Intro to Semiparametrics

Eric Slud

We define here basic ideas and notations from van der Vaart, *Asymptotic Statistics* (1998), Ch. 25, appealing also to the monograph *Efficient and Adaptive Estimation for Semiparametric Models* (1993), by Bickel et al.

Definitions.

- **Statistical model:** a family $\mathcal{P} = \{P_\vartheta\}_{\vartheta \in \Theta}$ of probability measures on some data space $\mathcal{X}$.
  (We usually take $\mathcal{X} = \mathcal{X}_n$ a product space with $n$ identical factors, $n$ called the *sample size*, and all $P_\vartheta$ iid product measures.)

- **Parameter** (finite-dimensional): a mapping $\psi : \mathcal{P} \to \mathbb{R}^k$ (below, will be assumed ‘smooth’).

- **Structural & Nuisance parameters:** if $\vartheta$ is in 1-to-1 correspondence with $(\beta, \eta)$, $\beta \in \mathbb{R}^k$, $\eta \in L$, with $\beta = \psi(P_\vartheta)$ a parameter vector of primary interest, then $\beta$ is called *structural* and $\eta$ *nuisance*. If $\vartheta$ and $L$ are infinite-dimensional, then the problem of estimating $\beta$ is called *semiparametric*.
• A smooth parametric submodel is a 1-dimensional curve $t \mapsto P_t$ mapping $[0, 1)$ to $\mathcal{P}$, smooth in the sense of quadratic mean differentiability: putting $P \equiv P_0$, there exists $g \in L_2(P)$ such that, for any measures $Q_t$ such that $P, P_t \ll Q_t$, as $t \to 0^+$

$$\int \left\{ \frac{1}{t} \left[ \left( \frac{dP_t}{dQ_t} \right)^{1/2} - \left( \frac{dP}{dQ_t} \right)^{1/2} \right] - \frac{g}{2} \left( \frac{dP}{dQ_t} \right)^{1/2} \right\}^2 dQ_t \to 0$$

$g$ is called the score function for $\{P_t\}_{t \geq 0}$.

• Tangent space $\dot{\mathcal{P}}_P$ is the set (necessarily a cone) of all score functions $g$ for smooth submodels.

• Smooth (at $P$) parameter mapping:

$\psi: \mathcal{P} \to \mathbb{R}^k$ satisfies: $\exists$ operator $\dot{\psi}_P$ on $\mathcal{P}_P$ so that

$\forall$ smooth submodel $\{P_t\}_{t \geq 0}$ with score $g$:

$$t^{-1}(\psi(P_t) - \psi(P)) \to \dot{\psi}_P g \quad \text{as} \quad t \to 0^+$$

Assume from now on (what must be checked in examples) that $\dot{\mathcal{P}}_P$ is a linear space and $\dot{\psi}_P$ a bounded linear operator on it, i.e. $\| \dot{\psi}_P g \| \leq C \| g \|_{L_2(P)}$ for a finite constant $C$, for all $g \in \dot{\mathcal{P}}_P$.

Then by Riesz Representation Theorem applied to the Hilbert space $\text{closure}(\dot{\mathcal{P}}_P) \subset L_2(P)$, $\exists$ unique element $\tilde{\psi}_P \in \text{closure}(\dot{\mathcal{P}}_P)$ called the efficient influence function such that

$$\dot{\psi}_P g = \int \tilde{\psi}_P g dP$$
**Remark 1** If the submodel family $P_t$ is absolutely continuous with respect to a fixed prob. measure $Q$ on $\mathcal{X}$, with densities $p(t,x)$ such that $(\partial/\partial t)p(0+,x)$ exists a.s. and

$$\exists \epsilon > 0, \sup_{t \in [0,\epsilon]} t^{-1} |p(t,x) - p(0,x)| \in L_2(P)$$

then the score function $g = (\partial/\partial t)p(0+,x)$.

**Definitions**, continued.

- **Fisher Information** for $t$ as (the only) unknown parameter in submodel $\{P_t\}$ is $\|g\|_{L_2(P)}^2 = \int g^2 dP$.

**Remark 2** If $a \in \mathbb{R}^k$ is arbitrary, and densities $p(t,x)$ are smooth with log-derivative dominated as in Remark 1, the Cramer-Rao lower variance bound for $a' \psi(P_t)$ within the submodel is

$$(a'(\psi_P g))^2 / \int g^2 dP = \left\{ \int (a'\tilde{\psi}_P) g dP \right\}^2 / \int g^2 dP$$

Taking sup via Cauchy-Schwarz over all $g \in \text{closure}(\mathcal{P}_P)$ is achieved when $g = a'\tilde{\psi}_P$, and gives $a' (\mathcal{I}(\beta))^{-1} a$ for $\beta \equiv \psi(P)$, $P \in \mathcal{P}$, where:

- **Semiparametric information bound** for $\beta$ is

$$\mathcal{I}(\beta) = \left\{ \int \left(\tilde{\psi}_P\right)^2 dP \right\}^{-1}$$
Next some definitions related to statistics and estimators.

- An estimator $T \equiv T(X)$ of $\psi(P) \in \mathbb{R}^k$ is a measurable function on $\mathcal{X}$.

- Estimator sequences $T_n = T_n(X)$ on $\mathcal{X}_n$ (data of sample size $n$, with $\psi(P)$ depending only on 1st factor of product-measure $P$) are **semiparametric consistent** if:

$$\forall P \in \mathcal{P}, \ T_n \rightarrow \psi(P) \text{ a.s. or in prob.}$$

- Estimators $T_n = T_n(X)$ are called **regular at $P$** if there exists a probability law $\mathcal{L}$ (on $\mathbb{R}^k$) s.t. $\forall$ submodel $\{P_t\}$ (with score $g \in \mathcal{P}_P$), as $n \rightarrow \infty$

$$\sqrt{n}(T_n - \psi(P_{1/\sqrt{n}})) \overset{D}{\rightarrow} \mathcal{L} \text{ under } P_{1/\sqrt{n}}$$
Results on Information Bounds & Estimators

Suppose \( \vartheta = (\beta, \eta) \), \( \beta \in \mathbb{R}^k \), \( \eta \in L \), \( \psi(P_{\beta,\eta}) \equiv \beta \). Generally, e.g. if logs of densities \( p(x, \beta, \eta) \) are smooth and moment-bounded w.r.t. \( \beta, \eta \) arguments, then the assumed-linear tangent space has the form

\[
\dot{P}_P = \beta \dot{\mathcal{P}}_P \oplus \eta \dot{\mathcal{P}}_P \equiv \{ g_1 + g_2 : g_1 = \text{score for } P_{\beta_t,\eta_0}, g_2 = \text{score for } P_{\beta_0,\eta_t} \}
\]

Define \( P = P_{\beta_0,\eta_0} \) and \( \dot{l}_{\beta} = \nabla_{\beta} \log p(x, \beta_0, \eta_0) \)

For submodel \( P_{\beta_0,\eta_t} \) with score \( g_2 \):

\[
\dot{\psi}_P g_2 \equiv 0 \Rightarrow \dot{\psi}_P \perp \eta \dot{\mathcal{P}}_P \Rightarrow \dot{\psi}_P \in \beta \dot{\mathcal{P}}_P \oplus \eta \dot{\mathcal{P}}_P
\]

For \( a \in \mathbb{R}^k \), submodel \( P_{\beta_0 + at,\eta_0} \) : score \( g_1 = a' \dot{l}_{\beta} \),

\[
\dot{\psi}_P g_1 = a \Rightarrow \int \dot{\psi}_P g_1 dP = \int \dot{\psi}_P (g_1 - \Pi_\eta g_1) dP = a
\]

where \( \Pi_\eta = \text{orthog. proj. onto } \eta \dot{\mathcal{P}}_P \). Put \( \tilde{l}_{\beta} \equiv \dot{l}_{\beta} - \Pi_\eta \dot{l}_{\beta} \).

\(
\forall a, \int \tilde{\psi}_P \tilde{l}_{\beta}^t a dP \equiv a \Rightarrow \tilde{\psi}_P = (\int \tilde{l}_{\beta} \tilde{l}_{\beta}^t dP)^{-1} \tilde{l}_{\beta}
\)

\( \tilde{l}_{\beta} \) is called efficient \textbf{score function} for \( \beta \).

\textbf{Note:} \( \mathcal{I}(\beta) = (\int \tilde{\psi}_P \tilde{\psi}_P^t dP)^{-1} = \int \tilde{l}_{\beta} \tilde{l}_{\beta}^t dP \)
Consequence of Projection Characterization of $\tilde{\psi}_P$

Suppose now that $\beta$ is 1-dimensional and that $\vartheta = (\beta, \eta)$ lies in a fixed finite-dimensional parameter set $\eta = (\lambda, \rho_0)$ where $\lambda \in \mathbb{R}^q$, $q < \infty$, and $\rho$ would in general be infinite-dimensional but is assumed known $= \rho_0$, and moreover that all components of $\tilde{\psi}_P$ lie in the span of $\hat{l}_\beta$ and of the components of $\hat{l}_\lambda \equiv \nabla_\lambda \log p(x, \beta_0, \lambda_0, \rho_0)$.

Then, since $\eta \hat{P}_P$ is exactly the subspace of $\hat{P}_P$ orthogonal to the single element $\tilde{\psi}_P$, and since $\hat{l}_\beta - \tilde{\psi}_P \perp \tilde{\psi}_P$, it follows that $\hat{l}_\beta - \tilde{\psi}_P \in \eta \hat{P}_P$. Since the components of $\hat{l}_\lambda$ lie in $\eta \hat{P}_P$, and since we have assumed $\tilde{\psi}_P \in \text{span}\{\hat{l}_\beta, \hat{l}_\lambda\}$, we conclude that for some $c \in \mathbb{R}^q$, $\tilde{\psi}_P = \hat{l}_\beta - c' \hat{l}_\lambda$. (A little further work shows that $c$ is uniquely determined as $(\int \hat{l}_\lambda \hat{l}_\lambda^r dP)^{-1} \int \hat{l}_\lambda \hat{l}_\beta dP$.)

Within the finite-dimensional model $(\beta, \lambda)$ reparameterized as $(\beta^*, \lambda) \equiv (\beta - c'(\lambda - \lambda_0), \lambda)$, it is easy to calculate that the information matrix is block-diagonal with upper-left element $\int \hat{l}_\beta \hat{l}_\beta^r dP$ and lower-right $q \times q$ block $\int \hat{l}_\lambda \hat{l}_\lambda^r dP$ and therefore that the asymptotic variance for ML estimators of $\beta$ is $(\mathcal{I}(\beta))^{-1} = \int \tilde{\psi}_P^2 dP$

Thus within finite-dimensional models of arbitrarily large but finite nuisance-parameter
dimension whose scores space $\tilde{\psi}_P$, the optimal asymptotic variance is the same $\int \tilde{\psi}^2_P dP$ whether the nuisance parameters are unknown as when they are known! (We already saw this same asymptotic variance from the Cramer-Rao bound in any 1-parameter submodel with score $\tilde{\psi}_P$.)

A deep Theorem provides a clearer view of $\mathcal{I}$ as the \textbf{semiparametric information bound} for estimates of $\psi(P) = \beta$ in the semiparametric setting $\mathcal{P} = \{P(\beta, \eta)\}$.

\textbf{Hajek Convolution Theorem}

The (generalized) \textit{Hajek Convolution Theorem}, van der Vaart p. 366, says when $\hat{\mathcal{P}}_P$ is linear space: every limit distribution $\mathcal{L}$ of regular estimator seq. $T_n$ is $\mathcal{N}(0, \int \tilde{\psi}_P \tilde{\psi}^*_P dP) * \mathcal{M}$ for some prob. law $\mathcal{M}$. 
Inverse Operators & Hilbert Nuisance Parameter

We continue with \( \vartheta = (\beta, \eta), \beta \in \mathbb{R}^k, \eta \in L, \) and now assume \( L \) a Hilbert space (or restrict attention to a neighborhood of nuisance parameters \( \eta_0 + tv, \ t \geq 0, \ v \in L \)). Define a nuisance score mapping

\[
s : L \rightarrow \dot{P}_P , \quad s(v) \equiv \left. \frac{\partial}{\partial t} \log p(x, \beta_0, \eta_0 + tv) \right|_{t=0}
\]

(or could replace \( tv \) in some problems by \( \kappa(\eta_0, t, v) \) with \( \kappa(\eta, 0) \equiv 0 \) and second partial \( \kappa_2(\eta_0, 0, v) \equiv v \)).

Assume that the covariance operator \( C : L \times L \rightarrow \mathbb{R} \) given by

\[
v'Cv = \int s(v) s(w) dP
\]

is a bounded nonsingular bilinear form, in which case the Riesz Representation Theorem implies that \( C : L \rightarrow L \) is a bounded (i.e., continuous) linear operator. Nonsingularity says that \( (\partial/\partial t) \log p(x, \beta_0, \eta_0 + tv)|_{t=0} \not\equiv 0 \) (which implies \( Cv \neq 0 \)) for \( v \neq 0 \), in which case \( C^{-1} \) exists as a mapping on \( L \).
Also define $B : L \rightarrow \mathbb{R}^k$ (where $k = \text{dim}(\beta)$) by

$$Bv = \int \dot{l}_\beta s(v) dP, \quad \forall \ v \in L$$

(Recall that both $\dot{l}_\beta$, $s(v)$ are measurable ($L_2$) real-valued functions on $\mathcal{X}$.) Cauchy-Schwarz implies $B$ is a bounded operator, and $B^* : \mathbb{R}^k \rightarrow L$ satisfies

$$\langle v, B^*a \rangle_L = a' Bv = \int a' \dot{l}_\beta s(v) dP$$

**Assume further** that $C^{-1}B^* : \mathbb{R}^k \rightarrow L$ is bounded. Then we check that $a'\dot{l}_\beta - s(C^{-1}B^* a) \perp \eta \tilde{\mathcal{P}}_P$, since

$$\int (a'\dot{l}_\beta - s(C^{-1}B^* a)) s(v) dP$$

$$= a' Bv - (C^{-1}B^* a)^{tr} Cv = 0$$

It follows in these circumstances that

$$\forall \ a \in \mathbb{R}^k, \ a'\tilde{l}_\beta = a'\dot{l}_\beta - s(C^{-1}B^* a)$$

and the semiparametric information bound is given by

$$a'\mathcal{I}(\beta)a = \int (a'\tilde{l}_\beta)^2 dP = \int (a'\dot{l}_\beta)^2 dP - \int (s(C^{-1}B^* a))^2 dP$$

or: $\mathcal{I}(\beta) = \int (\dot{l}_\beta)^2 dP - BC^{-1}B^*$ as in fin-dim case!