Submitted to The Annals of Statistics

Non-Parametric Estimators Which Can Be "Plugged-In"

By Peter J. Bickel and Ya'acov Ritov

University of California at Berkeley and The Hebrew University of Jerusalem

*†We consider nonparametric estimation of an object such as a probability density or a regression function. Can such an estimator achieve the 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, e.g., 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, i.e., distribution functions. We give appropriate constructions of such estimates.

1. Introduction We consider the following type of problem. Let $X_1, X_2, \ldots, X_n$ be i.i.d. , $X_1 \sim P_\vartheta$, $\vartheta \in \Theta$, a subset of a linear space of function. 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 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 best among all regular estimates of $\tau(\vartheta)$ converging at

---

AMS 1991 subject classifications. Primary 62G07; 62G30; 62F12

†*Key words and phrases.* Efficient estimator; density estimation, nonparametric regression.

rate $n^{-1/2}$ (Efficient)? 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 universal Donsker class. 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, ϑ is a density, $\mathcal{T}$ as above, then, if P_n is the empirical distribution, we would want

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

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

Our interest in this problem stems from the fairly well known fact that if one takes $\tilde{\vartheta}_n$ to be a standard minimax estimate such as a nonnegative kernel or wavelet estimate of appropriate bandwidth for the two derivative Sobolev spaces then, typically, $n^{1/2}|\tau(\tilde{\vartheta}_n) - \tau(\vartheta)| \xrightarrow{P} \infty$. Thus, if the density estimate $\tilde{\vartheta}_n$ is based on a nonnegative kernel with an optimal bandwidth, $\sigma_n = O(n^{-1/5})$, then $\int x^2\tilde{\vartheta} = n^{-1}\sum_{i=1}^n X_i^2 + \sigma_n^2$ which is not $\sqrt{n}$ 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 $l(\vartheta, \tilde{\vartheta}_n) = \left|\int h(\tilde{\vartheta}_n - \vartheta)\right|$ and more so for $l(\vartheta, \tilde{\vartheta}_n) = \sup_{\mathcal{H}} \left|\int h(\tilde{\vartheta}_n - \vartheta)\right|$.

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. We do argue in this paper that there is no free lunch, i.e., 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 (2000) 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 the PIP will be called a plug-in estimator, or PIE in short. A statistical problem which admits a PIE will be considered as having the 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. The PIP is a feature of a statistical problem with specified global loss function and family $\mathcal{T}$. Thus we shall speak of problems having the PIP (and show that there are problems which do not have it). 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.

Our paper is organized as follows. We begin in Section 2 by briefly discussing our motivating example of kernel density estimation in connection with the strongest version of PIP. Section 3 is conceptual and asks to what extent various PIP's are possible. The main result of this section, Theorem 3.2, is negative showing that if one takes $\mathcal{T}$ too big, e.g. the set of all bounded linear functions, then we cannot adapt uniformly as in (1.2). 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, e.g. a universal Donsker class, then plug-in is typically possible and verify the conditions in a number of important cases. Although the result of Section 4 suggests a possible PIE, this estimator is not 'natural'. In Section 5 we exhibit several specific and more plausible methods of estimation.

2. Kernel density estimation Consider a standard kernel estimator:

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

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_0 is assumed to have. Thus, if p_0 is assumed to have a Taylor expansion 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 minimaxity over Hölder balls for integrated square error and other global loss functions.) Consider now estimation of the cdf of P_0. The estimator which is based on integrating $\hat{p}$ is:

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

It is immediate that $n\ \mathrm{Var}(\hat{P}(y) - P_n(y)) \to 0$ (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,

$$\begin{aligned} &\sup_y \sqrt{n}\left|\hat{P}_n(y) - P_n(y) - E\ P_n(y) + P(y)\right| \\ &\qquad = \sup_y \left|\int \psi\big((y-x)/\sigma_n\big) E_n(x)\,dx - E_n(y)\right| = o_p(1), \end{aligned} \tag{2.3}$$

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

$$E\ \hat{P}_n(y) - P_0(y) = \int \psi(x)(P_0(y + \sigma x) - P_0(y))dx.$$

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_0 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* the bias will be of order

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

If this holds uniformly for $P \in \mathcal{P}$, then $\hat{P}$ is efficient. If $\mathcal{P}$ is the ball in a Sobolev space of order α, $\left\{p : \int |D^{\alpha}p|^2 \leq M\right\}$ and the loss is integrated squared error, then we can plug-in for the distribution function and hence also for all functions $h(x) = \int_{-\infty}^{x} d\mu(y)$ where μ is a finite signed measure. In this context we define PIP as minimaxity of $\hat{p}_n$ for integrated squared error loss and efficiency for $\tau_h(\hat{p}_n)$.

If $d > 1$ this argument fails for $\mathcal{P}$ as above, since the cdf does not necessarily have one derivative more than the density. However, it is still true that use of a kernel with more zero moments than needed for density estimation will enable the type of PIP just defined for some $\mathcal{P}$ and $\mathcal{H}$.

Assume that p has α derivatives, or more precisely, suppose $\mathcal{P} \subseteq \{p : \int |\omega|^{2\alpha} |\mathcal{F}p(\omega)|^2 d\omega < A\}$ 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 if we use a kernel ψ such that $|\mathcal{F}\psi(\omega) - 1| \leq B(1 \bigwedge |\omega|^{\alpha+\gamma})$.

To see this we argue again to establish (2.3) and consider

$$E_P \left(\int h\hat{p}_n - \int hp \right) = \int \mathcal{F}h \; \mathcal{F}p(\mathcal{F}\psi_\sigma - 1) = O_{\mathcal{P}}(\sigma^{\alpha+\gamma}) \tag{2.4}$$

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.

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 PIP's hold. Then we consider the Bayesian situation. If we define PIP in this context to require Bayes optimality rather than minimaxity we argue that for "quadratic" loss functions and τ which are linear functionals, PIP holds trivially. Then we turn to PIP's 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 to 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 non-singular 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

$$\mathcal{T} \iff \mathcal{H} = \{h : \mathcal{X} \to R,\ \sup_{\vartheta\in\Theta} E_\vartheta h^2(X) < \infty\} \tag{3.5}$$

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

$$\int hp_{\hat{\vartheta}} - \int hp_\vartheta = E\big(h(X)\dot{l}^{\mathcal{T}}(X;\vartheta)\big)I^{-1}(\vartheta)n^{-1}\sum_{i=1}^{n}\dot{l}(X_i;\vartheta) + o_{\mathcal{P}}(n^{-1/2}), \tag{3.6}$$

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

3.2. *A Bayesian perspective* We briefly discuss now the PIP from a Bayesian perspective. Suppose Θ belongs to a a bounded convex subset of a Hilbert space $\mathcal{L}$, then the Bayesian may wish to estimate Θ considering the squared Hilbert norm as his norm (note that in this sub-section we use Θ to denote the parameter, considered as a random variable).

Then the no data Bayes estimate is well defined by $E(\Theta)$ minimizing $E\|\Theta - c\|^2$ for $c \in \mathcal{L}$. Similarly, the Bayes estimate is $E_D(\Theta)$ where E_D is expectation with respect to the posterior. It follows immediately then that

$$\tau\big(E_D(\Theta)\big) = E_D\tau(\Theta) \tag{3.7}$$

for all bounded linear functionals τ. However, as soon as we consider nonlinear functionals, the situation becomes unclear.

One way of formalizing the notion of robustness against choice of loss functions for a Bayesian is to postulate that the loss function is selected at random, i.e., have a prior over potential decision theoretic goals the data is to be put to. If the prior beliefs on τ are independent of those on Θ, the Bayesian may wish to minimize over ϑ for some μ and π:

$$\iint E_\vartheta\{\|\hat\vartheta - \vartheta\|^2/r_n^2 + n\lambda\big(\tau(\hat\vartheta) - \tau(\vartheta)\big)^2\}d\mu(\tau)d\pi(\vartheta) < \infty. \tag{3.8}$$

Equivalently, the Bayesian minimizes a weighted average of the loss incurred from the different aspects of the problem. Implicitly, we assume that the parameter ϑ and the objective τ are selected independently and that the Bayes risk is uniformly bounded. The parameters r_n and λ define the relative weights the Bayesian gives to the two components of the problem. In the typical case, where the two aspects of the problem, estimating ϑ and its functionals, are given equivalent weights, then r_n will be of the same order as the Bayes risk for estimating ϑ. Formally,

DEFINITION 3.1. *The* independence Bayesian plug-in property *is satisfied if there exists an estimator $\hat\vartheta_n$ satisfying (3.8).*

The Bayesian does not lose much by minimizing the average risk by a single estimator instead of minimizing the different risks by an arsenal of estimators, each fitted for a specific task. In fact, the independence Bayesian plug-in risk is no more than twice the weighted average Bayesian risks for the different components. Let

ϑ_π and τ_π are the Bayesian estimators of ϑ and τ. Here is the proof:

$$\begin{aligned}
&\inf_{\hat\vartheta} E_D\left(\left\|\hat\vartheta-\vartheta\right\|^2/r_n^2+n\lambda\int(\tau(\hat\vartheta)-\tau(\vartheta))^2\,d\mu(\tau)\right)\\
&\quad=E_D\left(\left\|\hat\vartheta_\pi-\vartheta\right\|^2/r_n^2+n\lambda\int(\hat\tau_\pi-\tau(\vartheta))^2\,d\mu(\tau)\right)\\
&\qquad+\inf_{\hat\vartheta}\left(\left\|\hat\vartheta-\hat\vartheta_\pi\right\|^2/r_n^2+n\lambda\int(\tau(\hat\vartheta)-\hat\tau_\pi)^2\,d\mu(\tau)\right)\\
&\quad\le E_D\left(\left\|\hat\vartheta_\pi-\vartheta\right\|^2/r_n^2+n\lambda\int(\hat\tau_\pi-\tau(\vartheta))^2\,d\mu(\tau)\right)\\
&\qquad+E_D\left(\left\|\vartheta-\hat\vartheta_\pi\right\|^2/r_n^2+n\lambda\int(\tau(\vartheta)-\hat\tau_\pi)^2\,d\mu(\tau)\right)
\end{aligned}$$

3.3. *Non-parametric families* Here is a first non-Bayesian definition of PIP. Suppose that the minimax rate for estimating $\vartheta\in\Theta$ is $n^{\gamma/2}$. That is,

$$\inf_{\tilde\vartheta_n}\sup_{\vartheta\in\Theta} n^{\gamma+\varepsilon}\left\|\tilde\vartheta_n-\vartheta\right\|^2\xrightarrow{\text{p}}\infty$$

$$\inf_{\tilde\vartheta_n}\sup_{\vartheta\in\Theta} n^{\gamma}\left\|\tilde\vartheta_n-\vartheta\right\|^2< O_p(1)$$

for any $\varepsilon>0$, where the infimum is taken over all possible estimators based on $X_1,\dots,X_n$.

Definition 3.2. *An estimate $\hat\vartheta_n$ of ϑ is a uniform PIE for a set $\mathcal{T}$ of functionals if under any P_ϑ, $\vartheta\in\Theta$:*

$$\left\{n^\gamma\|\hat\vartheta_n-\vartheta\|_2^2+n\sup_{\tau\in\mathcal{T}}(\tau(\hat\vartheta_n)-\tau(\vartheta))^2\right\}=O_p(1). \tag{3.9}$$

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

Theorem 3.1. *Suppose there exists an estimated process $\chi_n(\tau)$, such that $n\sup_\tau(\chi_n(\tau) - \tau(\vartheta))^2 = O_P(1)$, i.e., $\chi_n(\tau)$ is a $\sqrt{n}$ consistent estimate of $\tau(\cdot)$. Then a uniform PIE exists.*

Proof. Define

$$S_n(\vartheta') \equiv n^\gamma \|\hat\vartheta - \vartheta'\|^2 + n \sup_{\tau\in\mathcal{T}} (\tau(\vartheta') - \tilde\tau_n)^2$$

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 PIE since $S_n(\vartheta) = O_p(1)$ where ϑ is the true value of the parameter.

A weaker requirement than (3.9) is that plug-in works for any parameter and functional (chosen a-priori and independently of the data). That is,

Definition 3.3. *An estimator $\hat\vartheta_n$ of ϑ is a weak PIE if*

$$\lim_{M\to\infty} \overline{\lim}_n \sup_{\Theta,\mathcal{T}} P_\vartheta\bigl(n^\gamma\|\hat\vartheta_n - \vartheta\|^2 + n(\tau(\hat\vartheta) - \tau(\vartheta))^2 \ge M\bigr) = 0. \tag{3.10}$$

The main result of this section suggests that even this PIP does not hold for non-parametric Θ and $\mathcal{T}$ large.

For simplicity we consider the Gaussian white noise model. Here $X_i \in l_2$, $X_i = (X_{i1}, \dots, \dots)$

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

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

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

This model is interesting in its own terms. In view of the work of Nussbaum (1994) Brown and Low (1996) it is equivalent in the sense of LeCam, for $\alpha > 1/2$, to more standard models 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 $\mathcal{T} = \{h : \|h\|_2 \leq 1\}$.

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:*

$$\sup_{\vartheta \in \Theta} E_P \|\hat{\vartheta}_n - \vartheta\|_2 = O(n^{-\alpha/(2\alpha+1)}). \tag{3.11}$$

However for any such $\hat{\vartheta}_n$, the equivalent of (3.10) *fails. In fact,*

$$\sup_{\vartheta \in \Theta, \tau \in \mathcal{T}} E\left[n^{2\alpha/(2\alpha+1)} \|\hat{\vartheta}_n - \vartheta\|_2^2 + n(\tau(\hat{\vartheta}_n) - \tau(\vartheta))^2\right] \to \infty. \tag{3.12}$$

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*

$$\max_{\vartheta \in [-a,a]} \{ Var_\vartheta(T) + \lambda^2 b_\vartheta^2(T)\} \geq \left(\frac{\lambda a}{1 + \lambda a}\right)^2 \tag{3.13}$$

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

PROOF. We can assume wlog that $\mathrm{Var}_\vartheta(T) < \infty$, 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 $\max_\vartheta \dot{b}_\vartheta(T) \leq -(1-\alpha)$. Hence,

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

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

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

By the information inequality that

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

Hence

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

Proof of Theorem 3.2. The rate stated in the theorem is achieved, for example by the estimator $\tilde{\vartheta}_n = (\tilde{\vartheta}_{n1}, \tilde{\vartheta}_{n2}, \dots)$ 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 $\hat{\vartheta}_n = (\vartheta_{n1}, \vartheta_{n2}, \dots)$ such that

$$\infty > \overline{\lim}_n \sup_{\vartheta\in\Theta,\tau\in\mathcal{T}} E_{\vartheta} n(\tau(\hat{\vartheta}_n)-\tau(\vartheta))^2$$

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

$$\begin{aligned}\infty &> \overline{\lim}_n \sup_{\vartheta\in\Theta} nE_{\vartheta}\left(\frac{\sum_{i=1}^\infty h_{ni}(\hat{\vartheta}_{ni}-\vartheta_i)}{(\sum_{i=1}^\infty h_{ni}^2)^{1/2}}\right)^2\\ &\geq \overline{\lim}_n n\sum_{i=1}^\infty h_{ni}^2(\hat{\vartheta}_n,\vartheta),\end{aligned} \tag{3.14}$$

by Cauchy-Schwarz.

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

$$\infty > \overline{\lim}_n n^{1-\frac{1}{\beta}} \sup_{\vartheta\in\Theta}\sum_{i=1}^\infty \left(\operatorname{Var}_{\vartheta}(\hat{\vartheta}_{ni}) + h_{ni}^2(\hat{\vartheta}_n,\vartheta)\right). \tag{3.15}$$

Combining (3.14) and (3.15) we obtain:

$$\infty > \overline{\lim}_n n^{1-1/\beta}\sup_{\Theta}\sum_{i=1}^\infty\left(\operatorname{Var}_{\vartheta}(\hat{\vartheta}_{ni}) + n^{1/\beta}h_{ni}^2(\hat{\vartheta}_n,\vartheta)\right). \tag{3.16}$$

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 $a = c i^{-\beta(1+\varepsilon)/2} n^{1/2}$ and $\lambda = n^{1/\beta}$:

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

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 (3.3) contradicts (3.16) 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.

DEFINITION 4.1. *Let $\|\tilde\vartheta_n - \vartheta\|^2 = O_p(r_n)$, r_n be the minimax estimation rate, and for each $\tau \in \mathcal{T}$, let $\tilde\tau_n$ be an efficient estimator of τ (i.e., an estimator that achieves the semiparametric information bound for estimation of $\tau(\vartheta)$). An estimator $\hat\vartheta_n$ is called an efficient PIE if $\|\hat\vartheta_n - \vartheta\|^2 = O_p(r_n)$ and $\sqrt{n}\sup_{\tau\in\mathcal{T}} |\tau(\hat\vartheta_n) - \tilde\tau| = o_p(1)$.*

Note that if $\mathcal{T}$ is a Donsker class being an efficient PIE implies that $\tau(\hat\vartheta_n)$ achieves the semiparametric information bound in the strong sense of Bickel, Klaassen, Ritov and Wellner (1998) 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 $g_{M1}, \dots, g_{MM}$ be an orthonormal basis of Θ_M. Let $h_{M1}, \dots, h_{MM}$ span $\Pi_M \mathcal{H}$, $\rho(g_{Mi}; h_{Mj}) = \delta_{ij}$, $i, j = 1, \dots, M$. All of these may depend on unknown parameters. Let ϑ_0 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 $M_n \to \infty$

$$\sup_{\mathcal{H}} n^{1/2} |\hat\rho(h - \Pi_{M_n} h) - \rho(\vartheta_0; h - \Pi_{M_n} h)| = o_p(1). \tag{4.17}$$

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

A3: For all $M < \infty$,

$$C(M) \equiv \sup_{\vartheta, n, j} n E_\vartheta \big(\hat\rho(h_{Mj}) - \rho(\vartheta_0; h_{Mj})\big)^2 < \infty.$$

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

$$\|\hat{\vartheta}_n - \vartheta_0\|_2 = O_p(r_n)$$

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

PROOF. Note that **A1** implies that if $M_n \to \infty$, then there exists a sequence $b_n \to 0$ (depending on M_n) such that

$$\sup_{\mathcal{H}} b_n^{-1} n^{1/2} |\hat{\rho}(h) - \hat{\rho}(\Pi_{M_n} h) - \rho(\vartheta_0; h - \Pi_{M_n} h)| = o_{\mathcal{P}}(1). \tag{4.18}$$

Let $M_n \to \infty$ but $C(M_n)M_n/nr_n^2 \to 0$, and let b_n be the sequence of (4.18). To simplify notation we occasionally drop the subscripts n and M_n. Next we consider the following problem:

$$\text{Minimize } \left\{ r_n^{-1} \|\vartheta - \tilde{\vartheta}_n\|_2 + b_n^{-1} n^{1/2} \sup_{\mathcal{H}} |\rho(\vartheta; h) - \hat{\rho}(h)|,\ \vartheta \in \Theta \right\} \tag{4.19}$$

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

$$\vartheta^* = \vartheta_0 + \sum_{j=1}^{M_n} \left(\hat{\rho}(h_j) - \rho(\vartheta_0; h_j)\right) g_j. \tag{4.20}$$

We claim that,

$$r_n^{-1} \|\vartheta^* - \tilde{\vartheta}_n\| = O_p(1) \tag{4.21}$$

and

$$b_n^{-1} n^{1/2} \sup_{\mathcal{H}} |\rho(\vartheta^*; h) - \hat{\rho}(h)| = o_p(1). \tag{4.22}$$

To see this compute first,

$$r_n^{-1} \|\vartheta^* - \tilde{\vartheta}_n\| \le r_n^{-1} \|\vartheta_0 - \tilde{\vartheta}_n\| + O_p\left(r_n^{-1} \left(\frac{C(M_n)M_n}{n} \right)^{1/2} \right) \tag{4.23}$$

since

$$E_P \|\vartheta^* - \vartheta_0\|_2^2 = E_P \left(\sum_{j=1}^{M_n} \left(\hat{\rho}(g_j) - \rho(\vartheta_0; g_j)\right)^2 \right) \le \frac{C(M_n)M_n}{n}$$

by **A3**. By definition of M_n, (4.21) follows. On the other hand, since $\hat\rho(h)$ is linear by assumption A1, and

$$\rho(\vartheta^*; h_j) = \rho(\vartheta_0; h_j) + \hat\rho(h_j) - \rho(\vartheta_0; h_j) = \hat\rho(h_j).$$

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

$$\rho(\vartheta^*; h) - \hat\rho(h) = \rho(\vartheta^*; h - \Pi_{M_n}h) - \hat\rho(h) + \hat\rho(\Pi_{M_n}h)$$

$$= \rho(\vartheta_0; h - \Pi_{M_n}h) - \hat\rho(h) + \hat\rho(\Pi_{M_n}h).$$

Hence, (4.22) follows from (4.17). But (4.21) and (4.22) imply that

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

Hence

$$\|\hat\vartheta_n - \tilde\vartheta_n\|_2 = O_p(r_n)$$

$$\sup_{\mathcal{H}}\left|\rho(\hat\vartheta_n; h) - \hat\rho(h)\right| = o_p(n^{-1/2})$$

and the theorem follows.

Note that although ϑ^* in the proof depends on the true ϑ_0, $\hat\vartheta_n$ doesn't.

We now give some simple conditions on the model for existence of efficient PIE. Suppose $\mathcal{X} = R^d$.

Let $\mathcal{B}_M = \{B_{M1}, \dots, B_{MM}\}$ be a partition of R^d, for instance, into rectangles. Let $\mathcal{S}_M = \operatorname{span}\{g_{M1}, \dots, g_{MM}\}$, $g_{Mj}(x) \equiv c_{Mj}p_0(x)\mathbf{I}(x \in B_{Mj})$, where $c_{Mj} = (\int_{B_{Mj}} p_0^2)^{-1/2}$ is a normalizing constant and $\mathbf{I}$ denotes an indicator. The projection operator is given by $\Pi_M h = \Pi(h \mid \mathcal{B}_M)$, where

$$\Pi(h \mid \mathcal{B}_M) \equiv p_0(x)\sum_{j=1}^{M}\frac{P_0(B_{Mj})}{\int_{B_{Mj}} p_0^2}E_0(h \mid B_{Mj})\mathbf{I}(x \in B_{Mj}).$$

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 $\{\mathcal{B}_i\}$ is a sequence of a nested partitions and that for $\alpha \le 2$ and all M

$$\frac{E(p_0^\alpha(X) \mid \mathcal{B}_M)}{\big(E(p_0(X) \mid \mathcal{B}_M)\big)^\alpha} \le C < \infty$$

and

$$\frac{E(p_0^\alpha(X) \mid \mathcal{B}_M)}{\big(E(p_0(X) \mid \mathcal{B}_M)\big)^\alpha} \xrightarrow{\text{a.s.}} 1, \quad \text{, as } M \to \infty$$

This condition is natural when p_0 is bounded and continuous (in particular, if the non-compact members of $\mathcal{B}_M$ are excluded).

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

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

We define the covering number with bracketing, $N_{[\,]}(\varepsilon, \mathcal{H}, D)$ as the minimal number of pairs (h_{1i}, h_{2i}), $i = 1, 2, \ldots, N$, such that $\big\| h_{21} - h_{1i} \big\|_D \le \varepsilon$, 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 $L_\alpha(P_0)$ (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 $\hat\rho(h) = P_n(h)$, i.e., 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 PIE's 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 a slight strengthening of the condition of Theorem* 2.5.6 *of van der Vaart and Wellner (1996),*

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon, \mathcal{H}, L_2(P))d\epsilon} < \infty$$

and has an envelope function H for $\mathcal{H} \in L_2(P)$. 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 $h_1 \leq h_2$ then $\Pi(h_1 \mid \mathcal{B}) \leq \Pi(h_2 \mid \mathcal{B})$.*

2. *Suppose **A4** holds. Then $E|\Pi(h \mid \mathcal{B})|^\alpha \leq C^{\alpha-2} E|h|^\alpha$ for any $h \in L_\alpha(P_0)$.*

PROOF. The first part of the Lemma is trivial. We proceed to prove the second part. For any $h \in l_\alpha(P_0)$:

$$\begin{aligned}
E|\Pi(h \mid \mathcal{B})|^\alpha &= \sum_{B \in \mathcal{B}} \frac{P_0^\alpha(B) \int_B p_0^{\alpha+1}}{\left(\int_B p_0^2\right)^\alpha} |E(h \mid B)|^\alpha \\
&\leq \sum_{B \in \mathcal{B}} \frac{P_0^{\alpha-1}(B) \int_B p_0^{\alpha+1}}{\left(\int_B p_0^2\right)^\alpha} \int_B |h|^\alpha p_0 \\
&= \sum_{B \in \mathcal{B}} \frac{E(p_0^\alpha(X) \mid B)}{E^\alpha(p_0(X) \mid B)} \int_B |h|^\alpha p_0 \\
&\leq \sum_{B \in \mathcal{B}} C \int_B |h|^\alpha p_0 = CE|h|^\alpha.
\end{aligned}$$

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

LEMMA 4.2. *Suppose **A4** holds, that $\mathcal{H}$ has an $L_2(P_0)$ envelope, uniform in P_0, 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* $\sup_{h\in\mathcal{H}}\left\|h-\Pi(h|\mathcal{B}))_M\right\|_{P_0}\to 0$.

PROOF. First note that since $\mathcal{H}$ has an $L_2(P_0)$ envelope,

$$\left\|h-E(h\mid\mathcal{B})\right\|_{P_0}\to 0 \tag{4.24}$$

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

Suppose now that $\left\|h_i-E(h_i\mid\mathcal{B}_i)\right\|_{P_0}\to 0$ for some sequence $h_i\in\mathcal{H}$. Then:

$$\begin{aligned}
&\lim_{i\to\infty}\left\|h_i-\Pi(h_i\mid\mathcal{B}_i)\right\|^2_{P_0}\\
&\qquad\le\lim_{i\to\infty}\left\|E(h_i\mid\mathcal{B}_i)-\Pi(h_i\mid\mathcal{B}_i)\right\|^2_{P_0}\\
&\qquad=\lim_{i\to\infty}\sum_{B\in\mathcal{B}_i}\int_B\left(E(h_i\mid B)\left(1-\frac{P_0(B)p_0(x)}{\int_B p_0^2}\right)\right)^2 p_0(x)dx\\
&\qquad\le\lim_{i\to\infty}\sum_{B\in\mathcal{B}_i}P_0(B)E(h_i^2\mid B)\left(\frac{P_0(B)\int_B p_0^3}{\left(\int_B p_0^2\right)^2}-1\right)=0
\end{aligned}\tag{4.25}$$

by assumption **A4** and bounded convergence.

If the conclusion of the proposition is not true, then there is $\varepsilon>0$ and a sequence $h_i\in\mathcal{H}$ such that $\left\|h_i-\Pi(h_i|\mathcal{B}_i)\right\|_{P_0}>2\varepsilon$. By (4.25) this implies that $\left\|h_i-E(h_i|\mathcal{B}_i)\right\|_{P_0}>2\varepsilon$ as well. the Let $i_1=1$ and define

$$i_j=\min\{i:\max_{k<j}\left\|h_{i_k}-\Pi(h_{i_k}|\mathcal{B}_i)\right\|_{P_0}\}<\varepsilon.$$

Note that i_j is finite by (4.24). Then

$$\begin{aligned}
\min_{k<j}\left\|h_{i_j}-h_{i_k}\right\|_{P_0}&\ge\min_{k<j}\left(\left\|h_{i_j}-E(h_{i_k}|\mathcal{B}_{i_j})\right\|_{P_0}-\left\|h_{i_k}-E(h_{i_k}|\mathcal{B}_{i_j})\right\|_{P_0}\right)\\
&\ge\min_{k<j}\left(\left\|h_{i_j}-E(h_{i_j}|\mathcal{B}_{i_j})\right\|_{P_0}-\left\|h_{i_k}-E(h_{i_k}|\mathcal{B}_{i_j})\right\|_{P_0}\right)\\
&\ge\varepsilon.
\end{aligned}$$

Hence there is no ε-net covering $\mathcal{H}$, contradicting the uniform pre-Gaussianity assumption, cf. van der Vaart and Wellner, Theorem 2,8.2).

Proof of the theorem. We need only establish A1. Since $\Pi(\cdot \mid \mathcal{B})$ is a conditional expectation it preserves order (Lemma 4.1) and also reduces $L_\alpha(P_0)$ norm, $\alpha \geq 1$, $E|\Pi(h \mid \mathcal{B})|^\alpha \leq E|h|^\alpha$. Therefore,

$$N_{[\ ]}(\epsilon, \Pi(\mathcal{H} \mid B_0), L_\alpha(P_0)) \leq N_{[\ ]}(\epsilon, \mathcal{H}, L_\alpha(P_0)).$$

(In fact only the usual $E(\Pi(h \mid B))^2 \leq Eh^2$ is needed.) Moreover if the envelope function H possesses a second moment so does the envelope to $\Pi(\mathcal{H} \mid B)$ since $E(\sup_{\mathcal{H}} |h(\cdot)| \mid \mathcal{B}) \geq \sup_{\mathcal{H}}(E(h \mid \mathcal{B}))^2$. 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 $\min_{B_i \in \mathcal{B}} P_0(B_i)$ can converge to infinity and 0 as slowly as needed. Therefore, we can have that $\max_{B \in \mathcal{B}_i} \int_B p_0^2 \, dP_n / \int_B p_0^3 \leq 2$ with probability converging to 1. Hence

$$\begin{aligned}
\left\|\Pi(h|\mathcal{B})\right\|_{P_n}^2 &= \sum_{B \in \mathcal{B}} \left(\frac{P_0(B)}{\int_B p_0^2}\right)^2 E^2(h \mid B) \int_B p_0^2 \, dP_n \\
&\leq 2 \sum_{B \in \mathcal{B}} \frac{P_0(B) \int_B p_0^3}{\left(\int_B p_0^2\right)^2} \int_B h^2 p_0 + o_{\mathcal{P}}(1) \\
&\leq 2CE(h^2) + o_{\mathcal{P}}(1),
\end{aligned}$$

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:** (i) *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 $\equiv \{a_i \leq x_i \leq b_i,\ 1 \leq i \leq d,\ \mathbf{a}, \mathbf{b} \in R^d\}$, $\mathcal{H} =$ Indicators of half spaces $\equiv \{\mathbf{a}^T\mathbf{x} \leq c : |\mathbf{a}| = 1, c \in R\}$

where $|\cdot|$ is the Euclidean norm, $\mathcal{H}$ = Fourier transforms restricted to a compact $\equiv \{h_{\mathbf{t}} : h_{\mathbf{t}}(\mathbf{x}) = \exp(i\mathbf{t}^T\mathbf{x}),\ \mathbf{t} \in K \text{ a compact}\}$, 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 $\mathcal{H} = \{\exp\{\mathbf{s}^T\mathbf{x}\} : |s| \le \epsilon\}$. 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

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

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_0 there exist functions $h_\tau(\cdot, p_0)$ such that

$$\tilde{\tau}_n = \tau(p_0) + \int h_\tau(x, p_0)dP_n(x) + o_{p_0}(n^{-1/2}).$$

This is just the statement that $\tau(p_0)$ is efficiently estimable over a nonparametric model $\mathcal{P}$.

(ii) Let $\tilde{\mathcal{T}} = \{\tau_h(p) = \int hp : h = h_\tau(\cdot, p_0) \text{ for some } \tau \in \mathcal{T},\ p_0 \in \mathcal{P}\}$. $\tilde{\mathcal{T}}$ satisfies the conditions of Theorem 4.1.

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

$$\sup\{|\tau(\hat{p}_n) - \tau(p_0) - \int h_\tau(\cdot, p_0)(\hat{p}_n - p_0)| : \tau \in \mathcal{T}\} = o_p(n^{-1/2})\ \forall\ P_0 \in \mathcal{P}_0.$$

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

$$\tau(p) = P^{-1}(s) : \epsilon \leq s \leq 1 - \epsilon,\ \epsilon > 0$$

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) if immediate with

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

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

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

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 $\epsilon \leq s \leq 1-\epsilon$ for n sufficiently large. Now uniform convergence of $\hat{P}_n$ and strict monotonicity of P implies

$$\sup_x\{|\hat{P}_n^{-1}(t) - P^{-1}(t)| : \epsilon \leq t \leq 1 - \epsilon\} \xrightarrow{P} 0.$$

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

4.3.1. *The density and cdf in the biased sample model* We consider the problem of density estimation with the cdf 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 Bickel et al. (1998) for a description of the efficient estimator of the cdf 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.

4.3.2. *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 cummulative integral, $\int_0^{\cdot} \lambda(t)\,dt$, both on a fixed interval $(0, a)$, such that with positive probability we observed uncensored values larger than a. Efficient estimation of the hazard function was discussed e.g. 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 density-cdf 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 cdf 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.

4.3.3. *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 $\tau_h = \int hf\vartheta$, $h \in \mathcal{H}$. 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** are 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 (4.19). We shall pursue this at the end of this section. A direct approach of modifying the kernel density estimator was already discussed. We next consider another example in which a "standard" estimator can be modified to obtain broad strong plug-in properties. Again, we concentrate on density estimators whose cdf are asymptotically equivalent to the empirical distribution function.

EXAMPLE 5.1. *Orthonormal and Log orthonormal series density estimators.* Another general class of density estimators is based on orthonormal bases $\psi_1, \psi_2, \ldots$. There are two main variants. The first is the sieve MLE based on the exponential family $c\exp\left(\sum_{j=1}^M \beta_j\psi_j(a\cdot)\right)$. The second is the density estimator given by

$$\sum_{j=1}^{M_n} P_n(\psi_j)\psi_j(\cdot). \tag{5.26}$$

If the ψ_j are splines the first is the log spline estimate (Kooperberg and Stone, 1992). Note that for both estimators:

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

We proceed for estimates of type (5.26). 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 $h \in \mathcal{H}$:

$$\begin{aligned}
&\int h(x)\hat{p}(x)dx - P_n(h) \\
&= \int (h(x) - h^*(x))\hat{p}(x)dx - P_n(h - h^*), \qquad (5.27)\\
&= \int (h(x) - h^*(x))(\hat{p}(x) - p_0(x))dx - (P_n(h - h^*) - P_0(h - h^*),
\end{aligned}$$

where h^* is some function approximating h in the span $\mathcal{S}_{M_n}$ of $h_1, \dots, h_{M_n}$ and $\psi_1, \dots, \psi_M$, say $h^* = \Pi_{M_n} h$ or $\Pi^P_{M_n} h$, the projection in $L_2(P)$. Suppose that the second term on the RHS is $o_p\left(n^{-1/2}\right)$. If so we need consider only the first term. Note that for the estimator given by (5.26), the first term is simply:

$$\int h^{\perp}(x)p^{\perp}(x) \tag{5.28}$$

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,

$$\int p^{\perp 2}(x)dx < CM_n/n \tag{5.29}$$

for some finite C. Hence we obtain the strong plug-in property for $\hat{p}_n$ if,

$$\sup_{h \in \mathcal{H}} \int h^{\perp 2}(x)dx = o(M_n^{-1}). \tag{5.30}$$

We proceed to a general theorem.

B1: The estimate $\tilde{p}_n$ of form (5.26) based on $\psi_1, \ldots, \psi_{M_n}$ satisfies (4.17).

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

$$\sup_{\mathcal{H}} \|h - \Pi_{M_n} h\|_2 = o_{\mathcal{P}}\left(n^{-1/2}\right). \tag{5.31}$$

B3: $p \leq C < \infty$.

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

Our discussion has established,

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 $\mathcal{Q}_\alpha = \{p \text{ on } I^d : \|D^\beta p\|_2 \leq C \text{ all } \beta \leq \alpha\}$ and $\alpha > \frac{d}{2}$ and the $\{\psi_1, \ldots, \psi_M, h_1, \ldots, h_{L_n}\}$ are a spline basis on the unit d-cube I^d.

Using the results and techniques of Stone(1991) one can show with some more labor that the log spline estimate also has this property if we also require that $p \geq c > 0$ 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 (5.26).

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

$$\sup_{\mathcal{P}} \|p - \Pi_{K_n} p\|_2 = O\left(b_n n^{-1/2}\right) \tag{5.32}$$

where b_n is given in (4.19).

B6: Let $\Pi^*_M h$ be defined by

$$\min{}^{-1} \left| \int h^* \; dP_n - \int h \; dP_n : h^* \in \mathcal{S}_M \right| .$$

Then, if K_n is as above,

$$\sup_{\mathcal{H}} |P_n(h) - P_n(h^*)| = o_{\mathcal{P}}\left(b_n n^{-1/2}\right) . \tag{5.33}$$

Define,

$$\hat{p}_n = \sum_{j=1}^{K_n} \hat{c}_j \psi_j \tag{5.34}$$

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

$$\begin{array}{ll} \text{Minimize} & \displaystyle\sum_{j=1}^{M_n} (\hat{c}_j - c_j)^2 + \sum_{j=M_n+1}^{K_n} c_j^2 \\ \text{subject to} & \displaystyle\left| \sum_{j=1}^{K_n} d_j c_j - P_n(h) \right| \leq n^{-1/2} b_n \end{array}$$

for all $\mathbf{d}$ 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 $\left(\frac{i_1}{K_n}, \dots, \frac{i_d}{K_n}\right)$ 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

[1] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist*, **10**, 1100-1120.

[2] 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.

[3] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998): *Efficient and adaptive estimation for semiparametric models*, Springer Verlag, New York. Boor

[4] Brown, L. D. and Low, M. G. (1996): Asymptotic equivalence of nonparametric regression and white noise. The Annals of Statistics **24**, pp. 2384–2398. Title : Author : .

[5] Cai, T. T. (2000): On Adaptive Wavelet Estimation of a Derivative and other related linear inverse problems. Unpublished.

[6] Csörgő, S. and Mielniczuk, J. (1988): Density estimation in the simple proportional hazards model. *tatistics & Probability Letters*, **6**, 419–426.

[7] Efron, B. and Tibshirani, R. (1996): Using specially designed exponential families for density estimation. *The Annals of Statistics* **24**. pp. 2431–2461.

[8] Kooperberg, C. ;Stone, C. J. (1992): A study of logspline density estimation. *Computational Statistics and Data Analysis* **12**, pp. 327–347

[9] 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.

[10] Ghorai, J. K. and Pattanaik, L. M (1993): Asymptotically optimal bandwidth selection of the kernel density estimator under the proportional hazards model. *Communications in Statistics*, **A22**, 1383–1401.

[11] Nussbaum, M. (1994): Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics* **24**, pp. 2399–2430.

[12] Shorack, G. R. (1969): Asymptotic normality of linear combinations of functions of order statistics, *The Annals of Mathematical Statistics*, **40**, 2041–2050.

[13] Stone, C. J. (1990): Large-sample inference for log-spline models *The Annals of Statistics* **18**, pp. 717–741.

[14] van der Vaart, A. and Wellner, J. A. (1996). "Weak Convergence and Empirical Processes", Springer, New York.

[15] Vardi, Y. (1985): Empirical distributions in selection bias model. *Ann. Statist.*, **13**, 178–205.