Introduction to Empirical Processes and Semiparametric Inference
Lecture 04: Overview Continued

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

We say that a map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$, is *Gâteaux-differentiable* at $\theta \in \mathbb{D}_\phi$

if there exists a map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\left\| \frac{\phi(\theta + th - \phi(\theta)}{t} - \phi'_\theta(h) \right\| \to 0 \,,$$

where $\theta + th \in \mathbb{D}_\phi$ for every $t > 0$ small enough.

We say that the map is *Hadamard-differentiable* if there exists a continuous linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\sup_{h \in K} \left\| \frac{\phi(\theta + th - \phi(\theta)}{t} \phi'_\theta(h) \right\| \to 0 \,, \tag{1}$$

for every compact set $K \subset \mathbb{D}$ where $\theta + th \in \mathbb{D}_\phi$ for every $t > 0$ small enough.

We say that the map is *Fréchet-differentiable* if the condition (1) holds for every bounded subset $K \subset \mathbb{D}$.

Semiparametrics: Models and Efficiency

A *statistical model* is a collection of probability measures $\{P \in \mathcal{P}\}$ that specify the distribution of a random observation X.

Consider the linear model $Y = \beta' Z + e$, where the observed data is $X = (Y, Z)$, $\beta \in \mathbb{R}^k$ is unknown, and the joint distribution of (Z, e) satisfies

- $E(e|Z) = 0$ almost surely,
- $E(e^2|Z) \leq K < \infty$ almost surely, and
- the distribution of (Z, e) is otherwise unspecified (mostly).

In this instance, the model $\mathcal{P}$ has

- an unkown parameter β of interest and
- some other unknown components of secondary interest but which allow flexibility (the joint distribution of (Z, e)).

In general, a model $\mathcal{P}$ has

- A parameter (often denoted ψ) of interest and
- several partially unspecified components not of interest.

For semiparametric models, $\psi = (\theta, \eta)$, where

- θ is finite dimensional and usually of primary interest, and
- η is infinite dimensional and usually of secondary interest.
- There may be other components not of interest which are not parameterized at all.
- θ and η may have multiple subcomponents.
- Choices of parameter names are contextual and vary in the literature, although there are some consistencies.

The goals of semiparametric inference are primarily to:

- Select an appropriate model for inference on X;
- Estimate one or more subcomponents of ψ, sometimes θ alone is the focus;
- Conduct inference (e.g., confidence intervals) for the parameters of interest;
- Try to obtain efficient estimators.

The standard set-up for inference we will use in this book is where estimation and inference is based on a sample, $X_1, \ldots, X_n$ of i.i.d. realizations of a distribution $P \in \mathcal{P}$.

The generic prameter of interest ψ can be viewed as a function of the distribution $\psi(P)$, since if we know the distribution, we know the parameter value.

An estimator T_n (based on the data $X_1, \ldots, X_n$) is *efficient* for $\psi(P)$ if the limiting variance V of $\sqrt{n}(T_n - \psi(P))$ is the smallest among all *regular* estimators of $\psi(P)$.

The inverse of V is the *information* for T_n.

The optimal efficiency of a parameter $\psi(P)$ depends on the complexity of the model $\mathcal{P}$.

Estimation under $\mathcal{P}$ is more taxing than estimation under any parametric submodel $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\} \subset \mathcal{P}$, where Θ_0 is finite-dimensional.

For example, in the linear regression model $\mathcal{P}$ setting above, if we assume Z and e are independent and e is $N(0, \sigma^2)$, where σ^2 is unknown, and call this model $\mathcal{P}_0$, $\mathcal{P}_0$ is a parametric submodel of $\mathcal{P}$.

Thus, and in general, information for $\mathcal{P}$ is worse (less) than information for any parametric submodel $\mathcal{P}_0 \subset \mathcal{P}$.

For a semiparametric model $\mathcal{P}$, if the information for the regular estimator T_n (to estimate $\psi(P)$) equals the minimum (infemum) of the Fisher informations for all parametric submodels of $\mathcal{P}$, then T_n is semiparametric efficient for $\psi(P)$, and the associated information for T_n is called the "efficient information" (also called the efficient information for $\mathcal{P}$).

This is because the only settings with more information than T_n are for parametric (not semiparametric) submodels.

Interestingly, when $\psi(P)$ is one-dimensional, it is always possible to identify a one-dimensional parametric submodel that has the same Fisher information as the efficient information for $\mathcal{P}$.

A parametric submodel that achieves the efficient information is called a *least favorable* or *hardest* submodel.

Fortunately, finding the efficient information for $\psi(P)$ only requires consideration of one-dimensional parametric submodels $\{P_t : t \in [0, \epsilon)\}$ surrounding representative (true) distributions $P \in \mathcal{P}$, where $P_0 = P$ and $P_t \in \mathcal{P}$ for all $t \in [0, \epsilon)$.

If $\mathcal{P}$ has a dominating measure μ, then each P can be expressed as a density p.

In this case, we require the densities around p to be smooth enough so that $g(x) = \partial p_t(x)/(\partial t)|_{t=0}$ exists with $\int_{\mathcal{X}} g^2(x)p(x)\mu(dx) < \infty$.

More generally, we say that a submodel is *differentiable in quadratic mean* with *score function* $g : \mathcal{X} \mapsto \mathbb{R}$ if

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

Sometimes it is important for us to consider a collection of many one-dimensional submodels surrounding a representative P, each submodel represented by a score function g: such a collection of score functions $\dot{\mathcal{P}}_P$ is called a *tangent set*.

Because, for any $g \in \dot{\mathcal{P}}_P$, $Pg = 0$ and $Pg^2 < \infty$, these tangent sets are subsets of $L_2^0(P)$, the space of all function $h : \mathcal{X} \mapsto \mathbb{R}$ with $Ph = 0$ and $Ph^2 < \infty$.

When the tangent set is closed under linear combinations, it is called a *tangent space*.

Consider a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^k$ and $\mathcal{P}$ is dominated by μ, such that the classical score function $\dot{\ell}_\theta(x) = \partial \log p_\theta(x)/(\partial\theta)$ exists with $P_\theta \|\dot{\ell}_\theta\|^2 < \infty$ (and a few other conditions hold).

One can show that each of the one-dimensional submodels $g(x) = h'\dot{\ell}_\theta(x)$, for any $h \in \mathbb{R}^k$, satisfies (2), forming the tangent space $\dot{\mathcal{P}}_{P_\theta} = \{h'\dot{\ell}_\theta : h \in \mathbb{R}^k\}$.

Thus there is a simple and direct connection between the classical score function and the more general tangent spaces.

Continuing with the parametric setting, we say that an estimator $\hat{\theta}$ of θ is efficient for estimating θ if

- it is regular and
- it's information achieves the Cramèr-Rao lower bound $P[\dot{\ell}_\theta \dot{\ell}'_\theta]$.

Thus the tangent set for the model contains information about the optimal efficiency.

This is also true for semiparametric models, although the relationship between tangents sets and the optimal information is more complex.

Consider estimation of the parameter $\psi(P) \in \mathbb{R}^k$ for the model $\mathcal{P}$:

For any estimator T_n of $\psi(P)$, if

$$\sqrt{n}(T_n - \psi(P)) = \sqrt{n}\mathbb{P}_n\check{\psi}_P + o_P(1),$$

then

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

For a given tangent set $\dot{\mathcal{P}}_P$, assume for each submodel $\{P_t : 0 \leq t < \epsilon\}$ satisfying (2) with some $g \in \dot{\mathcal{P}}_P$ and some $\epsilon > 0$, that

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

for some linear map $\dot{\psi}_P : L_2^0(P) \mapsto \mathbb{R}^k$:

In this setting, we say that ψ is differentiable at P relative to $\dot{\mathcal{P}}_P$.

When $\dot{\mathcal{P}}_P$ is a linear space, there exists a measurable function $\tilde{\psi}_P : \mathcal{X} \mapsto \mathbb{R}^k$ such that

$$\dot{\psi}_P(g) = P\left[\tilde{\psi}_P(X)g(X)\right],$$

for each $g \in \dot{\mathcal{P}}_P$.

The function $\tilde{\psi}_P \in \dot{\mathcal{P}}_P \subset L_2^0(P)$ is unique and is called the *efficient influence funtion* for the parameter ψ in the model $\mathcal{P}$.

The efficient influence function $\tilde{\psi}_P$ for ψ can usually be found by taking any influence function $\check{\psi}_P$ and projecting it onto $\dot{\mathcal{P}}_P$.

Moreover, we will see in Theorem 18.8 that full efficiency of an estimator T_n of $\psi(P)$ can be verified by checking that the influence function of T_n lies in $\dot{\mathcal{P}}_P$.

Consider a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$.

Suppose the parameter of interest is $\psi(P_\theta) = f(\theta)$, for $f : \mathbb{R}^k \mapsto \mathbb{R}^d$ with derivative $\dot{f}_\theta$ at θ, and that the Fisher information for θ, $I_\theta \equiv P\left[\dot{\ell}_\theta \dot{\ell}'_\theta\right]$ is invertible.

Recall that, in this case, $g(X) = h'\dot{\ell}_\theta(X)$, for some $h \in \mathbb{R}^k$, and thus $P_t = P_{\theta+th}$, which implies that $\dot{\psi}_P(g) = \dot{f}_\theta h$.

Since also we have

$$\begin{aligned} \dot{f}_\theta I_\theta^{-1} P\left[\dot{\ell}_\theta(X) g(X)\right] &= \dot{f}_\theta I_\theta^{-1} P\left[\dot{\ell}_\theta \dot{\ell}_\theta' h\right] \\ &= \dot{f}_\theta h, \end{aligned}$$

we have that $\dot{\psi}_P(g) = \dot{f}_\theta h = P\left[\tilde{\psi}_P g\right]$, provided

$$\tilde{\psi}_P(X) = \dot{f}_\theta I_\theta^{-1} \dot{\ell}_\theta(X).$$

Any estimator T_n for which $\sqrt{n}(T_n - \psi(P)) = \sqrt{n}\mathbb{P}_n\tilde{\psi}_P + o_P(1)$ has asymptotic variance $\dot{f}_\theta I_\theta^{-1} \dot{f}_\theta^T$ and thus achieves the Cramèr Rao lower bound and is efficient.

It turns out that one can show that, in the general semiparametric setting, $\tilde{\psi}_P$ achieve the semiparametric efficiency bound.

This requires the tangent set $\dot{\mathcal{P}}_P$ to be rich enough to correspond to all regular, one-dimensional parametric submodels of $\dot{\mathcal{P}}_P$ but not richer:

- If g is the score function for some one-dimensional parametric submodel of $\mathcal{P}$, then $g \in \dot{\mathcal{P}}_P$.
- If $g \in \dot{\mathcal{P}}_P$, then there exists a one-dimensional parametric submodel $\{P_t, 0 \leq t < \epsilon\} \subset \mathcal{P}$ for which g is the score function.

We have focussed on the setting where the parameter $\psi(P)$ is finite-dimensional.

One can also study efficiency when $\psi(P)$ is infinite-dimensional.

One way to express efficiency in this setting is through the convolution theorem which states that for any regular estimator T_n of $\psi(P)$, $\sqrt{n}(T_n - \psi(P))$ has a weak limit that is the convolution of a Gaussian process Z and an independent process M, where

- $Z = \mathbb{G}\tilde{\psi}_P$ and
- $\tilde{\psi}_P$ is the efficient influence function of $\psi(P)$.

A regular estimator T_n for which M is zero almost surely is an efficient estimator for $\psi(P)$.

Sometimes we will refer to efficiency in the infinite-dimensional setting as *uniform efficiency*.

Suppose $\psi(P) \in \ell^\infty(T)$; then (Theorem 18.9) T_n is uniformly efficient for $\psi(P)$ if

- $T_n(t)$ is efficient for $\psi(P)(t)$ for each $t \in T$, and
- $\sqrt{n}(T_n - \psi(P))$ converges weakly in $\ell^\infty(T)$ to a tight process.

A parameter $\psi(P)$ of particular interest is the parametric component θ of a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where

- Θ is an open subset of $\mathbb{R}^k$ and
- H is an arbitrary set that my be infinite dimensional.

Tangent sets (spaces) can be used to derive an efficient estimator for $\psi(P_{\theta,\eta}) = \theta$ through the formation of an *efficient score function*.

In this setting, we use submodels of the form

$$\{P_{\theta+ta,\eta_t}, 0 \leq t < \epsilon\}$$

that are differentiable in quadratic mean with score function

$$\left.\frac{\partial \log dP_{\theta+ta,\eta_t}}{\partial t}\right|_{t=0} = a'\dot{\ell}_{\theta,\eta} + g,$$

where

- $a \in \mathbb{R}^k$,
- $\dot{\ell}_{\theta,\eta} : \mathcal{X} \mapsto \mathbb{R}^k$ is the ordinary score for θ when η is fixed, and where
- $g : \mathcal{X} \mapsto \mathbb{R}$ is an element of the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ for the submodel $\mathcal{P}_\theta = \{P_{\theta,\eta} : \eta \in H\}$ (holding θ fixed).

The tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ (the *tangent set for* η) should be rich enough to reflect all parametric submodels of $\mathcal{P}_\theta$.

The tangent set for the full model is

$$\dot{\mathcal{P}}_{P_{\theta,\eta}} = \left\{ a'\dot{\ell}_{\theta,\eta} + g : a \in \mathbb{R}^k, g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}} \right\}.$$

The efficient score $\tilde{\ell}_{\theta,\eta}$ is computed by projecting $\dot{\ell}_{\theta,\eta}$ onto the orthocomplement of $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$.

The efficient information is

$$\tilde{I}_{\theta,\eta} = P\left[\tilde{\ell}_{\theta,\eta}\tilde{\ell}'_{\theta,\eta}\right]$$

and the efficient influence function is

$$\tilde{\psi}_{\theta,\eta} = \tilde{I}^{-1}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}.$$

If we can find an estimator T_n of θ such that

$$\sqrt{n}(T_n - \theta) = \sqrt{n}\mathbb{P}_n\tilde{\psi}_{\theta,\eta} + o_P(1),$$

then T_n is a semiparametrically efficient estimator of θ.

Much of the work in semiparametrics involves finding $\tilde{\psi}_{\theta,\eta}$ and finding estimators T_n which are asymptotically efficient.

Consider the Cox model $\Lambda(t; Z) = e^{\beta' Z}\Lambda_0(t)$ under right censoring:

- If $\hat{\beta}$ is the partial likelihood estimator of β, then $\hat{\beta}$ is a semiparametric efficient estimator for β.
- If $\hat{\Lambda}$ is the Breslow estimator

$$\int_0^{(\cdot)} \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(t) e^{\hat{\beta}' Z_i}},$$

  then $\hat{\Lambda}$ is uniformly efficient for Λ_0.

We will focus more in these ideas in Part III toward the end of the semester.