Model-Agnostic Berry-Esseen-Type Bounds for Augmented Inverse Probability Weighted Estimators in Randomized Controlled Trials

Hongxiang (David) Qiu

Department of Epidemiology and Biostatistics, Michigan State University

JSM 2025

Table of Contents

Motivation

- Modern non-/semi-parametric estimators have been increasingly popular in causal inference, machine learning, . . .
  - Augmented inverse probability weighting (AIPW)
  - Double/debiased machine leaning (DML)
  - Targeted minimum loss-based estimation (TMLE)
  - . . .
  - Variants: various nuisance estimators, sample-splitting/cross-fitting, calibration, . . .

Motivation

- Modern non-/semi-parametric estimators have been increasingly popular in causal inference, machine learning, . . .
  - Augmented inverse probability weighting (AIPW)
  - Double/debiased machine leaning (DML)
  - Targeted minimum loss-based estimation (TMLE)
  - . . .
  - Variants: various nuisance estimators, sample-splitting/cross-fitting, calibration, . . .
- These estimators share same asymptotic normal distribution under same/similar conditions, but may differ in moderate samples.

Example: Cross-fitting

- Cross-fitting is a technique applicable to many estimators

Example: Cross-fitting

- Cross-fitting is a technique applicable to many estimators
- Numerical simulations have shown that cross-fitting may improve moderate-sample performance (Li et al., 2022; Smith et al., 2024)

Example: Cross-fitting

- Cross-fitting is a technique applicable to many estimators
- Numerical simulations have shown that cross-fitting may improve moderate-sample performance (Li et al., 2022; Smith et al., 2024)
  - Simulations cannot cover all scenarios. . .

Example: Cross-fitting

- Cross-fitting is a technique applicable to many estimators
- Numerical simulations have shown that cross-fitting may improve moderate-sample performance (Li et al., 2022; Smith et al., 2024)
  - Simulations cannot cover all scenarios. . .
- Theoretically, it is widely accepted that cross-fitting improves the estimator by dropping *Donsker conditions* via sample splitting

Example: Cross-fitting

- Cross-fitting is a technique applicable to many estimators
- Numerical simulations have shown that cross-fitting may improve moderate-sample performance (Li et al., 2022; Smith et al., 2024)
  - Simulations cannot cover all scenarios...
- Theoretically, it is widely accepted that cross-fitting improves the estimator by dropping *Donsker conditions* via sample splitting
- What if Donsker conditions are known to hold? Is cross-fitting still better?

Example: Cross-fitting

- Cross-fitting is a technique applicable to many estimators
- Numerical simulations have shown that cross-fitting may improve moderate-sample performance (Li et al., 2022; Smith et al., 2024)
  - Simulations cannot cover all scenarios. . .
- Theoretically, it is widely accepted that cross-fitting improves the estimator by dropping *Donsker conditions* via sample splitting
- What if Donsker conditions are known to hold? Is cross-fitting still better?
- Generally, how can we theoretically compare these estimators and spot their differences in a meaningful way, given that they have the same asymptotic normal distribution?

Objective: Confidence interval (CI) coverage

Overarching goal:

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

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution
- Meaningful question: E.g., can we trust Wald-CIs based on asymptotic normality?

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution
- Meaningful question: E.g., can we trust Wald-CIs based on asymptotic normality?
- For simpler problems (e.g., sample mean), this rate (or its upper bound) is known (Berry-Esseen bound)

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution
- Meaningful question: E.g., can we trust Wald-CIs based on asymptotic normality?
- For simpler problems (e.g., sample mean), this rate (or its upper bound) is known (Berry-Esseen bound)
- To the best of my knowledge, no existing literature directly addresses this question for modern flexible non-/semi-parametric estimators

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution
- Meaningful question: E.g., can we trust Wald-CIs based on asymptotic normality?
- For simpler problems (e.g., sample mean), this rate (or its upper bound) is known (Berry-Esseen bound)
- To the best of my knowledge, no existing literature directly addresses this question for modern flexible non-/semi-parametric estimators
- In this work, consider a simple, yet practical, setting:

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution
- Meaningful question: E.g., can we trust Wald-CIs based on asymptotic normality?
- For simpler problems (e.g., sample mean), this rate (or its upper bound) is known (Berry-Esseen bound)
- To the best of my knowledge, no existing literature directly addresses this question for modern flexible non-/semi-parametric estimators
- In this work, consider a simple, yet practical, setting:
  - AIPW estimator in randomized controlled trials (RCTs)

Objective: Confidence interval (CI) coverage

Overarching goal:

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

- Distinct question from the estimator's convergence rate or asymptotic distribution
- Meaningful question: E.g., can we trust Wald-CIs based on asymptotic normality?
- For simpler problems (e.g., sample mean), this rate (or its upper bound) is known (Berry-Esseen bound)
- To the best of my knowledge, no existing literature directly addresses this question for modern flexible non-/semi-parametric estimators
- In this work, consider a simple, yet practical, setting:
  - AIPW estimator in randomized controlled trials (RCTs)
  - Wald-CI with plug-in influence function-based standard error (SE)

Table of Contents

Setup

- Data: n iid data drawn from P_*

Setup

- Data: n iid data drawn from P_*
  - X: baseline covariates

Setup

- Data: n iid data drawn from P_*
  - X: baseline covariates
  - A: randomized (may depend on X) binary treatment

Setup

- Data: n iid data drawn from P_*
  - X: baseline covariates
  - A: randomized (may depend on X) binary treatment
  - Y: real-valued outcome

Setup

- Data: n iid data drawn from P_*
  - X: baseline covariates
  - A: randomized (may depend on X) binary treatment
  - Y: real-valued outcome
- Estimand: mean counterfactual outcome $\psi_* := \mathbb{E}[Y^1]$ (average treatment effect is similar)

Setup

- Data: n iid data drawn from P_*
  - X: baseline covariates
  - A: randomized (may depend on X) binary treatment
  - Y: real-valued outcome
- Estimand: mean counterfactual outcome $\psi_* := \mathbb{E}[Y^1]$ (average treatment effect is similar)
- Propensity score $\pi_*(X) := \Pr(A = 1 \mid X)$ is known

Setup

- Data: n iid data drawn from P_*
  - X: baseline covariates
  - A: randomized (may depend on X) binary treatment
  - Y: real-valued outcome
- Estimand: mean counterfactual outcome $\psi_* := \mathbb{E}[Y^1]$ (average treatment effect is similar)
- Propensity score $\pi_*(X) := \Pr(A = 1 \mid X)$ is known
- Outcome model $Q_*(X) := \mathbb{E}[Y \mid X, A = 1]$ is unknown and may be estimated flexibly

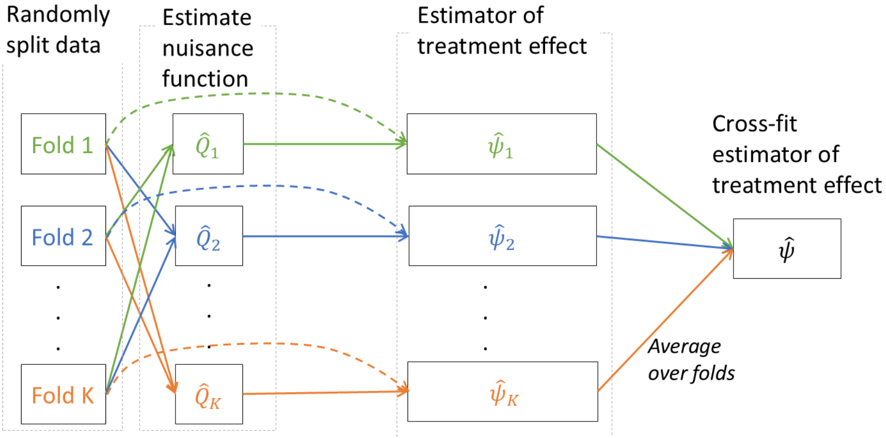
Review non-cross-fit AIPW estimator

Doubly-robust transformation (uncentered influence function):

$$\mathcal{T}(Q)(x, a, y) := \frac{a}{\pi_*(x)}(y - Q(x)) + Q(x)$$

1. Estimate Q_* with a flexible estimator $\hat{Q}$
2. $\tilde{\psi} := \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}(\hat{Q})(X_i, A_i, Y_i)$
3. Plug-in asymptotic variance estimator: $\tilde{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \{\mathcal{T}(\hat{Q})(X_i, A_i, Y_i) - \tilde{\psi}\}^2$
4. Nominal $(1-\alpha)$-level Wald-CI: $\tilde{\psi} \pm z_{\alpha/2}\tilde{\sigma}/\sqrt{n}$

Review cross-fit AIPW estimator



Review cross-fit AIPW estimator

1. Split data into K folds of equal size. Let I_k be the index set of fold k.
2. For each fold k,
   a) Estimate Q_* with a flexible estimator $\hat{Q}_k$ using data out of fold k
   b) $\hat{\psi}_k := \frac{1}{|I_k|} \sum_{i \in I_k} \mathcal{T}(\hat{Q}_k)(X_i, A_i, Y_i)$
   c) $\hat{\sigma}_k^2 := \frac{1}{|I_k|} \sum_{i \in I_k} \{\mathcal{T}(\hat{Q}_k)(X_i, A_i, Y_i) - \hat{\psi}_k\}^2$
3. Combine all folds: $\hat{\psi} := \frac{1}{K} \sum_{k=1}^{K} \hat{\psi}_k$, $\hat{\sigma}^2 := \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2$
4. Nominal $(1-\alpha)$-level Wald-CI: $\hat{\psi} \pm z_{\alpha/2} \hat{\sigma} / \sqrt{n}$

Review of asymptotic properties

Because of known propensity score (i.e., randomization), AIPW estimator is more robust than in observational settings (Rubin & Van Der Laan, 2008).

- If $\hat{Q}$ ($\hat{Q}_k$) converges to Q_* (regardless of rates), then $\tilde{\psi}$ ($\hat{\psi}$) is asymptotically efficient
- If $\hat{Q}$ ($\hat{Q}_k$) converges to some function Q_∞, then $\tilde{\psi}$ ($\hat{\psi}$) is asymptotically normal

(Assuming Donsker conditions for $\tilde{\psi}$)

Table of Contents

Notations

- Let $Q_{\#}$ be any fixed function that may depend on n that is close to $\hat{Q}$, e.g., $x \mapsto \mathbb{E}[\hat{Q}(x)]$.

Notations

- Let $Q_{\#}$ be any fixed function that may depend on n that is close to $\hat{Q}$, e.g., $x \mapsto \mathbb{E}[\hat{Q}(x)]$.
  - Better non-asymptotic approximation than the limit Q_∞ of $\hat{Q}$

Notations

- Let $Q_{\#}$ be any fixed function that may depend on n that is close to $\hat{Q}$, e.g., $x \mapsto \mathbb{E}[\hat{Q}(x)]$.
  - Better non-asymptotic approximation than the limit Q_{∞} of $\hat{Q}$
- Approximate scaled variance based on $Q_{\#}$:

$$\sigma_{\#}^2 := \mathbb{E}[\{\mathcal{T}(Q_{\#})(X, A, Y) - \psi_*\}^2]$$

Notations

- Let $Q_{\#}$ be any fixed function that may depend on n that is close to $\hat{Q}$, e.g., $x \mapsto \mathbb{E}[\hat{Q}(x)]$.
  - Better non-asymptotic approximation than the limit Q_{∞} of $\hat{Q}$
- Approximate scaled variance based on $Q_{\#}$:

$$\sigma_{\#}^2 := \mathbb{E}[\{\mathcal{T}(Q_{\#})(X, A, Y) - \psi_*\}^2]$$

- Expectation of asymptotic variance estimator:

$$\sigma_{\dagger}^2 := \begin{cases} \mathbb{E}[\tilde{\sigma}^2] & \text{non-cross-fit} \\ \mathbb{E}[\hat{\sigma}^2] & \text{cross-fit} \end{cases}$$

Notations

- Let $Q_{\#}$ be any fixed function that may depend on n that is close to $\hat{Q}$, e.g., $x \mapsto \mathbb{E}[\hat{Q}(x)]$.
  - Better non-asymptotic approximation than the limit Q_{∞} of $\hat{Q}$
- Approximate scaled variance based on $Q_{\#}$:

$$\sigma_{\#}^2 := \mathbb{E}[\{\mathcal{T}(Q_{\#})(X, A, Y) - \psi_*\}^2]$$

- Expectation of asymptotic variance estimator:

$$\sigma_{\dagger}^2 := \begin{cases} \mathbb{E}[\tilde{\sigma}^2] & \text{non-cross-fit} \\ \mathbb{E}[\hat{\sigma}^2] & \text{cross-fit} \end{cases}$$

- ϕ: standard Gaussian density

Cross-fit

$$
\begin{aligned}
&\Pr(\hat{\psi} - z_{\alpha/2}\hat{\sigma}/\sqrt{n} \leq \psi_* \leq \hat{\psi} + z_{\alpha/2}\hat{\sigma}/\sqrt{n}) \\
&= 1 - \alpha + 2\phi(z_{\alpha/2}) z_{\alpha/2} \frac{\sigma_\dagger - \sigma_\#}{\sigma_\#} + \mathrm{O}\left(\sqrt{\frac{\log n}{n}} + \left\{ \mathbb{E} \| \hat{Q}_k - Q_\# \|^2_{P_*,2} \right\}^{1/3} \right)
\end{aligned}
$$

The constants in the O-term depend on P_* and are omitted here.

Cross-fit

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

The constants in the O-term depend on P_* and are omitted here.

- This form is somewhat deceiving: The green rate is the slowest

Cross-fit

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

The constants in the O-term depend on P_* and are omitted here.

- This form is somewhat deceiving: The green rate is the slowest
- Under subgaussian assumptions on $\{\hat{Q}_k(X) - Q_\#(X)\}/\|\hat{Q}_k - Q_\#\|_{P_*,2}$ (given $\hat{Q}_k$) and $\|\hat{Q}_k - Q_\#\|_{P_*,2}/\sqrt{\mathbb{E}\|\hat{Q}_k - Q_\#\|^2_{P_*,2}}$ etc., the green rate can be replaced by a faster rate $\sqrt{\mathbb{E}\|\hat{Q}_k - Q_\#\|^2_{P_*,2} \log \|\hat{Q}_k - Q_\#\|^{-2}_{P_*,2}}$, comparable to the rate of $\sigma_\dagger - \sigma_\#$ except for a log factor.

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)
- Assume $\|\hat{Q} - Q_{\#}\|_{P_*,2} = o_p(n^{-1/4})$

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)
- Assume $\|\hat{Q} - Q_{\#}\|_{P_*,2} = o_p(n^{-1/4})$
  - A common rate requirement for AIPW estimator

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)
- Assume $\|\hat{Q} - Q_{\#}\|_{P_*,2} = o_p(n^{-1/4})$
  - A common rate requirement for AIPW estimator
  - Often satisfied if $\hat{Q}$ minimizes an empirical risk over a Donsker class

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)
- Assume $\|\hat{Q} - Q_{\#}\|_{P_*,2} = o_p(n^{-1/4})$
  - A common rate requirement for AIPW estimator
  - Often satisfied if $\hat{Q}$ minimizes an empirical risk over a Donsker class
- If using a VC-hull-type class, let ν be the VC-dimension of the associated VC-class

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)
- Assume $\|\hat{Q} - Q_{\#}\|_{P_*,2} = o_p(n^{-1/4})$
  - A common rate requirement for AIPW estimator
  - Often satisfied if $\hat{Q}$ minimizes an empirical risk over a Donsker class
- If using a VC-hull-type class, let ν be the VC-dimension of the associated VC-class
- Let $\delta \lesssim n^{-1/4}$

Non-cross-fit: more on Donsker condition

- Donsker conditions are needed for asymptotic normality without cross-fitting.
- Assume that Donsker conditions are satisfied by a VC-hull-type or a VC-type class with a constant envelope M
  - Highly Adaptive Lasso (HAL)
- Assume $\|\hat{Q} - Q_{\#}\|_{P_*,2} = o_p(n^{-1/4})$
  - A common rate requirement for AIPW estimator
  - Often satisfied if $\hat{Q}$ minimizes an empirical risk over a Donsker class
- If using a VC-hull-type class, let ν be the VC-dimension of the associated VC-class
- Let $\delta \lesssim n^{-1/4}$
- Used the concentration inequality for suprema of empirical processes in Chernozhukov et al. (2014)

Non-cross-fit

$$\Pr(\tilde{\psi} - z_{\alpha/2}\tilde{\sigma}/\sqrt{n} \leq \psi_* \leq \tilde{\psi} + z_{\alpha/2}\tilde{\sigma}/\sqrt{n})$$
$$= 1 - \alpha + 2\phi(z_{\alpha/2}) z_{\alpha/2} \frac{\sigma_\dagger - \sigma_\#}{\sigma_\#} + \mathrm{O}\left(\sqrt{\frac{\log n}{n}} + \left\{ \mathbb{E} \| \hat{Q} - Q_\# \|^2_{P_*,2} \right\}^{1/3} \right)$$
$$+ \underbrace{\mathrm{O}(R(\delta, \nu, n)) + \Pr(\| \hat{Q} - Q_\# \|_{P_*,2} > \delta M)}_{\text{additional terms compared to cross-fitting}}$$

where

$$R(\delta, \nu, n) = \begin{cases} \delta^{2/(\nu+2)} + n^{-1/2} \delta^{4/(\nu+2)-2} & \text{VC-hull-type class} \\ \delta \sqrt{\log \delta^{-1}} + n^{-1/2} \log \delta^{-1} & \text{VC-type class} \end{cases}$$

The green rate can be replaced by a faster rate $\sqrt{\mathbb{E} \| \hat{Q} - Q_\# \|^2_{P_*,2} \log \| \hat{Q} - Q_\# \|^{-2}_{P_*,2}}$ under similar subgaussian assumptions.

Non-cross-fit

Explicit effect of function class complexity:

- If the function class is rich (VC-hull-type with moderate-to-large ν), $R(\delta, \nu, n)$ and the green rate are the slowest
- If the function class is not as rich (VC-type), then $R(\delta, \nu, n)$ does not dominate
- Note that $R(\delta, \nu, n)$ might have room for improvement
  - $R(\delta, \nu, n)$ arises from empirical processes, the deviation of $\tilde{\psi}$ from a sample mean
  - Simulations suggest that non-cross-fit estimator can be fairly close to normal

Non-cross-fit

Explicit effect of function class complexity:

- If the function class is rich (VC-hull-type with moderate-to-large ν), $R(\delta, \nu, n)$ and the green rate are the slowest
- If the function class is not as rich (VC-type), then $R(\delta, \nu, n)$ does not dominate
- Note that $R(\delta, \nu, n)$ might have room for improvement
  - $R(\delta, \nu, n)$ arises from empirical processes, the deviation of $\tilde{\psi}$ from a sample mean
  - Simulations suggest that non-cross-fit estimator can be fairly close to normal

Implicit effect of function class complexity: The L_2-convergence rate may depend on the function class complexity.

Non-cross-fit

Explicit effect of function class complexity:

- If the function class is rich (VC-hull-type with moderate-to-large ν), $R(\delta, \nu, n)$ and the green rate are the slowest
- If the function class is not as rich (VC-type), then $R(\delta, \nu, n)$ does not dominate
- Note that $R(\delta, \nu, n)$ might have room for improvement
    - $R(\delta, \nu, n)$ arises from empirical processes, the deviation of $\tilde{\psi}$ from a sample mean
    - Simulations suggest that non-cross-fit estimator can be fairly close to normal

Implicit effect of function class complexity: The L_2-convergence rate may depend on the function class complexity.

Effect of asymptotic variance estimator's bias $\sigma_{\dagger} - \sigma_{\#}$: It could systematically affect Wald-CI coverage, especially if

- $R(\delta, \nu, n)$ can be improved with sharper empirical process bounds, and
- subgaussian assumptions are satisfied so that the green term is somewhat comparable to the rate of $\sigma_{\dagger} - \sigma_{\#}$

Table of Contents

Cross-fit

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

If we use flexible $\hat{Q}_k$, we often anticipate $\mathbb{E}\|\hat{Q}_k - Q_\#\|_{P_*,2}^2$ to be much slower than n^{-1}, so we anticipate $\sigma_\dagger^2 - \sigma_\#^2 > 0$, i.e., increased coverage.

Non-cross-fit

$$\sigma_\dagger^2 - \sigma_\#^2 = \underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\pi_*(X_i)^2}(Y_i - \hat{Q}(X_i))^2\right] - \mathbb{E}\left[\frac{A}{\pi_*(X)^2}(Y - Q_\#(X))^2\right]}_{(I)}$$

$$+ \underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{Q}(X_i)^2\right] - \mathbb{E}[Q_\#(X)^2]}_{(II)}$$

$$+ \underbrace{2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\pi_*(X_i)}(Y_i - \hat{Q}(X_i))\hat{Q}(X_i)\right]}_{(III)} - \underbrace{2\mathbb{E}\left[\frac{A}{\pi_*(X)}(Y - Q_\#(X))Q_\#(X)\right]}_{(IV)}$$

$$- \underbrace{\mathrm{Var}(\tilde{\psi})}_{\text{order } n^{-1}}$$

Non-cross-fit

Analysis of each term:

(I) Anticipated to be ≤ 0 and of order $\mathbb{E}\|\hat{Q} - Q_\#\|_{P_*,2}$: When π_* is a constant and $\hat{Q}$ is an empirical MSE minimizer over a function class containing $Q_\#$,

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

(II) Anticipated to be ≤ 0 if $\hat{Q}$ is shrunk towards 0 or smoothed; otherwise, no obvious bias

(III) & (IV) Anticipated to be ≈ 0: If π_* is a constant, and $\hat{Q}$ and $Q_\#$ are projections, then $(III) = (IV) = 0$.

Non-cross-fit

Analysis of each term:

(I) Anticipated to be ≤ 0 and of order $\mathbb{E}\|\hat{Q} - Q_{\#}\|_{P_*,2}$: When π_* is a constant and $\hat{Q}$ is an empirical MSE minimizer over a function class containing $Q_{\#}$,

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

(II) Anticipated to be ≤ 0 if $\hat{Q}$ is shrunk towards 0 or smoothed; otherwise, no obvious bias

(III) & (IV) Anticipated to be ≈ 0: If π_* is a constant, and $\hat{Q}$ and $Q_{\#}$ are projections, then $(III) = (IV) = 0$.

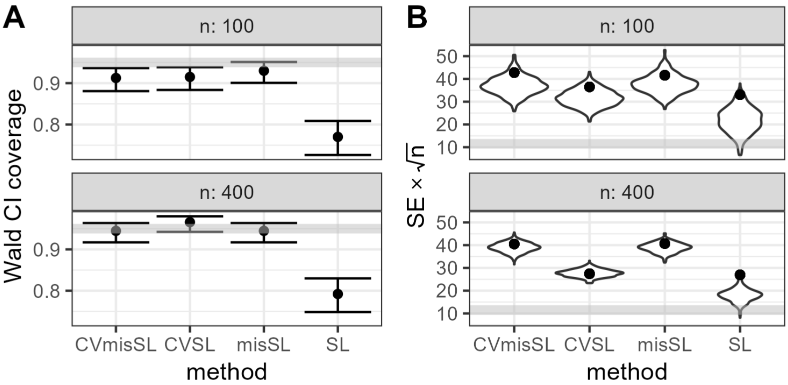
If we use flexible $\hat{Q}$, we often anticipate $\mathbb{E}\|\hat{Q} - Q_{\#}\|_{P_*,2}^2$ to be much slower than n^{-1}, so we might anticipate $\sigma_{\dagger}^2 - \sigma_{\#}^2 < 0$, i.e., decreased coverage.

Table of Contents

Setup

- Estimate average treatment effect in RCT with 7 covariates
- Very complicated true outcome model Q_*
- Small to moderate samples: $n = 100, 400$
- CV: 20-fold cross-fitting
- $\hat{Q}$: SL = Super Learner + GLM-type + HAL; misSL = Super Learner + GLM-type

Results



(B): Dots are Monte Carlo estimates of estimators' standard deviations. Thick gray line is efficient standard deviation.

Interpretations

- With $\hat{Q}$ closer to the truth Q_*, we gain more efficiency.
- Cross-fitting or simple $\hat{Q}$ yields better Wald-CI coverage
- Non-cross-fitting and flexible $\hat{Q}$ (SL): underestimate $\sigma^2_{\#} \implies$ undercoverage
- Cross-fitting and flexible $\hat{Q}$ (CVSL): overestimate $\sigma^2_{\#} \implies$ overcoverage (?)
- Efficient asymptotic variance is a poor approximation to the variance of SL/CVSL for moderate n

Discussion

- **These bounds might not be tight**

Discussion

- **These bounds might not be tight**
  - A smaller upper bound does not imply actual faster rate

Discussion

- **These bounds might not be tight**
  - A smaller upper bound does not imply actual faster rate
  - The bounds might be improved with more information on $\hat{Q}$ and better proof techniques

Discussion

- **These bounds might not be tight**
  - A smaller upper bound does not imply actual faster rate
  - The bounds might be improved with more information on $\hat{Q}$ and better proof techniques
- **A spectrum of complexity**: not just "Donsker vs. non-Donsker"

Discussion

- **These bounds might not be tight**
  - A smaller upper bound does not imply actual faster rate
  - The bounds might be improved with more information on $\hat{Q}$ and better proof techniques
- **A spectrum of complexity**: not just "Donsker vs. non-Donsker"
- Cross-fitting can outperform non-cross-fitting, even if Donsker conditions hold

Discussion

- **These bounds might not be tight**
  - A smaller upper bound does not imply actual faster rate
  - The bounds might be improved with more information on $\hat{Q}$ and better proof techniques
- **A spectrum of complexity**: not just "Donsker vs. non-Donsker"
- Cross-fitting can outperform non-cross-fitting, even if Donsker conditions hold
- Potential **trade-off** between efficiency and Wald-CI coverage in RCT

Discussion

- **These bounds might not be tight**
  - A smaller upper bound does not imply actual faster rate
  - The bounds might be improved with more information on $\hat{Q}$ and better proof techniques
- **A spectrum of complexity**: not just "Donsker vs. non-Donsker"
- Cross-fitting can outperform non-cross-fitting, even if Donsker conditions hold
- Potential **trade-off** between efficiency and Wald-CI coverage in RCT
  - For more efficiency,

$$\begin{aligned} & \text{more flexible } \hat{Q} \text{ to approximate complicated truth } Q_* \\ \Longrightarrow\ & \text{slower } \mathbb{E}\|\hat{Q} - Q_\#\|_{P_*,2}^2 \\ \Longrightarrow\ & \text{slower convergence of Wald-CI coverage to its nominal coverage} \end{aligned}$$

References I

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

Li, H., Rosete, S., Coyle, J., Phillips, R. V., Hejazi, N. S., Malenica, I., Arnold, B. F., Benjamin-Chung, J., Mertens, A., Colford, J. M., van der Laan, M. J., & Hubbard, A. E. (2022). Evaluating the robustness of targeted maximum likelihood estimators via realistic simulations in nutrition intervention trials. *Statistics in Medicine*, 41(12), 2132–2165.

Rubin, D. B. & Van Der Laan, M. J. (2008). Covariate Adjustment for the Intention-to-Treat Parameter with Empirical Efficiency Maximization. *UCB Division of Biostatistics Working Paper*, 229.

Smith, M. J., Phillips, R. V., Maringe, C., & Fernandez, M. A. L. (2024). Performance of Cross-Validated Targeted Maximum Likelihood Estimation. *arXiv preprint arXiv:2409.11265v1*.