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_ϑ *iid* product measures.)
- **Parameter** (finite-dimensional): a mapping $\psi : \mathcal{P} \to \mathbf{R}^k$ (below, will be assumed 'smooth').
- **Structural & Nuisance** parameters: if ϑ is in 1-to-1 correspondence with (β, η), $\beta \in \mathbf{R}^k$, $\eta \in L$, with $\beta = \psi(P_\vartheta)$ a parameter vector of primary interest, then β is called *structural* and η *nuisance*. If ϑ and L are infinite-dimensional, then the problem of estimating β 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 differentiablility*: 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 $t \to 0+$
$$\int \Big\{ \frac{1}{t} \Big[(\frac{dP_t}{dQ_t})^{1/2} - (\frac{dP}{dQ_t})^{1/2} \Big] - \frac{g}{2} (\frac{dP}{dQ_t})^{1/2} \Big\}^2 dQ_t \ \to \ 0$$
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 :** $\psi : \mathcal{P} \to \mathbf{R}^k$ satisfies: $\exists$ operator $\dot{\psi}_P$ on $\dot{\mathcal{P}}_P$ so that $\forall$ smooth submodel $\{P_t\}_{t \geq 0}$ with score g :
$$t^{-1} (\psi(P_t) - \psi(P)) \ \to \ \dot{\psi}_P \, g \qquad \text{as} \quad t \to 0+$$

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 $closure(\dot{\mathcal{P}}_P) \subset L_2(P)$, $\exists$ unique element $\tilde{\psi}_P \in closure(\dot{\mathcal{P}}_P)$ 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*

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

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 $\|g\|^2_{2,P} \;=\; \int g^2\, dP$.

Remark 2 *If* $a \in \mathbf{R}^k$ *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*

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

Taking sup via Cauchy-Schwarz over all $g \in closure(\dot{\mathcal{P}}_P)$ *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 β 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 $\psi(P) \in \mathbf{R}^k$ 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:

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

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

$$\sqrt{n}\,(T_n - \psi(P_{1/\sqrt{n}})) \xrightarrow{\mathcal{D}} \mathcal{L} \quad \text{under} \quad P_{1/\sqrt{n}}$$

Results on Information Bounds & Estimators

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

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

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

For submodel P_{β_0,η_t} with score g_2 :

$$\dot{\psi}_P\, g_2 \equiv 0 \quad \Rightarrow \quad \tilde{\psi}_P \perp {}_\eta\dot{\mathcal{P}}_P \quad \Rightarrow \quad \tilde{\psi}_P \in {}_\beta\dot{\mathcal{P}}_P \ominus {}_\eta\dot{\mathcal{P}}_P$$

For $a \in \mathbf{R}^k$, submodel P_{β_0+at,η_0} : score $g_1 = a'\,\dot{l}_\beta$,

$$\dot{\psi}_P\, g_1 = a \Rightarrow \int \tilde{\psi}_P\, g_1\, dP = \int \tilde{\psi}_P\, (g_1 - \Pi_\eta g_1)\, dP = a$$

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

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

$\tilde{l}_\beta$ is called **efficient score function** for β.

Note: $\mathcal{I}(\beta) = \Big(\int \tilde{\psi}_P\, \tilde{\psi}_P^{tr}\, dP\Big)^{-1} = \int \tilde{l}_\beta \tilde{l}_\beta^{tr}\, dP$

Consequence of Projection Characterization of $\tilde{\psi}_P$

Suppose now that β is 1-dimensional and that $\vartheta = (\beta, \eta)$ lies in a fixed finite-dimensional parameter set $\eta = (\lambda, \rho_0)$ where $\lambda \in \mathbf{R}^q$, $q < \infty$, and ρ 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 $\dot{l}_\beta$ and of the components of $\dot{l}_\lambda \equiv \nabla_\lambda \log p(x, \beta_0, \lambda_0, \rho_0)$.

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

Within the finite-dimensional model (β, λ) 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 $\int \tilde{l}_\beta \tilde{l}_\beta^{tr} dP$ and lower-right $q \times q$ block $\int \tilde{l}_\lambda \tilde{l}_\lambda^{tr} dP$ and therefore that the asymptotic variance for ML estimators of β 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 $\int \tilde{\psi}_P^2 dP$ whether the nuisance parameters are unkown 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 **semiparametric information bound** for estimates of $\psi(P) = \beta$ in the semiparametric semtting $\mathcal{P} = \{P_{(\beta,\eta)}\}$.

Hajek Convolution Theorem

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

Inverse Operators & Hilbert Nuisance Parameter

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

$$s : L \to_\eta \dot{P}_P \quad , \qquad s(v) \equiv \frac{\partial}{\partial t} \log p(x, \beta_0, \eta_0 + tv) \Big|_{t=0}$$

(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 $C : L \times L \to \mathbf{R}$ given by

$$v' C v = \int s(v)\, s(w)\, dP$$

is a bounded nonsingular bilinear form, in which case the Riesz Representation Theorem implies that $C : L \to L$ 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 $B : L \to \mathbf{R}^k$ (where $k = \dim(\beta)$) by

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

(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 $B^* : \mathbf{R}^k \to L$ satisfies

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

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

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

$$= \; a' B v \, - \, (C^{-1} B^* \, a)^{tr} \, C \, v \; = \; 0$$

It follows in these circumstances that

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

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