Semiparametric estimation: efficiency bounds

Paul Delatte
delatte@usc.edu
University of Southern California

Last updated: 4 November, 2022

1 Semiparametric models

A semiparametric model is a family of Borel probability distributions of the form

$$\mathscr{P} = \{\mathbb{P}_{\theta,G} : \theta \in \Theta, G \in \mathscr{G}\}$$

where $\Theta \subseteq \mathbb{R}^d$ is some finite dimensional space and $\mathscr{G}$ is some infinite dimensional space (generally some space of distributions). Within the limit of this definition, semiparametric models seem to be completely equivalent to nonparametric models (taking $\Theta = \{0\}$); the difference lies in the inferential procedure. In general, in semiparametric models, the parameter of interest is only the finite dimensional one θ while the infinite dimensional parameter G is considered as a nuisance parameter. In nonparametric models, the parameter of interest is the infinite dimensional parameter G. Semiparametric models appear naturally by simple relaxation of parametric models.

Example 1 (Symmetric Location Model)**.** We observe $X = \theta + \varepsilon$ where $\varepsilon \sim G \in \mathscr{G}_s$ and $\mathscr{G}_s$ is the set of distributions on $\mathbb{R}$ with density g with respect to the Lebesgue measure λ which is symmetric around 0. This defines a semiparametric model

$$\mathscr{P} = \left\{\mathbb{P}_{\theta,G} : \frac{d\mathbb{P}_{\theta,G}}{d\lambda}(x) = g(x-\theta),\ \theta \in \mathbb{R},\ G \in \mathscr{G}_s\right\}$$

The parametric normal location-scale model with $\varepsilon \sim N(0, \sigma^2)$ belongs to $\mathscr{P}$.

Example 2 (Regression)**.** We observe $X = (Y, Z) \sim \mathbb{P}_{\theta,G}$ where

$$Y = \mu(Z, \theta) + \sigma(Z, \theta)\varepsilon,$$

Z and ε are independent, the functions μ and σ are known up to the finite dimensional parameter θ, and $\varepsilon \sim G \in \mathscr{G}$ where $\mathscr{G}$ is the collection of all absolutely continuous distributions on $\mathbb{R}$. This defines a semiparametric model

$$\mathscr{P} = \left\{\mathbb{P}_{\theta,G} : \frac{d\mathbb{P}_{\theta,G}}{d\lambda \times m}(y, z) = g\left(\frac{y - \mu(z,\theta)}{\sigma(z,\theta)}\right)h(z),\ \theta \in \mathbb{R}^d,\ G \in \mathscr{G}\right\}$$

where m is a σ-finite measure. The parametric normal (homoskedastic) linear regression model belongs to $\mathscr{P}$ with $\varepsilon \sim N(0,1)$, $\theta = (\beta, \eta) \in \mathbb{R}^d \times \mathbb{R}_+$, $\sigma(Z, \theta) = \eta$, and $\mu(Z, \theta) = \beta^T Z$.

Compared to parametric settings, finding "good" estimators in semiparametric models is not as straightforward. There is no direct intuitive method like MLE which delivers estimators as "good" as one can hope (even before considering what good means). For this reason, it is best to start by clearly defining a notion of efficiency in semiparametric models. In parametric models, efficiency is defined by considering all $n^{1/2}$-consistent (regular) estimators[1] for a given problem and finding a lower bound on their asymptotic variances. The extension to semiparametric and nonparametric models is based on the idea that estimation cannot be better in those models than efficient estimation in their (one-dimensional) parametric submodel in which estimation is the most difficult (and so efficient estimation is the least favorable).

2 Parametric efficiency bounds

Consider i.i.d. random variables $X_1, \dots, X_n$ with finite dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Assume the models to be "regular enough" (see regularity conditions for CAN of MLE): in general, we can assume local asymptotic normality (LAN) of the model

$$\log \frac{d\mathbb{P}_{\theta+n^{-1/2}v,n}}{d\mathbb{P}_{\theta,n}} = v^T \Delta_n - \frac{1}{2} v^T I(\theta) v + o_{\mathbb{P}_{\theta,n}}(1),$$

where $I(\theta) > 0$ and $\Delta_n \xrightarrow[\mathbb{P}_{\theta,n}]{d} N(0, I(\theta))$, which holds, e.g., under quadratic mean differentiability (QMD)

$$\int \left(\sqrt{p_{\theta+v}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} v^T \dot{l}_\theta(x) \sqrt{p_\theta(x)} \right)^2 d\mu(x) = o\left(\|v\|^2 \right) \quad \text{as} \quad \|v\| \to 0$$

with $\Delta_n = n^{-1/2} \sum_{i=1}^n \dot{l}_\theta(x)$ and $I(\theta) = \mathbb{E}\,(\dot{l}_\theta \dot{l}_\theta^T)$. Note that under some regularity conditions (including the (a.e.) existence of the pointwise derivative of p_θ with respect to θ), we have that $\dot{l}_\theta = \partial \log p_\theta / \partial\theta = \dot{p}_\theta / p_\theta$ and we find back the standard definition of the score function $\dot{l}_\theta$ and of the Fisher information $I(\theta)$. Assume further that the estimators are (locally) regular (see previous footnote) in the sense that for every sequence $(\theta_n)_{n\in\mathbb{N}}$ in Θ with $\lim_{n\to\infty} \sqrt{n}(\theta_n - \theta) = v \in \mathbb{R}^k$,

$$\sqrt{n}(T_n - \theta_n) \rightsquigarrow_{\mathbb{P}_{\theta_n}} L_\theta,$$

where L_θ is a distribution that depends on θ but not on h, or equivalently, for every $v \in \mathbb{R}^k$,

$$\sqrt{n}\left(T_n - \theta - \frac{v}{\sqrt{n}} \right) \rightsquigarrow_{\mathbb{P}_{\theta+v/\sqrt{n}}} L_\theta.$$

Under these conditions, we have efficiency results: the convolution theorem of Kaufman and Hajek, which shows that among $n^{1/2}$-consistent regular estimators, those with normal limit and inverse Fisher asymptotic variance are "best".

[1] Recall that: $n^{1/2}$-consistent means that $n^{1/2}(T_n - \theta) = O_p(1)$; regularity means that the estimators have "limit distribution [that] does not depend on the direction of approach of θ to θ_0 (see Delatte HDstats). Regularity is used for the convolution theorem to hold. If regularity is not assumed, one can consider LAM theorems, but in semiparametric modelling it is standard to assume regularity (see comment S.2.3. p.27 in Bickel et al.)

Proposition 3 (**Convolution Theorem**). *Let* $(\mathbb{P}_\theta : \theta \in \Theta)$ *be Borel probability measures where* $\Theta \subseteq \mathbb{R}^d$ *is open. If* $(\mathbb{P}_\theta : \theta \in \Theta)$ *is differentiable in quadratic mean at* $\theta_0 \in \Theta$ *with invertible Fisher information matrix* $I(\theta_0)$ *and* T_n *is a regular estimator of* θ *at* θ_0 *with scaled limit distribution* L_{θ_0}*, then there exists a probability measure* M_{θ_0} *such that*

$$L_{\theta_0} = N(0, I^{-1}(\theta_0)) * M_{\theta_0}.$$

In particular, if L_{θ_0} *has variance* $V(\theta_0)$*, then* $V(\theta_0) - I^{-1}(\theta_0)$ *is positive semidefinite.*

Proof. T.8.8. in vdV AS p.115 or T.4.1. in Wellner notes p.110 or T.2.3.1. in Bickel&Wellner EAESM p.24. The conditions can be relaxed, in particular to LAN. □

This equivalently says, if we denote $\mathbb{Z}_{\theta_0} \sim L_{\theta_0}$ the scaled weak limit of T_n, that there are random variables Z_{θ_0} and Δ_{θ_0}, independent of one another, such that $\mathbb{Z}_{\theta_0} = Z_{\theta_0} + \Delta_{\theta_0}$ where $Z_{\theta_0} \sim N(0, I^{-1}(\theta_0))$ and $\Delta_{\theta_0} \sim M_{\theta_0}$. "In words, [this] says that the [scaled] limiting distribution of any regular estimator T_n of θ must be at least as "spread out" as the $N(0, I^{-1}(\theta_0))$ distribution of Z_{θ_0}." (Wellner p.110) Thus among regular estimators we could legitimately define an (asymptotically) efficient estimator as one for which the scaled limiting distribution is exactly $N(0, I^{-1}(\theta_0))$. This can be restated in terms of asymptotic optimality with respect to bowl-shaped loss functions, by applying Anderson's lemma.

Corollary 4 (Hajek, 1970)). *Suppose the conditions of the convolution theorem hold for* $(\mathbb{P}_\theta : \theta \in \Theta)$ *and* T_n*. If* $l \colon \mathbb{R}^d \to \mathbb{R}^+$ *is bowl-shaped, then*

$$\liminf_{n\to\infty} \mathbb{E}_{\theta_0}[l(\sqrt{n}(T_n - \theta_0))] \geq \mathbb{E}_{\theta_0}[l(Z_{\theta_0})]$$

where $Z_{\theta_0} \sim N(0, I^{-1}(\theta_0))$.

Proof. C.1. in Wellner p.110 and "Asymptotic optimality theorem" in Bickel&Wellner p.26. This obtains by combining the convolution theorem with Anderson's lemma. □

The estimators achieving the efficiency bounds $N(0, I^{-1}(\theta))$ can be completely characterized (see L.8.14. in vdV AS p.120 or T.1. in BKRW EAESM p.24): we have that

$$\sqrt{n}(T_n - \theta_n) \rightsquigarrow_{\mathbb{P}_{\theta_n}} N(0, I^{-1}(\theta))$$

if and only if

$$\sqrt{n}\left(T_n - \theta_n - \frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta_n)\dot{l}_{\theta_n}(X_i)\right) \xrightarrow{\mathbb{P}_{\theta_n}} 0,$$

that is, T_n is asymptotically linear in the efficient influence function $\tilde{l}_\theta(x) = I^{-1}(\theta)\dot{l}_\theta(x)$.

The restriction to regular estimators (which excludes some important estimators) can be relaxed by considering a minimax framework. This gives the important LAM theorems which deliver approximately the same results: the best estimators are those with limiting scaled distribution $N(0, I^{-1}(\theta))$.

Proposition 5 (**LAM Theorem** (Hajek, 1972)). *Let* $(\mathbb{P}_\theta : \theta \in \Theta)$ *be Borel probability measures where* $\Theta \subseteq \mathbb{R}^d$ *is open. Let* T_n *be any estimator. If* $(\mathbb{P}_\theta : \theta \in \Theta)$ *is differentiable*

in quadratic mean at $\theta_0 \in \Theta$ with invertible Fisher information matrix $I(\theta_0)$ and $l\colon \mathbb{R}^d \to \mathbb{R}^+$ is a bowl-shaped function, then for any $\delta > 0$,

$$\liminf_{n\to\infty} \sup_{\{\theta:\|\theta-\theta_0\|<\delta\}} \mathbb{E}_\theta[l(\sqrt{n}(T_n-\theta))] \geq \mathbb{E}[l(Z_{\theta_0})]$$

where $Z_{\theta_0} \sim N(0, I^{-1}(\theta_0))$.

Proof. T.12.1. in Ibragimov&Has'minskii SE p.162 for $\varphi(\varepsilon) = n^{-1/2}$. Then R.2.12.2. in Ibragimov&Has'minskii SE p.168 allows to restate the inequality as in T.4.2. in Wellner notes p.110 or "Locally asymptotic minimax theorem" in Bickel&Wellner p.27 or T.16.25. in Keener TS p.340, that is,

$$\lim_{t\to+\infty} \liminf_{n\to\infty} \sup_{\{\theta:\sqrt{n}\|\theta-\theta_0\|\leq t\}} \mathbb{E}_\theta[l(\sqrt{n}(T_n-\theta))] \geq \mathbb{E}[l(Z_{\theta_0})]$$

Another more refined version is T.8.11. in vdW AS p.117-118, which yields

$$\sup_I \liminf_{n\to\infty} \sup_{v\in I} \mathbb{E}_{\theta_0+v/\sqrt{n}}[l(\sqrt{n}(T_n-\theta_0+v/\sqrt{n}))] \geq \mathbb{E}[l(Z_{\theta_0})]$$

where the first supremum runs over all finite subsets I of $\mathbb{R}^d$. The result in vdV is proved using the weak topology for experiments. (It is proved in greater generality in 3.11.5. in WCEP vdW&Wellner p.417.) The condition of QMD with invertible Fisher can again be weakened to LAN. The finer result as in vdV can be extended to other limit experiments (see e.g. T.5. in Pollard's thoughts (2000), Lecture 7 in Pollard's Paris lectures (2001), S.7.4. in Torgersen (1991), S.62. in Strasser (1985), or S.5. in vdV (2002)). □

Remark. All the results extend to estimation of $q(\theta)$ instead of θ where $q\colon \mathbb{R}^d \to \mathbb{R}^k$ is differentiable. In this case, the information bound becomes $\dot{q}(\theta)I^{-1}(\theta)\dot{q}(\theta)^T$ and the efficient influence function $\tilde{l}_\theta(x) = \dot{q}(\theta)I^{-1}(\theta)\dot{l}_\theta(x)$.

Addendum. Nuissance parameters. We derive parametric efficiency bounds in presence of (finite-dimensional) nuissance parameters because the same ideas will be of use in the semiparametric case. Consider $\theta = (\theta_1^T, \theta_2^T)^T$ where $\theta_1 \in \mathbb{R}^d$ is the parameter of interest and $\theta_2 \in \mathbb{R}^k$ is the nuisance parameter. Denote

$$I(\theta) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \quad \text{and} \quad I^{-1}(\theta) = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix}.$$

By taking $q(\theta) = \theta_1$ and using the formulas derived in the remark above, we find that the information bound for estimating θ_1 is

$$I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$$

and the efficient influence function for estimating θ_1 is

$$\tilde{l} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2)$$

where $\dot{l} = (\dot{l}_1^T, \dot{l}_2^T)^T$.

These results can be reinterpreted geometrically in terms of the Hilbert space $L_2(\mathbb{P}_\theta)$.

If we denote the linear span of x by $[x]$, then the standard formulas[2] for the projection matrix yields that the projection $\Pi(\dot{l}_1 \mid [\dot{l}_2])$ of $\dot{l}_1$ onto $[\dot{l}_2]$ is

$$\Pi(\dot{l}_1 \mid [\dot{l}_2]) = I_{12} I_{22}^{-1} \dot{l}_2.$$

Then we get that

$$\tilde{l} = (\mathbb{E}\,(l_1^* l_1^{*T}))^{-1} l_1^*$$

where

$$l_1^* = \dot{l}_1 - \Pi(\dot{l}_1 \mid [\dot{l}_2])$$

can be interpreted as the efficient score function for the estimation of θ_1. Similarly, we can get that

$$\Pi(\tilde{l} \mid [\dot{l}_1]) = (\mathbb{E}\,(\dot{l}_1 \dot{l}_1^T))^{-1} \dot{l}_1.$$

This means that the projection of the efficient influence function for estimating θ_1 onto the span of $\dot{l}_1$ is equal to the efficient influence function for estimating θ_1 when θ_2 is known. Moreover, if $\dot{l}_1 \perp \dot{l}_2$, then $l_1^* = \dot{l}_1$ and so

$$\tilde{l} = (\mathbb{E}\,(\dot{l}_1 \dot{l}_1^T))^{-1} \dot{l}_1 = \Pi(\tilde{l} \mid [\dot{l}_1]),$$

that is, under orthogonality of the score function components, we have that "estimation of θ_1 is asymptotically as difficult when θ_2 is unknown as when θ_2 is known".

3 Semiparametric efficiency bounds

There are two equivalent methods to derive efficiency bounds for semiparametric models: one that applies when the parameters are defined as functions of $\mathscr{P}$ (and thus also works for nonparametric models); one that applies when the parameters are defined via implicit parametrizations (and thus applies more directly to semiparametric models).

The (standard) inner product on $L_2(\mu)$ is simply denoted $\langle \cdot\,, \cdot \rangle$, that is, if $g, h \in L_2(\mu)$, then

$$\langle g, h \rangle = \int g h \, d\mu.$$

If there is risk a confusion, we may write $\langle g, h \rangle_{L_2(\mu)}$ to make clear the measure with respect to which integrals are taken.

3.1 Submodels and tangent spaces

Before introducing the two methods, we need to present the general idea of these methods which accommodate infinite-dimensional parameters by considering parametric one-dimensional submodels of the original model. A central object fo this approach is the tangent space (something we already encountered in a different context when considering the spans $[\dot{l}_1]$ and $[\dot{l}_2]$ in the parametric case).

For parametric regular models $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$, we know that score functions (or their generalization under QMD) play an important role. Suppose that $\mathscr{P}$ is dominated by a measure μ so that $\mathbb{P}_\theta \in \mathscr{P}$ has μ-density p_θ. Recall that under quadratic mean differentiability (QMD) of $\mathscr{P}$ at θ (which is equivalent to Fréchet differentiability

[2] $A(A^TA)^{-1}A^T$ is the projection matrix onto the space spanned by the columns of A.

of the map $\theta \mapsto \sqrt{p_\theta} =: s_\theta$ at θ)

$$\int \left(\sqrt{p_{\theta+v}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} v^T \dot{l}_\theta(x) \sqrt{p_\theta(x)} \right)^2 d\mu(x) = o\left(\|v\|^2 \right) \quad \text{as} \quad \|v\| \to 0,$$

the score function is taken to be $\dot{l}_\theta$ in the expression above. Then we can define the **tangent space** $\dot{\mathscr{P}}$ at $\mathbb{P}_\theta \in \mathscr{P}$ as the linear span of the components of $\dot{l}_\theta$ (which is closed since finitely generated), that is, $\dot{\mathscr{P}} := [\dot{l}_\theta] = [\dot{l}_{\theta,1}, \ldots, \dot{l}_{\theta,d}]$. Since score functions have mean zero $\mathbb{E}_\theta(\dot{l}_{\theta,i}) = 0$, we have that $\dot{\mathscr{P}} \subseteq L_2^0(\mathbb{P}_\theta) = \{h \in L_2(\mathbb{P}_\theta) : \mathbb{E}_\theta h = 0\}$ (see E.2.4. for a proof).

Let us now consider the semiparametric and nonparametric cases. Let $\mathscr{P}$ be some nonparametric model (dominated by a σ-finite measure μ) and let $\mathbb{P}_* \in \mathscr{P}$ be the true distribution. We identify $\mathscr{P}$ with the subset $\mathscr{S}$ of the Hilbert space $L_2(\mu)$ through the correspondence $\mathbb{P} \mapsto \sqrt{\frac{d\mathbb{P}}{d\mu}} =: s$. We then define a **one-dimensional submodel** of $\mathscr{P}$ (for $\mathbb{P}_*$) as any (one-dimensional) continuously differentiable curve $\mathscr{S}^1$ of $\mathscr{S}$ such that $s_* \in \mathscr{S}^1$, that is, there exists a continuously Fréchet differentiable map $C\colon B_1(0;\mathbb{R}) \to \mathscr{S}$ with rank 1 derivative such that $C(t_0) = s_*$ for some $t_0 \in B_1(0;\mathbb{R})$. That is,

$$\mathscr{S}^1 = C(B_1(0;\mathbb{R})) = \{s_t : t \in C^{-1}(\mathscr{S})\}.$$

For a given submodel, the parametrization C may not be unique, but under continuous differentiability and the rank condition the linear space spanned by the tangent vector is invariant under equivalent parametrizations (see BKRW p.49). For our purpose, we can thus restrict attention to curves such that $C(0) = s_*$, abuse terminology by calling curve passing through $s_0 = s_*$ (not only the image of any such C but) the function C itself, and abuse notation by writing for a given submodel $\mathscr{S}_1$ parametrized by any such curve C

$$\mathscr{S}^1 = \mathscr{S}^C.$$

In what follows, we directly write $\mathbb{P}_0$ for $\mathbb{P}_*$ and s_0 for s_*. Moreover, since the assumption that C is Fréchet differentiable at s_0 is equivalent to the fact that the model $\mathscr{P}^C = \{\mathbb{P}_t : t \in C^{-1}(\mathscr{S})\}$ corresponding to $\mathscr{S}^C = \{s_t : t \in C^{-1}(\mathscr{S})\}$ is QMD at 0, we can consider the score function of $\mathscr{P}^C$ at $\mathbb{P}_0$ which we denote $\dot{l}^C$. Then we define the **tangent set** $\dot{\mathscr{P}}^0$ of $\mathscr{P}$ at $\mathbb{P}_0$ as the union of score functions $\dot{l}^C$ for all one-dimensional submodels $\mathscr{P}^C$ of $\mathscr{P}$ (with $C(\mathbb{P}_*) = 0$), or equivalently, for all curves C passing trough s_0. That is,

$$\dot{\mathscr{P}}^0 = \bigcup \{\dot{l}^C : C \text{ is a curve of } \mathscr{S} \text{ passing through } s_0\}$$

We then define the **tangent space** $\dot{\mathscr{P}}$ of $\mathscr{P}$ at $\mathbb{P}_0$ as the closed linear span of $\dot{\mathscr{P}}^0$, that is,

$$\dot{\mathscr{P}} = [\dot{\mathscr{P}}^0].$$

Naturally, we can define $\dot{\mathscr{S}}^0$ and $\dot{\mathscr{S}}$ by

$$\dot{\mathscr{S}}^0 = \bigcup \{\dot{s}^C : C \text{ is a curve of } \mathscr{S} \text{ passing through } s_0\}$$

and

$$\dot{\mathscr{S}} = [\dot{\mathscr{S}}^0]$$

where $\dot{s}^C$ is the Fréchet derivative of $C(= t \mapsto s_t)$ at 0 which relates to $\dot{l}^C$ by

$$\dot{l}^C = 2\frac{\dot{s}^C}{s_0}.$$

By definition and some convergence results (see E.2.4.), we have the following inclusions:

$$\dot{\mathscr{P}}^0 \subseteq \dot{\mathscr{P}} \subseteq L_2^0(\mathbb{P}_0) \quad \text{and} \quad \dot{\mathscr{S}}^0 \subseteq \dot{\mathscr{S}} \subseteq L_2(\mu).$$

Example 6 (Dominated Family of Probability Measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$)**.** Let μ be a fixed σ-finite measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\mathscr{P}$ the set of all probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ dominated by μ, that is,

$$\mathscr{P} = \{\mathbb{P} \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) : \mathbb{P} << \mu\}.$$

Then we claim that

$$\dot{\mathscr{P}} = L_2^0(\mathbb{P}_0) = \{g \in L_2(\mathbb{P}_0) : \int g \, d\mathbb{P}_0 = 0\},$$

or equivalently that

$$\dot{\mathscr{S}} = \{g \in L_2(\mu) : g = g\mathbb{1}_{[s_0>0]}, \langle g, s_0\rangle = \int g s_0 \, d\mu = 0\}.$$

We should actually show that $\dot{\mathscr{P}}^0 = \dot{\mathscr{P}} = L_2^0(\mathbb{P}_0)$.

We start with the direct inclusion $\dot{\mathscr{P}}^0 \subseteq L_2^0(\mathbb{P}_0)$. Let $g \in \dot{\mathscr{P}}^0$. By definition (see equivalence QMD and Fréchet differentiability – (1) in T.75.2. in Strasser), $g \in L_2(\mathbb{P}_0)$ and QMD at $t = 0$ implies that

$$\frac{1}{t^2}\int \left(\sqrt{\frac{d\mathbb{P}_t}{d\mathbb{P}_0}} - 1 - \frac{t}{2}g\right)^2 d\mathbb{P}_0 = o(1) \quad \text{as} \quad t \to 0+,$$

i.e.,

$$\frac{1}{t}\left(\sqrt{\frac{d\mathbb{P}_t}{d\mathbb{P}_0}} - 1\right) \xrightarrow[t\to 0+]{L_2(\mathbb{P}_0)} \frac{1}{2}g.$$

Since L_2 convergence implies L_1 convergence, we have

$$\begin{aligned}\frac{1}{2}\int g \, d\mathbb{P}_0 &= \lim_{t\to 0+}\frac{1}{t}\int\left(\sqrt{\frac{d\mathbb{P}_t}{d\mathbb{P}_0}} - 1\right)d\mathbb{P}_0 \\ &= -\lim_{t\to 0+}\frac{2}{t}d_2^2(\mathbb{P}_0, \mathbb{P}_t),\end{aligned}$$

where the second inequality follows from the properties of the Hellinger distance. Moreover, QMD implies

$$\lim_{t\to 0+}\frac{1}{t^2}d_2^2(\mathbb{P}_0, \mathbb{P}_t) < +\infty,$$

hence

$$\int g \, d\mathbb{P}_0 = 0.$$

We now show the reverse implication $L_2^0(\mathbb{P}_0) \subseteq \dot{\mathscr{P}}^0$. To show this, we show that $DL_2^0(\mathbb{P}_0) = \{h \in L_2^0(\mathbb{P}_0) : h \text{ is bounded}\}$ is dense in $L_2^0(\mathbb{P}_0)$ and is contained in $\dot{\mathscr{P}}^0$. To see that $DL_2^0(\mathbb{P}_0)$ is dense, consider $g_n = g\mathbb{1}_{\{|g|<n\}}$, $n \in \mathbb{N}$, for $g \in L_2^0(\mathbb{P}_0)$. Then $((g_n - \int g_n \, d\mathbb{P}_0))_{n\in\mathbb{N}}$ is a sequence in $DL_2^0(\mathbb{P}_0)$ that approximates g since for any $n \in \mathbb{N}$,

$$\begin{aligned}\int \left(\left(g_n - \int g_n \, d\mathbb{P}_0\right) - g\right)^2 d\mathbb{P}_0 &= \int (g_n - g)^2 \, d\mathbb{P}_0 - \left(\int g_n \, d\mathbb{P}_0\right)^2 \\ &\leq \int (g_n - g)^2 \, d\mathbb{P}_0 \\ &= \int_{|g|>n} g^2 \, d\mathbb{P}_0,\end{aligned}$$

where the first equality follows from $\int g \, d\mathbb{P}_0 = 0$ after expanding the square. To see that $DL_2^0(\mathbb{P}_0)$ is in $\dot{\mathscr{P}}^0$, take $h \in DL_2^0(\mathbb{P}_0)$ and define for any $t \in \mathbb{R}$,

$$p(t) = \exp(th - b(t))p_0$$

with

$$b(t) = \log \int \exp(th)p_0 \, d\mu.$$

Then $t \mapsto p(t) \in \mathscr{P}$ is an exponential family, and for $h \neq 0$, regular with $\dot{s}(0) = hs_0/2$, since $\dot{b}(0) = \int h \, d\mathbb{P}_0 = 0$. Therefore, $h \in \dot{\mathscr{P}}^0$.

Since $\text{span}(\dot{\mathscr{P}}^0) \subseteq L_2^0(\mathbb{P}_0)$ by linearity of the integral and $L_2^0(\mathbb{P}_0)$ is closed, we have that $\dot{\mathscr{P}}^0 = L_2^0(\mathbb{P}_0)$ implies $\dot{\mathscr{P}} = L_2^0(\mathbb{P}_0)$. This concludes the proof.

Reference. E.3.2.1. in BKRW p.52 and L.B.1. in S&T p.334. The proof that QMD with score g implies that $\int g \, d\mathbb{P}_0 = 0$ can be found in Strasser (1985) T.75.2. p.383 (see also Appendix B for further references). For the proof that $DL_2^0(\mathbb{P}_0)$ is dense in $L_2^0(\mathbb{P}_0)$, see L.75.5. in Strasser (1985).

These definitions also allow us to recast the definitions of semiparametric and nonparametric models. A nonparametric model $\mathscr{P}$ is a model that has maximal tangent space $\dot{\mathscr{P}} = L_2^0(\mathbb{P}_0)$. A semiparametric model $\mathscr{P}$ is a model that has tangent space $\dot{\mathscr{P}}$ an infinite-dimensional proper subset of $L_2^0(\mathbb{P}_0)$.

3.2 Parameters defined via implicit parametrization

Consider a semiparametric model

$$\mathscr{P} = \{\mathbb{P}_{\theta,G} : \theta \in \Theta, G \in \mathscr{G}\}$$

where $\Theta \subseteq \mathbb{R}^d$ and $\mathscr{G}$ is an infinite-dimensional space. Fix the true distribution to be $\mathbb{P}_0 = \mathbb{P}_{\theta_0,G_0}$. Then define the submodels

$$\mathscr{P}_1 = \{\mathbb{P}_{\theta,G_0} : \theta \in \Theta\}$$

and

$$\mathscr{P}_2 = \{\mathbb{P}_{\theta_0,G} : G \in \mathscr{G}\}.$$

By definition, $\dot{\mathscr{P}} \supseteq \dot{\mathscr{P}}_1 + \dot{\mathscr{P}}_2$. In general, the reverse inclusion holds, and we have an equality $\dot{\mathscr{P}} = \dot{\mathscr{P}}_1 + \dot{\mathscr{P}}_2$. The potential issue is that $\dot{\mathscr{P}}_1 + \dot{\mathscr{P}}_2$ might not be closed when both $\mathscr{P}_1$ and $\mathscr{P}_2$ are infinite-dimensional.

Under regularity conditions, the Hajek–Le Cam convolution theorem with information bounds derived in the case of (finite-dimensional) nuisance parameters can be extended to models of the form of $\mathscr{P}$ (see C.3.4.1. in BKRW EAESM p.72). If T_n is an estimator of θ_0 locally regular at $\mathbb{P}_0$ for any parametric submodel, then the limit distribution of $\sqrt{n}(T_n - \theta_0)$ is the convolution of some distribution and the normal distribution with mean 0 and covariance matrix (**information bound**)

$$(\mathbb{E}_0(l_1^* l_1^{*T}))^{-1} \quad \text{with} \quad l_1^* = \dot{l}_1 - \Pi(\dot{l}_1 \mid \dot{\mathscr{P}}_2).$$

Moreover, T_n is efficient (in the sense of having scaled limit distribution the normal distribution so defined) if it is asymptotically linear with efficient influence function

$$\tilde{l} = (\mathbb{E}_0(l_1^* l_1^{*T}))^{-1} l_1^*.$$

The computation of the information bound seems to call for the determination of $\dot{\mathscr{P}}_2$ which is not a trivial task. Often, the exact determination of $\dot{\mathscr{P}}_2$ may be avoided (as well as the computation of $\Pi(\dot{l}_1 \mid \dot{\mathscr{P}}_2)$). For instance, if θ is one-dimensional, one may want to maximize the parametric bounds $(\mathbb{E}_0((l_1 - \Pi(\dot{l}_1|[\dot{l}_2])^2))^{-1}$ over $[\dot{l}_2]$ where $\dot{l}_2$ runs over the score functions for (one-dimensional) parametric submodels of $\mathscr{P}_2$. If the components of $\Pi(\dot{l}_1 \mid \dot{\mathscr{P}}_2)$ can be written as the $L_2^0(\mathbb{P}_0)$ limits of sequences of $\dot{l}_2$'s, then we obtain the information bound in the extended convolution theorem. (See BKRW EAESM p.79-80 for a more general discussion of possible methods.)

If $\dot{l}_1 \perp \dot{\mathscr{P}}_2$, then $l_1^* = \dot{l}_1$, and so it is possible to estimate θ_0 as well as if the nuisance parameter G_0 was known. In this case, an efficient estimator is called adaptive since it adapts to the unknown nuisance parameter (Stein, 1956).

3.3 Parameters defined via functions on $\mathscr{P}$

Consider an arbitrary model $\mathscr{P}$ (dominated by a σ-finite measure μ). We identify $\mathscr{P}$ with the subset $\mathscr{S}$ of the Hilbert space $L_2(\mu)$ through the correspondence $\mathbb{P} \mapsto \sqrt{\frac{d\mathbb{P}}{d\mu}} =: s$. Suppose the parameter of interest is defined via the function $\psi : \mathscr{S} \to \mathbb{R}^d$.

Assume first that $d = 1$. The parameter (function) $\psi : \mathscr{S} \to \mathbb{R}$ is said to be **pathwise differentiable** at $s_0 \in \mathscr{S}$ if there exists a bounded linear functional $\tilde{\psi} : \dot{\mathscr{S}} \to \mathbb{R}$ defined on the tangent space $\dot{\mathscr{S}}$ such that for any one-dimensional submodel $\mathscr{S}^C = \{s_t : t \in C^{-1}(\mathscr{S})\}$ with $C(s_0) = 0$ and tangent vector $\dot{s} \in \dot{\mathscr{S}}$,

$$\psi(s_t) = \psi(s_0) + t\tilde{\psi}(\dot{s}) + o(t) \quad \text{as} \quad t \to 0. \tag{$\dagger$}$$

This means that the real-valued map $\lambda : t \mapsto \psi(s_t)$ is differentiable in the ordinary sense at $t = 0$ and that the derivative $\frac{d\lambda(t)}{dt}\Big|_{t=0} = \tilde{\psi}(\dot{s})$ has a special representation as a continuous linear functional on the tangent space $\dot{\mathscr{S}}$. By the Riesz representation theorem, there exists a unique $\dot{\psi} \in \dot{\mathscr{S}}$ such that

$$\tilde{\psi}(\dot{s}) = \int \dot{\psi} \dot{s} \, d\mu = \langle \dot{\psi}, \dot{s} \rangle$$

for all $\dot{s} \in \dot{\mathscr{S}}$. We should call $\tilde{\psi}$ the **pathwise derivative** of ψ at s_0 and $\dot{\psi}$ the **Riesz representer** of (the pathwise derivative of) ψ at s_0. (Our notations should be dependent on s_0 or $t = 0$, but we do not do it for simplicity.) The notion of pathwise differentiability corresponds to the notion of Hadamard differentiability tangentially to $\dot{\mathscr{S}}$ (see A.5. in BKRW p.453-464).

Remark. The pathwise derivative is a linear functional on the tangent space $\dot{\mathscr{S}}$ but the condition for pathwise differentiability to hold is only to be verified on the generating family $\dot{\mathscr{S}}^0$ (since the defining property is only to be checked for curves $t \mapsto s_t$ passing through s_0 with tangent $\dot{s}$, that is, for $\dot{s}$ in $\dot{\mathscr{S}}^0$). The continuous linear functional derived for $\dot{\mathscr{S}}^0$ then extends (uniquely) to $\text{span}(\dot{\mathscr{S}}^0)$ by linearity and to $\dot{\mathscr{S}}$ by the Hahn–Banach theorem. Moreover, for the application of the Riesz representation theorem, if the representer $\dot{\psi}$ is guessed, it only needs to be checked on $\dot{\mathscr{S}}^0$ (and not on all of $\dot{\mathscr{S}}$, for it then extends uniquely), that is, it must be verified that $\tilde{\psi}(\dot{s}) = \langle \dot{\psi}, \dot{s} \rangle$ for all $\dot{s} \in \dot{\mathscr{S}}^0$. However, it should be clear (and also verified) that $\dot{\psi}$ lies in $\dot{\mathscr{S}}$ (and not necessarily in $\dot{\mathscr{S}}^0$).

Remark. Suppose we define the parameter via a function on $\mathscr{P}$, that is, via a function $\nu \colon \mathscr{P} \to \mathbb{R}^d$ such that $\nu(\mathbb{P}) = \psi(s)$. Define the continuous linear functional $\tilde{\nu} \colon \dot{l} \to \mathbb{R}$ by

$$\tilde{\nu}(\dot{l}) = \tilde{\psi}\left(\frac{s_0}{2}\dot{l}\right).$$

Then (†) holds if and only if

$$\nu(\mathbb{P}_t) = \nu(\mathbb{P}_0) + t\tilde{\nu}(\dot{l}) + o(t) \quad \text{as} \quad t \to 0$$

for any one-dimensional submodel $\mathscr{P}^C$ in $\mathscr{P}$ corresponding to the submodel $\mathscr{S}^C$ where $\dot{l} = 2\frac{\dot{s}}{s_0}$ is the score function of $\mathscr{P}^C$. We should call this condition pathwise differentiability of $\mathscr{P}$ at $\mathbb{P}_0$. By the Riesz representation theorem, there also exists a unique $\dot{\nu} \in \dot{\mathscr{P}}$ such that

$$\tilde{\nu}(\dot{l}) = \langle \dot{\nu}, \dot{l} \rangle = \int \dot{\nu}\dot{l}\, d\mathbb{P}_0$$

for all $\dot{l} \in \dot{\mathscr{P}}$. From $\tilde{\nu}(\dot{l}) = \tilde{\psi}\left(\frac{s_0}{2}\dot{l}\right)$, it directly follows that $\dot{\nu} = \frac{\dot{\psi}}{2s_0}$ and so

$$\langle \dot{\nu}, \dot{\nu} \rangle = \int \dot{\nu}^2\, d\mathbb{P}_0 = \frac{1}{4}\int \dot{\psi}^2\, d\mu = \frac{1}{4}\langle \dot{\psi}, \dot{\psi} \rangle.$$

For reasons that will become clear later, the Riesz representer $\dot{\nu}$ of the pathwise derivative $\tilde{\nu}$ is alternatively called the efficient influence function for $\mathbb{P}_0$ and is denoted $\dot{\nu} := \tilde{l}$.

Assume now that $d \geq 1$. The parameter (function) $\psi = (\psi_1, \ldots, \psi_d) \colon \mathscr{S} \to \mathbb{R}^d$ is said to be **pathwise differentiable** at $s_0 \in \mathscr{S}$ if each real-valued function ψ_i is pathwise differentiable at s_0. The **pathwise derivative** of ψ is simply $\tilde{\psi} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_d)$ and its **Riesz representer** $\dot{\psi} = (\dot{\psi}_1, \ldots, \dot{\psi}_d)$. If we abuse notation and denote $\langle g, h \rangle := \int gh\, d\mu \in \mathbb{R}^d$ and $\langle g, g^T \rangle := \int gg^T\, d\mu \in \mathbb{R}^{d\times d}$ for $g \in (L_2(\mu; \mathbb{R}))^d$ and $h \in L_2(\mu; \mathbb{R})$, then the pathwise derivative and representer can be written as

$$\tilde{\psi}(\dot{s}) = \langle \dot{\psi}, \dot{s} \rangle.$$

Moreover, the relations we derived in Section 3.3 for parameters functions defined on $\mathscr{P}$ extend similarly for $d \geq 1$ with $\nu \colon \mathscr{P} \to \mathbb{R}^d$. In particular, we have with the extended

notations

$$\langle \dot{\nu}, \dot{\nu}^T \rangle = \frac{1}{4} \langle \dot{\psi}, \dot{\psi}^T \rangle.$$

In light of the definition of pathwise differentiability, there should exist a close correspondence between pathwise derivative and information bounds. To see this, suppose for simplicity that ν is real-valued, i.e., $d = 1$. By Section 2 with $q = (t \mapsto \nu(\mathbb{P}_t))$, we know that the information bound for a submodel $\{\mathbb{P}_t : t \in C^{-1}(\mathscr{S})\}$ with score function $\dot{l} \in \dot{\mathscr{P}}$ at $t = 0$ is given by

$$\left(\frac{dq(t)}{dt}\bigg|_{t=0} \right)^2 (E(\dot{l}^2))^{-1} = \frac{\langle \dot{\nu}, \dot{l} \rangle^2}{\langle \dot{l}, \dot{l} \rangle^2}.$$

By taking an infimum over all one-dimensional submodels (or equivalently over all elements $\dot{l}$ of the tangent space $\dot{\mathscr{P}}$), we have a lower bound for estimating $\nu(\mathbb{P}_0)$. By Cauchy—Schwarz's inequality and the fact that $\dot{\nu} \in \dot{\mathscr{P}}$, we have that

$$\sup_{\dot{l} \in \dot{\mathscr{P}}} \frac{\langle \dot{\nu}, \dot{l} \rangle^2}{\langle \dot{l}, \dot{l} \rangle^2} = \langle \dot{\nu}, \dot{\nu} \rangle.$$

Thus $\langle \dot{\nu}, \dot{\nu} \rangle$ plays the role of the information bound. This result can be generalized formally to $d > 1$. We then obtain an extension of the Hajek–Le Cam convolution theorem (see T.25.20. in VdV AS p.366 or T.3.2.2. in BKRW p.63). If $\nu \colon \mathscr{P} \to \mathbb{R}^d$ is pathwise differentiable at $\mathbb{P}_0$ with Riesz representer $\dot{\psi}$, $\overline{\dot{\mathscr{P}}^0} = \dot{\mathscr{P}}$, and T_n is an estimator of $\nu(\mathbb{P}_0)$ locally regular at $\mathbb{P}_0$ for any parametric submodel, then the limit distribution of $\sqrt{n}(T_n - \nu(\mathbb{P}_0))$ is the convolution of some distribution and the normal distribution with mean 0 and covariance matrix (**information bound**)

$$\langle \dot{\nu}, \dot{\nu}^T \rangle = \int \dot{\nu} \dot{\nu}^T \, d\mathbb{P}_0.$$

Moreover, T_n is efficient (in the sense of having scaled limit distribution the normal distribution so defined) if it is asymptotically linear in $\dot{\nu}$.

The result can then be applied to semiparametric models where the parameters are defined implicitly; in which case, we naturally find the correspondence $\dot{\nu} = \tilde{l}$ where $\tilde{l}$ is the efficient influence function derived in the previous section.

Example 7 (Parametric Models (see BKRW p.60-61))**.** Let $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ be a μ-dominated model with densities p_θ that is QMD at $\mathbb{P}_\theta$ with score $\dot{l}_\theta$. Then $\dot{\mathscr{P}} = \{h^T \dot{l}_\theta : h \in \mathbb{R}^d\}$. Define $\nu \colon \mathscr{P} \to \mathbb{R}^k$ the parameter of interest and consider the map $q \colon \mathbb{R}^d \to \mathbb{R}^k$ given by $q(\theta) = \nu(\mathbb{P}_\theta)$. Then ν is pathwise differentiable if q is differentiable in the ordinary sense. Indeed, the submodel $\{\mathbb{P}_{\theta + th} : t \in B_1(0; \mathbb{R})\}$ has score $h^T \dot{l}_\theta$. Since by definition $\langle I^{-1}(\theta) \dot{l}_\theta, \dot{l}_\theta^T \rangle$ is the identity matrix, we have

$$\begin{aligned} \nu(\mathbb{P}_{\theta + th}) &= q(\theta + th) \\ &= q(\theta) + t \dot{q}(\theta) h + o(t) \\ &= \nu(\theta) + t \langle \dot{q}(\theta) I^{-1}(\theta) \dot{l}_\theta, \dot{l}_\theta^T h \rangle + o(t). \end{aligned}$$

By identification, we find that

$$\dot{\nu} = \dot{q}(\theta) I^{-1}(\theta) \dot{l}_\theta,$$

which is nothing more than the efficient influence function for the estimation of $\nu(\mathbb{P}_\theta) = q(\theta)$. The standard information bound can then be naturally expressed as

$$\langle \dot{\nu}, \dot{\nu}^T \rangle = \dot{q}(\theta) I^{-1}(\theta) \dot{q}(\theta)^T.$$

Another method to derive bounds in semiparametric models with this approach consists in embedding $\mathscr{P}$ in a larger model $\mathscr{P}_e$. Consider ν_e an extension of ν on $\mathscr{P}_e$. In general, $\dot{\nu}_e$ is easy to compute when (nonparametric) $\mathscr{P}_e$ is sufficiently large. Then pathwise differentiability for both ν and ν_e yields $\dot{\nu}_e - \dot{\nu} \perp \dot{\mathscr{P}}$, that is,

$$\dot{\nu} = \Pi(\dot{\psi}_e \mid \dot{\mathscr{P}}).$$

Viewed as an element of $L_2(\mathbb{P}_0)$, $\dot{\nu}_e$ is said to a gradient of ν and $\dot{\nu}$ the canonical gradient of ν. This gives another method to derive information bounds.

3.4 Examples: some efficiency bounds

There are two methods to derive bounds:

1. efficient influence function (with computation of projectors or not)

2. pathwise derivative: 2.a. find dual norm of pathwise derivative; or 2.b. find Riesz representer (by guessing).

To prove pathwise differentiability, there are two methods:

A. guess representer and verify definition of pathwise differentiability;

B. a. assume pathwise differentiability and compute $\tilde{\psi}$ by Leibniz's rule; b. verify that $\tilde{\psi}$ is a bounded linear map on the tangent space; c. guess the representer $\dot{\phi}$; d. verify that $\dot{\phi}$ lies in the tangent space.

3.4.1 Mean

We find the efficiency bound for mean estimation by guessing the Riesz representer and verifying that the definition of pathwise differentiability holds (method 2.b.A).

Let $\mathscr{P}$ be a model dominated by a measure μ which concentrates on $[-M, M]$. The parameter of interest is

$$\nu(\mathbb{P}) = \int x \, d\mathbb{P}(x).$$

Take $s = \sqrt{d\mathbb{P}/d\mu}$ and define

$$\nu(\mathbb{P}) = \psi(s) = \int x s^2(x) \, d\mu(x).$$

Fix s_0. We claim that ψ is pathwise differentiable at s_0 with derivative

$$\dot{\psi}(s_0)(x) = 2 s_0(x)(x - \mathbb{E}_0(X)) \quad \text{a.e. } \mu.$$

We first verify that $\dot{\psi} \in \dot{\mathscr{P}}$ using the characterization of $\dot{\mathscr{P}}$ derived in Example 2.4.: indeed, we have that $\int \dot{\psi} s_0 \, d\mu = 2 \int s_0^2(x)(x - \mathbb{E}_0(X)) \, d\mu(x) = 0$. Then we verify that the definition of pathwise differentiability holds for $\dot{\psi}$. Since $\int s^2(x) \, d\mu(x) =$

$\int s_0^2(x)\,d\mu(x) = 1$, we have

$$\begin{aligned}\psi(s) - \psi(s_0) - \langle \dot{\psi}, s - s_0\rangle &= \int (x - \mathbb{E}_0(X))(s - s_0)^2(x)\,d\mu(x) \\ &= O(\|s - s_0\|^2),\end{aligned}$$

where the second inequality follows from the fact that $\int (x - \mathbb{E}_0(X))(s - s_0)^2(x)\,d\mu(x) \leq c\int (s - s_0)^2(x)\,d\mu(x)$ with $c = M - \mathbb{E}_0(X)$. Therefore,

$$\lim_{\|s-s_0\|\to 0} \frac{\psi(s) - \psi(s_0)}{\|s - s_0\|} = \lim_{\|s-s_0\|\to 0} \left\langle \dot{\psi}, \frac{s - s_0}{\|s - s_0\|} \right\rangle + C\|s - s_0\| = \langle \dot{\psi}, \dot{s}\rangle.$$

From this, it follows that the efficient influence function is

$$\dot{\nu}(x) = x - \mathbb{E}_0(X)$$

and the information bound for mean estimation is

$$\langle \dot{\nu}, \dot{\nu}\rangle = \mathrm{Var}_0(X).$$

Since the asymptotic variance of the sample mean $\overline{X}$ is $\mathrm{Var}_0(X)$, it follows that the sample mean is an efficient estimator for the mean.

Reference. E.3.3.2. in BKRW EAESM p.67-68. See also S.3. in SEBMM Severini&Tripathi p.191 where the guess for the representer is more natural as it is derived from the computation of the linear functional $\tilde{\psi}$ by Leibniz's rule. Moreover, the method of Severini&Tripathi to prove pathwise differentiability is generally simpler than using the definition as above: 1. assume pathwise differentiability and compute $\tilde{\psi}$ by Leibniz's rule; 2. verify that $\tilde{\psi}$ is a bounded linear map on the tangent space; 3. guess the representer $\dot{\phi}$; 4. verify that $\dot{\phi}$ lies in the tangent space.

3.4.2 Distribution function

We find the efficiency bound for estimating a distribution function method 2.b.B.

Let $\mathscr{P}$ be the set of Borel probability measures on $\mathbb{R}$ dominated by the Lebesgue measure μ. Consider the correspondence between $\mathscr{P}$ and $\mathscr{S}$ by the standard mapping $\mathbb{P} \mapsto s = \sqrt{d\mathbb{P}/d\mu}$. Assume $\mathbb{P}_0$ is the true distribution. Define for any $\mathbb{P} \in \mathscr{P}$ and any $\xi \in \mathbb{R}$, the parameter function

$$\psi(s)(\xi) = \nu(\mathbb{P})(\xi) = \mathbb{P}((-\infty, \xi]) = \int_{\mathbb{R}} \mathbb{1}_{(-\infty,\xi]}(x)s^2(x)\,dx.$$

We start by characterizing the tangent spaces $\dot{\mathscr{P}}$ and $\dot{\mathscr{S}}$. Since $\mathscr{P}$ is the set of all Borel probability measures dominated by μ, we have by E.2.4. that $\dot{\mathscr{P}} = L_2^0(\mu)$.

Fix $\xi \in \mathbb{R}$. Let s_t be an element of a one-dimensional curve of $\mathscr{S}$ passing through s_0. Assume $t \mapsto \psi(s_t)(\xi)(=: F_t(\xi))$ can be differentiated at $t = 0$ and denote $\tilde{\psi}$ the derivative. By Leibniz's rule, we have

$$\tilde{\psi}(\dot{s}) = 2\int_{\mathbb{R}} \mathbb{1}_{(-\infty,\xi]}(x)s_0(x)\dot{s}(x)\,dx$$

where $\dot{s} \in \dot{\mathscr{S}}^0$ is the derivative of $t \mapsto s_t$ at $t = 0$. Since $\int_{\mathbb{R}} s_0(x)\dot{s}(x)\,dx = 0$, we can rewrite

$$\tilde{\psi}(\dot{s}) = 2\int_{\mathbb{R}} \mathbb{1}_{(-\infty,\xi]}(x)s_0(x)\dot{s}(x)\,dx - 2F_0(\xi)\int_{\mathbb{R}} s_0(x)\dot{s}(x)\,dx.$$

The map $\tilde{\psi}$ is a bounded linear map on $\dot{\mathscr{S}}^0$. Then we guess that the Riesz representer is

$$\dot{\psi}(x) = 2\Big(\mathbb{1}_{(-\infty,\xi]}(x) - F_0(\xi)\Big)s_0(x).$$

We verify our guess by making sure that $\dot{\psi}$ is a representer by verifying that for all $\dot{s} \in \dot{\mathscr{S}}^0$,

$$\langle \dot{\psi}, \dot{s} \rangle = \int \dot{\psi}(x)\dot{s}(x)\,dx = \tilde{\psi}(\dot{s})$$

and by verifying that $\dot{\psi} \in \dot{\mathscr{S}}$ through the characterization

$$\langle \dot{\psi}, s_0 \rangle = \int_{\mathbb{R}} \dot{\psi}(x)s_0(x)\,dx = 2(F_0(\xi) - F_0(\xi)) = 0.$$

Then the efficient influence function for the estimation of $F_0(\xi)$ is

$$\dot{\nu}(x) = \mathbb{1}_{(-\infty,\xi]}(x) - F_0(\xi)$$

and so the efficiency bound is

$$\frac{1}{4}\langle \dot{\psi}, \dot{\psi} \rangle = \langle \dot{\nu}, \dot{\nu} \rangle = \int \Big(\mathbb{1}_{(-\infty,\xi]}(x) - F_0(\xi)\Big)^2 s_0^2(x)\,dx = F_0(\xi)(1 - F_0(\xi)).$$

Since the empirical distribution function $\mathbb{F}_n(\xi) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,\xi]}(X_i)$ at ξ has asymptotic variance $F_0(\xi)(1 - F_0(\xi))$, it is an efficient estimator for $F_0(\xi)$.

Reference. E.5.3.1. in BKRW EAESM p.191-192 or S.3. in SEBMM Severini&Tripathi p.196-197.

A Dictionary between notations: ours to BKRW EAESM to SEBMM Severini&Tripathi

We state the correspondences in the order of the title:

$$
\begin{array}{rcccccc}
\text{pdf}^{1/2}:\ \sqrt{p_\theta} = \sqrt{\frac{d\mathbb{P}_\theta}{d\mu}} & =: & s(\theta) & \leftrightarrow & s(\theta) & \leftrightarrow & \phi(\theta) \\
\text{Score}:\ \frac{\partial \log L_\theta}{\partial \theta} = 2\frac{\dot{s}(\theta)}{s(\theta)}\mathbb{1}_{[s(\theta)>0]} & =: & \dot{l}(\theta) & \leftrightarrow & \dot{l}(\theta) & \leftrightarrow & U = \dot{S}_{\dot{\phi}} \\
\text{Parameter via } \mathscr{S}:\ (\mapsto: \mathscr{S} \to \mathbb{R}^d) & =: & \psi(s) & \leftrightarrow & \nu(s) & \leftrightarrow & \rho(s) \\
\text{Pathwise derivative}:\ (\mapsto: \dot{\mathscr{S}} \to \mathbb{R}^d) & =: & \tilde{\psi}(\dot{s}) = \langle \dot{\psi}, \dot{s} \rangle & \leftrightarrow & \dot{\nu}(\dot{s}) & \leftrightarrow & \nabla\rho(\dot{\phi}) \\
\text{Riesz representer}:\ (\in \dot{\mathscr{S}}) & =: & \dot{\psi} & \leftrightarrow & \dot{\nu} & \leftrightarrow & 4\phi* \\
(L_2(\mu) \supseteq \dot{\mathscr{S}}) - \text{inner product}:\ (\dot{\mathscr{S}} \times \dot{\mathscr{S}} \to \mathbb{R}) & =: & \langle \cdot, \cdot \rangle & \leftrightarrow & \_ & \leftrightarrow & \frac{1}{4}\langle \cdot, \cdot \rangle_F \\
\text{Parameter via } \mathscr{P}:\ (\mapsto: \mathscr{P} \to \mathbb{R}^d) & =: & \nu(\mathbb{P}) & \leftrightarrow & \nu(\mathbb{P}) & \leftrightarrow & \_
\end{array}
$$

References

Bickel, P. J., C. Klaassen, Y. Ritov, and J. A. Wellner (2005): "Semiparametric inference and models," *Mimeo, University of California, Berkeley*.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4. Springer.

Groeneboom, P., and J. A. Wellner (1992): *Information bounds and nonparametric maximum likelihood estimation*, vol. 19. Springer.

Ibragimov, I. A., and R. Z. Has' Minskii (1981): *Statistical estimation: asymptotic theory*. Springer.

Pollard, D. (2001): "Lecture in Paris on Le Cam's theory," *Lecture notes*.

Severini, T. A., G. Tripathi, et al. (2013): "Semiparametric efficiency bounds for microeconometric models: A survey," *Foundations and Trends® in Econometrics*, 6(3–4), 163–397.

Strasser, H. (1985): *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*, vol. 7. Walter de Gruyter.

Torgersen, E. (1991): *Comparison of statistical experiments*, vol. 36. Cambridge University Press.

van der Vaart, A., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.

Van der Vaart, A. W. (1998): *Asymptotic statistics*. Cambridge university press.

WELLNER, J. A. (2005): "Advanced theory of statistical inference," *Lecture notes for STAT580 at UWash*.