ t(z) &= c\Delta_i \end{aligned}\right\}, \qquad i(m+1)^{-1} \le z < (i + \tfrac{1}{2})/(m+1) \tag{2.12}$$

$$\left.\begin{aligned} p(z|w) &= 1 + \Delta_i w \\ t(z) &= -c\Delta_i \end{aligned}\right\}, \qquad (i + \tfrac{1}{2})(m+1)^{-1} \le z < (i+1)/(m+1).$$

Again with probability $1 - O(n^{-1})$ the posterior of $\Delta_1, \ldots, \Delta_m$ is the same as the prior distribution. Note also by construction that $\beta W + t(Z) \equiv 0$. So, with probability $1 - O(n^{-1})$

$$P[\alpha = 1 | W_i, Z_i, Y_i, i = 1, \ldots, n] = P[\alpha = 1 | W_i, Z_i, i = 1, \ldots, n]$$

is proportional to,

$$E\{\Pi_{i=1}^{n} (1 - \Delta_i W_i)^{\delta_i} (1 + \Delta_i W_i)^{1-\delta_i}\} \tag{2.13}$$

where $W_1, Z_1, \ldots, W_n, Z_n$ are fixed. If Z_i falls in $[j_i/(m+1), (j_i+1)/(m+1))$, we define $\delta_i = 1$ if Z_i is in the first half of that interval and 0 if it is in the second. The expectation in (2.13) is again 1 and we conclude that the posterior distribution of α is the same as its prior and hence that the Bayes risk of π_n is bounded away from 0. (2.11) follows.

b) If $\sigma = 1$ (say) proceed as follows. Let $\alpha, \Delta_1, \ldots, \Delta_m$ be as above. Suppose $P[W = 0] = P[W = 1] = \frac{1}{2}$ and that the conditional distribution of Z given $W = 0$ is $\mathbf{U}(0,1)$. Under π_n if $\alpha = -1$, $\beta = 0$ and Z given $W = 1$ is also $\mathbf{U}(0,1)$. Let

$$\begin{aligned} t_n(z) &= a_n \Delta_i, && i/(m+1) \le z < (i + \tfrac{1}{2})/(m+1) \\ &= -a_n \Delta_i, && (i + \tfrac{1}{2})/(m+1) \le z < i/(m+1). \end{aligned}$$

If $\alpha = 1$, $\beta = c_n$ and

$$\begin{aligned} p(z|W=1) &= 1 - b_n \Delta_i, && i/(m+1) \le z < (i + \tfrac{1}{2})/(m+1) \\ &= 1 + b_n \Delta_i, && (i + \tfrac{1}{2})/(m+1) \le z < (i+1)/(m+1). \end{aligned} \tag{2.14}$$

With probability $1 - O(n^{-1})$ there is at most one Z_i in each interval $[i(m+1)^{-1}, (i+1)(m+1)^{-1})$. Conditional on that event, being given (W_i, Z_i, Y_i) is the same as being given (W_i, V_i, Y_i) where V_i is the fractional part of $(m+1)Z_i$. Further, the posterior distribution of β is the same as the conditional distribution of β given, $\{(V_i, Y_i): W_i = 1\}$. Given $W_i = 1$, V_i is $U(0,1)$ by (2.14) since the conditional distribution of Δ_{j_i} given $W_i = 1$, where $Z_i \in (j_i/(m+1), (j_i+1)/(m+1))$ is the same as its prior.

Finally, the conditional density of Y_i given $W_i = 1$, V_i, $\alpha = 1$, is

$$\frac{1}{2}(1 - b_n)\phi(y - c_n - a_n) + \frac{1}{2}(1 + b_n)\phi(y - c_n + a_n)$$

$$= \phi(y) + y\phi(y)(c_n - a_n b_n) + O(c_n^2 + a_n^2).$$

If $a_n = c_n^{1-\delta}$, $b_n = c_n^{\delta}$, $\delta > 0$ the density of Y_i given $W_i = 1$, V_i, $\alpha = 1$ is $\phi(y)(1 + c_n^{2-2\delta} h(y) + O(c_n^3 + a_n^3))$ where $\int \phi(y) h(y)\, dy = 0$. One can show the joint distribution of $\{(V_i, Y_i): W_i = 1\}$ under $\alpha = 1$ is contiguous to that under $\alpha = 0$ provided $c_n^{2-2\delta} = O(n^{-1/2})$. Hence by taking $c_n = n^{-1/4+\varepsilon}$, $\varepsilon > 0$ arbitrary we can deduce that β cannot be estimated at a rate better than $n^{-1/4+\varepsilon}$. □

3. Validity of the bounds for a class of models.

We consider semiparametric models with the following structure:

(3.1) $$\mathbf{P} = \bigcup_{m=1}^{\infty} \mathbf{P}_m, \quad \mathbf{P}_m \subset \mathbf{P}_{m+1}, \quad \forall m$$

and $\mathbf{P}_m$ regular parametric. That is, we can write

$$\mathbf{P}_m = \{P_{(\theta,\eta^m)}: \theta \in \Theta, \ \eta^m = (\eta_1, \ldots, \eta_{d-1}), \text{ with } d = d(m)$$

$$\text{and } \eta_j \in E_j,\ j = 1, \ldots, d-1,\ E_j, \Theta \text{ open subsets of } R\}$$

i) $\mathbf{P} \ll \mu$.

ii) The maps $(\theta, \eta^m) \to P_{(\theta,\eta^m)}$ are 1-1 for all m. Further if $P \in \mathbf{P}_m = \mathbf{P}_m \cap \mathbf{P}_{m'}$, $m' > m$ then the first $d(m)$ coordinates of $\eta^{m'}$ agree with η^m.

iii) The maps $(\theta, \eta^m) \to s(\theta, \eta^m) \equiv (\frac{dP_{(\theta,\eta^m)}}{d\mu})^{1/2} \in L_2(\mu)$ are continuously Fréchet differentiable with derivative $\dot{s}(\theta, \eta^m) = (\dot{s}_1, \ldots, \dot{s}_d)(\theta, \eta^m)$, $\dot{s}_j \in L_2(\mu)$, $j = 1, \ldots, d$.

iv) The information matrix,

$$I(\theta, \eta^m) \equiv 4[\int \dot{s}_i \dot{s}_j (\theta, \eta^m)\, d\mu]_{d\times d} = [E_{(\theta,\eta^m)} \dot{l}_i \dot{l}_j (\theta, \eta^m)]_{d\times d}$$

is nonsingular for all (θ, η^m) where $\dot{l}(\theta, \eta^m) = 2\frac{\dot{s}}{s}(\theta, \eta^m)$ is the derivative of the log likelihood.

In words, every member of $\mathbf{P}$ belongs to a nice parametric model whose dimension d can however be arbitrarily large. A moment's thought will show that most if not all semiparametric models proposed in the literature can be thought of as the closures (for weak convergence) of such $\mathbf{P}$. For example, the symmetric location model $\{P: P$ is absolutely continuous on R, symmetric about some $\theta \in R\}$ is the closure of $\mathbf{P}$ as in (3.1) where $P_{(\theta,\eta^m)}$, for example, has

$$\log p_{(\theta,\eta^m)}(x) = h(x - \theta, \eta^m)$$

where

$$h''(x, \eta^m) = \Sigma_{k=1}^{d-1} \eta_k 1(|x| < b_{km})$$

where $d = 2^m + 1$, $b_{km} = mk2^{-m}$, $k = 1, \ldots, d-1$. That is we assume that the log density of $X - \theta$ is a symmetric quadratic spline with knots at $\pm b_{km}$ which is constant for $|x| > m$. Such models have been considered by Faraway (1987) and Stone (1986) among others. It is well known, see Le Cam (1956), Bickel (1982), that there exist estimates $\hat{\theta}_{mn}, \hat{\eta}_{mn}$ which are efficient on $\mathbf{P}_m$. In particular,

(3.2) $$\hat{\theta}_{mn} - \theta_0 = n^{-1} \Sigma_{i=1}^{n} \tilde{l}_{0m}(X_i) + o_{P_0}(n^{-1/2})$$

where

$$\tilde{l}_{0m} = \frac{s^{-1}}{2} \frac{s_1^*}{\|s_1^*\|^2}$$

and

$$s_1^* = \dot{s}_1 - \Pi(\dot{s}_1 | [\dot{s}_2, \ldots, \dot{s}_d]),$$

$\Pi(h|L)$ denotes the projection of $h \in L_2(\mu)$ on the closed linear subspace L in the $L_2(\mu)$ norm, $\|\cdot\|$, and $[\dot{s}_2, \ldots, \dot{s}_d]$ is the linear span of $\{\dot{s}_2, \ldots, \dot{s}_d\}$. $\hat{\eta}_{mn} - \eta_0$ has a similar expansion but we only note that

$$\hat{\eta}_{mn} - \eta_0 = O_{P_0}(n^{-1/2}). \tag{3.3}$$

These relations hold for each m fixed, all $P_0 \in \mathbf{P}_m$, as $n \to \infty$. Frequently, we achieve (3.2), (3.3) using the maximum likelihood estimates of θ, η^m under $\mathbf{P}_m$. For any $P \in \mathbf{P}$ let $\eta = (\eta_1, \ldots, \eta_{d(P)})$ where $d(P)$ is the smallest m such that $P \in \mathbf{P}_m$. For the model $\mathbf{P}$, the information bound in estimating θ at $P_0 = P_{(\theta_0, \eta_0)}$ is given by:

$$I^{-1}(P_0; \theta) = \frac{1}{4} \|\dot{s}_1 - \Pi(\dot{s}_1 | \dot{\zeta}_2(\theta_0, \eta_0)\|^{-2}$$

where

$$\dot{\zeta}_2(\theta_0, \eta_0) = \text{closure of the linear span of } \{\dot{s}_2(\theta_0, \eta_0), \ldots, \dot{s}_l(\theta_0, \eta_0), \ldots\}.$$

Here, for $m \geq m(P_0)$ we consider P_0 as a member of $\mathbf{P}_m$ i.e. corresponding to (θ_0, η_0^m) such that $P_0 = P_{(\theta_0, \eta_0^m)}$.

Suppose $I(P_0; \theta) > 0$ for all $P_0 \in \mathbf{P}$. Let

$$\tilde{l}(\theta_0, \eta_0) = 2 s^{-1}(\theta_0, \eta_0)(\dot{s}_1(\theta_0, \eta_0) - \Pi(\dot{s}_1(\theta_0, \eta_0) | \dot{\zeta}_2(\theta_0, \eta_0))) / I(P_0; \theta) \tag{3.4}$$

be the efficient influence function for estimating θ in $\mathbf{P}$ at P_0. $\tilde{l}$ depends on (θ_0, η_0).

Theorem 4: Suppose that if $P_{(\theta_k, \eta_k^m)} \in \mathbf{P}_m$, $\theta_k \to \theta_0$, $\eta_k^m \to \eta_0^m$ then

$$\Pi(v | \dot{\zeta}_2(\theta_k, \eta_k^m)) \to \Pi(v | \dot{\zeta}_2(\theta_0, \eta_0)) \tag{3.5}$$

for all $v \in L_2(\mu)$ and

$$\overline{\lim}_k \|\tilde{l}(\theta_k, \eta_k^m)\|_\infty < \infty \tag{3.6}$$

where $\|\cdot\|_\infty$ is the sup norm.

Then there exists $\hat{\theta}_n$ such that,

$$\hat{\theta}_n = \theta_0 + n^{-1} \Sigma_{i=1}^n \tilde{l}_0(X_i) + o_{P_0}(n^{-1/2})$$

where $\tilde{l} = \tilde{l}(\theta_0, \eta_0)$.

Moreover, the $\hat{\theta}_n$ are at least locally regular. That is, for all $P_0 \in \mathbf{P}$, $\{P_\tau : |\tau| < 1\}$ is a regular parametric submodel of $\mathbf{P}$, $\tau_n = O(n^{-1/2})$ we have $\mathbf{L}_{\tau_n}(n^{1/2}(\hat{\theta}_n - \theta(P_{\tau_n})))$ tending to a limit law independent of $\{P_\eta\}$.

The construction is essentially to pick the lowest dimensional submodel $\mathbf{P}_{\hat{m}_n}$ which is close enough to the empirical distribution, then treat $\hat{m}_n$ as fixed, compute the efficient estimate $\hat{\eta}_{\hat{m}_n n}$ of $\eta_{\hat{m}_n}$ in that model and then "solve the equation".

$$\sum_{i=1}^{n} \tilde{l}(\theta, \hat{\eta}_{m_n n}) = 0. \tag{3.7}$$

The resulting estimate is well behaved if $P \in \mathbf{P}$. However, if $P \in \overline{\mathbf{P}} - \mathbf{P}$, we necessarily have $\hat{m}_n \to \infty$ and no guarantee that the solution of (3.7) is even consistent, much less efficient. In fact, the examples of the previous section make it clear that there is no hope for such a general consistency theorem. The question remains whether one can formulate reasonable conditions on the structure of $\tilde{l}$ and the behaviour of the distance in suitable metrics between $\mathbf{P}_m$ and members of $\overline{\mathbf{P}} - \mathbf{P}$ as a function of m which will yield the validity of the information bounds for members of $\mathbf{P}$. An attempt in this direction is the work of Severini and Wong (1987). However, we do not pursue this, in part, because we believe that the checking of any such conditions in models of interest will be at least as difficult as the construction of efficient estimates by one of a number of heuristic methods which have been developed — see BKRW, Ch. 7 for a discussion.

Proof: Let d_K be the Kolmogorov distance between distributions. Let $\hat{\theta}_{mn}$, $\hat{\eta}_{mn}$ be as in (3.2), (3.3) and

$\tilde{P}_m$ be the corresponding member of $\mathbf{P}_m$.

Let $\hat{m}_n$ be the first m such that $d_K(\hat{P}_m, P_n) \le \varepsilon_n$ where $\varepsilon_n \to 0$, $n^{1/2}\varepsilon_n \to \infty$, P_n is the empirical distribution. Evidently, if $m_0 = m(P_{(\theta_0, \eta_0)})$,

$$P_0[\hat{m}_n = m_0] \to 1.$$

Moreover, $\hat{P}_{\hat{m}_n} \longleftrightarrow (\hat{\theta}_{\hat{m}_n n}, \hat{\eta}_{\hat{m}_n n}) = (\theta_0, \eta_0) + O_{p_0}(n^{-1/2})$. Therefore, by (3.5)

$$\int (\tilde{l}(\theta_n, \hat{\eta}_{m_n n}) - \tilde{l}(\theta_n, \eta_n))^2 s^2(\theta_n, \eta_n)\, d\mu = o_{p_0}(1) \tag{3.8}$$

for all sequences $P_{(\theta_n, \eta_n)} \in \mathbf{P}_{m_0}$ with $|\theta_n - \theta_0| = O(n^{-1/2})$, $|\eta_n - \eta_0| = O(n^{-1/2})$.

Moreover, using (3.6), we see that,

$$\begin{aligned} \int \tilde{l}(\theta_n, \hat{\eta}_{\hat{m}_n n}) s^2(\theta_n, \eta_n)\, d\mu &= 2\int \tilde{l}(\theta_n, \hat{\eta}_{m_0 n})(s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n})) \\ &\quad s(\hat{\theta}_n, \hat{\eta}_{m_0 n})\, d\mu + O_{p_0}(\| s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n}) \|^2) \\ &= 2\int \tilde{l}(\theta_n, \hat{\eta}_{m_0 n})(\dot{s}_2(\theta_n, \hat{\eta}_{m_0 n}), \ldots, \dot{s}_{m_0}(\theta_m, \hat{\eta}_{m_0 n}))(\eta_n - \hat{\eta}_{m_0 n})' s(\hat{\theta}_n, \hat{\eta}_{m_0 n})\, d\mu \end{aligned} \tag{3.9}$$

$$+ o_{p_0}(|\eta_n - \hat{\eta}_{m_0 n}|) + O_{p_0}(\|s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n})\|^2)$$

The first term on the right in (3.9) is 0 by (3.4). The last two terms are $o_{p_0}(n^{-1/2})$ by (3.2), (3.3) so

$$\int \tilde{l}(\theta_n, \hat{\eta}_{\hat{m}_n n}) s^2(\theta_n, \eta_n) du = o_{p_0}(n^{-1/2}). \tag{3.10}$$

Together (3.8) and (3.10) yield the existence of $\hat{\theta}_n$ — see Klaassen (1986) for example. □

Thus the $\hat{\theta}_n$ are at least locally regular and $n^{1/2}(\hat{\theta}_n - \theta_0)$ is asymptotically normal $(0, I^{-1}(P_0; \theta))$ i.e. achieves the information bound.

Note: 1) Conditions (3.5) and (3.6) are trivially satisfied by the symmetric location example. Condition (3.6) can be interpreted as a robustness condition for efficient estimates in $\mathbf{P}_m$. That is, on the model $\mathbf{P}_m$, efficient influence functions are bounded and bounded uniformly in small Hellinger neighbourhoods of any P.

2) It is easy to check that if in the Engle et al. model we, for instance, let $\mathbf{P}_m$ be such that $t(Z)$ and $\log P(W = 1 | Z)$ are representable as splines with $d(m)$ knots, condition (3.5) is satisfied. Although condition (3.6) fails for ε Gaussian, $\tilde{l}$ is of the form ε times functions which are uniformly $\|\cdot\|_\infty$ bounded and (3.7) continues to hold.

3) A further peculiarity of these models is that, if we only consider the asymptotic behaviour of $\hat{\theta}_n$ at fixed (θ, η), it is asymptotically inadmissible. However, when we consider its behaviour over "contiguous" neighbourhoods in $\mathbf{P}$ it is uniquely asymptotically minimax. More precisely let $\{P_t, |t| < 1\}$ be a regular parametric submodel of $\mathbf{P}$ passing through $P_0 = P_{(\theta_0, \eta_0)}$. Corresponding to this model is its score function at (θ_0, η_0) given by (say) $s_0^{-1} v$ where $v \in \dot{\zeta}_2(\theta_0, \eta_0)$. Consider $\hat{\hat{\theta}} \equiv \hat{\theta}_{\hat{m}_n n}$. By LeCam's third lemma, if $\theta_n \equiv \theta_n(t) = \theta(P_{tn^{-1/2}})$, $\eta_n \eta_n(t) = \eta(P_{tn^{-1/2}})$, then

$$L_{(\theta_n, \eta_n)} \sqrt{n}(\hat{\hat{\theta}} - \theta_n) \to \mathbf{N}(2t \int v s_1^* d\mu, \frac{1}{4} \|s_1^*\|^{-2}). \tag{3.11}$$

On the other hand, by the same argument,

$$L_{(\theta_n, \eta_n)} \sqrt{n}(\hat{\theta} - \theta_n) \to \mathbf{N}(0, I^{-1}(P_0; \theta)).$$

Now,

$$\begin{aligned} I(P_0; \theta) &= \frac{1}{4} \|\dot{s}_1 - \Pi(\dot{s}_1 | \dot{\zeta}_2(\theta_0, \eta_0))\|^2 \\ &\leq \frac{\|s_1^*\|^2}{4}. \end{aligned}$$

So at (θ_0, η_0), i.e. $t = 0$, both $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\sqrt{n}(\hat{\hat{\theta}} - \theta_0)$ are asymptotically normal with mean 0 and the asymptotic variance of $\sqrt{n}\,\hat{\hat{\theta}}$ is smaller than that of $\hat{\theta}$. However, evidently, on each parametric submodel, for any bounded bowl shaped loss function l,

$$\underline{\lim}_M \underline{\lim}_n \sup \{E_{(\theta_n(t), \eta_n(t))}\, l\,(n^{1/2}(\hat{\hat{\theta}} - \theta_n) : |t| \le M n^{-1/2}\} \;=\; \sup_d l\,(d)$$

higher than the comparable asymptotic minimax risk for $\hat{\theta}$.

This is a superefficiency phenomenon. The estimator $\hat{\hat{\theta}}$ is, in view of (3.11), not locally regular i.e. the limit of $\mathbf{L}_{(\theta_n, \eta_n)}\,(\sqrt{n}(\hat{\hat{\theta}} - \theta_n))$ is not independent of t.

References

Begun, J.M., Hall, W.J., Huang, W.M. and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric - nonparametric models. *Ann. Statist.*, **11**, 432-452.

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

Bickel, P.J., Klaassen C.A.J., Ritov, Y. and Wellner, J.A. (1989). "Efficient and Adaptive Inference in Semiparametric Models" Forthcoming monograph, Johns Hopkins University Press, Baltimore.

Bickel, P.J. and Ritov J. (1988). Estimating integrated squared derivatives. To appear in *Sankhyā*

Bretagnolle, J. and Huber C. (1979). Estimation des densites: risque minimax. *Z. Warsch. verw. Gebiete*, **47**, 119-137.

Chen, H. (1988). Convergence rates for the parametric component in a partially linear model. *Ann. of Statist.*, **16**, 136-146.

Cuzick, J. (1987). Semiparametric additive regression. *Tech. Report*, Imperial Cancer Research Laboratories, London.

Donoho, D.L. and Liu, R.C. (1988). Geometrizing rates of convergence. *Tech. Report*, Statistics Dept., U. of California at Berkeley.

Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310-320.

Faraway, J.J. (1987). Smoothing in Adaptive Estimation. (Ph.D. Thesis, University of California at Berkeley)

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. 6th Berkeley Symp. on Math Statist. and Prob.*, **1**, 175-194.

Hasminskii, R. and Ibragimov, I.A. (1979). On the nonparametric estimation of functionals. *Proc. 2nd Prague Symp. Asymptotic Statist.* (P. Mandl and M. Huskova eds.) 41-51, North Holland Amsterdam.

LeCam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. 3rd Berkeley Symp. on Math Statist. and Prob.*, **1**, 129-156.

Levit, B.Y. (1978). Infinite dimensional information bounds. *Theor. Prob. Applic.*, **20**, 723-740.

Pfanzagl, J. (1982). Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statistics **13**. Springer Verlag.

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

Severini, T.A. and Wong, W-H (1987). Profile likelihood and semiparametric models. *Tech. Report*, U. of Chicago.

Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Stat. Probab.*, **1** 187-195, Univ. of California Press.

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

Stone, C.J. (1986). A nonparametric framework for statistical modeling. *Tech. Report*, Dept. of Statistics, University of California at Berkeley.