Discussion of "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges"

Panagiotis (Panos) Toulis
University of Chicago, Booth School

January 6, 2017

Notation and problem

- Units indexed by i.
- $W_i \in \{0, 1\}$, binary treatment status.
- $Y_i(1)$ = potential outcome under treatment;
  $Y_i(0)$ = potential outcome under control.
- Causal effect of treatment relative to control:

$$\tau = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(\tau_i).$$

Notation and problem

- Units indexed by i.
- $W_i \in \{0, 1\}$, binary treatment status.
- $Y_i(1)$ = potential outcome under treatment;
  $Y_i(0)$ = potential outcome under control.
- Causal effect of treatment relative to control:

$$\tau = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(\tau_i).$$

- Estimand τ encodes that main difference with machine learning is missing data.

Models

- **Conditional expectation of outcome**:

$$\mu(w, x) = \mathbb{E}(Y_i(w)|X_i = x).$$

- **Propensity score**:

$$e(x) = p(W_i = 1|X_i = x).$$

Models

- **Conditional expectation of outcome**:

$$\mu(w, x) = \mathbb{E}(Y_i(w)|X_i = x).$$

- **Propensity score**:

$$e(x) = p(W_i = 1|X_i = x).$$

- Paper insights: focus on diagnostics and robustness.
- Don't rely on fixed models. Use **supplementary analyses** with any estimator for τ.
- Aim for **credible** estimate that adjusts for differences in covariates X (*..but what does this mean, esp. in high-dimensional problems?*).

Double robustness

- Best estimators in practice combine both models.
- **Doubly robust** estimators adjust directly for $Y \sim X$ and $W \sim X$. If either is consistent, then DR is consistent.

Double robustness

- Best estimators in practice combine both models.
- **Doubly robust** estimators adjust directly for $Y \sim X$ and $W \sim X$. If either is consistent, then DR is consistent.
- DR methods usually do better than simple methods (e.g., inverse weighting) but their performance depends crucially on the missing data mechanism—this issue remains open.
- "Two wrong models are not necessarily better than one" (Kang and Schafer, 2007).

Choice of estimand

- If τ is hard to estimate one idea is to move the goal post and estimate:

$$\tau_\omega = \frac{\mathbb{E}(\omega(X_i) \cdot \tau_i)}{\mathbb{E}(\omega(X_i))}.$$

- If estimates for τ and τ_ω are very different, then report both results.

Choice of estimand

- If τ is hard to estimate one idea is to move the goal post and estimate:

$$\tau_\omega = \frac{\mathbb{E}(\omega(X_i) \cdot \tau_i)}{\mathbb{E}(\omega(X_i))}.$$

- If estimates for τ and τ_ω are very different, then report both results.

- Perhaps implies a diagnostic:

$$\max_{||\omega|| \leq 1} |\widehat{\tau_\omega} - \widehat{\tau}|/\text{se}.$$

Weighting versus balancing

- Weighting by propensity score balances only in expectation. We can also balance in sample:

$$g(\{X_i : W_i = 1\}) = g(\{X_i : W_i = 0\}).$$

Weighting versus balancing

- Weighting by propensity score balances only in expectation. We can also balance in sample:

$$g(\{X_i : W_i = 1\}) = g(\{X_i : W_i = 0\}).$$

- Balancing and propensity scores can be nicely combined under the same framework (Imai and Ratkovic, 2013) since fitting a PS model essentially balances the score function of the model between treatment groups.

Sensitivity

- Mainly need to know which X are important for both W and Y (as in Belloni (2013)). Way to measure such association:

$$B = \mathbb{E}(\textit{sample\_mean\_diff}) - \tau.$$

Sensitivity

- Mainly need to know which X are important for both W and Y (as in Belloni (2013)). Way to measure such association:

$$B = \mathbb{E}(\textit{sample\_mean\_diff}) - \tau.$$

- Can extend this idea further. Suppose we are using a misspecified PS model $e(x)$. Then, imbalance in summary $b(X)$ based on sublassification is:

$$\textit{imbalance} = \mathbb{E}_s\{\text{Cov}(W, b(X)|e(X) = s)\}.$$

Can use such diagnostic to assess PS model, and also optimize marginal balance in **unobserved** covariates. (Toulis & Volfovsky, working paper).

Other considerations

- Most methods depend on *sparsity*, but the L1-loss is not a silver bullet.

Other considerations

- Most methods depend on *sparsity*, but the L1-loss is not a silver bullet.

- 'Tipping-point' ideas in sensitivity analysis may be useful as well (Rosenbaum, 2005).

Other considerations

- Most methods depend on *sparsity*, but the L1-loss is not a silver bullet.
- 'Tipping-point' ideas in sensitivity analysis may be useful as well (Rosenbaum, 2005).
- Some recommendations on checking model misspecification (Athey & Imbens, 2015) are akin to ideas that connect causality with (prediction) **invariance** (Peters et.al., 2015).

Other considerations

- Most methods depend on *sparsity*, but the L1-loss is not a silver bullet.

- 'Tipping-point' ideas in sensitivity analysis may be useful as well (Rosenbaum, 2005).

- Some recommendations on checking model misspecification (Athey & Imbens, 2015) are akin to ideas that connect causality with (prediction) **invariance** (Peters et.al., 2015).

  -Analyze huge observational data (ML is critical).
  -Reduce to logically minimal model (Causal model is critical).
  -Experiment with model and repeat if necessary.