Introduction to Empirical Processes and Semiparametric Inference Lecture 22: Semiparametric Models and Efficiency

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Tangent Sets

For a statistical model $\{P \in \mathcal{P}\}$ on a sample space $\mathcal{X}$, a one-dimensional model $\{P_t\}$ is a *smooth submodel at* P if $P_0 = P$,

$$\{P_t : t \in N_\epsilon \equiv (-\epsilon, \epsilon)\} \subset \mathcal{P}$$

for some $\epsilon > 0$, and the following holds for some measurable "tangent" function $g : \mathcal{X} \mapsto \mathbb{R}$:

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

as $t \to 0$.

In the previous slide, P is usually the true but unknown distribution of the data.

Note that Lemma 11.11 forces the g in (1) to be contained in $L_2^0(P)$.

A tangent set $\dot{\mathcal{Q}}_P \subset L_2^0(P)$ *represents* a submodel $\mathcal{Q} \subset \mathcal{P}$ at P if the following hold:

(i) For every smooth one-dimensional submodel $\{P_t\}$ for which

$$P_0 = P \quad \text{and} \quad \{P_t : t \in N_\epsilon\} \subset \mathcal{Q} \text{ for some } \epsilon > 0, \qquad (2)$$

and for which (1) holds for some $g \in L_2^0(P)$, we have $g \in \dot{\mathcal{Q}}_P$; and

(ii) For every $g \in \dot{\mathcal{Q}}_P$, there exists a smooth one-dimensional submodel $\{P_t\}$ such that (2) and (1) both hold.

An appropriate question to ask at this point is why the focus on one-dimensional submodels?

The basic reason is that score functions for finite dimensional submodels can be represented by tangent sets corresponding to one-dimensional submodels.

To see this, let

$$\mathcal{Q} \equiv \{P_\theta : \theta \in \Theta\} \subset \mathcal{P},$$

where $\Theta \subset \mathbb{R}^k$.

Let $\theta_0 \in \Theta$ be the true value of the parameter, i.e. $P = P_{\theta_0}$. Suppose that the members P_θ of $\mathcal{Q}$ all have densities p_θ dominated by a measure μ, and that

$$\dot{\ell}_{\theta_0} \equiv \frac{\partial}{\partial \theta} \log p_\theta \bigg|_{\theta=\theta_0},$$

where $\dot{\ell}_{\theta_0} \in L_2^0(P)$, $P\|\dot{\ell}_\theta - \dot{\ell}_{\theta_0}\|^2 \to 0$ as $\theta \to \theta_0$, and the meaning of the extension of $L_2^0(P)$ to vectors of random variables is obvious.

The tangent set

$$\dot{\mathcal{Q}}_P \equiv \{h'\dot{\ell}_{\theta_0} : h \in \mathbb{R}^k\}$$

contains all the information in the score $\dot{\ell}_{\theta_0}$, and, moreover, it is not hard to verify that $\dot{\mathcal{Q}}_P$ represents $\mathcal{Q}$.

Thus one-dimensional submodels are sufficient to represent all finite-dimensional submodels.

Moreover, since semiparametric efficiency is assessed by examining the information for the worst finite-dimensional submodel, one-dimensional submodels are sufficient for semiparametric models in general, including models with infinite-dimensional parameters.

Now if $\{P_t : t \in N_\epsilon\}$ and $g \in \dot{\mathcal{P}}_P$ satisfy (1), then for any $a \geq 0$, everything will also hold when ϵ is replaced by ϵ/a and g is replaced by ag.

Thus we can usually assume, without a significant loss in generality, that a tangent set $\dot{\mathcal{P}}_P$ is a *cone*, i.e., a set that is closed under multiplication by nonnegative scalars.

We will also frequently find it useful to replace a tangent set with its closed linear span, or to simply assume that the tangent set is closed under limits of linear combinations, in which case it becomes a tangent space.

For an arbitrary model parameter $\psi : \mathcal{P} \mapsto \mathbb{D}$, consider the fairly general setting where $\mathbb{D}$ is a Banach space $\mathbb{B}$.

In this case, ψ is *differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$* if, for every smooth one-dimensional submodel $\{P_t\}$ with tangent $g \in \dot{\mathcal{P}}_P$,

$$\left.\frac{d\psi(P_t}{dt}\right|_{t=0} = \dot{\psi}_P(g)$$

for some bounded linear operator $\dot{\psi}_P : \dot{\mathcal{P}}_P \mapsto \mathbb{B}$.

When $\dot{\mathcal{P}}_P$ is a linear space, it is a subspace of the Hilbert space $L_2^0(P)$ and some additional results follow.

To begin with, the Riesz representation theorem yields that for every $b^* \in \mathbb{B}^*$,

$$b^* \dot{\psi}_P(g) = P\left[\tilde{\psi}_P(b^*)g\right]$$

for some operator $\tilde{\psi}_P : \mathbb{B}^* \mapsto \overline{\text{lin}}\, \dot{\mathcal{P}}_P$.

Note also that for any $g \in \dot{\mathcal{P}}_P$ and $b^* \in \mathbb{B}^*$, we also have

$$b^* \dot{\psi}_P(g) = \langle g, \dot{\psi}_P^*(b^*) \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product on $L_2^0(P)$ and $\dot{\psi}_P^*$ is the adjoint of $\dot{\psi}_P$.

Thus the operator $\tilde{\psi}_P$ is precisely $\dot{\psi}_P^*$.

In this case, $\tilde{\psi}_P$ is the *efficient influence function*.

Regularity

The estimator T_n is *regular* at P relative to $\dot{\mathcal{P}}_P$ if for every smooth one-dimensional submodel $\{P_t\} \subset \mathcal{P}$ and every sequence t_n with $t_n = O(n^{-1/2})$,

$$\sqrt{n}(T_n - \psi(P_{t_n})) \overset{P_n}{\rightsquigarrow} Z,$$

for some tight Borel random element Z, where $P_n \equiv P_{t_n}$.

An estimator sequence $\{T_n\}$ for a parameter $\psi(P)$ is *asymptotically linear* if there exists an *influence function* $\check{\psi}_P : \mathcal{X} \mapsto \mathbb{B}$ such that

$$\sqrt{n}(T_n - \psi(P)) - \sqrt{n}\mathbb{P}_n\check{\psi}_P \stackrel{\mathsf{P}}{\to} 0.$$

There are a number of ways to establish regularity, but when $\mathbb{B} = \ell^\infty(\mathcal{H})$, for some set $\mathcal{H}$, and T_n is asymptotically linear, the conditions for regularity can be expressed as properties of the influence function.

Fortunately, we only need to consider the influence function as evaluated for the subset of linear functionals $\mathbb{B}' \subset \mathbb{B}^*$ that are coordinate projections.

These projections are defined for each $h \in \mathcal{H}$ as

$$\dot{\psi}_P(g)(h) = P[\tilde{\psi}_P(h)g].$$

Theorem 18.1 *Assume*

- T_n *and* $\psi(P)$ *are in* $\ell^\infty(\mathcal{H})$,
- ψ *is differentiable at* P *relative to the tangent space* $\dot{\mathcal{P}}_P$ *with efficient influence function*

$$\tilde{\psi}_P : \mathcal{H} \mapsto L_2^0(P),$$

and

- T_n *is asymptotically linear for* $\psi(P)$*, with influence function* $\check{\psi}_P$.

For each $h \in \mathcal{H}$, let $\check{\psi}^{\bullet}_P(h)$ be the projection of $\check{\psi}_P(h)$ onto $\dot{\mathcal{P}}_P$. TFAE:

(i) *The class $\mathcal{F} \equiv \{\check{\psi}_P(h) : h \in \mathcal{H}\}$ is P-Donsker and $\check{\psi}^{\bullet}_P(h) = \tilde{\psi}_P(h)$ almost surely for all $h \in \mathcal{H}$.*

(ii) T_n is regular at P.

We now present a way of verifying an efficient influence function:

Proposition 18.2 *Assume* $\psi : \mathcal{P} \mapsto \ell^\infty(\mathcal{H})$ *is differentiable at* P *relative to the linear tangent set* $\dot{\mathcal{P}}_P$*, with bounded linear derivative* $\dot{\psi}_P : \dot{\mathcal{P}}_P \mapsto \ell^\infty(\mathcal{H})$*.*

Then $\check{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$ *is an efficient influence function if and only if the following both hold:*

(i) $\check{\psi}_P(h)$ *is in the closed linear span of* $\dot{\mathcal{P}}_P$ *for all* $h \in \mathcal{H}$*, and*

(ii) $\dot{\psi}_P(g)(h) = P[\check{\psi}_P(h)g]$ *for all* $h \in \mathcal{H}$ *and* $g \in \dot{\mathcal{P}}_P$*.*

This simple proposition is a primary ingredient in a “calculus” of semiparametric efficient estimators.

A second important ingredient is making sure that $\dot{\mathcal{P}}_P$ is rich enough to represent all smooth finite-dimensional submodels of $\mathcal{P}$.

Consequences of Non-Regularity

Let T_n be asymptotically linear for $\psi(P)$, with influence function $\check{\psi}_P$.

If T_n is not regular, we know from the proof of Theorem 18.1 that there exists a function $\tilde{g} \in \dot{\mathcal{P}}_P$ such that $P\tilde{g}^2 > 0$ and, for each $a \in \mathbb{R}$, there exists a sequence of contiguous, one-dimensional submodels P_n, for which

$$\sqrt{n}(T_n(h) - \psi(P_n)(h)) \quad \overset{P_n}{\rightsquigarrow} \quad \mathbb{G}\check{\psi}_P(h) + aP\tilde{g}^2. \qquad (3)$$

This means that T_n has a serious defect.

Specifically, we see from (3) that for any $M < \infty$ and $\epsilon > 0$, we can alter a to generate a one-dimensional submodel $\{P_n\}$ for which

$$\text{pr}\left(\left\|\sqrt{n}(T_n - \psi(P_n))\right\|_{\mathcal{H}} > M\right) > 1 - \epsilon.$$

Thus the estimator T_n has arbitrarily poor performance for certain submodels which are represented by $\dot{\mathcal{P}}_P$.

Hence regularity is not just a mathematically convenient definition, but it reflects, even in infinite-dimensional settings, a certain intuitive reasonableness about T_n.

This does not mean that nonregular estimators are never useful, because they can be, especially when the parameter $\psi(P)$ is not $\sqrt{n}$-consistent.

Nevertheless, regular estimators are very appealing when they are available.

Efficiency

We now turn our attention to the question of efficiency in estimating general Banach-valued parameters.

We first present general optimality results and characterize efficient estimators in the special Banach space $\ell^\infty(\mathcal{H})$.

We then

- consider efficiency of Hadamard-differentiable functionals of efficient parameters,
- show how to establish efficiency of estimators in $\ell^\infty(\mathcal{H})$ from efficiency of all one-dimensional components, and
- examine the related issue of efficiency in product spaces.

The next two theorems, which characterize optimality in Banach spaces, are the key results of this section.

For a Borel random element Y, let $L(Y)$ denote the law of Y (as in Section 7.1), and let $*$ denote the convolution operation.

Define a function $u : \mathbb{B} \mapsto [0, \infty)$ to be *subconvex* if, for every $b \in \mathbb{B}$, $u(b) \geq 0 = u(0)$ and $u(b) = u(-b)$, and also, for every $c \in \mathbb{R}$, the set $\{b \in \mathbb{B} : u(b) \leq c\}$ is convex and closed.

A simple example of a subconvex function is the norm $\| \cdot \|$ for $\mathbb{B}$.

Here are the theorems:

Theorem 18.3 (Convolution theorem) *Assume that* $\psi : \mathcal{P} \mapsto \mathbb{B}$ *is differentiable at* P *relative to the tangent space* $\dot{\mathcal{P}}_P$*, with efficient influence function* $\tilde{\psi}_P$.

Assume that T_n *is regular at* P *relative to* $\dot{\mathcal{P}}_P$*, with* Z *being the tight weak limit of* $\sqrt{n}(T_n - \psi(P))$ *under* P*.*

Then

$$L(Z) = L(Z_0) * L(M),$$

where M *is some Borel random element in* $\mathbb{B}$, *and* Z_0 *is a tight Gaussian process in* $\mathbb{B}$ *with covariance*

$$P[(b_1^* Z_0)(b_2^* Z_0)] = P\left[\tilde{\psi}_P(b_1^*)\tilde{\psi}_P(b_2^*)\right]$$

for all $b_1^*, b_2^* \in \mathbb{B}^*$.

Theorem 18.4 *Assume the conditions of Theorem 18.3 hold and that* $u : \mathbb{B} \mapsto [0, \infty)$ *is subconvex.*

Then

$$\limsup_{n\to\infty} \mathsf{E}_* u\left(\sqrt{n}(T_n - \psi(P))\right) \geq \mathsf{E}u(Z_0),$$

where Z_0 *is as defined in Theorem 18.3.*

The previous two theorems characterize optimality of regular estimators in terms of the limiting process Z_0, which is a tight, mean zero Gaussian process with covariance obtained from the efficient influence function.

This can be viewed as an asymptotic generalization of the Cramér-Rao lower bound.

We say that an estimator T_n is *efficient* if it is regular and the limiting distribution of $\sqrt{n}(T_n - \psi(P))$ is Z_0, i.e., T_n achieves the optimal lower bound.

The following proposition assures us that Z_0 is fully characterized by the distributions of $b^* Z_0$ for b^* ranging over all of $\mathbb{B}^*$:

Proposition 18.5 *Let X_n be an asymptotically tight sequence in a Banach space $\mathbb{B}$ and assume $b^* X_n \rightsquigarrow b^* X$ for every $b^* \in \mathbb{B}^*$ and some tight, Gaussian process X in $\mathbb{B}$.*

Then $X_n \rightsquigarrow X$.

Proof. Let

$$\mathbb{B}_1^* \equiv \{b^* \in \mathbb{B}^* : \|b^*\| \leq 1\}$$

and

$$\tilde{\mathbb{B}} \equiv \ell^\infty(\mathbb{B}_1^*).$$

Note that

$$(\mathbb{B}, \|\cdot\|) \subset (\tilde{\mathbb{B}}, \|\cdot\|_{\mathbb{B}_1^*})$$

by letting $x(b^*) \equiv b^* x$ for every $b^* \in \mathbb{B}^*$ and all $x \in \mathbb{B}$ and recognizing that

$$\|x\| = \sup_{b^* \in \mathbb{B}_1^*} |b^* x| = \|x\|_{\mathbb{B}_1^*}$$

by the Hahn-Banach theorem.

Thus weak convergence of X_n in $\tilde{\mathbb{B}}$ will imply weak convergence in $\mathbb{B}$ by Lemma 7.8.

Since we already know that X_n is asymptotically tight in $\tilde{\mathbb{B}}$, we are done if we can show that all finite-dimensional distributions of X_n converge.

Accordingly, let

$$b_1^*, \ldots, b_m^* \in \mathbb{B}_1^*$$

be arbitrary and note that for any $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$,

$$\sum_{j=1}^m \alpha_j X_n(b_j^*) = \tilde{b}^* X_n$$

for

$$\tilde{b}^* \equiv \sum_{j=1}^m \alpha_j b_j^* \in \mathbb{B}^*.$$

Since we know that $\tilde{b}^* X_n \rightsquigarrow \tilde{b}^* X$, we now have that

$$\sum_{j=1}^m \alpha_j b_j^* X_n \rightsquigarrow \sum_{j=1}^m \alpha_j b_j^* X.$$

Thus

$$(X_n(b_1^*), \ldots, X_n(b_m^*))^T \rightsquigarrow (X(b_1^*), \ldots, X(b_m^*))^T$$

since $(\alpha_1, \ldots, \alpha_j) \in \mathbb{R}^m$ was arbitrary and X is Gaussian.

Since $b_1^*, \ldots, b_m^*$ and m were also arbitrary, we have the result that all finite-dimensional distributions of X_n converge, and the desired conclusion now follows.$\square$