Introduction to Empirical Processes and Semiparametric Inference Lecture 11: Empirical Process Methods, Continued

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Empirical Process Methods, Continued

Today, we will discuss the following basic concepts:

- The Rademacher Process
- The Symmetrization Inequality
- Measurability

The Rademacher Process

We now consider an important sub-Gaussian process: the *Rademacher process*

$$X(a) = \sum_{i=1}^{n} \epsilon_i a_i, \quad a \in \mathbb{R}^n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. *Rademacher random variables* satisfying $P(\epsilon = -1) = P(\epsilon = 1) = 1/2$.

We will verify shortly that this is indeed a sub-Gaussian process with respect to the Euclidean distance $d(a, b) = \|a - b\|$ (which obviously makes $T = \mathbb{R}^n$ into a metric space).

This process will emerge in our development of Donsker results based on uniform entropy.

The following lemma verifies that Rademacher processes are sub-Gaussian:

Lemma 8.7 (Hoeffding's inequality). Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ *and* $\epsilon_1, \ldots, \epsilon_n$ *be independent Rademacher random variables. Then*

$$\text{pr}\left(\left|\sum_{i=1}^{n} \epsilon_i a_i\right| > x\right) \leq 2e^{-\frac{1}{2}x^2/\|a\|^2},$$

for the Euclidean norm $\|\cdot\|$.

Hence $\|\sum \epsilon a\|_{\psi_2} \leq \sqrt{6}\|a\|$.

Proof. For any λ and Rademacher variable ϵ, one has

$$\mathsf{E}e^{\lambda\epsilon} = (e^{\lambda} + e^{-\lambda})/2 = \sum_{i=0}^{\infty} \lambda^{2i}/(2i)! \leq e^{\lambda^2/2},$$

where the last inequality follows from the relation $(2i)! \geq 2^i i!$ for all nonnegative integers.

Hence Markov's inequality gives for any $\lambda > 0$

$$\mathsf{pr}\left(\sum_{i=1}^{n} \epsilon_i a_i > x\right) \leq e^{-\lambda x}\mathsf{E}\exp\left\{\lambda \sum_{i=1}^{n} \epsilon_i a_i\right\} \leq \exp\{(\lambda^2/2)\|a\|^2 - \lambda x\}.$$

Setting $\lambda = x/\|a\|^2$ yields the desired upper bound.

Since multiplying $\epsilon_1, \ldots, \epsilon_n$ by -1 does not change the joint distribution, we obtain

$$\text{pr}\left(-\sum_{i=1}^{n} \epsilon_i a_i > x\right) = \text{pr}\left(\sum_{i=1}^{n} \epsilon_i a_i > x\right),$$

and the desired upper bound for the absolute value of the sum follows.

The bound on the ψ_2-norm follows directly from Lemma 8.1.$\square$

The Symmetrization Inequality

We now discuss a powerful technique for empirical processes called *symmetrization*.

We begin by defining the "symmetrized" empirical process

$$f \mapsto \mathbb{P}_n^\circ f \equiv n^{-1} \sum_{i=1}^{n} \epsilon_i f(X_i),$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables which are also independent of $X_1, \ldots, X_n$.

The basic idea behind symmetrization is to replace supremums of the form $\|(\mathbb{P}_n - P)f\|_{\mathcal{F}}$ with supremums of the form $\|\mathbb{P}_n^\circ f\|_{\mathcal{F}}$.

This replacement is very useful in Glivenko-Cantelli and Donsker theorems based on uniform entropy.

Note that the processes $(\mathbb{P}_n - P)f$ and $\mathbb{P}_n^\circ f$ both have mean zero.

A deeper connection between these two processes is that a Donsker theorem or Glivenko-Cantelli theorem holds for one of these processes if and only if it holds for the other.

One potentially troublesome difficulty is that the supremums involved may not be measurable.

In this setting, we will assume that $X_1, \ldots, X_n$ are the coordinate projections of the product space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$, where $(\mathcal{X}, \mathcal{A}, P)$ is the probability space for a single observation and $\mathcal{A}^n$ is the product σ-field generated from the sets $A_1 \times \cdots \times A_n$, where $A_1, \ldots, A_n \in \mathcal{A}$.

In some settings, an additional source of randomness, independent of $X_1, \ldots, X_n$, will be involved which we will denote Z.

If we let the probability space for Z be $(\mathcal{Z}, \mathcal{D}, Q)$, we will assume that the resulting underlying joint probability space has the form

$$(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{D}, Q) = (\mathcal{X}^n \times \mathcal{Z}, \mathcal{A}^n \times \mathcal{D}, P^n \times Q),$$

where we define the product σ-field $\mathcal{A}^n \times \mathcal{D}$ in the same manner as before.

Now $X_1, \ldots, X_n$ are equal to the coordinate projections onto the first n coordinates, while Z is equal to the coordinate projection onto the $(n+1)$st coordinate.

Theorem 8.8 (Symmetrization). For every nondecreasing, convex $\phi : \mathbb{R} \mapsto \mathbb{R}$ *and class of measurable functions* $\mathcal{F}$*,*

$$\begin{aligned} \mathsf{E}^* \phi\left(\frac{1}{2}\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) &\leq \mathsf{E}^* \phi\left(\|\mathbb{P}_n^\circ\|_{\mathcal{F}}\right) \\ &\leq \mathsf{E}^* \phi\left(2\|\mathbb{P}_n - P\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}}\right), \end{aligned}$$

where $R_n \equiv \mathbb{P}_n^\circ 1 = n^{-1} \sum_{i=1}^n \epsilon_i$ *and the outer expectations are computed based on the product* σ*-field described above.*

Before giving the proof, we make a few observations.

Firstly, the constants $1/2$, 1 and 2 appearing in front of the three respective supremum norms in the chain of inequalities can all be replaced by $c/2$, c and $2c$, respectively, for any positive constant c.

This follows trivially since, for any positive c, $x \mapsto \phi(cx)$ is nondecreasing and convex whenever $x \mapsto \phi(x)$ is nondecreasing and convex.

Secondly, we note that most of our applications of this theorem will be for the setting $\phi(x) = x$.

Thirdly, we note that the first inequality in the chain of inequalities will be of greatest use to us.

However, the second inequality in the chain can be used to establish the following Glivenko-Cantelli result.

Proposition 8.9. For any class of measurable functions $\mathcal{F}$, TFAE:

(i) $\mathcal{F}$ *is* P*-Glivenko-Cantelli and* $\|P\|_{\mathcal{F}} < \infty$.

(ii) $\|\mathbb{P}_n^\circ\|_{\mathcal{F}} \overset{\text{as}*}{\to} 0$.

There is also a similar equivalence involving Donsker results which we will discuss in Chapter 10.

Proof of Theorem 8.8. Let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$.

Formally, $Y_1, \ldots, Y_n$ are the coordinate projections on the last n coordinates in the product space

$$(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{D}, Q) \times (\mathcal{X}^n, \mathcal{A}^n, P^n).$$

Here, $(\mathcal{Z}, \mathcal{D}, Q)$ is the probability space for the n-vector of independent Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$ used in $\mathbb{P}_n^\circ$.

By Lemma 6.13 (coordinate projections are perfect maps), the outer expectations in the theorem are unaffected by the enlarged product probability space.

For fixed $X_1, \ldots, X_n$,

$$\begin{aligned} \|\mathbb{P}_n - P\|_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} [f(X_i) - \mathsf{E} f(Y_i)] \right| \\ &\leq \mathsf{E}_Y^* \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right|, \end{aligned}$$

where E_Y^* is the outer expectation with respect to $Y_1, \ldots, Y_n$ computed by treating the $X_1, \ldots, X_n$ as constants and using the probability space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$.

Applying Jensen's inequality, we obtain

$$\phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathsf{E}_Y \phi\left(\left\|\frac{1}{n}\sum_{i=1}^{n}[f(X_i) - f(Y_i)]\right\|_{\mathcal{F}}^{*Y}\right),$$

where $*Y$ denotes the minimal measurable majorant of the supremum with respect to $Y_1, \ldots, Y_n$ and holding $X_1, \ldots, X_n$ fixed.

Because ϕ is nondecreasing and continuous, the $*Y$ inside of the ϕ in the above can be removed after replacing E_Y with E_Y^*, as a consequence of Lemma 6.8 (part A(i)).

Now take the expectation of both sides with respect to $X_1, \ldots, X_n$ to obtain

$$\mathsf{E}^* \phi \left(\|\mathbb{P}_n - P\|_{\mathcal{F}} \right) \leq \mathsf{E}_X^* \mathsf{E}_Y^* \phi \left(\frac{1}{n} \left\| \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \right).$$

The repeated outer expectation can now be bounded above by the joint outer expectation E^* by Lemma 6.14 (Fubini's theorem for outer expectations).

By the product space structure of the underlying probability space, the outer expectation of any function $g(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ remains unchanged under permutations of its $2n$ arguments.

Since

$$-[f(X_i) - f(Y_i)] = [f(Y_i) - f(X_i)],$$

we have for any n-vector $(e_1, \ldots, e_n) \in \{-1, 1\}^n$, that

$$\left\| n^{-1} \sum_{i=1}^{n} e_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}$$

is just a permutation of

$$h(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \equiv \left\| n^{-1} \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}.$$

Hence

$$\mathsf{E}^*\phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathsf{E}_\epsilon \mathsf{E}^*_{X,Y}\phi\left\|\frac{1}{n}\sum_{i=1}^n e_i[f(X_i) - f(Y_i)]\right\|_{\mathcal{F}}.$$

Now the triangle inequality combined with the convexity of ϕ yields

$$\mathsf{E}^*\phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathsf{E}_\epsilon \mathsf{E}^*_{X,Y}\phi\left(2\|\mathbb{P}^\circ_n\|_{\mathcal{F}}\right),$$

Since $\phi(a+b) \leq (\phi(2a) + \phi(2b))/2$.

By the perfectness of coordinate projections, $\mathrm{E}^*_{X,Y}$ can be replaced by $\mathrm{E}^*_X \mathrm{E}^*_Y$.

Now $\mathrm{E}_\epsilon \mathrm{E}^*_X \mathrm{E}^*_Y$ is bounded above by the joint expectation E^* by reapplication of Lemma 6.14 (outer Fubini's).

This proves the first inequality.

For the second inequality, let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$ as before.

Holding $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$ fixed, we have

$$\begin{aligned}
\|\mathbb{P}_n^\circ f\|_{\mathcal{F}} &= \|\mathbb{P}_n^\circ (f - Pf) + \mathbb{P}_n^\circ Pf\|_{\mathcal{F}} \\
&= \|\mathbb{P}_n^\circ (f - \mathsf{E} f(Y)) + R_n Pf\|_{\mathcal{F}} \\
&\leq \mathsf{E}_Y^* \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}}.
\end{aligned}$$

Applying Jensen's inequality, we now have

$$\phi\left(\|\mathbb{P}_n^\circ\|_{\mathcal{F}}\right) \leq \mathsf{E}_Y^* \phi\left(\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i[f(X_i) - f(Y_i)]\right\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}}\right).$$

Using the permutation argument we used previously, we can replace the $\epsilon_1, \ldots, \epsilon_n$ in the summation with all 1's, and take expectations with respect to $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$ (which are still present in R_n).

This gives us

$$\mathsf{E}^* \phi \left(\|\mathbb{P}_n^\circ\|_{\mathcal{F}} \right) \leq \mathsf{E}_\epsilon \mathsf{E}_X^* \mathsf{E}_Y^* \phi \left(\left\| \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}} \right).$$

After adding and subtracting Pf in the summation and applying the convexity of ϕ, we can bound the right-hand-side by

$$\frac{1}{2}\mathsf{E}_\epsilon \mathsf{E}_X^* \mathsf{E}_Y^* \phi\left(2\left\|\frac{1}{n}\sum_{i=1}^n [f(X_i) - Pf]\right\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}}\right)$$
$$+\frac{1}{2}\mathsf{E}_\epsilon \mathsf{E}_X^* \mathsf{E}_Y^* \phi\left(2\left\|\frac{1}{n}\sum_{i=1}^n [f(Y_i) - Pf]\right\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}}\right).$$

By reapplication of the permutation argument and Lemma 6.14 (outer Fubini's), we obtain the desired upper bound.□

Measurability

The above symmetrization results will be most useful when the supremum $\|\mathbb{P}_n^\circ\|_{\mathcal{F}}$ is measurable and Fubini's theorem permits taking the expectation

- first with respect to $\epsilon_1, \ldots, \epsilon_n$ given $X_1, \ldots, X_n$ and
- secondly with respect to $X_1, \ldots, X_n$.

Without this measurability, only the weaker version of Fubini's theorem for outer expectations applies (Lemma 6.14), and thus the desired reordering of expectations may not be valid.

To overcome this difficulty, we will assume that the class $\mathcal{F}$ is a *P-measurable class*.

A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, on the probability space $(\mathcal{X}, \mathcal{A}, P)$, is P-measurable if

$$(X_1, \ldots, X_n) \mapsto \left\| \sum_{i=1}^{n} e_i f(X_i) \right\|_{\mathcal{F}}$$

is measurable on the completion of $(\mathcal{X}^n, \mathcal{A}^n, P^n)$ for every constant vector $(e_1, \ldots, e_n) \in \mathbb{R}^n$.

It is possible to weaken this condition, but at least some measurability assumptions will usually be needed.

In the Donsker theorem for uniform entropy, it will be necessary to assume that several related classes of $\mathcal{F}$ are also P-measurable:

- $\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$, for all $\delta > 0$, and
- $\mathcal{F}_\infty^2 \equiv \{(f - g)^2 : f, g \in \mathcal{F}\}$ (recall that $\|f\|_{P,2} \equiv (Pf^2)^{1/2}$).

Another assumption on $\mathcal{F}$ that is stronger than P-measurability but often easier to verify in statistical applications is *pointwise measurability*.

A class $\mathcal{F}$ of measurable functions is pointwise measurable if there exists a countable subset $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$, there exists a sequence $\{g_m\} \in \mathcal{G}$ with $g_m(x) \to f(x)$ for every $x \in \mathcal{X}$.

Since, by Exercise 8.5.6,

$$\|\sum e_i f(X_i)\|_{\mathcal{F}} = \|\sum e_i f(X_i)\|_{\mathcal{G}}$$

for all $(e_1, \ldots, e_n) \in \mathbb{R}^n$, pointwise measurable classes are P-measurable for all P.

Consider, for example, the class $\mathcal{F} = \{\mathbf{1}\{x \leq t\} : t \in \mathbb{R}\}$ where the sample space $\mathcal{X} = \mathbb{R}$.

Let $\mathcal{G} = \{\mathbf{1}\{x \leq t\} : t \in \mathbb{Q}\}$, and fix the function

$$x \mapsto f(x) = \mathbf{1}\{x \leq t_0\}$$

for some $t_0 \in \mathbb{R}$.

Note that $\mathcal{G}$ is countable.

Let $\{t_m\}$ be a sequence of rationals with $t_m \geq t_0$, for all $m \geq 1$, and with $t_m \downarrow t_0$.

Then $x \mapsto g_m(x) = \mathbf{1}\{x \leq t_m\}$ satisfies $g_m \in \mathcal{G}$, for all $m \geq 1$, and $g_m(x) \to f(x)$ for all $x \in \mathbb{R}$.

Since t_0 was arbitrary, we have just proven that $\mathcal{F}$ is pointwise measurable (and hence also P-measurable for all P).

Hereafter, we will use the abbreviation PM as a shorthand for denoting pointwise measurable classes.

Another nice feature of PM classes is that they have a number of useful preservation features.

An obvious example is that when $\mathcal{F}_1$ and $\mathcal{F}_2$ are PM classes, then so is $\mathcal{F}_1 \cup \mathcal{F}_2$.

The following lemma provides a number of additional preservation results:

Lemma 8.10. Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be PM classes of real functions on $\mathcal{X}$, and let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ be continuous.

Then the class $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is PM, where $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ denotes the class

$$\{\phi(f_1, \ldots, f_k) : (f_1, \ldots, f_k) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k\}.$$

Lemma 8.10 automatically yields many other useful PM preservation results, including the following for PM classes $\mathcal{F}_1$ and $\mathcal{F}_2$:

- $\mathcal{F}_1 \wedge \mathcal{F}_2$ (all possible pairwise minimums) is PM.
- $\mathcal{F}_1 \vee \mathcal{F}_2$ (all possible pairwise maximums) is PM.
- $\mathcal{F}_1 + \mathcal{F}_2$ is PM.
- $\mathcal{F}_1 \cdot \mathcal{F}_2 \equiv \{f_1 f_2 : f_1 \in \mathcal{F}_1,\ f_2 \in \mathcal{F}_2\}$ is PM.

These are useful to establish Donsker properties for important statistical settings.

The following proposition shows an additional property of PM classes that potentially simplifies the measurability requirements of the Donsker theorem for uniform entropy, Theorem 8.19, given in Section 8.4:

Proposition 8.11. Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ on the probability space $(\mathcal{X}, \mathcal{A}, P)$.

Provided $\mathcal{F}$ is PM with envelope F such that $P^* F^2 < \infty$, then $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ are PM for all $0 < \delta \leq \infty$.

We next consider establishing P-measurability for the class

$$\left\{ \mathbf{1}\{Y - \beta^T Z \leq t\} : \beta \in \mathbb{R}^k, t \in \mathbb{R} \right\},$$

where

$$X \equiv (Y, Z) \in \mathcal{X} \equiv \mathbb{R} \times \mathbb{R}^k$$

has distribution P, for arbitrary P.

This class was considered in the linear regression example of Section 4.1.

The desired measurability result is stated in the following lemma:

Lemma 8.12. Let

$$\mathcal{F} \equiv \left\{ \mathbf{1}\{Y - \beta^T Z \leq t\} : \beta \in \mathbb{R}^k, t \in \mathbb{R} \right\}.$$

Then the classes $\mathcal{F}$*,*

$$\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F},\ \|f - g\|_{P,2} < \delta\},$$

and

$$\mathcal{F}_\infty^2 \equiv \left\{ (f - g)^2 : f, g \in \mathcal{F} \right\}$$

are all P*-measurable for* $0 < \delta \leq \infty$ *and any probability measure on* $\mathcal{X}$*.*