ORIE 6745: Causality and Learning for Intelligent Decision Making September 20, 2017

Lecture 9-10

Lecturer: Nathan Kallus *Scribe: Nirvan Tyagi*

1 Covariate-Balancing Designs

Denote baseline covariates (e.g., subjects or units) as $x_i \in \mathcal{X}$.

1. **Block-randomized / Stratified design**:
   Coarsen covariates into J *strata*.
   $$C : \mathcal{X} \to \{1, \dots, J\} = [J]$$
   Completely randomize treatment for units within each stratum independently and separately. If there exist uneven number of units in a stratum, collect and randomize "leftovers" after. We can consider the process of assigning treatments as drawing $z_{1:n}$ uniformly at random from a restricted label set $BR_n \subseteq CR_n$ (denoting block-randomized and complete-randomized).
   $$BR_n = \left\{ z_{1:n} : \sum_i z_i = \frac{n}{2};\ \forall j \in [J], \left| \sum_i (-1)^{z_i} \cdot \mathbb{I}[C(x_i) = j] \right| \leq 1 \right\}$$
   $$CR_n = \left\{ z_{1:n} : \sum_i z_i = \frac{n}{2} \right\}$$

2. **Pair-matched design**:
   Similar to block-randomized design, except each strata contains exactly two units and assigning units to strata is done on the fly dependent on the sampled data. Units are paired into $n/2$ pairs and for each pair, one unit is randomly assigned as control and the other treatment.

   Choosing pairs: The intuition behind pair matching is to pair similar points so as to balance out their treatment effects between control and treatment. Given a distance metric $\delta(x, x')$, create pairs using a non-bipartite graph matching algorithm minimizing the sum of within-pair distances (e.g., poly-time solvable using Edmond's algorithm).

3. **"Rerandomization" design** (Morgan & Rubin '12):
   Denote $\Delta(z_{1:n})$ as the difference in mean vectors of covariates x_i between treatment and control groups, where $\mathcal{X} \subseteq \mathbb{R}^d$. The goal of rerandomization is to pick a treatment assignment $z_{1:n}$ such that $\Delta(z_{1:n})$ is "small".
   $$\Delta(z_{1:n}) = \frac{2}{n} \sum_i (-1)^{1+z_i} x_i$$
   Fix a positive semidefinite matrix $V \geq 0$, then assign treatments $z_{1:n}$ by drawing randomly from RR_n.
   $$RR_n = \left\{ z_{1:n} : \sum_i z_i = \frac{n}{2};\ \Delta(z_{1:n})^\top V \Delta(z_{1:n}) \leq a \right\}$$
   Choosing V: Often set as the inverse sample covariance matrix, $V = \hat{\Sigma}^{-1}$. This helps normalize over the d dimensions of $\mathcal{X}$.

   Choosing a: The value $\Delta(z_{1:n})^\top V \Delta(z_{1:n})$ distributes close to a χ^2 distribution. Typically, a is chosen as some desired quantile of χ^2. Drawing is done using rejection sampling, i.e. resampling $z_{1:n}$ from CR_n until RR_n property is achieved.

2 Inference with Covariate Balancing Designs

Recall the mean difference estimator as one estimate for effect.

$$\hat{\tau} = \frac{2}{n}\sum_i (-1)^{1+z_i} Y_i$$

Performing inference using covariate balancing designs uses the same techniques as before (e.g., permutation test, randomization test), except instead of drawing new assignments from CR_n, draw from the covariate balancing assignments.

3 What is Balance?

The following definition gives properties that pre-treatment balancing designs must abide to in order to maintain useful inference properties. The definition does not say anything about what a "good" balancing design should look like.

Definition 12 (a priori balance)**.** A design is *a priori balancing* if:

(a) $z_{1:n} \perp\!\!\!\perp Y_{1:n} \,|\, x_{1:n}$: label assignments depend only on data $x_{1:n}$ and not on outcomes $Y_{1:n}$.

(b) $\mathbb{P}[Z_{1:n} = z_{1:n} \,|\, x_{1:n}] = \mathbb{P}[Z_{1:n} = 1 - z_{1:n} \,|\, x_{1:n}]$: the opposite label assignments are equally as likely (*i.e.*blinding the identity of treatment).

(c) $\sum_i z_i = \frac{n}{2}$

3.1 Properties of Mean Difference Estimator Under A Priori Balancing

1. **Expectation of $\hat{\tau}$**:
   First we show $\hat{\tau}$ is an unbiased estimator of SATE.

$$\mathbb{E}[\hat{\tau} \,|\, x_{1:n}, Y_{1:n}(0,1)] = \frac{2}{n}\sum_i \mathbb{E}[z_i Y_i(1) - (1-z_i)Y_i(0) \,|\, x_{1:n}, Y_{1:n}(0,1)]$$

   We can say the following about $z_{1:n}$ is labels were picked with a priori balancing:

$$\begin{aligned} \mathbb{E}[z_i \,|\, x_{1:n}, Y_{1:n}(0,1)] &= \mathbb{E}[z_i \,|\, x_{1:n}] && \text{(by (a))} \\ &= \frac{1}{2} && \text{(by (b))} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\hat{\tau} \,|\, x_{1:n}, Y_{1:n}(0,1)] &= \frac{2}{n}\sum_i \mathbb{E}[z_i Y_i(1) - (1-z_i)Y_i(0) \,|\, x_{1:n}, Y_{1:n}(0,1)] \\ &= \frac{2}{n}\sum_i (\frac{1}{2}Y_i(1) - \frac{1}{2}Y_i(0)) && \text{(by a priori balancing)} \\ &= \frac{1}{n}\sum_i Y_i(1) - Y_i(0) = SATE \end{aligned}$$

2. **Variance of $\hat{\tau}$:**

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}) &= \mathbb{E}[\mathrm{Var}[\hat{\tau} \mid x_{1:n}, Y_{1:n}(0,1)]] + \mathrm{Var}\left[\underbrace{\mathbb{E}[\hat{\tau} \mid x_{1:n}, Y_{1:n}(0,1)]}_{SATE}\right] && \text{(Law of Total Variance)} \\
&= \mathbb{E}\Big[(\hat{\tau} - SATE)^2 \mid x_{1:n}, Y_{1:n}(0,1)\Big] + \mathrm{Var}[SATE] \\
&= \mathbb{E}\Big[(\hat{\tau} - SATE)^2\Big] + \mathrm{Var}[SATE]
\end{aligned}$$

The above expression is composed of two terms. The second term, $\mathrm{Var}(SATE)$, is a constant term across any balancing strategy. Consider the following rewriting of the first term to gain more insight into the effect of different balancing strategies.

$$\begin{aligned}
\hat{\tau} &= \frac{2}{n}\sum_{i:z_i=1} Y_i(1) - \frac{2}{n}\sum_{i:z_i=0} Y_i(0) \\
SATE &= \frac{1}{n}\sum_{i:z_i=1} (Y_i(1) - Y_i(0)) + \frac{1}{n}\sum_{i:z_i=0} (Y_i(1) - Y_i(0)) \\
\hat{\tau} - SATE &= \frac{2}{n}\sum_{i:z_i=1}\left(\frac{Y_i(0)+Y_i(1)}{2}\right) - \frac{2}{n}\sum_{i:z_i=0}\left(\frac{Y_i(0)+Y_i(1)}{2}\right) \\
&= \frac{2}{n}\sum_i (-1)^{1+z_i}\underbrace{\left(\frac{Y_i(0)+Y_i(1)}{2}\right)}_{\hat{Y}_i}
\end{aligned}$$

4 Effects of Balancing

Is balancing always a good idea? Consider the following example:

$x_{1:n}$:	$-\frac{n}{2}$	$\dots$	-1	1	$\dots$	$\frac{n}{2}$	$\in \mathbb{R}$
$\hat{Y}_i = Y_i(0) = Y_i(1)$:	1	0	$\dots$	$\dots$	1	0	

Balancing design	$\mathrm{Var}(\hat{\tau} \mid x_{1:n}, Y_{1:n}(0,1))$
Complete randomization	$\frac{4}{n-1}$
Blocking (by sign)	$\frac{4}{n-6}$
Pair matching (by absolute value)	$\frac{8}{n}$
Rerandomization ($a = 0$)	4

When we apply balancing designs, a choice is made on how to balance. We try to balance by grouping units that we think will have similar outcomes together and then splitting them between treatment and control. However, if our heuristic for grouping similar points together is wrong, then we find that balancing can actually lead to increased variance of $\hat{\tau}$ over simple complete randomization.

In fact, there always exists a set of potential outcomes for any balancing design where complete randomization would give better variance.

Theorem 13. *Among all a priori balancing designs, complete randomization minimizes*

$$\max_{\|\hat{Y}\|_2 \le 1} \mathrm{Var}(\hat{\tau} \mid x_{1:n}, Y_{1:n}(0,1))$$

Corollary 14. *If* $\mathbb{P}(z_i = z_j \mid x_{1:n}) \neq \frac{n-2}{2n}$ *(as it is for complete randomization), then* $\exists\, Y_{1:n}(0,1)$ *such that*

$$\mathrm{Var}(\hat{\tau} \mid x_{1:n}, Y_{1:n}(0,1)) > \mathrm{Var}(\hat{\tau}_{CR} \mid x_{1:n}, Y_{1:n}(0,1))$$

5 Reinterpreting Balance

We need some way to measure how much our balancing heuristic based on observations x are correct in grouping together inputs with similar potential outcomes Y. We try to capture this using the following $\hat{f}$ definition.

$$\begin{aligned}
\hat{Y}_i &= \frac{Y_i(0) + Y_i(1)}{2} \\
\hat{f}(x) &= \mathbb{E}[\hat{Y}_i \mid x] \\
\epsilon_i &= \hat{Y}_i - \hat{f}(x_i) \\
u_i &= (-1)^{1+z_i} \\
B(z_{1:n}; \hat{f}) &= \frac{2}{n} \sum_i u_i \epsilon_i
\end{aligned}$$

Recall $\mathbb{E}[(\hat{\tau} - SATE)^2]$ was the term in the variance that could be improved through balancing. We can then break down $\hat{\tau} - SATE$ with these new definitions. The ϵ_i term captures the part that cannot be improved by balancing due to inability of x to predict Y, while the $\hat{f}$ term captures the part where balancing can help.

$$\begin{aligned}
\hat{\tau} - SATE &= \frac{2}{n} \sum_i (-1)^{1+z_i} \hat{Y}_i \\
&= \frac{2}{n} \sum_i u_i \hat{Y}_i \\
&= \frac{2}{n} \sum_i u_i \hat{f}(x_i) + \frac{2}{n} \sum_i u_i \epsilon_i
\end{aligned}$$

Before we are ready to look closer at $\mathbb{E}[(\hat{\tau} - SATE)^2]$, let's make a few observations.

$$\begin{aligned}
\mathbb{E}[\epsilon_i \mid x_{1:n}] &= \mathbb{E}[\hat{Y}_i \mid x_i] - \mathbb{E}[\hat{f}(x_i) \mid x_i] \\
&= \hat{f}(x_i) - \hat{f}(x_i) = 0
\end{aligned}$$

For arbitrary function of x_i:

$$\mathbb{E}[\epsilon_i h(x_i) \mid x_{1:n}] = h(x_i) \underbrace{\mathbb{E}[\epsilon_i \mid x_{1:n}]}_{0} = 0 \qquad (\epsilon_i \text{ uncorrelated to } h)$$

Now let's break down $\mathbb{E}[(\hat{\tau} - SATE)^2]$. The square of $\hat{\tau} - SATE$ has three types of terms in the product that we analyze separately.

$$\begin{aligned}
\mathbb{E}\Big[u_i u_j \epsilon_i \hat{f}(x_j)\Big] &= \mathbb{E}\Big[\hat{f}(x_j) \mathbb{E}[u_i u_j \epsilon_i \mid x_{1:n}]\Big] && \text{(iterated expectation)} \\
&= \mathbb{E}\Big[\hat{f}(x_j) (\mathbb{E}[u_i u_j \mid x_{1:n}] \cdot \mathbb{E}[\epsilon_i \mid x_{1:n}])\Big] \\
&= 0 && (\mathbb{E}[\epsilon_i \mid x_{1:n}] = 0)
\end{aligned}$$

$$\mathbb{E}[u_i u_j \epsilon_i \epsilon_j] = \mathbb{E}[\mathbb{E}[u_i u_j \mid x_{1:n}] \cdot \mathbb{E}[\epsilon_i \epsilon_j \mid x_{1:n}]]$$
$$= \begin{cases} \mathbb{E}[\mathbb{E}[u_i u_j \mid x_{1:n}] \cdot \mathbb{E}[\epsilon_i \mid x_{1:n} \cdot \mathbb{E}[\epsilon_j \mid x_{1:n}]]] = 0 & i \neq j \\ \mathbb{E}[1 \cdot \epsilon_i^2] = \sigma_\epsilon^2 & i = j \end{cases}$$

$$\mathbb{E}\Big[u_i u_j \hat{f}(x_i)\hat{f}(x_j)\Big] \text{ captured by } B^2(z_{1:n}; \hat{f})$$

Combining these results gives us:

$$\mathbb{E}\Big[(\hat{\tau} - SATE)^2\Big] = \mathbb{E}[B^2(z_{1:n}; \hat{f})] + \frac{4}{n}\sigma_\epsilon^2$$

$$\mathrm{Var}(\hat{\tau}) = \underbrace{\mathbb{E}[B^2(z_{1:n}; \hat{f})]}_{\text{affected by balancing!}} + \frac{4}{n}\sigma_\epsilon^2 + \mathrm{Var}(SATE)$$

What observations can we make about this deconstruction?

1. **What if** $\mathbb{E}[B^2(z_{1:n}; \hat{f})] = 0$**?**

   $$\mathrm{Var}(\hat{\tau}) = \frac{4}{n}\sigma_\epsilon^2 + \mathrm{Var}(SATE)$$

   Recall the variance of $\hat{\tau}_{CR}$ under complete randomization.

   $$\mathrm{Var}(\hat{\tau}) = \frac{4}{n}\sigma_{\hat{Y}}^2 + \mathrm{Var}(SATE)$$

   When x_i is trivial and gives no information about Y_i, then these two variance expressions are equal to each other.

   $$\hat{f}(x_i) = \mathbb{E}[\hat{Y}_i \mid x] = \mathbb{E}[\hat{Y}_i]$$
   $$B(z_{1:n}; \hat{f}) = \sum_i u_i \hat{f}(x_i) = \sum_i u_i \mathbb{E}[\hat{Y}_i] = 0$$
   $$\epsilon_i = \hat{Y}_i - \mathbb{E}[\hat{Y}_i]$$
   $$\sigma_\epsilon^2 = \sigma_{\hat{Y}}^2$$

2. **What if** $\mathrm{Var}(SATE) = 0$**?** (e.g., constant effect $Y_i(1) = Y_i(0) = \tau$)

   $$1 - \frac{\mathrm{Var}(\hat{\tau})}{\mathrm{Var}(\hat{\tau}_{CR}} = 1 - \frac{\frac{4}{n}\sigma_\epsilon^2 + \mathbb{E}[B^2(z_{1:n}; \hat{f})]}{\frac{4}{n}\sigma_{\hat{Y}}^2}$$
   $$= 1 - \frac{\sigma_\epsilon^2}{\sigma_{\hat{\tau}}^2} \qquad \text{(if } \mathbb{E}B^2 = 0\text{)}$$
   $$= R^2 \qquad \text{(coefficient of determination)}$$

   $\sigma_{\hat{Y}}^2 \geq \sigma_\epsilon^2$:

   $$\sigma_{\hat{Y}}^2 = \mathrm{Var}(\hat{Y}_i) = \mathrm{Var}\Big(\mathbb{E}[\hat{Y}_i \mid x_i]\Big) + \mathbb{E}\Big[\mathrm{Var}(\hat{Y}_i \mid x_i)\Big] \qquad \text{(law of total variance)}$$
   $$= \mathrm{Var}\Big(\mathbb{E}[\hat{Y}_i \mid x_i]\Big) + \mathbb{E}\Big[(\hat{Y}_i - \mathbb{E}[\hat{Y}_i \mid x_i])^2\Big]$$
   $$= \mathrm{Var}\Big(\mathbb{E}[\hat{Y}_i \mid x_i]\Big) + \mathbb{E}\big[\epsilon^2\big]$$
   $$= \mathrm{Var}(\hat{f}(x_i) + \sigma_\epsilon^2$$

3. **How can we minimize $|B(z_{1:n}; \hat{f})|$?**
Notice that B is linear in f:

$$B(z_{1:n}; \alpha f + \beta g) = \alpha B(z_{1:n}; f) + \beta B(z_{1:n}; g)$$

This means we need to limit the magnitude of $\hat{f}$, else it can just be scaled up. We consider minimum worst-case $|B|$ over $\hat{f}$ relative to its magnitude, where $\|f\| \in [0, \infty]$:

$$\mathbb{B}(z_{1:n}; \|\cdot\|) = \sup_{\|f\| \leq 1} B(z_{1:n}; f) = \sup_{f} \frac{B(z_{1:n}; f)}{\|f\|}$$

The following properties for norm must hold:

(a) $\|\alpha f\| = |\alpha| \|f\|$

(b) $\|f + g\| \leq \|f\| + \|g\|$

(c) $|B(z_{1:n}; f)| \leq M_{z_{1:n}} \cdot \|f\|$ for some $M_{z_{1:n}}$ else $\mathbb{B}(z_{1:n}; \|\cdot\|) = \infty$

By definition, $|B(z_{1:n}; \hat{f})| \leq \|\hat{f}\| \cdot \mathbb{B}(z_{1:n}; \|\cdot\|)$, giving us:

$$\mathrm{Var}(\hat{\tau}) \leq \|\hat{f}\|^2 \cdot \mathbb{E}\big[\mathbb{B}^2(z_{1:n}; \|\cdot\|)\big] + \frac{4}{n}\sigma_\epsilon^2 + \mathrm{Var}(SATE)$$

Define the following:

$$f \sim g \Leftrightarrow B(z_{1:n}; f) = B(z_{1:n}; g) \; \forall \; z_{1:n} \in CR_n$$
$$[f] = \{g : f \sim g\}$$
$$\|\{f\}\| = \inf_{g \in \{f\}} \|g\|$$

Then,

(a) $\|[\cdot]\|$ is a norm for the vector space $\mathcal{F} = \{[f] : \|[f]\| < \infty\}$
(b) $B(z_{1:n}; [f])$ well defined
(c) $B(z_{1:n}; [f])$ continuous in $[f]$

$$\begin{aligned} |B(z_{1:n}; [f]) - B(z_{1:n}, [g])| &= |B(z_{1:n}, [f - g])| \\ &\leq M_{z_{1:n}} \|[f - g]\| \\ &= M_{z_{1:n}} \|[f] - [g]\| \end{aligned}$$

Given a normed vector space $(V, \|\cdot\|)$, the dual space $(V^*, \|\cdot\|_*)$:

$$\begin{aligned} V^* &= \text{ all bounded linear operators} \\ &= \{A : V \to \mathbb{R} : A(\alpha v + \beta u) = \alpha A v + \beta A u, \exists M : |Av| \leq M\|v\|\} \\ &= \text{ all continuous linear operators} \\ \|A\|_* &= \sup_{\|v\| \leq 1} Av \end{aligned}$$

(Note: if A is linear and not bounded, then $\sup_{\|v\| \leq 1} Av = \infty$)

$$\mathbb{B}(z_{1:n}; \|\cdot\|) = \|B(z_{1:n}; \cdot)\|_*$$

Definition 15 (pure-strategy optimal design (PSOD))**.** Given $\|\cdot\|$, then the PSOD chooses assignments at random from

$$\operatorname*{argmin}_{z_{1:n} \in CR_n} \mathbb{B}(z_{1:n}; \|\cdot\|)$$

5.1 PSOD for Balancing

1. **Block-randomized design**:

$$\|f\|_{L_\infty(C)} = \begin{cases} \sup_{x\in X} f(x) & \underbrace{|f(C^{-1}(j))| = 1 \ \forall j}_{f \text{ is piecewise constant on strata}} \\ \infty & \text{otherwise} \end{cases}$$

Theorem 16. *The PSOD wrt $\|f\|_{L_\infty(C)}$ is the block-randomized design.*

Proof.

$$\begin{aligned}
\mathbb{B}(z_{1:n}; \|\cdot\|_{L_\infty(C)}) &= \frac{2}{n} \sup_{\|v\|_\infty \le 1} \sum_i u_i \sum_j \mathbb{I}[C(x_i) = j] v_j \\
&= \frac{2}{n} \sup_{\|v\| \le 1} v^\top \left(\sum_i u_i e_{C(x_i)} \right) \\
&= \frac{2}{n} \left\| \sum_i u_i e_{C(x_i)} \right\|_1 \\
&= \frac{2}{n} \sum_j \left| \sum_i u_i \mathbb{I}[C(x_i) = j] \right| \\
&= \frac{2}{n} \sum_j \left| \sum_{i: C(x_i) = j} u_i \right|
\end{aligned}$$

□

2. **Optimal pair matching design**: Next time