The Annals of Statistics
2003, Vol. 31, No. 4, 1033–1053
© Institute of Mathematical Statistics, 2003

NONPARAMETRIC ESTIMATORS WHICH CAN BE "PLUGGED-IN"

BY PETER J. BICKEL AND YA'ACOV RITOV

University of California, Berkeley and Hebrew University

We consider nonparametric estimation of an object such as a probability density or a regression function. Can such an estimator achieve the ratewise minimax rate of convergence on suitable function spaces, while, at the same time, when "plugged-in," estimate efficiently (at a rate of $n^{-1/2}$ with the best constant) many functionals of the object? For example, can we have a density estimator whose definite integrals are efficient estimators of the cumulative distribution function? We show that this is impossible for very large sets, for example, expectations of all functions bounded by $M < \infty$. However, we also show that it is possible for sets as large as indicators of all quadrants, that is, distribution functions. We give appropriate constructions of such estimates.

1. Introduction.

We consider the following type of problem. Let $X_1, X_2, \dots, X_n$ be i.i.d., $X_1 \sim P_\vartheta$, $\vartheta \in \Theta$, a subset of a linear space of functions. Suppose the minimax rate for estimating ϑ with some global loss function, for instance, a Banach norm on Θ, is slower than the parametric $n^{1/2}$ rate. Let $\mathcal{T}$ be a collection of functionals from Θ to $\mathbb{R}$. Suppose that for each $\tau \in \mathcal{T}$, $\tau(\vartheta)$ can be estimated at the $n^{1/2}$ rate. Is there an estimator $\hat{\vartheta}$ of ϑ which achieves the minimum nonparametric rate above while at the same time, for all $\tau \in \mathcal{T}$, $\tau(\hat{\vartheta})$ converges to $\tau(\vartheta)$ at rate $n^{-1/2}$? Even better, can we have $\tau(\hat{\vartheta})$ be efficient, that is, best among all regular estimates of $\tau(\vartheta)$ converging at rate $n^{-1/2}$? Even more, can we have the $n^{-1/2}$ convergence be suitably uniform on $\mathcal{T}$?

For instance, and this is the prototypical example, let Θ be a ball in a Sobolev or Hölder space of densities or regression functions on R^d. Let the set of functionals be

$$\mathcal{T} = \left\{\tau_h,\ h \in \mathcal{H},\ \tau_h(\vartheta) = \int h(x)\vartheta(x)\,dx\right\}, \tag{1.1}$$

where $\mathcal{H}$ is a reasonably large class of functionals. We want to find an estimate $\hat{\vartheta}_n$ that achieves the minimax rate for integrated square error and, at the same time, can be "plugged-in" to estimate all functionals (parameters) $\tau(\vartheta)$ with $\tau_h \in \mathcal{T}$ efficiently. For instance, if P_n is the empirical distribution, we would want

$$\tau_h(\hat{\vartheta}_n) \equiv \int h\hat{\vartheta}_n = \int h\,dP_n + o_p(n^{-1/2}) \tag{1.2}$$

uniformly for $\vartheta \in \Theta$ and $h \in \mathcal{H}$. (By convention, $\int h$ will be an integral with respect to Lebesgue measure.)

Received August 1999; revised August 2002.
AMS 2000 subject classifications. 62G07, 62G30, 62F12.
Key words and phrases. Efficient estimator, density estimation, nonparametric regression.

Our interest in this problem stems from the fairly well-known fact that if one takes $\tilde{\vartheta}_n$ to be a standard asymptotically rate-minimax estimate such as a non-negative kernel of appropriate bandwidth for the two derivative Sobolev spaces then, typically, $n^{1/2}|\tau(\tilde{\vartheta}_n) - \tau(\vartheta)| \to_P \infty$. Thus, if the density estimate $\tilde{\vartheta}_n$ is based on a non-negative symmetric kernel, $k(\cdot)$, $\int v^2 k(v)\,dv = 1$, and an optimal bandwidth, $\sigma_n \approx n^{-1/5}$, then $\int x^2 \tilde{\vartheta}(x)\,dx = n^{-1}\sum_{i=1}^n X_i^2 + \sigma_n^2$ which is not a $\sqrt{n}$ consistent estimator of EX^2.

This failure can be seen as a lack of robustness against choice of loss function. Such $\tilde{\vartheta}_n$ behave well for $l(\vartheta, \tilde{\vartheta}_n) = \int(\vartheta - \hat{\vartheta}_n)^2$ but poorly for $l(\vartheta, \tilde{\vartheta}_n) = |\int h(\tilde{\vartheta}_n - \vartheta)|$. This bad behavior can become even worse if we take $l(\vartheta, \tilde{\vartheta}_n) = \sup_{\mathcal{H}} |\int h(\tilde{\vartheta}_n - \vartheta)|$.

If this lack of robustness can be remedied there are practical consequences. It is often the case that one wants to use the density estimate for inference about specific features like skewness and kurtosis or other aspects of shape. Failure to have the plug-in property means that for these purposes every subsequent user must return to the empirical distribution for such estimates and, for the inexperienced user, these appear to be inconsistent. We do argue in this paper that one cannot hope to efficiently plug-in for all regular parameters. But we also show that rather broad prior ideas of what one may need to plug-in for can be accommodated.

Of equal interest is the fact that shape estimation of the density may itself be qualitatively improved by "getting the functionals in $\mathcal{T}$ right." Efron and Tibshirani (1996) provide one method for getting a finite number of functionals right and thereby improving an overly rough estimate. We go in the other direction. Start with an oversmooth estimate and roughen it using the requirement that it has to do well on $\mathcal{T}$.

Cai (2002) establishes another plug-in property. He considers the white noise model $dY(t) = g(t)\,dt + \sigma n^{-1/2}\,dZ(t)$, $t \in R$, and suggests an estimator $\hat{g}$ of g, such that for a wide range of linear operators, $K^{-1}\hat{g}$ is an almost rate efficient estimator of $K^{-1}g$. His main example for the operator K is the derivative.

We use the acronym PIP to denote plug-in properties of the type we have described. An estimator with a PIP will be called a plug-in estimator, or a PIE for short. A statistical problem which admits a PIE will be considered as having the appropriate PIP. As we have noted there are potentially many notions of plug-in. We will define them completely as we discuss them in what follows. A PIP is a feature of a statistical problem with specified global loss function and family $\mathcal{T}$. Thus we shall speak of problems having a PIP (and also show that there are problems which do not admit any reasonable PIP). On the other hand we will focus on particular classes of estimates which are well known and/or attractive computationally and see if they can be modified to have a PIP for natural sets of functionals. PIPs will sometimes be formulated in terms of quadratic loss, but most often in terms of weaker stochastic order properties; see the following.

Our paper is organized as follows. We begin in Section 2 by briefly discussing two methods of density estimation in connection with our strongest version of PIP. Section 3 is conceptual and asks to what extent various PIPs are possible. The main result of this section, Theorem 3.2, is negative showing that if one takes $\mathcal{T}$ too big, for example, the set of all bounded linear functions, then we cannot adapt uniformly at least for quadratic risk. On the other hand, in Section 4 we provide an existence theorem which shows that in an important special case, if $\mathcal{T}$ is a reasonably small class, for example, a universal Donsker class, then plug-in is typically possible and we verify the conditions in a number of important cases. The existence theorem of Section 4 suggests a possible PIE. In Section 5 we further discuss one of the methods of estimation of Section 2 and show how the method suggested by Section 4 can be implemented.

Notational convention. We adopt the following uniform definitions. We write $V_n(\vartheta) = O_p(c_n)$ and $v_n(\vartheta) = o_p(c_n)$, for any sequence $\{c_n\}$, if

$$\limsup_{n,M} \sup_{\vartheta\in\Theta} P_\vartheta(|V_n(\vartheta)| > M) = 0,$$

$$\limsup_{n} \sup_{\vartheta\in\Theta} P_\vartheta(|V_n(\vartheta)| > \varepsilon) = 0 \qquad \text{for all } \varepsilon > 0.$$

We also define $V_n(\vartheta) = \Omega_p(c_n)$ if $V_n(\vartheta) = O_p(c_n)$ and $V_n^{-1}(\vartheta) = O_p(c_n^{-1})$.

In this paper we also state many results in terms of the minimax rate of convergence with respect to quadratic loss. However, the results can be generalized as follows. Given a nonparametric estimator with whatever rate properties (which may be uniform or not, minimax or not), there is a PIE with the same nonparametric properties and the prescribed uniform parametric properties.

2. Modifying standard methods of density estimation.

Consider first a standard kernel estimator on R:

$$\hat{p}_n(x) = \frac{1}{n\sigma}\sum_{i=1}^{n} \psi\left(\frac{x - X_i}{\sigma}\right),$$

where ψ is the density of a (not necessarily positive) distribution function Ψ, and σ is a bandwidth that depends on n. The kernel and the bandwidth are usually selected depending on how many derivatives (to α terms) p is assumed to have. Thus, if p is assumed to have a Taylor expansion with a uniformly bounded derivative of order α, then ψ is selected such that its first $\alpha - 1$ moments are 0, and $\sigma = n^{-1/(2\alpha+1)}$ to balance the bias and the standard error of the estimator. (This achieves rate-minimaxity over Sobolev balls of order $1 \le \alpha \le \infty$ for integrated square error and other global loss functions.) Consider now estimation of the c.d.f. of p. The estimator which is based on integrating $\hat{p}$ is

$$\hat{P}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\frac{y - X_i}{\sigma}\right),$$

where $\Psi(\cdot) = \int_{\infty}^{\cdot} \psi$. It is immediate that $n \operatorname{Var}(\hat{P}(y) - P_n(y)) \to 0$, where, with some natural abuse of notation, we use $P(y)$ to denote $P((-\infty, y])$. Moreover, denote the empirical process by $E_n = \sqrt{n}(P_n - P)$. Then

$$\begin{aligned}\sup_y \sqrt{n}\,|\hat{P}_n(y) - P_n(y) - E\hat{P}_n(y) + P(y)| \\ = \sup_y \left| \sigma_n^{-1} \int \psi\big((y-x)/\sigma_n\big) E_n(x)\,dx - E_n(y) \right| = o_p(1),\end{aligned} \tag{2.1}$$

since the empirical process converges to a uniformly bounded and continuous random process uniformly in P. Now

$$E\hat{P}_n(y) - P(y) = \int \psi(x)\big(P(y+\sigma x) - P(y)\big)\,dx.$$

If ψ is selected as above, this term is of order $\sigma^{\alpha} = n^{-\alpha/(2\alpha+1)}$, an order larger than $n^{-1/2}$, and $\hat{P}_n$ has no conceivable plug-in property for this problem. On the other hand if p has a Taylor expansion of order α, then P_o, which is one order smoother, has an expansion of order $\alpha + 1$. Hence, *if one starts with a kernel which has one more zero moment than needed for density estimation for a given bounded derivative*, the bias will be of order

$$\sigma^{\alpha+1} = n^{-(\alpha+1)/(2\alpha+1)} = o(n^{-1/2}).$$

If this order property of the bias holds uniformly for suitably shrinking neighborhoods of P in a model $\mathcal{P}$, then $\hat{P}(\cdot)$ is efficient for estimating P; see Bickel, Klaassen, Ritov and Wellner (1998) (BKRW, hereafter). If $\mathcal{P}$ is the ball in an L_∞ Sobolev space of order α, $\{p : \sup|D^{\alpha}p| \le M\}$, and the loss is integrated squared error, then we can plug-in for the distribution function and hence also for all functions $h(x) = \int_{-\infty}^{x} p(y)\,d\mu(y)$ where μ is a finite signed measure. In this context we define PIP as rate-minimaxity of $\hat{p}_n$ for integrated squared error loss and efficiency in the sense of BKRW for $\tau_h(\hat{p}_n)$.

If $d > 1$ this argument fails for $\mathcal{P}$ as above, since the c.d.f. does not necessarily have enough derivatives. Thus, if $d = 3$ and $\alpha = 2$, P needs to have more than three derivatives. If $P(x, y, z) \equiv H_1(x)H_2(y)H_3(z)$, and p has two derivatives, then the H_i's necessarily have three derivatives, but P may fail to have anything more than that.

Assume that p has α derivatives, or more precisely, suppose $\mathcal{P} \subseteq \{p : \int |\omega|^{2\alpha} \times |\mathcal{F}p(\omega)|^2\,d\omega < A\}$ where $\mathcal{F}$ is the Fourier transform operator. Now if $\mathcal{H} \subseteq \{h : \sup_{\omega} |w|^{\gamma} |\mathcal{F}h(\omega)| < C\}$ for some $\gamma > d/2$, then PIP holds for the natural multivariate extension of $\hat{P}_n(\cdot)$ and minimaxity defined above, if we use a kernel ψ such that $|\mathcal{F}\psi(\omega) - 1| \le B(1 \wedge |\omega|^{\alpha+\gamma})$.

To see this we argue again to establish (2.1) and consider

$$E_P\Big(\int h\hat{p}_n - \int hp\Big) = \int \mathcal{F}h\mathcal{F}p(\mathcal{F}\psi_\sigma - 1) = O(\sigma^{\alpha+\gamma}), \tag{2.2}$$

where $\psi_\sigma(x) \equiv \sigma^{-1}\psi(\sigma^{-1})$.

We should remark that the strong smoothness requirement imposed on $\mathcal{H}$ is needed only for having a kernel estimator with PIP. It is not needed in general, as we show in Section 4.

Our second example is sketchier but the ideas raised in our discussion presage in some respects the general existence theorem of Section 4. A formal statement of the theorem it suggests is given in Section 5. This method can also be shown to work for distribution functions in R, and more general Sobolev spaces than the L_∞ space—see Section 5. However, it does not generalize to the general distribution function and related functionals for R^d, $d > 1$.

Another general class of density estimators is based on orthonormal bases $\psi_1, \psi_2, \dots$. There are two main variants. The first is the sieve MLE based on the exponential family $c(\beta)\exp(\sum_{j=1}^{M}\beta_j\psi_j(\cdot))$. The second is the density estimator given by

$$\sum_{j=1}^{M_n} P_n(\psi_j)\psi_j(\cdot), \tag{2.3}$$

where $P(h) = \int h\,dP$. If the ψ_j are splines the first is the log spline estimate [Kooperberg and Stone (1992)]. Note that for both estimators

$$\int \psi_j(x)\hat{p}(x)\,dx = P_n(\psi_j), \qquad j = 1, 2, \dots, M_n.$$

We proceed for estimates of type (2.3). Suppose that the "natural" density estimator is based on M_n base functions, so that if $0 \le c \le p \le C$, $r_n = M_n/n$. Add to them M_n functions, $h_1, \dots, h_{M_n}$ that approximate $\mathcal{H}$ and proceed as above. The resultant estimator, call it $\hat{p}_n$, will have twice the variance and less bias so it will achieve the same convergence rate as the original density estimator and it will yield an efficient estimator of $h_1, \dots, h_{M_n}$. Now, for a general function $h \in \mathcal{H}$:

$$\begin{aligned}
&\int h(x)\hat{p}(x)\,dx - P_n(h)\\
&\quad = \int \big(h(x) - h^*(x)\big)\hat{p}(x)\,dx - P_n(h - h^*), \qquad (2.4)\\
&\quad = \int \big(h(x) - h^*(x)\big)\big(\hat{p}(x) - p(x)\big)\,dx - P_n(h - h^*) - P(h - h^*),
\end{aligned}$$

where h^* is some function approximating h in the span $\mathcal{S}_{M_n}$ of $h_1, \dots, h_{M_n}$ and $\psi_1, \dots, \psi_{M_n}$, say the projections in $L_2(P)$ or L_2 (*Lebesgue*). Suppose that the second term on the right-hand side is $o_p(n^{-1/2})$. If so, we need consider only the first term. Note that for the estimator given by (2.3), the first term is simply

$$\int h^\perp(x)p^\perp(x), \tag{2.5}$$

where the $\perp$ denotes the projection on the orthocomplement of $\mathcal{S}_M$.

Now, in the common cases, the estimator has bias and random error of the same order. That is,

$$\int p^{\perp 2}(x)\,dx < CM_n/n \tag{2.6}$$

for some finite C. Hence we expect to obtain a plug-in property for $\hat{p}_n$ and this $\mathcal{H}$ if

$$\sup_{h\in\mathcal{H}} \int h^{\perp 2}(x)\,dx = o(M_n^{-1}). \tag{2.7}$$

3. Feasibility of "plug-in." In this section we investigate different perspectives on the plug-in property. We start by reminding the reader that this problem is nonparametric and frequentist in nature: if Θ is Euclidean, the parameterization is regular and $\mathcal{T}$ includes only smooth functions, then the strongest possible PIPs hold. Then we turn to PIPs in frequentist nonparametric models. The main result of the section is that unless the class $\mathcal{T}$ of functionals is restricted, PIP defined in various ways is not possible. Even the class of all bounded linear functionals, as in (1.2), may be too big for PIP.

3.1. *Regular parametric families.* If $\mathcal{P}$ is regular parametric, $\mathcal{P} = \{p_\vartheta : \vartheta \in \Theta \subseteq R^d\}$, p_ϑ the density of $X \in \mathcal{X}$, $\vartheta \to p_\vartheta$ is 1–1, continuously Hellinger differentiable and the Fisher information matrix $I(\vartheta)$ is nonsingular for all ϑ then an efficient estimate $\hat{\vartheta}$, often the maximum likelihood estimator, exists and $\mathcal{L}_\vartheta\{\sqrt{n}(\hat{\vartheta} - \vartheta)\} \to \mathcal{N}(0, I^{-1}(\vartheta))$ uniformly on compacts. For any differentiable τ, $\mathcal{L}_\vartheta\{\sqrt{n}(\tau(\hat{\vartheta}) - \tau(\vartheta))\} \to \mathcal{N}(0, I^{-1}(X; \tau(\vartheta)))$, where $I(X; \tau(\vartheta))$ is the Fisher information bound for estimating $\tau(\vartheta)$ when observing X. The efficient estimate of the density p_ϑ is the PIE $p_{\hat{\vartheta}}(\cdot)$ which converges (e.g., in the Hellinger distance) at rate $n^{-1/2}$, and if

$$\mathcal{T} \Leftrightarrow \mathcal{H} = \Bigl\{h : \mathcal{X} \to R,\ \sup_{\vartheta\in\Theta} E_\vartheta h^2(X) < \infty\Bigr\} \tag{3.1}$$

with correspondence given by (1.1), then plug-in also works again in the sense that $\int h\ p_{\hat{\vartheta}}$ is efficient and so

$$\begin{aligned} &\int hp_{\hat{\vartheta}} - \int hp_\vartheta \\ &\quad = E\bigl(h(X)\dot{l}^{\mathcal{T}}(X; \vartheta)\bigr)I^{-1}(\vartheta)n^{-1}\sum_{i=1}^{n}\dot{l}(X_i; \vartheta) + o_{\mathcal{P}}(n^{-1/2}), \end{aligned} \tag{3.2}$$

where $\dot{l}(\cdot; \cdot)$ is the Hellinger derivative of the log-likelihood function.

3.2. *Nonparametric families.* Here is a first definition of PIP. Suppose that the stochastic minimax rate for estimating $\vartheta \in \Theta$ is r_n^{-1}. By this we mean

$$\inf_{\tilde{\vartheta}_n} r_n^{-2} \|\tilde{\vartheta}_n - \vartheta\|^2 = \Omega_p(1), \tag{3.3}$$

where the infimum is taken over all possible estimators based on $X_1, \dots, X_n$. Measure theoretic difficulties can be handled by considering outer measures.

DEFINITION 3.1. An estimate $\hat{\vartheta}_n$ of ϑ is a *uniform PIE* for a set $\mathcal{T}$ of functionals and a model $\mathcal{P}$, if

$$\left\{ r_n^{-2} \|\hat{\vartheta}_n - \vartheta\|_2^2 + n \sup_{\tau \in \mathcal{T}} \bigl(\tau(\hat{\vartheta}_n) - \tau(\vartheta)\bigr)^2 \right\} = O_p(1). \tag{3.4}$$

In general no such $\hat{\vartheta}_n$ exists. For instance, if Θ is a subset of an inner-product space, and $\mathcal{T}$ is indexed by $\{h : \|h\|_2 \le 1\}$, with $\tau_h(\vartheta) \equiv \langle h, \vartheta \rangle$, then $\sup_{\tau \in \mathcal{T}} (\tau(\hat{\vartheta}) - \tau(\vartheta))^2 = \|\hat{\vartheta} - \vartheta\|_2^2$. But if $\mathcal{H}$ is not too large (e.g., finite) then a PIE exists.

THEOREM 3.1. *Suppose there exists a purely data dependent process $\chi_n(\tau)$ on $\mathcal{T}$, such that $n \sup_{\mathcal{T}} (\chi_n(\tau) - \tau(\vartheta))^2 = O_P(1)$, that is, $\chi_n(\tau)$ is a uniformly $\sqrt{n}$ consistent estimate of $\tau(\cdot)$. Then a uniform PIE exists.*

PROOF. Define, for $\tilde{\vartheta}_n$ as in (3.3),

$$S_n(\vartheta') \equiv r_n^{-2} \|\tilde{\vartheta}_n - \vartheta'\|^2 + n \sup_{\tau \in \mathcal{T}} \bigl(\tau(\vartheta') - \chi_n(\tau)\bigr)^2$$

and $\hat{\vartheta}_n$ such that $S_n(\hat{\vartheta}_n) \le \inf_{\vartheta} S_n(\vartheta) + n^{-1}$. Then $\hat{\vartheta}_n$ is well defined and a uniform PIE since $S_n(\vartheta) = O_p(1)$ where ϑ is the true value of the parameter. □

A weaker requirement than (3.4) is that plug-in works for any parameter and functional (chosen a priori and independently of the data). We consider the following definition:

DEFINITION 3.2. An estimator $\hat{\vartheta}_n$ of ϑ is a *weak PIE* if

$$\sup_{\Theta} \sup_{\mathcal{T}} E_{\vartheta} \bigl(r_n^{-2} \|\hat{\vartheta}_n - \vartheta\|^2 + n\bigl(\hat{\tau}(\hat{\vartheta}) - \tau(\vartheta)\bigr)^2\bigr) < \infty. \tag{3.5}$$

We consider in this definition the usual quadratic loss function notions. Evidently, PIP and minimaxity in this sense imply the corresponding stochastic proporties. However, failure of PIP in this sense is weaker than failure in the stochastic sense. The main result of this section is that weak PIE and minimaxity in this sense are incompatible for nonparametric Θ and $\mathcal{T}$ large.

We consider the Gaussian white noise model. Here Θ is a subset of ℓ_2 and $X_i \in l_2$, $X_i = (X_{i1}, \dots, \dots)$,

$$X_{ij} = \vartheta_j + \varepsilon_{ij}, \qquad 1 \le i \le n,$$

where ε_{ij} are i.i.d. $\mathcal{N}(0,1)$. Our parameter set is given by

$$\Theta = \Big\{\vartheta : \sum i^{2\alpha}\vartheta_i^2 \le 1\Big\}.$$

This model or experiment when endowed with a space of loss functions and decision procedures (in the sense of Le Cam) is interesting in its own terms. In view of the work of Nussbaum (1996) and Brown and Low (1996) it is equivalent in the sense of Le Cam, for $\alpha > 1/2$, to more standard experiments embodying of nonparametric density and regression estimation when suitably described. For simplicity we do not go beyond the white noise model and this Θ, but some extension is clearly possible.

A linear functional τ on Θ can be identified with $h \in l_2$ with $\tau_h(\vartheta) = \sum \vartheta_j h_j$. Let $\mathcal{T} = \{h : \|h\|_2 \le 1\}$.

THEOREM 3.2. *If the white noise model holds and Θ and $\mathcal{T}$ are given as above, then there exists $\hat{\vartheta}_n$ which achieves the minimax rate for squared error,*

$$\sup_{\vartheta\in\Theta} E_P \|\hat{\vartheta}_n - \vartheta\|_2^2 = O\big(n^{-2\alpha/(2\alpha+1)}\big). \tag{3.6}$$

However, for any such $\hat{\vartheta}_n$,

$$\sup_{\vartheta\in\Theta,\tau\in\mathcal{T}} E\big[n^{2\alpha/(2\alpha+1)}\|\hat{\vartheta}_n - \vartheta\|_2^2 + n\big(\tau(\hat{\vartheta}_n) - \tau(\vartheta)\big)^2\big] \to \infty. \tag{3.7}$$

The proof is based on the following elementary lemma.

LEMMA 3.1. *Suppose that $X \sim N(\vartheta, 1)$, $\vartheta \in [-a, a]$ and let $\lambda > 0$. Let T be any estimator of ϑ. Then*

$$\max_{\vartheta\in[-a,a]} \{\operatorname{Var}_\vartheta(T) + \lambda^2 b_\vartheta^2(T)\} \ge \left(\frac{\lambda a}{1+\lambda a}\right)^2, \tag{3.8}$$

where $b_\vartheta(T) = E_\vartheta T - \vartheta$.

PROOF. We can assume without loss of generality that $\operatorname{Var}_\vartheta(T) < \infty$, as otherwise the result is trivial. Moreover, the bias function has a well-defined derivative by the Hellinger differentiability of the normal density. Denote $\max_\vartheta(1 + \dot{b}_\vartheta(T))^2 = \alpha^2$, $\alpha > 0$. Then $\max_\vartheta \dot{b}_\vartheta(T) \le -(1-\alpha)$. Hence,

$$b_a(T) - b_{-a}(T) \le -2(1-\alpha)a.$$

Therefore, either $b_a(T) \le -(1-\alpha)a$ or $b_{-a}(T) \ge (1-a)a$. It follows that

$$\max_{\vartheta} b_{\vartheta}^2(T) \ge \max\{b_{-a}^2(T), b_a^2(T)\} \ge (1-\alpha)^2 a^2.$$

By the information inequality,

$$\max_{\vartheta} \operatorname{Var}_{\vartheta}(T) \ge \max_{\vartheta} \bigl(1+\dot{b}_{\vartheta}(T)\bigr)^2 = \alpha^2.$$

Hence

$$\begin{aligned}\max_{\vartheta\in[-a,a]}\{\operatorname{Var}_{\vartheta}(T)+\lambda^2 b_{\vartheta}^2(T)\} &\ge \min_{\alpha>0}\max\{\alpha^2, \lambda^2(1-\alpha)^2a^2\}\\ &= \left(\frac{\lambda a}{1+\lambda a}\right)^2. \qquad \square\end{aligned}$$

PROOF OF THEOREM 3.2. The rate stated in the theorem is achieved, for example, by the estimator $\tilde{\vartheta}_n = (\tilde{\vartheta}_{n1}, \tilde{\vartheta}_{n2}, \ldots)$ with $\tilde{\vartheta}_{ni} = n^{-1}\sum_{j=1}^n X_{ji}$ for $i < n^{1/(1+2\alpha)}$ and $\tilde{\vartheta}_{ni} = 0$ otherwise.

Suppose there exists an estimator $\hat{\vartheta}_n = (\vartheta_{n1}, \vartheta_{n2}, \ldots)$ that satisfies (3.5), in particular,

$$\infty > \limsup_n \sup_{\vartheta\in\Theta, \tau\in\mathcal{T}} E_{\vartheta} n\bigl(\tau(\hat{\vartheta}_n) - \tau(\vartheta)\bigr)^2.$$

Let $h_{ni} = h_{ni}(\hat{\vartheta}_n, \vartheta) = E_{\vartheta}(\hat{\vartheta}_{ni} - \vartheta_i)$. We obtain that, in particular,

$$\begin{aligned}\infty &> \limsup_n \sup_{\vartheta\in\Theta} nE_{\vartheta}\left(\frac{\sum_{i=1}^{\infty} h_{ni}(\hat{\vartheta}_{ni} - \vartheta_i)}{(\sum_{i=1}^{\infty} h_{ni}^2)^{1/2}}\right)^2\\ &\ge \limsup_n n\sum_{i=1}^{\infty} h_{ni}^2(\hat{\vartheta}_n, \vartheta),\end{aligned} \tag{3.9}$$

by Cauchy–Schwarz.

Let $\beta = 2\alpha + 1$. Since $\hat{\vartheta}_n$ achieves the optimal nonparametric rate

$$\infty > \limsup_n n^{1-1/\beta} \sup_{\vartheta\in\Theta} \sum_{i=1}^{\infty} \bigl(\operatorname{Var}_{\vartheta}(\hat{\vartheta}_{ni}) + h_{ni}^2(\hat{\vartheta}_n, \vartheta)\bigr). \tag{3.10}$$

Combining (3.9) and (3.10) we obtain

$$\infty > \limsup_n n^{1-1/\beta} \sup_{\Theta} \sum_{i=1}^{\infty} \bigl(\operatorname{Var}_{\vartheta}(\hat{\vartheta}_{ni}) + n^{1/\beta} h_{ni}^2(\hat{\vartheta}_n, \vartheta)\bigr). \tag{3.11}$$

Consider now the set $\Theta^* = \{\vartheta : |\vartheta_i| \le c i^{-\beta(1+\varepsilon)/2}\} \subset \Theta$ for some small c and

$\varepsilon \in (0, \beta^{-1})$. Using the lemma, with $a = ci^{-\beta(1+\varepsilon)/2}n^{1/2}$ and $\lambda = n^{1/\beta}$,

$$
\begin{aligned}
&n^{1-1/\beta}\sup_{\Theta^*}\left(\sum_i \mathrm{Var}_\vartheta(\hat{\vartheta}_{ni}) + n^{1/\beta}h_{ni}^2(\hat{\vartheta}_n, \vartheta)\right)\\
&\quad = n^{-1/\beta}\sup_{\Theta^*}\left(\sum_i \mathrm{Var}_\vartheta(n^{1/2}\hat{\vartheta}_{ni}) + n^{1/\beta}nh_{ni}^2(\hat{\vartheta}_n, \vartheta)\right)\\
&\quad \geq n^{-1/\beta}\sum_{i=1}^{\infty}\left(\frac{cn^{1/2\beta+1/2}i^{-\beta(1+\varepsilon)/2}}{1+cn^{1/2\beta+1/2}i^{-\beta(1+\varepsilon)/2}}\right)^2 \qquad (3.12)\\
&\quad \geq n^{-1/\beta}\sum_{i=1}^{\lfloor n^{(1/\beta+1/\beta^2)/(1+\varepsilon)}\rfloor}\left(\frac{n^{1/2\beta+1/2}i^{-\beta(1+\varepsilon)/2}}{1+n^{1/2\beta+1/2}i^{-\beta(1+\varepsilon)/2}}\right)^2\\
&\quad \geq \left(\frac{c}{1+c}\right)^2 n^{(1-\varepsilon\beta)/\beta^2(1+\varepsilon)}.
\end{aligned}
$$

Note that we have converted estimation of ϑ_i with error variance $1/n$ to estimation of $\sqrt{n}\vartheta_i$ with error variance 1.

But (3.12) contradicts (3.11) and hence $\hat{\vartheta}_n$ does not exist. □

4. Minimaxity and efficient plug-in.

4.1. *Main results.* We will now define the statistically most interesting and strongest version of a PIP which, in fact, is the one regular parametric families possess.

Recall that an estimator $\tilde{\tau}_n$ of $\tau(\vartheta)$ is said to be efficient if $n^{1/2}(\tilde{\tau}_n - \tau(\vartheta))$ converges to a Gaussian distribution with mean 0 and variance the semiparametric information bound uniformly on regular submodels of Θ—see BKRW, Definition 5.2.7, page 182.

DEFINITION 4.1. Let $\|\tilde{\vartheta}_n - \vartheta\|^2 = O_p(r_n^2)$, where r_n is the (stochastic) minimax estimation rate, and for each $\tau \in \mathcal{T}$, let $\tilde{\tau}_n$ be an efficient estimator of τ. An estimator $\hat{\vartheta}_n$ is called an efficient PIE if $\|\hat{\vartheta}_n - \vartheta\|^2 = O_p(r_n^2)$ and $\sqrt{n}\sup_{\tau\in\mathcal{T}}|\tau(\hat{\vartheta}_n) - \tilde{\tau}| = o_p(1)$.

Note that if $\mathcal{T}$ is a universal Donsker class (in the sense of being Donsker uniformly for all $P \in \mathcal{P}$) being an efficient PIE implies that $\tau(\hat{\vartheta}_n)$, viewed as a process on $\mathcal{T}$, achieves the semiparametric information bound in the strong sense of BKRW, Definition 5.2.7, page 182.

We will now discuss the possibility of the efficient PIP in the special context of linear functionals. We consider Θ to be a subspace of some Hilbert space $\mathcal{S}$, and consider $\mathcal{T} = \{\rho(\cdot; h) : h \in \mathcal{H}\}$, where $\mathcal{H} \subset \mathbb{H}$, $\mathbb{H}$ some linear space, and $\rho : \Theta \times \mathbb{H} \to R$ is a bilinear function.

Let $\{\Theta_M\}$, $M \geq 1$, be a sequence of finite dimensional linear subspaces of Θ, where M is the dimension of $\{\Theta_M\}$. Let $\Pi_M : \mathcal{H} \to \mathbb{H}$ be a projection operator, defined by $\rho(\vartheta; h - \Pi_M h) = 0$, for all $\vartheta \in \Theta_M$ and $h \in \mathcal{H}$, and let $g_{M1}, \ldots, g_{MM}$ be an orthonormal basis of Θ_M. Let $h_{M1}, \ldots, h_{MM}$ span $\Pi_M \mathcal{H}$, $\rho(g_{Mi}; h_{Mj}) = \delta_{ij}$, $i, j = 1, \ldots, M$. All of these may depend on unknown parameters. Let ϑ be the true value of the parameter. We make the following assumptions:

A1. Let $\hat{\rho}(h)$ be an efficient estimator of $\rho(\vartheta; h)$, $h \in \mathbb{H}$. We assume that $\hat{\rho}(h)$ is linear, $\hat{\rho}(h_1 + h_2) = \hat{\rho}(h_1) + \hat{\rho}(h_2)$, and can be approximated uniformly by $\hat{\rho}(\Pi_{M_n} h)$ in the sense that for any $M_n \to \infty$,

$$\sup_{\mathcal{H}} n^{1/2} \bigl|\hat{\rho}\bigl(h - \Pi_{M_n} h\bigr) - \rho\bigl(\vartheta; h - \Pi_{M_n} h\bigr)\bigr| = o_p(1). \tag{4.1}$$

A2. There exists an estimator $\tilde{\vartheta}_n$ such that $\|\tilde{\vartheta}_n - \vartheta\|_2 = O_p(r_n)$.

A3. For all $M < \infty$,

$$C(M) \equiv \sup_{\vartheta, n, j} n E_{\vartheta} \bigl(\hat{\rho}(h_{Mj}) - \rho\bigl(\vartheta; h_{Mj}\bigr)\bigr)^2 < \infty.$$

THEOREM 4.1. *Under* A1–A3 *there exists an estimate* $\hat{\vartheta}_n$ *which is an efficient PIE for* $\Theta, \mathcal{T}$. *That is*,

$$\|\hat{\vartheta}_n - \vartheta\|_2 = O_p(r_n),$$
$$\sup_{\mathcal{H}} \bigl|\rho(\hat{\vartheta}_n; h) - \hat{\rho}(h)\bigr| = o_{\mathcal{P}}(n^{-1/2}).$$

PROOF. Note that A1 implies that if $M_n \to \infty$, then there exists a sequence $b_n \to 0$ (depending on M_n, but not on ϑ) such that

$$\sup_{\mathcal{H}} b_n^{-1} n^{1/2} \bigl|\hat{\rho}(h) - \hat{\rho}\bigl(\Pi_{M_n} h\bigr) - \rho\bigl(\vartheta; h - \Pi_{M_n} h\bigr)\bigr| = o_{\mathcal{P}}(1). \tag{4.2}$$

Let $M_n \to \infty$ but $C(M_n) M_n / n r_n^2 \to 0$, and let b_n be the sequence of (4.2). To simplify notation we occasionally drop the subscripts n and M_n. Next we consider the following problem:

$$\text{Minimize } \Bigl\{ r_n^{-1} \|\vartheta - \tilde{\vartheta}_n\|_2 + b_n^{-1} n^{1/2} \sup_{\mathcal{H}} \bigl|\rho(\vartheta; h) - \hat{\rho}(h)\bigr|,\ \vartheta \in \Theta \Bigr\} \tag{4.3}$$

and let $\hat{\vartheta}_n$ be an (approximate) minimizer. Define

$$\vartheta^* = \vartheta^*(\vartheta) = \vartheta + \sum_{j=1}^{M_n} \bigl(\hat{\rho}(h_j) - \rho(\vartheta; h_j)\bigr) g_j. \tag{4.4}$$

We claim that

$$r_n^{-1}\|\vartheta^* - \tilde{\vartheta}_n\| = O_p(1) \tag{4.5}$$

and

$$b_n^{-1}n^{1/2}\sup_{\mathcal{H}}\bigl|\rho(\vartheta^*; h) - \hat{\rho}(h)\bigr| = o_p(1). \tag{4.6}$$

To see this, first compute

$$r_n^{-1}\|\vartheta^* - \tilde{\vartheta}_n\| \le r_n^{-1}\|\vartheta - \tilde{\vartheta}_n\| + O_p\left(r_n^{-1}\left(\frac{C(M_n)M_n}{n}\right)^{1/2}\right), \tag{4.7}$$

since for all ϑ,

$$E_\vartheta\|\vartheta^* - \vartheta\|_2^2 = E_\vartheta\left(\sum_{j=1}^{M_n}\bigl(\hat{\rho}(g_j) - \rho(\vartheta; g_j)\bigr)^2\right) \le \frac{C(M_n)M_n}{n},$$

which does not depend on ϑ by A3. By definition of M_n, (4.5) follows. On the other hand, since $\hat{\rho}(h)$ is linear by assumption A1, and

$$\rho(\vartheta^*; h_j) = \rho(\vartheta; h_j) + \hat{\rho}(h_j) - \rho(\vartheta; h_j) = \hat{\rho}(h_j),$$

then $\rho(\vartheta^*; h) = \hat{\rho}(h)$ for all $h \in \Pi_{M_n}\mathcal{H}$. Therefore, for $h \in \mathcal{H}$,

$$\begin{aligned}\rho(\vartheta^*; h) - \hat{\rho}(h) &= \rho\bigl(\vartheta^*; h - \Pi_{M_n}h\bigr) - \hat{\rho}(h) + \hat{\rho}\bigl(\Pi_{M_n}h\bigr)\\ &= \rho\bigl(\vartheta; h - \Pi_{M_n}h\bigr) - \hat{\rho}\bigl(h - \Pi_{M_n}h\bigr).\end{aligned}$$

Hence, (4.6) follows from (4.1). But (4.5) and (4.6) imply that

$$\min_{\vartheta\in\Theta}\left\{r_n^{-1}\|\tilde{\vartheta} - \vartheta\|_2 + b_n^{-1}n^{1/2}\sup_{\mathcal{H}}|\rho(\vartheta; h) - \hat{\rho}(h)|,\ \vartheta \in \Theta\right\} = O_p(1).$$

Hence

$$\|\hat{\vartheta}_n - \tilde{\vartheta}_n\|_2 = O_p(r_n),$$

$$\sup_{\mathcal{H}}\bigl|\rho(\hat{\vartheta}_n; h) - \hat{\rho}(h)\bigr| = o_p(n^{-1/2}),$$

and the theorem follows. □

Note that $\hat{\vartheta}_n$ does not depend on ϑ, although ϑ^*, Θ_M and M_n may depend on the unknown parameter.

We now give some simple conditions on the model and $\mathcal{T}$ for existence of efficient PIEs. Suppose $\mathcal{X} = R^d$. In the following discussion ϑ is identified, as usual, with the density $p = p_\vartheta$, and with the c.d.f., $P = P_\vartheta$.

Let $\mathcal{B}_M = \{B_{M1}, \dots, B_{MM}\}$ be a partition of R^d, for instance, into rectangles. Let $\mathcal{S}_M = \operatorname{span}\{g_{M1}, \dots, g_{MM}\}$, $g_{Mj}(x) \equiv c_{Mj}p(x)\mathbb{1}(x \in B_{Mj})$, where

$c_{Mj} = (\int_{B_{Mj}} p^2)^{-1/2}$ is a normalizing constant and $\mathbb{1}$ denotes an indicator. The projection operator is given by $\Pi_M h = \Pi(h|\mathcal{B}_M)$, where

$$\Pi(h|\mathcal{B}_M)(x) \equiv p(x) \sum_{j=1}^{M} \frac{P(B_{Mj})}{\int_{B_{Mj}} p^2} E_0(h|B_{Mj}) \mathbb{1}(x \in B_{Mj}).$$

Assumption A1 has two aspects. The first is that the members of $\mathcal{H}$ can be approximated uniformly by their projections on $\mathcal{S}_M$, and the second is that the empirical process results can be applied to this projection. We deal with the two aspects separately.

A4. Suppose $\{P_\vartheta\}$ is dominated by a measure ν. $\{\mathcal{B}_i\}$ is a sequence of nested partitions and for $\alpha \le 2$ and all M,

$$\sup_{\Theta} \frac{E_\vartheta(p_\vartheta^\alpha(X) \mid \mathcal{B}_M)}{(E_\vartheta(p_\vartheta(X) \mid \mathcal{B}_M))^\alpha} \le C < \infty$$

and

$$\sup_{\Theta} \left| \frac{E_\vartheta(p_\vartheta^\alpha(X) \mid \mathcal{B}_M)}{(E_\vartheta(p_\vartheta(X) \mid \mathcal{B}_M))^\alpha} - 1 \right| \to 0 \qquad \text{as } M \to \infty \text{ a.e. } \nu.$$

This condition is natural when p_ϑ is bounded and continuous (in particular, if the noncompact members of $\mathcal{B}_M$ are excluded).

Typically one proves tightness or weak convergence of an empirical process indexed by a set of functions by proving some bound on a covering number for this set. We define the covering number $N(\varepsilon, \mathcal{H}, D)$ to be the smallest number of functions $h_1, \dots, h_N$ such that

$$\sup_{h \in \mathcal{H}} \min_{1 \le i \le N} \|h - h_i\|_D \le \varepsilon.$$

We define the covering number with bracketing, $N_{[\,]}(\varepsilon, \mathcal{H}, D)$, as the minimal number of pairs (h_{1i}, h_{2i}), $i = 1, 2, \dots, N$, such that $\|h_{21} - h_{1i}\|_D \le \varepsilon$, and for every $h \in \mathcal{H}$ there is $1 \le i \le N$ such that $h_{1i} \le h \le h_{2i}$. The metric D is typically either $L_\alpha(P)$ (or an equivalent measure like the uniform) or $L_\alpha(P_n)$, where P_n is the empirical distribution function.

We now argue that if $\hat{\rho}(h) = P_n(h)$, that is, the model is nonparametric, the usual conditions for $\mathcal{H}$ to be a P Donsker class carry over under A4 so that assumption A1 is satisfied and hence efficient PIEs can be constructed for broad classes of examples. Note that assumption A3 is automatically satisfied for this choice of $\hat{\rho}$.

THEOREM 4.2. *Suppose that $\mathcal{H}$ satisfies the following slight strengthening of the condition of Theorem* 2.5.6 *of van der Vaart and Wellner* (1996),

$$\sup_{\Theta} \int_0^\infty \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{H}, L_2(P_\vartheta))\, d\varepsilon} < \infty,$$

and has an envelope function H *for* $\mathcal{H}$, $\sup_\Theta \int H^2\, dP_\vartheta < \infty$. *Then, under* A1–A4 *an efficient PIE can be constructed.*

The proof uses two lemmas which are of independent use in semiparametric models where $\hat{\rho}$ is more complicated.

The following lemma describes some of the properties of the projection.

LEMMA 4.1.

1. *If* $h_1 \le h_2$ *then* $\Pi(h_1|\mathcal{B}) \le \Pi(h_2|\mathcal{B})$.
2. *Suppose* A4 *holds. Then* $E|\Pi(h|\mathcal{B})|^\alpha \le C^{\alpha-2} E|h|^\alpha$ *for any* $h \in L_\alpha(P)$.

PROOF. The first part of the lemma is trivial. We proceed to prove the second part. For any $h \in L_\alpha(P)$

$$\begin{aligned} E|\Pi(h|\mathcal{B})|^\alpha &= \sum_{B\in\mathcal{B}} \frac{P^\alpha(B)\int_B p^{\alpha+1}}{(\int_B p^2)^\alpha} |E(h|B)|^\alpha \\ &\le \sum_{B\in\mathcal{B}} \frac{P^{\alpha-1}(B)\int_B p^{\alpha+1}}{(\int_B p^2)^\alpha} \int_B |h|^\alpha p \\ &= \sum_{B\in\mathcal{B}} \frac{E(p^\alpha(X)\mid B)}{E^\alpha(p(X)\mid B)} \int_B |h|^\alpha p \\ &\le \sum_{B\in\mathcal{B}} C \int_B |h|^\alpha p \\ &= CE|h|^\alpha. \qquad \square \end{aligned}$$

We now prove that $\mathcal{H}$ can be approximated by its projections.

LEMMA 4.2. *Suppose* A4 *holds, that* $\mathcal{H}$ *has an* $L_2(P)$ *envelope, uniform in* P, *and that the empirical process indexed by* $\mathcal{H}$ *is uniformly pre-Gaussian uniformly in* P [*see van der Vaart and Wellner* (1996), *page* 169 *for the definition*]. *Then* $\sup_\vartheta \sup_{h\in\mathcal{H}} \|h - \Pi(h|\mathcal{B}))_M\|_{L_2(P_\vartheta)} \to 0$.

PROOF. First note that since $\mathcal{H}$ has an $L_2(P)$ envelope,

$$\|h - E(h|\mathcal{B})\|_{L_2(P)} \to 0 \tag{4.8}$$

for any $h \in \mathcal{H}$.

Suppose now that $\|h_i - E(h_i|\mathcal{B}_i)\|_{L_2(P_i)} \to 0$ for some sequence $\{(h_i, \mathcal{B}_i, P_i)\}$. Let E_i be the expectation under P_i. Then

$$\begin{aligned}
&\|h_i - \Pi(h_i|\mathcal{B}_i)\|^2_{L_2(P_i)}\\
&\quad \le 2\|E_i(h_i|\mathcal{B}_i) - \Pi(h_i|\mathcal{B}_i)\|^2_{L_2(P_i)} + 2\|h_i - E_i(h_i|\mathcal{B}_i)\|^2_{L_2(P_i)}\\
&\quad = 2\sum_{B\in\mathcal{B}_i}\int_B \biggl(E_i(h_i|B)\biggl(1 - \frac{P_i(B)p_i(x)}{\int_B p_i^2}\biggr)\biggr)^2 p(x)\,dx\\
&\qquad + 2\|h_i - E_i(h_i|\mathcal{B}_i)\|^2_{L_2(P_i)}\\
&\quad \le 2\sum_{B\in\mathcal{B}_i} P_i(B)E_i(h_i^2|B)\biggl(\frac{P_i(B)\int_B p_i^3}{(\int_B p_i^2)^2} - 1\biggr)\\
&\qquad + 2\|h_i - E_i(h_i|\mathcal{B}_i)\|^2_{L_2(P_i)} \to 0
\end{aligned} \tag{4.9}$$

by assumption A4 and bounded convergence.

If the conclusion of the lemma is not true, then there is $\varepsilon > 0$ and a sequence $h_i \in \mathcal{H}$, such that $\|h_i - \Pi(h_i|\mathcal{B}_i)\|_{L_2(P_i)} > 2\varepsilon$. By (4.9), this implies that $\|h_i - E_i(h_i|\mathcal{B}_i)\|_{L_2(P_i)} > 2\varepsilon$ as well. Let $i_1 = 1$ and define

$$i_j = \min\Bigl\{i : \max_{k<j} \|h_{i_k} - \Pi(h_{i_k}|\mathcal{B}_i)\|_{L_2(P_i)}\Bigr\} < \varepsilon.$$

Note that i_j is finite by (4.8). Then

$$\begin{aligned}
&\min_{k<j} \|h_{i_j} - h_{i_k}\|_{L_2(P_i)}\\
&\quad \ge \min_{k<j}\bigl(\|h_{i_j} - E_i(h_{i_k}|\mathcal{B}_{i_j})\|_{L_2(P_i)} - \|h_{i_k} - E_i(h_{i_k}|\mathcal{B}_{i_j})\|_{L_2(P_i)}\bigr)\\
&\quad \ge \varepsilon.
\end{aligned}$$

Hence there is no ε-net covering $\mathcal{H}$, contradicting the uniform pre-Gaussianity assumption; cf. van der Vaart and Wellner [(1996), Theorem 2.8.2]. □

PROOF OF THEOREM 4.2. We need only establish A1. Since $\Pi(\cdot|\mathcal{B})$ is a conditional expectation it preserves order (Lemma 4.1) and also reduces $L_\alpha(P)$ norm, $\alpha \ge 1$, $E|\Pi(h|\mathcal{B})|^\alpha \le E|h|^\alpha$. Therefore,

$$N_{[\,]}\bigl(\varepsilon, \Pi(\mathcal{H}|B_0), L_\alpha(P)\bigr) \le N_{[\,]}\bigl(\varepsilon, \mathcal{H}, L_\alpha(P)\bigr).$$

[In fact only the usual $E(\Pi(h|B))^2 \le Eh^2$ is needed.] Moreover, if the envelope function H possesses a second moment so does the envelope to $\Pi(\mathcal{H}|B)$ since $E(\sup_{\mathcal{H}} |h(\cdot)||\mathcal{B}) \ge \sup_{\mathcal{H}}(E(h|\mathcal{B}))^2$. The result follows. □

Similar arguments can be applied to the uniform entropy Theorem 2.5.1 of van der Vaart and Wellner (1996). Recall that $\mathcal{B}$ is not related to the estimator but only to the proof of its existence, hence the number of sets in the partition $\mathcal{B}$

and $\min_{B_i \in \mathcal{B}} P(B_i)$ can converge to infinity and 0 as slowly as needed. Therefore, we can have that $\max_{B \in \mathcal{B}_i} \int_B p^2 \, dP_n / \int_B p^3 \le 2$ with probability converging to 1. Hence

$$\begin{aligned}\|\Pi(h|\mathcal{B})\|_{P_n}^2 &= \sum_{B \in \mathcal{B}} \left(\frac{P(B)}{\int_B p^2}\right)^2 E^2(h|B) \int_B p^2 \, dP_n \\ &\le 2 \sum_{B \in \mathcal{B}} \frac{P(B) \int_B p^3}{(\int_B p^2)^2} \int_B h^2 p + o_{\mathcal{P}}(1) \\ &\le 2CE(h^2) + o_{\mathcal{P}}(1),\end{aligned}$$

by A4. Hence establishing a bound on $N(\varepsilon, \Pi(\mathcal{H}|\mathcal{B}), L_2(P_n))$ will be relatively straightforward [if one has it for $N(\varepsilon, \mathcal{H}, L_2(P_n))$].

4.2. *Examples.*

Nonparametric: *Linear* $\mathcal{T}$. In view of Theorem 4.1 existence of an efficient PIE for a number of important examples of linear $\mathcal{T}$ is immediate. We mention the empirical d.f. $\mathcal{H} =$ Indicators of rectangles $\equiv \{a_i \le x_i \le b_i,\ 1 \le i \le d,\ \mathbf{a}, \mathbf{b} \in R^d\}$, $\mathcal{H} =$ Indicators of half spaces $\equiv \{\mathbf{a}^T \mathbf{x} \le c : |\mathbf{a}| = 1, c \in R\}$ where $|\cdot|$ is the Euclidean norm, $\mathcal{H} =$ Fourier transforms restricted to a compact $\equiv \{h_{\mathbf{t}} : h_{\mathbf{t}}(\mathbf{x}) = \exp(i\mathbf{t}^T\mathbf{x}),\ \mathbf{t} \in K$ a compact$\}$, all sets of inferential interest. Here is a more surprising example.

PIE for all moments and cumulants. Suppose $\mathcal{X} = I^d$, the unit cube. Let $\mathcal{H} = \{\exp\{\mathbf{s}^T\mathbf{x}\} : |s| \le 1\}$. Let $\hat{p}_n$ be a PIE for $\mathcal{H}$ which is evidently a Donsker class. We claim that $\hat{p}_n$ is a simultaneous PIE for all moments and hence all cumulants. To see this, note that

$$\sup_{|\mathbf{s}| \le 1} \left| n^{1/2} \biggl(\int \exp\{\mathbf{s}^T\mathbf{x}\} \hat{p}_n(\mathbf{x}) \, d\mathbf{x} - \int \exp\{\mathbf{s}^T\mathbf{x}\} \, dP_n(\mathbf{x}) \biggr) \right| \xrightarrow{P} 0.$$

The expression within "| |" is an analytic function of $\mathbf{s}$. Since it converges uniformly to 0 on a compact with nonempty interior, all its derivatives which are also analytic must similarly converge to 0 and our claim follows.

Nonlinear functionals. In the usual way we can get results for nonlinear $\mathcal{T}$ from linear ones. Suppose $\mathcal{T}$, P are such that for suitable $\tilde{\tau}_n$:

(i) For all τ, P there exist functions $h_\tau(\cdot, p)$ such that

$$\tilde{\tau}_n = \tau(p) + \int h_\tau(x, p) \, dP_n(x) + o_p(n^{-1/2}).$$

This is just the statement that $\tau(p)$ is efficiently estimable over a nonparametric model $\mathcal{P}$.

(ii) Let $\tilde{\mathcal{T}} = \{\tau_h(p) = \int hp : h = h_\tau(\cdot, p)$ for some $\tau \in \mathcal{T}$, $p \in \mathcal{P}\}$. $\tilde{\mathcal{T}}$ satisfies the conditions of Theorem 4.1.

(iii) Let $\hat{p}_n$ be an efficient PIE for $\tilde{\mathcal{T}}$. Then,

$$\sup\left\{\left|\tau(\hat{p}_n) - \tau(p) - \int h_\tau(\cdot, p)(\hat{p}_n - p)\right| : \tau \in \mathcal{T}\right\} = o_p(n^{-1/2}).$$

Then, $\hat{p}_n$ is an efficient PIE for τ. As an example of a nonlinear process satisfying these conditions consider, $d = 1$,

$$\tau(p) = P^{-1}(s) : \varepsilon \le s \le 1 - \varepsilon, \qquad \varepsilon > 0,$$

where $P(x) = \int_{-\infty}^{x} p(u)\,du$, $\mathcal{P}$ is the set of all p in a compact subset of an L_2 Sobolev ball with $\inf p > 0$. Then, the PIE for the d.f. is a PIE for $\mathcal{T}$. To see this note that (i) is immediate with

$$h_\tau(x, p) = -\big(1(-\infty, P^{-1}(s)) - s\big)/p\big(P^{-1}(s)\big)$$

and (ii) is easy to check since the Sobolev metric is stronger than L_2. Finally, write

$$\hat{P}_n^{-1}(s) - P^{-1}(s) = \frac{-(\hat{P}_n^{-1}(s) - P^{-1}(s))}{(P(\hat{P}_n^{-1}(s)) - s)}\big(\hat{P}_n\big(\hat{P}_n^{-1}(s)\big) - P\big(\hat{P}_n^{-1}(s)\big)\big)$$

in a form first proposed by Shorack (1969). We define $\hat{P}_n^{-1}(s)$ as the smallest x such that $\hat{P}_n(x) = s$. Since

$$\sup_x |\hat{P}_n - P|(x) \xrightarrow{P} 0,$$

such an x exists for all $\varepsilon \le s \le 1 - \varepsilon$ for n sufficiently large. Now uniform convergence of $\hat{P}_n$ and strict monotonicity of P imply

$$\sup_x \{|\hat{P}_n^{-1}(t) - P^{-1}(t)| : \varepsilon \le t \le 1 - \varepsilon\} \xrightarrow{P} 0$$

and tightness of $n^{1/2}(\hat{P}_n(\cdot) - P(\cdot))$ inherited from PIE can be used to complete the proof in a standard fashion.

4.3. *Semiparametric examples.* We now give a brief description of three further examples where our result can be used.

The density and c.d.f. in the biased sample model. We consider the problem of density estimation with the c.d.f. as our collection of functionals but we have a biased sample model [Vardi (1985)]. In this model we observe (X, Δ) where $\Delta \in \{d_1, \dots, d_k\}$, and the conditional density of X given $\Delta = \delta$ is $w(x;\delta)f(x)/\int w(x';\delta)f(x')\,dx'$ with w known and f completely unknown. We want to estimate f and its cumulative integral. See Gill, Vardi and Wellner (1988) and BKRW for a description of the efficient estimator of the c.d.f. and its linear functionals. Suppose that $0 < \inf w < \sup w < \infty$. Suppose, for simplicity, that $\sum_{i=1}^{k} w(\cdot; d_i)$ is at least as smooth as the density f. Then

$$\tilde{f} = \frac{\tilde{g}_n}{\sum_{i=1}^{k} \tilde{p}_i \int w(x'; d_i)\, d\tilde{F}_n(x')}$$

is a rate optimal density estimator, where $\tilde{g}_n$ is a rate optimal estimator of the density estimator of the marginal density of X (based only on the marginal empirical distribution of X), $\tilde{p}_i$, $i = 1, \dots, k$, are the empirical probabilities of the strata, and $\tilde{F}_n$ is an efficient estimator of the distribution of F. Note that the estimator in the denominator is bounded away from 0 and infinity, and is efficient. Hence A2 is satisfied. It is easy to check A1 and A3 directly. We conclude that there is a PIE of the density f.

The hazard rate and the hazard function of the Cox model. Consider the Cox model with hazard function $\lambda(t)\exp(\beta' z)$, where t is the time and z is a vector of covariates. We may consider estimating the nonparametric $\lambda(\cdot)$ [Csörgő and Mielniczuk (1988) and Ghorai and Pattanaik (1993)] and its cumulative integral, $\int_0^{\cdot} \lambda(t)\, dt$, both on a fixed interval $(0, a)$, if we observe uncensored values larger than a with positive probability. Efficient estimation of the hazard function was discussed, for example, by Andersen and Gill (1982) and Tsiatis (1981). See Begun, Hall, Huang and Wellner (1983) for discussion of the information bound. Note that verifying the conditions A1–A3 is not much different in this example than it is in the density-c.d.f. example, since the functionals are of the same type, and their efficient estimators are linear. This is so, even though in this case the efficient estimator (Nelson–Aalen) is not linear in the observations, as the c.d.f. is. An extension of this example which is only partially covered by Theorem 4.1 is to the time-dependent covariate case, and to functionals of the form $\int_0^t \exp(\beta' z(s))\lambda(s)\, ds$. Extending the result to cover this case seems to be straightforward.

Functionals of a nonparametric regression function. Suppose $Y = \vartheta(X) + \varepsilon$, where X and ε are independent, $\varepsilon \sim N(0, 1)$ and ϑ belongs to some smoothness set Θ. We can consider now a set of functionals $\mathcal{T}$ of the form $\tau_h = \int h f \vartheta$, $h \in \mathcal{H}$. These functionals can be estimated efficiently by $n^{-1} \sum_{i=1}^{n} h(X_i) Y_i$, and this can be done uniformly if $\mathcal{H}$ is some VC class with an envelope H, $EH^2(X) < \infty$. For Θ_M we can consider any increasing sieve whose limit is Θ. Verifying the conditions is simple (note that conditions A2 and A3 impose hardly any difficulty). Our main result shows that there exists an estimator of the regression function, achieving the minimax rate, that yields efficient estimators of all members of $\mathcal{T}$ at the same time.

5. Construction of estimates.

The method underlying the proof of Theorem 5.1 can be implemented by solving the optimization problem (4.3). We shall pursue this at the end of this section. Two approaches of modifying the kernel density estimate and orthogonal series estimates were discussed in Section 2. We begin this section by a formal statement of the conditions needed for the series method of Section 2 to work.

Here are the conditions.

B1. The estimate $\tilde{p}_n$ of form (2.3) based on $\psi_1, \dots, \psi_{M_n}$ satisfies (4.1).

B2. Let $\mathcal{S}_{M_n}$ be the linear span of $\psi_1, \dots, \psi_{M_n}$ and an additional set $h_1, \dots, h_{L_n}$ of orthonormal functions where $L_n = \Omega(M_n)$ and Π_{M_n} is a projection on $\mathcal{S}_{M_n}$. Suppose

$$\sup_{\mathcal{H}} \|h - \Pi_{M_n} h\|_2 = o_{\mathcal{P}}(n^{-1/2}). \tag{5.1}$$

B3. $p \le C < \infty$.

B4. $\sup_{\mathcal{P}} \|p - \Pi_{M_n} p\|_2^2 = O(\frac{M_n}{n})$.

Our discussion has established the following:

THEOREM 5.1. *If $\mathcal{P}$, $\mathcal{H}$ are such that* B1–B4 *and* A2 *hold then $\hat{p}_n$ is a strong plug-in estimate.*

REMARK. The conditions are easily seen to hold if, for instance, $\mathcal{P} = \mathcal{Q}_\alpha$ where $\mathcal{Q}_\alpha = \{p \text{ on } I^d : \|D^\beta p\|_2 \le C \text{ all } \beta \le \alpha\}$ and $\alpha > d/2$ and the $\{\psi_1, \dots, \psi_M, h_1, \dots, h_{L_n}\}$ are a spline basis on the unit d-cube I^d.

Using the results and techniques of Kooperberg and Stone (1991) one can show with some more labor that the log spline estimate also has this property if we also require that $p \ge c > 0$ on I^d. In fact, it is possible for $d = 1$ as was conjectured by Stone (1990) by taking $M_n = n^{1/2+\varepsilon}$ to obtain the strong PIP for the distribution function as well.

For $d > 1$ we have the same difficulties with $\mathcal{Q}_\alpha$ as we do for kernel density estimates.

We finally study the method implicitly suggested by the existence theorems.

We consider $\tilde{p}_n$ of the form (2.3).

B5. There exists K_n such that if $\mathcal{S}_{K_n}$ is the linear span of $\psi_1, \dots, \psi_{K_n}$ then

$$\sup_{\mathcal{P}} \|p - \Pi_{K_n} p\|_2 = O(b_n n^{-1/2}), \tag{5.2}$$

where b_n is given in (4.3).

B6. Let $\Pi_M^* h$ be defined by

$$\arg\min \left| \int h^* \, dP_n - \int h \, dP_n : h^* \in \mathcal{S}_M \right|.$$

Then, if K_n is as above,

$$\sup_{\mathcal{H}} |P_n(h) - P_n(h^*)| = o_{\mathcal{P}}(b_n n^{-1/2}). \tag{5.3}$$

Define

$$\hat{p}_n = \sum_{j=1}^{K_n} \hat{c}_j \psi_j, \tag{5.4}$$

where $\hat{c}_j = P_n(\psi_j)$, $1 \le j \le M_n$ and $\hat{c}_{M_n+1}, \dots, \hat{c}_{K_n}$ is obtained as the solution of the quadratic programming problem

$$\text{Minimize } \sum_{j=1}^{M_n} (\hat{c}_j - c_j)^2 + \sum_{j=M_n+1}^{K_n} c_j^2$$

$$\text{subject to } \left| \sum_{j=1}^{K_n} d_j c_j - P_n(h) \right| \le n^{-1/2} b_n$$

for all **d** such that $\Pi^{\alpha}_{K_n}(h) = \sum_{j=1}^{K_n} d_j \psi_j$ for some $h \in \mathcal{H}$. If the conditions of Theorem 4.1 are satisfied and B5 and B6 hold, then $\hat{p}_n$ clearly will have the efficient PIP. Thus to obtain an estimate which has the strong PIP for $\mathcal{P} =$ Sobolev ball and $\mathcal{H} =$ Indicators of cubes in R^d we can take $(\psi_1, \dots, \psi_{K_n})$ to be, say, an orthogonal basis for the space generated by all splines of order $1 \le \beta \le \alpha$ with knots at $(\frac{i_1}{K_n}, \dots, \frac{i_d}{K_n})$ and $0 \le i_j \le K_n$, $1 \le j \le d$ and $K_n = n^{1/2+\varepsilon}$.

This formulation makes it clear that efficient plug-in is achieved by only approximately matching $\int h\, dP_n$ for all $h \in \mathcal{H}$ rather than exactly as we have done up to now.

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120.

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., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

CAI, T. T. (2002). On adaptive wavelet estimation of a derivative and other related linear inverse problems. *J. Statist. Plann. Inference* **108** 325–349.

CSÖRGŐ, S. and MIELNICZUK, J. (1988). Density estimation in the simple proportional hazards model. *Statist. Probab. Lett.* **6** 419–426.

EFRON, B. and TIBSHIRANI, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24** 2431–2461.

GILL, R. D., VARDI, Y. and WELLNER, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16** 1069–1112.

GHORAI, J. K. and PATTANAIK, L. M. (1993). Asymptotically optimal bandwidth selection of the kernel density estimator under the proportional hazards model. *Comm. Statist. Theory Methods* **22** 1383–1401.

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

SHORACK, G. R. (1969). Asymptotic normality of linear combinations of functions of order statistics. *Ann. Math. Statist.* **40** 2041–2050.

STONE, C. J. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.

STONE, C. J. (1991). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

VARDI, Y. (1985). Empirical distributions in selection bias models (with discussion). *Ann. Statist.* **13** 178–205.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: bickel@stat.berkeley.edu

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM 91905
ISRAEL
E-MAIL: yaacov@mscc.huji.ac.il