Introduction to Empirical Processes and Semiparametric Inference Lecture 03: Overview Continued

Yair Goldberg, Ph.D.

Postdoctoral Fellow, Biostatistics

University of North Carolina-Chapel Hill

Review

Let $X_1, \ldots, X_n$ be an i.i.d. sample drawn from a probability measure P on an arbitrary sample space $\mathcal{X}$.

Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$.

We define the empirical process as $\{\mathbb{P}_n f, f \in \mathcal{F}\}$ where $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical measure.

More specifically, we have the empirical process $\{\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(X_i), f \in \mathcal{F}\}$.

We say that a class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is P-Glivenko-Cantelli if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{\text{as}*}{\to} 0,$$

where $Pf = \int_{\mathcal{X}} f(s)P(dx)$.

The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the minimum number of ϵ-brackets in $L_r(P)$ needed in order to ensure that every $f \in \mathcal{F}$ is contained in at least one bracket.

If $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$ then $\mathcal{F}$ is P-Glivenko-Cantelli.

Define the random measure

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P),$$

and define $\mathbb{G}$ to be a mean zero Gaussian process indexed by $\mathcal{F}$,

- with covariance $E\left[f(X)g(X)\right] - E\left[f(X)\right]E\left[g(X)\right]$, for all $f, g \in \mathcal{F}$,
- and having appropriately continuous sample paths (almost surely).

We say that $\mathcal{F}$ is P-Donsker if

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \text{ in } \ell^\infty(\mathcal{F}).$$

The bracketing entropy integral is defined as

$$J_{[]}(\delta, \mathcal{F}, L_r(P)) \equiv \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_r(P))}\, d\epsilon.$$

We saw that when $\mathcal{F}$ is a class of measurable functions with $J_{[]}(\infty, \mathcal{F}, L_2(P)) < \infty$, $\mathcal{F}$ is P-Donsker.

The Functional Delta Method

Let X_n be a sequence of random vectors such that

$$\sqrt{n}(X_n - \theta) \rightsquigarrow X$$

where $\theta \in \mathbb{R}^p$. Let the function $\phi : \mathbb{R}^p \mapsto \mathbb{R}^q$ has a derivative $\phi'(\theta)$.

Then

$$\sqrt{n}(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi(\theta)'X\ .$$

This multivariate delta method can be generalized to random processes.

Quantile example (or what are we missing?)

Define $\xi_p = F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$ for some $p \in (0, 1)$.
$\xi(p)$ is the p-th quantile of the distribution function F.

When F is strictly monotonically increasing and continuous at ξ_p we have $F(\xi_p) = p$.

Can we use the delta method here?

Quantile example (or what are we missing?)

We have $\sqrt{n}\,(\mathbb{F}_n(t) - F(t)) \rightsquigarrow G(t) \equiv \mathbb{B}(F(t))$.

Define $\phi(F)(p) = F^{-1}(p)$ for all $p \in [a,b] \subset (0,1)$.

We hope that

$$\sqrt{n}\Big(\phi(\mathbb{F}_n) - \phi(F)\Big) \rightsquigarrow \phi'(\mathbb{B}(F))\,.$$

Note that $\phi : \mathbb{D} \mapsto \mathbb{E}$ where $\mathbb{D}$ is the space of distribution functions, and $\mathbb{E}$ is the space of monotonic functions on $[0,1]$.

We need to

- Define derivatives.
- Generalize the delta method.
- Validate the delta method for bootstrapping.

Normed spaces

A normed space is a metric space $(\mathbb{D}, d)$, where $d(x, y) = \|x - y\|$ for every $x, y \in \mathbb{D}$ where $\| \cdot \|$ is a norm. A norm satisfies

- $\|x + y\| \leq \|x\| + \|y\|$
- $\|\alpha x\| = |\alpha| \cdot \|x\|$
- $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$

for all $x, y \in \mathbb{D}$ and $\alpha \in \mathbb{R}$.

We say that $\| \cdot \|$ is a semi-norm if $\|x\| = 0$ does not necessarily mean that $x = 0$.

Normed spaces

Examples of normed spaces:

- For $1 \leq r < \infty$, $L_r(P)$ is a normed space of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_{P,r} \equiv [Pf^r(X)]^{1/r} < \infty$.
- $\ell^\infty(T)$ is the collection of all bounded functions $f : T \mapsto \mathbb{R}$ with the norm $\|f\|_\infty = \sup_{t \in T} f$.
- The cadlag space $D[0,1]$ with the sup norm, where $D[0,1]$ is the space of right continuous with left-hands limits real functions.
- Any linear subspace of a normed space.

Differentiability in normed spaces

We say that a map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$, $\mathbb{D}_\phi \subset \mathbb{D}$, is *Gâteaux-differentiable*
at $\theta \in \mathbb{D}_\phi$
if for every fixed $h \in \mathbb{D}$
with $\theta + th \in \mathbb{D}_\phi$ for every $t > 0$ small enough,
there exists an element $\phi'_\theta(h) \in \mathbb{E}$ such that

$$\frac{\phi(\theta + th) - \phi(\theta)}{t} \to \phi'_\theta(h)$$

as $t \downarrow 0$.

Differentiability in normed spaces

We say that a map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$,

is *Hadamard-differentiable* at $\theta \in \mathbb{D}_\phi$

tangentially to $\mathbb{D}_0 \subset \mathbb{D}$

if there exists continuous linear map

$\phi'_\theta : \mathbb{D}_0 \mapsto \mathbb{E}$ such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h)$$

for all converging sequences $t_n \downarrow 0$ and $h_n \to h \in \mathbb{D}_0$

with $h_n \in \mathbb{D}$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n large enough.

Quantile example (revisited)

Recall that $\phi(F)(p) = F^{-1}(p)$ for all $p \in [a,b] \subset (0,1)$.
Let $[u,v] = [F^{-1}(a) - \varepsilon, F^{-1}(b) + \varepsilon]$.
Define $\mathbb{D} = D[u,v]$, the space of cadlag functions on $[u,v]$.
Define $\mathbb{D}_\phi$, the space of distribution functions restricted to $[u,v]$.
Define $\mathbb{D}_0 = C[u,v]$, the space of continuous functions on $[u,v]$.

Assume that F has continuous density f such that $f(t) > 0$ for all $t \in [u,v]$.
Then ϕ is Hadamard differentiable with derivative

$$\phi_F(h)'(p) = \frac{-h(F^{-1}(p))}{f(F^{-1}(p))} \text{ for all } p \in [a,b] \, .$$

Weak Convergence

Theorem 2.8.

Let $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$ be Hadamard-differentiable at $\theta \in \mathbb{D}_\phi$, tangentially to $\mathbb{D}_0 \subset \mathbb{D}$.

Assume that

$$r_n(X_n - \theta) \rightsquigarrow X$$

for some sequence $r_n \to \infty$, where X_n takes its values in $\mathbb{D}_\phi$, and X is a tight process taking its values in $\mathbb{D}_0$.

Then

$$r_n\big(\phi(X_n) - \phi(\theta)\big) \rightsquigarrow \phi'_\theta(X)\,.$$

Quantile example (revisited)

We so that ϕ is Hadamard differentiable with derivative

$$\phi_F(h)'(p) = \frac{-h(F^{-1}(p))}{f(F^{-1}(p))}$$

By Theorem 2.8,

$$\begin{aligned}
\sqrt{n}\left(\mathbb{F}_n^{-1}(p) - F^{-1}(p)\right) &= \sqrt{n}\left(\phi(\mathbb{F}_n)(p) - \phi(F)(p)\right) \\
&= \frac{-\mathbb{B}(F(F^{-1}(p)))}{f(F^{-1}(p))} + o_P(1) \\
&= \frac{-\sqrt{n}\left(\mathbb{F}_n(F^{-1}(p)) - F(F^{-1}(p))\right)}{f(F^{-1}(p))} + o_P(1) \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbf{1}\{X_i \le F^{-1}(p)\} - p}{f(F^{-1}(p))} + o_P(1)
\end{aligned}$$

Bootstrapping

Define $\hat{\mathbb{P}}_n(f) = \frac{1}{n}\sum_{i=1}^n W_{ni} f(X_i)$, where the weights $(W_{n1}, \ldots, W_{nn})$ are independent of X_i.

We saw that whenever $\mathcal{F}$ is P-Donsker

$$\hat{\mathbb{G}}_n = \sqrt{n}c\left(\hat{\mathbb{P}}_n - \mathbb{P}_n\right) \underset{W}{\overset{\mathsf{P}}{\rightsquigarrow}} \mathbb{G}\,.$$

Theorem 2.9. Let $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$ be Hadamard-differentiable at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$ with derivative ϕ'_θ. If $\hat{\mathbb{P}}_n$ and $\mathbb{P}_n$ take values at $\mathbb{D}_\phi$ and $\mathbb{G}$ takes values at $\mathbb{D}_0$, then

$$\sqrt{n}c\left(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)\right) \underset{W}{\overset{\mathsf{P}}{\rightsquigarrow}} \phi'_\theta(\mathbb{G})\,.$$

Z-estimators

Many statistics can be written as zero, or approximate-zeros, of estimating equations based on empirical processes: these are called “Z-estimators”.

An example is $\hat{\beta}$ from linear regression which can be written as a zero of $U_n(\beta) = \mathbb{P}_n [X(Y - X'\beta)]$.

We would like to generalize Z-estimator results to processes.

Let $\Psi_n : \Theta \mapsto \mathbb{L}$ be data-dependent functions where Θ and $\mathbb{L}$ are normed spaces.

We say that $\hat{\theta}_n$ is a *Z-estimator* if $\|\Psi_n(\hat{\theta}_n)\| \xrightarrow{\mathsf{P}} 0$.

The main statistical issues for Z-estimators are

- consistency
- asymptotic normality
- the validity of the bootstrap

A non-trivial example

Let $(U_1, \delta_1), \ldots, (U_n, \delta_n)$ be a sample of right-censored failure time observations where

$U_i = T_i \wedge C_i, \delta_i = \mathbf{1}\{T_i \leq C_i\}$

where T_i are failure times and C_i are censoring times.

The Kaplan-Meier estimator of the survival function $S \equiv 1 - F$ is

$$S_n(t) = \prod_{i:U_i \leq t} \left(1 - \frac{\delta_i}{\#\{U_j \geq U_i\}}\right)$$

Let Θ be the space of all survival functions S restricted to the segment $[0, \tau]$.

Efron (1967) showed that the Kaplan-Meier estimator is the solution of $\Psi_n(\hat{S}_n) = 0$ where $\Psi_n : \Theta \mapsto \Theta$ is defined as

$$\Psi_n(S)(t) = \mathbb{P}_n \psi_{S,t}$$

where

$$\psi_{S,t} = \mathbf{1}\{U > t\} + (1 - \delta)\mathbf{1}\{U \leq t\}\mathbf{1}\{S(U) > 0\}\frac{S(t)}{S(U)} - S(t) \, .$$

Consistency

Usually $\Psi_n : \Theta \mapsto \mathbb{L}$ which can be data-dependent is an estimator of a fixed function $\Psi : \Theta \mapsto \mathbb{L}$ for which $\Psi(\theta_0) = 0$.

Theorem 2.10. Let $\Psi(\theta_0) = 0$. Assume that if $\|\Psi(\theta_n)\| \stackrel{\text{P}}{\to} 0$ then $\|\theta_n - \theta_0\| \to 0$. Then

1. If $\|\Psi_n(\hat{\theta}_n)\| \stackrel{\text{P}}{\to} 0$, and $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \stackrel{\text{P}}{\to} 0$, then $\|\hat{\theta}_n - \theta_0\| \stackrel{\text{P}}{\to} 0$.
2. If $\|\Psi_n(\hat{\theta}_n)\| \stackrel{\text{as}*}{\to} 0$, and $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \stackrel{\text{as}*}{\to} 0$, then $\|\hat{\theta}_n - \theta_0\| \stackrel{\text{as}*}{\to} 0$.

Consistency

Back to the Kaplan-Meier example.

$$
\begin{aligned}
\Psi_n(S)(t) &= \mathbb{P}_n\Big(\mathbf{1}\{U>t\} \\
&\quad +(1-\delta)\mathbf{1}\{U\le t\}\mathbf{1}\{S(U)>0\}\frac{S(t)}{S(U)} - S(t)\Big) \\
\Psi(S)(t) &= P\Big(\mathbf{1}\{U>t\} \\
&\quad +(1-\delta)\mathbf{1}\{U\le t\}\mathbf{1}\{S(U)>0\}\frac{S(t)}{S(U)} - S(t)\Big) = 0
\end{aligned}
$$

We need to show that the identifiability condition $\|\Psi(\theta_n)\| \to 0$ then $\|\theta_n - \theta_0\| \to 0$ holds.

We also need to show that $\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \stackrel{\text{as}*}{\to} 0$.

Weak Convergence

Let $\Psi(\theta_0) = 0$ for some θ_0 in the interior of θ.

Let $\hat{\theta}_n$ be a sequence of estimators such that
$\sqrt{n}\|\Psi_n(\hat{\theta}_n)\| \xrightarrow{\mathsf{P}} 0$ and $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathsf{P}} 0$.

Let $\mathbb{G}_n(\theta) = \sqrt{n}\,(\Psi_n(\theta) - \Psi(\theta))$.

Weak Convergence

Theorem 2.11 If

1. $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta_0$.
2. $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ where $\mathbb{G}$ is a tight process.
3. $(1+\sqrt{n}\|\hat{\theta}_n - \theta_0\|)^{-1}\|\mathbb{G}_n(\hat{\theta}_n) - \mathbb{G}_n(\theta_0)\| \xrightarrow{\mathsf{P}} 0$
4. Ψ is Fréchet-differentiable at θ_0 with continuous inverse $\dot{\Psi}_{\theta_0}^{-1}$.

Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}(\mathbb{G})$.

When $\Psi_n(\theta) = \mathbb{P}_n\psi_\theta$ and $\Psi(\theta) = P\psi_\theta$, then under some conditions on the class $\{\psi_\theta\}$, a bootstrap version of this theorem can be proved.

Differentiability in normed spaces

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

$$\frac{\|\phi(\theta + h_n) - \phi(\theta) - \phi'_\theta(h_n)\|}{\|h_n\|} \to 0$$

for all sequences h_n such that $\|h_n\| \to 0$ and $\theta + h_n \in \mathbb{D}_\phi$ for every $n \geq 1$

M-estimators

Many statistics can be written as maxima or minima of objective functions based on empirical processes. These are called “M-estimators”.

Let $M_n : \Theta \mapsto \mathbb{R}$ be data-dependent functions where (Θ, d) is a metric space.

We say that $\hat{\theta}_n$ is an *M-estimator* if
$M_n(\hat{\theta}_n) - \sup_{\theta \in \Theta} M_n(\theta) \overset{\mathsf{P}}{\to} 0$.

Examples include least-squares, maximum likelihood and minimum penalized likelihoods.

The main statistical issues for M-estimators are

- consistency
- asymptotic normality,
- the validity of the bootstrap, and
- convergence rates

Consistency

Assume that the following identifiability condition holds:

For some $\theta_0 \in \Theta$, $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ implies $d(\theta_n, \theta_0) \to 0$.

Theorem 2.12. Let $\hat{\theta}_n$ be a sequence of estimators. Then

1. If $M_n(\hat{\theta}_n) - \sup_{\theta\in\Theta} M_n(\theta) \stackrel{\mathsf{P}}{\to} 0$, and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \stackrel{\mathsf{P}}{\to} 0$, then $d(\hat{\theta}_n, \theta_0) \stackrel{\mathsf{P}}{\to} 0$.
2. If $M_n(\hat{\theta}_n) - \sup_{\theta\in\Theta} M_n(\theta) \stackrel{\mathsf{as*}}{\to} 0$, and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \stackrel{\mathsf{as*}}{\to} 0$, then $d(\hat{\theta}_n, \theta_0) \stackrel{\mathsf{as*}}{\to} 0$.

Partly Linear Regression

Suppose we observe the random triplet $X = (Y, Z, U)$, where $Z \in \mathbb{R}^p$ and $U \in \mathbb{R}$ are covariates that are not linearly dependent, and Y is a dichotomous outcome with

$$E\{Y|Z, U\} = \nu[\beta' Z + \eta(U)],$$

where $\beta \in \mathbb{R}^p$, Z is restricted to a bounded set, and $U \in [0, 1]$, $\nu(t) = 1/(1 + e^{-t})$, and $\eta : [0, 1] \mapsto \mathbb{R}$ is an unknown smooth function.

We assume that the first $k - 1$ derivatives of η exist and are absolutely continuous, with

$$J^2(\eta) \equiv \int_0^1 \left[\eta^{(k)}(t)\right]^2 dt < \infty$$

We defined the the penalized log-likelihood

$$\tilde{L}_n(\beta, \eta) = n^{-1} \sum_{i=1}^{n} \log p_{\beta,\eta}(X_i) - \hat{\lambda}_n^2 J^2(\eta) \, .$$

It can be shown that when the smoothing parameter $\hat{\lambda}_n$ is chosen wisely

- $\sqrt{n}(\hat{\beta} - \beta)$ converges to a mean zero Gaussian vector.
- $\sup_{u \in [0,1]} |\hat{\eta}(u) - \eta(u)| \xrightarrow{\mathsf{P}} 0$.
- $n^{k/(2k+1)} P \left[(\hat{\eta}(U) - \eta(U))^2 \right] \xrightarrow{\mathsf{P}} 0$.