Berry-Esseen-Type Bound for Nonparametric Average Treatment Effect Estimator in Randomized Trials

Hongxiang (David) Qiu

Department of Epidemiology and Biostatistics, Michigan State University

Motivation

- Nonparametric average treatment effect estimators based on semiparametric efficiency theory have been increasingly popular.
  - Use flexible data-adaptive machine learning methods to estimate nuisances such as the outcome model and the propensity score.
  - ***Approximately normal in large samples*** under minimal assumptions on the data-generating process and relatively mild assumptions on nuisance estimators.
  - Asymptotically valid statistical inference follows from, e.g., Wald confidence intervals (CIs).
  - ***Is my sample size large enough (for reliable inference)?***
- Cross-fitting has been increasingly popular.
  - Technically, by splitting the data, cross-fitting drops the so-called ***"Donsker"/"entropy" condition***, which essentially restricts the flexibility of nuisance estimators.
  - Allows generalizable nuisance estimators that might almost interpolate the training data (e.g., deep neural networks).
  - ***Is cross-fitting useless when Donsker condition holds?***
- Many existing methodological frameworks to construct nonparametric estimators of causal effects, e.g., estimating equation, one-step correction, double machine learning, TMLE. Potentially multiple methods to construct CIs.
  - All these estimators are asymptotically normal with the same asymptotic variance under similar conditions.
  - All yield asymptotically valid inference under similar conditions.
  - ***Can we theoretically show that one is better than another?***

Objective

Overarching goal: What is the convergence rate of CI coverage to its nominal coverage?

- A ***distinct question*** from the convergence rate or asymptotic distribution of estimators.
- Concerns the convergence rate of the ***sampling distribution*** to the ***asymptotic distribution***.
- Since statistical inference is a main usage of asymptotic normality, CI coverage is a ***natural follow-up question*** to asymptotic normality.

In this study, we focus on the following simpler (standard) setting:

- Observe n i.i.d. data points consisting of covariate X, binary treatment A, and outcome Y, drawn from true distribution P_*.
- Estimate mean counterfactual outcome $\psi_* \coloneqq E[Y(1)]$. Use Wald CI for statistical inference. Similar for ATE.
- RCT (allowing randomization based on covariate)
  - Standard G-formula identification based on ignorability
  - Known propensity score $\pi_*(x) = P_*(A = 1 \mid X = x)$
- Augmented inverse probability weighted (AIPW) estimator, with or without cross-fitting. Need to estimate the outcome model:

$$Q_*(x) \coloneqq E[Y \mid X = x, A = 1]$$

  The estimator $\hat{Q}$ can be flexible.
- AIPW estimator is asymptotically efficient if $\hat{Q} \to Q_*$.
- AIPW estimator is asymptotically normal as long as $\hat{Q} \to Q_\infty$ for some function Q_∞.

Review of AIPW estimators

Define doubly-robust transformation with known propensity score:

$$\mathcal{T}(Q)(x,a,y) \coloneqq \frac{I(a=1)}{\pi_*(x)}\big(y - Q(x)\big) + Q(x)$$

Non-cross-fit AIPW estimator

$$\tilde{\psi} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}\big(\hat{Q}\big)(X_i, A_i, Y_i)$$

Associated influence function-based asymptotic variance estimator:

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\big\{\mathcal{T}\big(\hat{Q}\big)(X_i, A_i, Y_i) - \tilde{\psi}\big\}^2$$

Nominal $(1-\alpha)$-level Wald CI: $\tilde{\psi} \pm z_{\alpha/2}\,\tilde{\sigma}/\sqrt{n}$.

Cross-fit AIPW estimator (double machine learning)

Split data into K folds of equal size. Let $I_k \subseteq \{1,2,\dots,n\}$ be the index set of fold k, and $\hat{Q}_k$ be the estimator of Q_* based on data out of fold k.

$$\hat{\psi}_k = \frac{1}{|I_k|}\sum_{i\in I_k} \mathcal{T}\big(\hat{Q}_k\big)(X_i, A_i, Y_i)\,, \qquad \hat{\psi} = \frac{1}{K}\sum_{k=1}^{K}\hat{\psi}_k$$

Associated influence function-based asymptotic variance estimator:

$$\hat{\sigma}_k^2 = \frac{1}{|I_k|}\sum_{i\in I_k}\big\{\mathcal{T}\big(\hat{Q}_k\big)(X_i, A_i, Y_i) - \hat{\psi}_k\big\}^2\,, \qquad \hat{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2$$

Nominal $(1-\alpha)$-level Wald CI: $\hat{\psi} \pm z_{\alpha/2}\,\hat{\sigma}/\sqrt{n}$.

Berry-Esseen-type bound

Let $Q_\#$ be any fixed function close to $\hat{Q}$ (e.g., $x \mapsto E[\hat{Q}(x)]$ or limit of $\hat{Q}$) or $\hat{Q}_k$. Define approximate scaled variance of estimator based on $Q_\#$:

$$\sigma_\#^2 \coloneqq E[\{\mathcal{T}(Q_\#)(x,a,y) - \psi_*\}^2]$$

and the mean of asymptotic variance estimator:

$$\sigma_@^2 \coloneqq \begin{cases} E[\tilde{\sigma}^2] & \text{without cross-fitting} \\ E[\hat{\sigma}^2] & \text{with cross-fitting} \end{cases}$$

Let ϕ denote the density of standard Gaussian.

Non-cross-fitting

- Donsker condition: Assume satisfied by VC-hull class with constant envelope M. E.g., $\hat{Q}$ obtained by Highly Adaptive Lasso (HAL).
- Assume $\big\|\hat{Q} - Q_\#\big\|_{P_*,2} = o_p(n^{-1/4})$.
- Used concentration inequality for suprema of empirical processes [1].

$$P\big(\tilde{\psi} - z_{\alpha/2}\,\tilde{\sigma}/\sqrt{n} \le \psi_* \le \tilde{\psi} + z_{\alpha/2}\,\tilde{\sigma}/\sqrt{n}\big)$$
$$= 1 - \alpha + 2\phi\big(z_{\alpha/2}\big)z_{\alpha/2}\frac{\sigma_@ - \sigma_\#}{\sigma_\#} + O\left(\sqrt{\log n\,/n} + \left\{E\big\|\hat{Q} - Q_\#\big\|_{P_*,2}^2\right\}^{1/3}\right)$$
$$+ \underbrace{O\big(R(\delta,\nu,n)\big) + P\left(\big\|\hat{Q} - Q_\#\big\|_{P_*,2} > \delta M\right)}_{\text{additional terms vs. cross-fitting}}$$

where $R(\delta,\nu,n) = \delta^{2/(\nu+2)} + n^{-1/2}\delta^{4/(\nu+2)-2}$, ν is the VC-dimension of the associated VC-class, and $\delta \lesssim n^{-1/4}$. $R(\delta,\nu,n)$ can be replaced by $\delta\sqrt{\log(1/\delta)} + n^{-1/2}\log(1/\delta)$ for VC-type classes.

Cross-fitting

$$P\big(\hat{\psi} - z_{\alpha/2}\hat{\sigma}/\sqrt{n} \le \psi_* \le \hat{\psi} + z_{\alpha/2}\hat{\sigma}/\sqrt{n}\big)$$
$$= 1 - \alpha + 2\phi\big(z_{\alpha/2}\big)z_{\alpha/2}\frac{\sigma_@ - \sigma_\#}{\sigma_\#} + O\left(\sqrt{\log n\,/n} + \left\{E\big\|\hat{Q}_k - Q_\#\big\|_{P_*,2}^2\right\}^{1/3}\right)$$

Green terms can be replaced by e.g. $\sqrt{E\big\|\hat{Q} - Q_\#\big\|_{P_*,2}^2 \log E\big\|\hat{Q} - Q_\#\big\|_{P_*,2}^{-2}}$ under subgaussian assumptions on e.g. $\{\hat{Q}_k(X) - Q_\#(X)\}/\big\|\hat{Q}_k - Q_\#\big\|_{P_*,2}$ (given $\hat{Q}_k$) and $\big\|\hat{Q} - Q_\#\big\|_{P_*,2}/\sqrt{E\big\|\hat{Q} - Q_\#\big\|_{P_*,2}^2}$.

Heuristics on variance estimators' bias

Non-cross-fitting

$$\sigma_@^2 - \sigma_\#^2$$
$$= \underbrace{E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\pi_*(X_i)^2}\big(Y_i - \hat{Q}(X_i)\big)^2\right] - E\left[\frac{I(A=1)}{\pi_*(X)^2}\big(Y - Q_\#(X)\big)^2\right]}_{\text{(I)}}$$
$$+ \underbrace{E\left[\frac{1}{n}\sum_{i=1}^{n}\hat{Q}(X_i)^2\right] - E[Q_\#(X)^2]}_{\text{(II)}} + \underbrace{2E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\pi_*(X_i)}\big(Y_i - \hat{Q}(X_i)\big)\hat{Q}(X_i)\right]}_{\text{(III)}}$$
$$\underbrace{- 2E\left[\frac{I(A=1)}{\pi_*(X)}\big(Y - Q_\#(X)\big)Q_\#(X)\right]}_{\text{(IV)}} - \underbrace{Var\big(\tilde{\psi}\big)}_{\text{order } 1/n}$$

- (I) anticipated to be ≤ 0: When π_* is a constant and $\hat{Q}$ is an empirical MSE minimizer over a function class containing $Q_\#$,

$$\text{(I)} \le E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\pi_*(X_i)^2}\big(Y_i - Q_\#(X_i)\big)^2\right] - E\left[\frac{I(A=1)}{\pi_*(X)^2}\big(Y - Q_\#(X)\big)^2\right] = 0$$

  Also anticipated to be of order $E\big\|\hat{Q} - Q_\#\big\|_{P_*,2}$.
- (II) anticipated to be ≤ 0 if $\hat{Q}$ is shrunk towards 0 or smoothed; otherwise, no clear bias.
- (III) & (IV) anticipated to be ≈ 0: If π_* is a constant, and $\hat{Q}$ and $Q_\#$ are projections, then $\text{(III)} = \text{(IV)} = 0$.

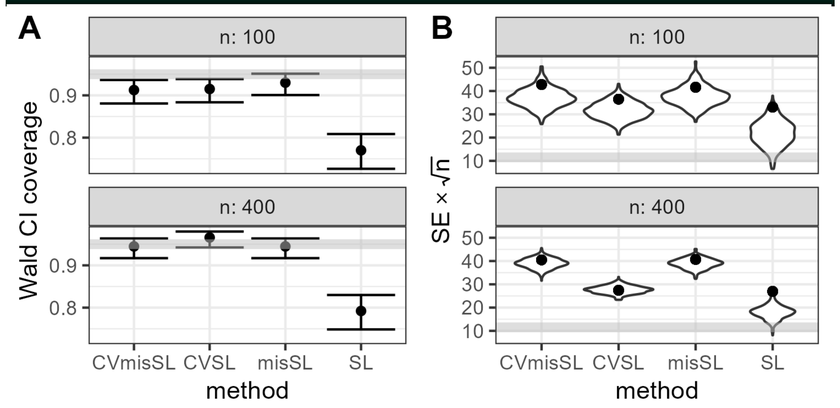
Conclusion: We might anticipate $\sigma_@ < \sigma_\#$ $\Longrightarrow$ Decreased coverage!

Cross-fitting

$$\sigma_@^2 - \sigma_\#^2 = \underbrace{E\int\frac{1-\pi_*(x)}{\pi_*(x)}\big\{\hat{Q}_k(x) - Q_\#(x)\big\}^2 dP_*(x)}_{\text{order } E\|\hat{Q}_k - Q_\#\|_{P_*,2}^2} - \underbrace{Var\big(\hat{\psi}\big)}_{\text{order } 1/n}$$

Conclusion: $E\big\|\hat{Q}_k - Q_\#\big\|_{P_*,2}^2 \gg 1/n \Longrightarrow \sigma_@ > \sigma_\#$ $\Longrightarrow$ Increased coverage!

Simulation & discussion



Estimate ATE in RCT with 7 covariates. CV=20-fold cross-fitting.
$\hat{Q}$: SL=Super Learner+GLM-type+HAL; misSL=Super Learner+GLM-type.
A. Wald CI coverage with 95% CI. Thick gray line: 95% nominal coverage
B. Distribution of scaled standard error. Black dots: Scaled Monte Carlo standard deviation estimate. Thick gray line: efficient asymptotic standard deviation.

- Cross-fitting or simple $\hat{Q}$ has better coverage.
- Non-cross-fitting + complex $\hat{Q}$ ⇒ underestimate $\sigma_\#^2$ ⇒ undercoverage
- Simple misspecified $\hat{Q}$ $\Longrightarrow$ large variance
- Efficient asymptotic variance is poor approximation for moderate n.
- ***Our bound might not be tight.***
- ***A spectrum of complexity***, not just "Donsker vs. non-Donsker"
- Potential ***trade-off*** between efficiency and Wald CI coverage in RCT.

References

[1] Chernozhukov V., Chetverikov D., & Kato K. (2014). Gaussian approximation of suprema of empirical processes. *AoS*, *42*(4), 1564–1597.