Specification-robust Causal Inference

Aditya Ghosh

Statistics Ph.D. student, Stanford University

Joint work with Dominik Rothenhäusler

Introduction

- Consider estimation and inference on the average treatment effect (ATE) in observational studies
- For causal identification we often need covariate adjustment to remove confounding

Introduction

- Consider estimation and inference on the average treatment effect (ATE) in observational studies
- For causal identification we often need covariate adjustment to remove confounding
- A subset of covariates X_S (with columns in S) is a *valid adjustment set* if

$$\underbrace{Y(1), Y(0)}_{\text{Potential outcomes}} \perp\!\!\!\perp \underbrace{A}_{\text{Treatment}} \mid \underbrace{X_S}_{\text{Covariates}} \qquad \text{(Ignorability)}$$

Introduction

- Consider estimation and inference on the average treatment effect (ATE) in observational studies
- For causal identification we often need covariate adjustment to remove confounding
- A subset of covariates X_S (with columns in S) is a *valid adjustment set* if

$$\underbrace{Y(1), Y(0)}_{\text{Potential outcomes}} \perp\!\!\!\perp \underbrace{A}_{\text{Treatment}} \mid \underbrace{X_S}_{\text{Covariates}} \qquad \text{(Ignorability)}$$

> Valid causal inference is *challenging* when *multiple adjustment sets* appear plausible, relying on different assumptions on the causal structure

⚠ Cannot detect which adjustment set is valid (ignorability is untestable)

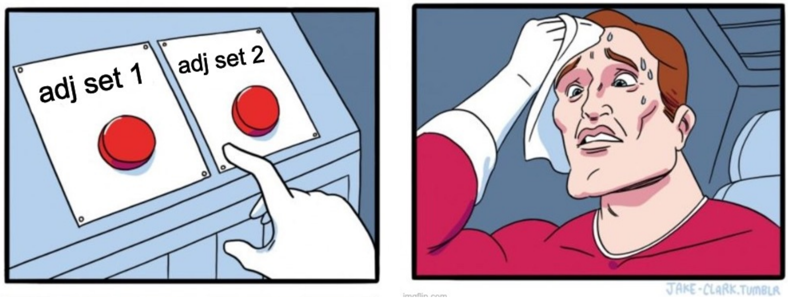


Figure 1: All candidate adjustment sets appear plausible; which one to use?

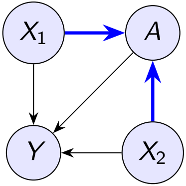
[1] VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4), 1406-1413.

Figure 1: All candidate adjustment sets appear plausible; which one to use?

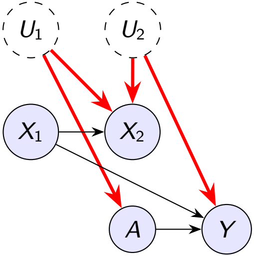
⚠ Our premise is different from standard confounder selection problems, e.g., VanderWeele and Shpitser (2011)[1].

[1]VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4), 1406-1413.

Uncertainty in adjustment sets cannot be ignored

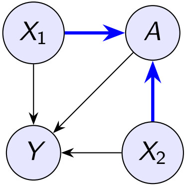


Example 1: Two confounders

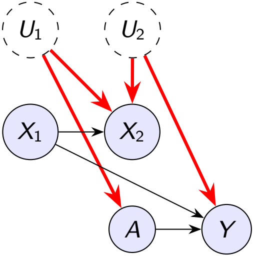


Example 2: X_2 is a non-confounder

Uncertainty in adjustment sets cannot be ignored



Example 1: Two confounders



Example 2: X_2 is a non-confounder

Table 1: Invalid adj. set can lead to zero coverage instead of nominal coverage

	Adjust for $\{X_1\}$		Adjust for $\{X_1, X_2\}$	
	coverage	width	coverage	width
Example 1	0.0%	0.518	95.2%	0.395
Example 2	94.9%	0.351	0.0%	0.328

How to make specification-robust inference?

Goal: Valid inference as long as at least one of the adjustment sets is valid

How to make specification-robust inference?

Goal: Valid inference as long as at least one of the adjustment sets is valid

- Invalid approach 1: Use the **intersection** of the candidate adjustment sets
  — Fails in absence of ignorability (Example 1)
- Invalid approach 2: Use the **union** of all candidate adjustment sets
  — Fails when we have non-confounders (Example 2)

Table 2: Union/intersection of candidate adj. sets might fail

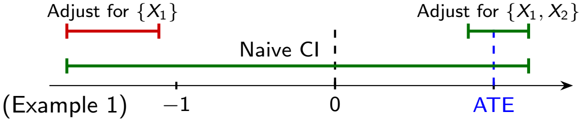
	Adjust for $\{X_1\}$		Adjust for $\{X_1, X_2\}$		
	coverage	width	coverage	width	
Example 1	0.0%	0.518	95.2%	0.395	intersection does not work
Example 2	94.9%	0.351	0.0%	0.328	union does not work

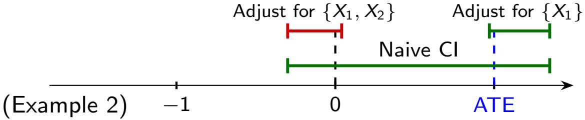
How to make specification-robust inference?

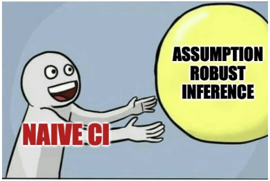
Naive approach: Take convex hull of all the CIs.

How to make specification-robust inference?

Naive approach: Take convex hull of all the CIs.



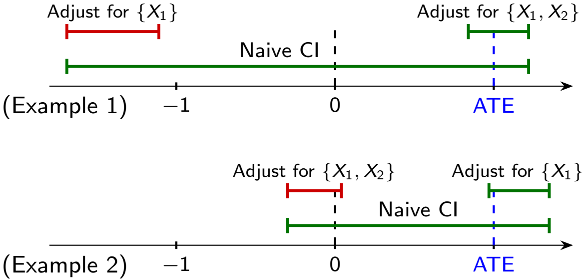




Naive CI is valid as long as one of the candidate adjustment sets is valid.

How to make specification-robust inference?

Naive approach: Take convex hull of all the CIs.



⚠ Even in large samples, width of Naive CI can be $O(1)$

Why are Naive CIs so wide?

- Assume that we have K adjustment sets $S_1, \ldots, S_K$
- For each k, off-the-shelf CATE estimators identify the contrasts

$$\tau(X_{S_k}) := \mathbb{E}_P[Y \mid A = 1, X_{S_k}] - \mathbb{E}_P[Y \mid A = 0, X_{S_k}]$$
$$\neq \mathbb{E}_P[Y(1) - Y(0) \mid X_{S_k}] \qquad \text{(unless } S_k \text{ is a valid adj. set)}$$

Why are Naive CIs so wide?

- Assume that we have K adjustment sets $S_1, \ldots, S_K$
- For each k, off-the-shelf CATE estimators identify the contrasts

$$\tau(X_{S_k}) := \mathbb{E}_P[Y \mid A = 1, X_{S_k}] - \mathbb{E}_P[Y \mid A = 0, X_{S_k}]$$
$$\neq \mathbb{E}_P[Y(1) - Y(0) \mid X_{S_k}] \qquad \text{(unless } S_k \text{ is a valid adj. set)}$$

- The CI for adj. set S_k is usually centered around $\overline{\tau}_k = \mathbb{E}_P[\tau(X_{S_k})]$
  ⚠ Note, $\overline{\tau}_k$ has no causal interpretation for an invalid adjustment set S_k

Why are Naive CIs so wide?

- Assume that we have K adjustment sets $S_1, \ldots, S_K$
- For each k, off-the-shelf CATE estimators identify the contrasts

$$\tau(X_{S_k}) := \mathbb{E}_P[Y \mid A = 1, X_{S_k}] - \mathbb{E}_P[Y \mid A = 0, X_{S_k}]$$

$$\neq \mathbb{E}_P[Y(1) - Y(0) \mid X_{S_k}] \qquad \text{(unless } S_k \text{ is a valid adj. set)}$$

- The CI for adj. set S_k is usually centered around $\overline{\tau}_k = \mathbb{E}_P[\tau(X_{S_k})]$
  ⚠ Note, $\overline{\tau}_k$ has no causal interpretation for an invalid adjustment set S_k
- The naive CI shrinks to $\sim \mathrm{Conv}\{\overline{\tau}_1, \ldots, \overline{\tau}_K\}$ in large samples.
  Hence the naive CI, though valid, might be of constant order!

Why are Naive CIs so wide?

- Assume that we have K adjustment sets $S_1, \ldots, S_K$
- For each k, off-the-shelf CATE estimators identify the contrasts

$$\tau(X_{S_k}) := \mathbb{E}_P[Y \mid A = 1, X_{S_k}] - \mathbb{E}_P[Y \mid A = 0, X_{S_k}]$$
$$\neq \mathbb{E}_P[Y(1) - Y(0) \mid X_{S_k}] \qquad \text{(unless } S_k \text{ is a valid adj. set)}$$

- The CI for adj. set S_k is usually centered around $\overline{\tau}_k = \mathbb{E}_P[\tau(X_{S_k})]$
  ⚠ Note, $\overline{\tau}_k$ has no causal interpretation for an invalid adjustment set S_k
- The naive CI shrinks to $\sim \mathrm{Conv}\{\overline{\tau}_1, \ldots, \overline{\tau}_K\}$ in large samples.
  Hence the naive CI, though valid, might be of constant order!

We present a novel solution to obtain a single CI which:

- ✔ is valid as long as one of the adjustment sets is valid, and
- ✔ has width shrinking at fast $n^{-1/2}$ rate with sample size

Our approach

Unable to pick a valid adjustment set $\implies$ cannot do better than $O(1)$ for the ATE of the original population

Unable to pick a valid adjustment set $\implies$ cannot do better than $O(1)$ for the ATE of the original population

We find a new target population *as close as possible* to the original population, for which all candidate adjustment sets target the same estimand

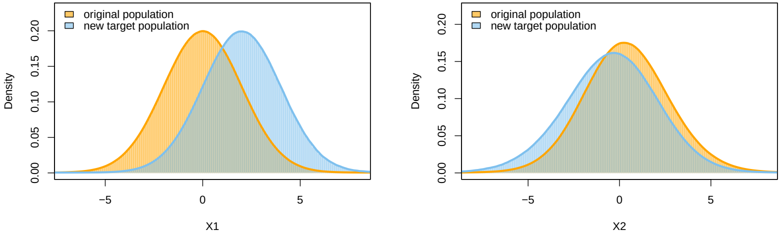


Figure 2: Visualizing the reweighting through the distributions of covariates.

Key idea (with notation)

- Assume that we have K adjustment sets $S_1, \ldots, S_K$, at least one valid
- For each j, we can identify the contrasts

$$\tau(X_{S_j}) := \mathbb{E}_P[Y \mid A = 1, X_{S_j}] - \mathbb{E}_P[Y \mid A = 0, X_{S_j}]$$

Key idea (with notation)

- Assume that we have K adjustment sets $S_1, \ldots, S_K$, at least one valid
- For each j, we can identify the contrasts

$$\tau(X_{S_j}) := \mathbb{E}_P[Y \mid A = 1, X_{S_j}] - \mathbb{E}_P[Y \mid A = 0, X_{S_j}]$$

We find a population $w^* P$ *as close as possible* to the original population P such that all adjustment sets target the *same estimand*:

Key idea (with notation)

- Assume that we have K adjustment sets $S_1, \ldots, S_K$, at least one valid
- For each j, we can identify the contrasts

$$\tau(X_{S_j}) := \mathbb{E}_P[Y \mid A = 1, X_{S_j}] - \mathbb{E}_P[Y \mid A = 0, X_{S_j}]$$

We find a population $w^* P$ *as close as possible* to the original population P such that all adjustment sets target the *same estimand*:

$$w^* = \operatorname*{argmin}_{w = w(X_{S_\cap})} D_{\mathsf{KL}}(wP \,\|\, P) \qquad (w \text{ is the likelihood ratio})$$

$$\text{subject to } \mathbb{E}_{wP}[\tau(X_{S_1})] = \mathbb{E}_{wP}[\tau(X_{S_2})] = \cdots = \mathbb{E}_{wP}[\tau(X_{S_k})].$$

Note, the weights only depend on the common covariates $X_{S_\cap}$ (when one of the adjustment sets is valid, this set does not contain any non-confounder)

Key idea (with notation)

- Assume that we have K adjustment sets $S_1, \ldots, S_K$, at least one valid
- For each j, we can identify the contrasts

$$\tau(X_{S_j}) := \mathbb{E}_P[Y \mid A = 1, X_{S_j}] - \mathbb{E}_P[Y \mid A = 0, X_{S_j}]$$

We find a population $w^* P$ *as close as possible* to the original population P such that all adjustment sets target the *same estimand*:

$$w^* = \underset{w = w(X_{S_\cap})}{\operatorname{argmin}} \; D_{\mathrm{KL}}(wP \,\|\, P) \qquad (w \text{ is the likelihood ratio})$$

$$\text{subject to} \quad \mathbb{E}_{wP}[\tau(X_{S_1})] = \mathbb{E}_{wP}[\tau(X_{S_2})] = \cdots = \mathbb{E}_{wP}[\tau(X_{S_k})].$$

⚠ Does this new estimand has a causal interpretation?

Key idea (with notation)

- Assume that we have K adjustment sets $S_1, \ldots, S_K$, at least one valid
- For each j, we can identify the contrasts

$$\tau(X_{S_j}) := \mathbb{E}_P[Y \mid A = 1, X_{S_j}] - \mathbb{E}_P[Y \mid A = 0, X_{S_j}]$$

We find a population w^*P *as close as possible* to the original population P such that all adjustment sets target the *same estimand*:

$$w^* = \underset{w = w(X_{S_\cap})}{\text{argmin}} \; D_{\mathrm{KL}}(wP \,\|\, P) \qquad (w \text{ is the likelihood ratio})$$

$$\text{subject to} \quad \mathbb{E}_{wP}[\tau(X_{S_1})] = \mathbb{E}_{wP}[\tau(X_{S_2})] = \cdots = \mathbb{E}_{wP}[\tau(X_{S_k})].$$

The common estimand is the ATE of the new target population!

An illustration using real-world data

401(k) data[2]; ATE: effect of 401(k) eligibility on net financial assets

Example. Consider the following candidate adjustment sets

$$S_1 = \{\text{age, education}\},$$
$$S_2 = S_1 \cup \{\text{family size, income}\},$$
$$S_3 = S_2 \cup \{\text{marital status, two-earner household}\},$$
$$S_4 = S_3 \cup \{\text{home-ownership, defined pension, IRA participation}\}.$$

[2]Chernozhukov et al. (2017, Double/Debiased machine learning of treatment and causal parameters); Abadie (2003, Semiparametric instrumental variable estimation of treatment response models)

An illustration using real-world data

401(k) data[2]; ATE: effect of 401(k) eligibility on net financial assets

Example. Consider the following candidate adjustment sets

$$S_1 = \{\text{age, education}\},$$
$$S_2 = S_1 \cup \{\text{family size, income}\},$$
$$S_3 = S_2 \cup \{\text{marital status, two-earner household}\},$$
$$S_4 = S_3 \cup \{\text{home-ownership, defined pension, IRA participation}\}.$$

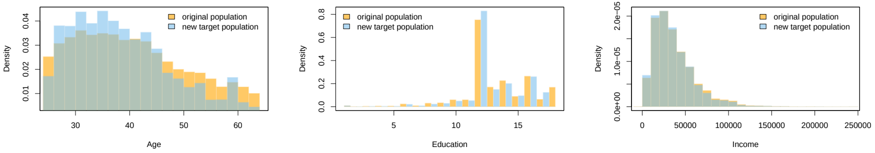


Figure 3: Distribution of the covariates age, income and education for the original population versus the new target population for the 401(k) dataset.

[2] Chernozhukov et al. (2017, Double/Debiased machine learning of treatment and causal parameters); Abadie (2003, Semiparametric instrumental variable estimation of treatment response models)

Confidence intervals for the 401(k) example

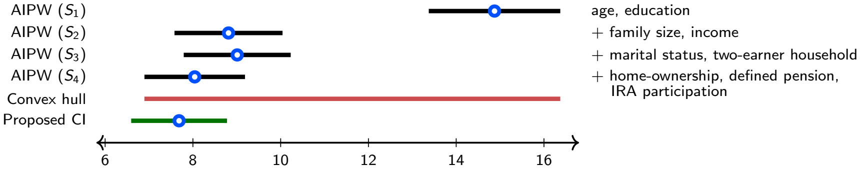


Figure 4: 95% CIs for the ATE of 401(k) eligibility on net financial assets (in 1000 USD). The black lines show AIPW intervals, the red line shows the convex hull, and the green line shows our specification-robust CI, which is about 77% narrower than the convex hull. The blue circles indicated the corresponding point estimates.

No free lunch

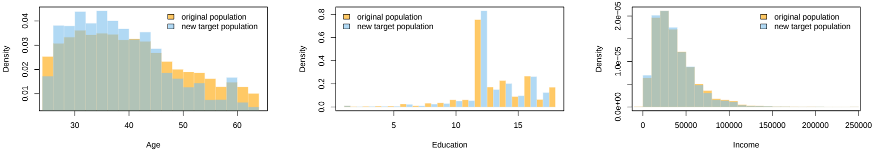


Figure 5: Original vs reweighted distribution of age, income and education in the 401(k) example. Here the new target population is visibly close to the original one.

- In general, the new target population is close to, but different from, the original population.

No free lunch

- In general, the new target population is close to, but different from, the original population.
- We need enough heterogeneity to guarantee a solution exists.

Define $g_k(X_{S_\cap}) = \mathbb{E}[\tau(X_{S_1}) - \tau(X_{S_k}) \mid X_{S_\cap}]$, $k = 2, \ldots, K$. We need $\lambda \mapsto \mathbb{E}_P[\exp(\lambda^\top g(X_{S_\cap}))]$ to be finite in an open nbd of $\mathbf{0}$, and coercive (goes to ∞ as $\lambda \to 0$) for existence. $\mathbb{E}_P[gg^\top] \succ 0$ guarantees uniqueness.

(Recall the def of contrasts: $\tau(X_{S_j}) = \mathbb{E}_P[Y \mid A = 1, X_{S_j}] - \mathbb{E}_P[Y \mid A = 0, X_{S_j}]$. $X_{S_\cap}$ denotes the covariates shared by all candidate adj sets.)

A sufficient heterogeneity condition is that all possible sign patterns for $\mathbb{E}[\tau(X_{S_1}) - \tau(X_{S_k}) \mid X_{S_\cap}]$ appear with positive probability.

Solving the optimization problem

With $g_k(X_{S_\cap}) := \mathbb{E}[\tau(X_{S_1}) - \tau(X_{S_k}) \mid X_{S_\cap}]$ $(k = 2, \ldots, K)$,

$$w^* = \operatorname*{argmin}_{w = w(X_{S_\cap})} D_{\mathsf{KL}}(wP \,\|\, P) \quad \text{subject to} \quad \mathbb{E}_P[w(X_{S_\cap}) g(X_{S_\cap})] = 0.$$

Solving the optimization problem

With $g_k(X_{S_\cap}) := \mathbb{E}[\tau(X_{S_1}) - \tau(X_{S_k}) \mid X_{S_\cap}]$ $(k = 2, \ldots, K)$,

$$w^* = \operatorname*{argmin}_{w = w(X_{S_\cap})} D_{\mathsf{KL}}(wP \,\|\, P) \quad \text{subject to} \quad \mathbb{E}_P[w(X_{S_\cap}) g(X_{S_\cap})] = 0.$$

Despite the apparent complexity, the above problem is tractable:

The transfer weights w^* are given by

$$w^*(X_{S_\cap}) = \frac{\exp\left(g(X_{S_\cap})^\top \lambda^*\right)}{\mathbb{E}_P\left[\exp\left(g(X_{S_\cap})^\top \lambda^*\right)\right]},$$

where

$$\lambda^* = \operatorname*{argmin}_{\lambda \in \mathbb{R}^{K-1}} \mathbb{E}_P\left[\exp\left(g(X_{S_\cap})^\top \lambda\right)\right].$$

Solving the optimization problem

With $g_k(X_{S_\cap}) := \mathbb{E}[\tau(X_{S_1}) - \tau(X_{S_k}) \mid X_{S_\cap}]$ $(k = 2, \ldots, K)$,

$$w^* = \underset{w = w(X_{S_\cap})}{\operatorname{argmin}} \; D_{\mathsf{KL}}(wP \,\|\, P) \quad \text{subject to} \quad \mathbb{E}_P[w(X_{S_\cap}) g(X_{S_\cap})] = 0.$$

Despite the apparent complexity, the above problem is tractable:

The transfer weights w^* are given by

$$w^*(X_{S_\cap}) = \frac{\exp\left(g(X_{S_\cap})^\top \lambda^*\right)}{\mathbb{E}_P\left[\exp\left(g(X_{S_\cap})^\top \lambda^*\right)\right]},$$

where

$$\lambda^* = \underset{\lambda \in \mathbb{R}^{K-1}}{\operatorname{argmin}} \; \mathbb{E}_P\left[\exp\left(g(X_{S_\cap})^\top \lambda\right)\right].$$

Finding weights w^* reduces to a standard finite-dimensional convex problem!

Main algorithm

Train black-box regressors:

1. Use any 'CATE' estimator to obtain $\widehat{\tau}(X_{S_k})$
2. Regress $\widehat{\tau}(X_{S_1}) - \widehat{\tau}(X_{S_k})$ on $X_{S_\cap}$ to obtain $\widehat{g}_k(X_{S_\cap})$

Main algorithm

1. Use any 'CATE' estimator to obtain $\widehat{\tau}(X_{S_k})$
2. Regress $\widehat{\tau}(X_{S_1}) - \widehat{\tau}(X_{S_k})$ on $X_{S_\cap}$ to obtain $\widehat{g}_k(X_{S_\cap})$

Estimate transfer weights:

3. Solve the convex problem: $\widehat{\lambda} := \mathrm{argmin}_{\lambda \in \mathbb{R}^{K-1}} \ \mathbb{P}_n \exp\left(\widehat{g}(X_{S_\cap})^\top \lambda\right),$
4. Compute transfer weights $\widehat{w}(X_{i,S_\cap}) \propto \exp(\widehat{g}(X_{i,S_\cap})^\top \widehat{\lambda})$

(We use cross-fitting in our implementation)

Main algorithm

1. Use any 'CATE' estimator to obtain $\widehat{\tau}(X_{S_k})$
2. Regress $\widehat{\tau}(X_{S_1}) - \widehat{\tau}(X_{S_k})$ on $X_{S_\cap}$ to obtain $\widehat{g}_k(X_{S_\cap})$

3. Solve the convex problem: $\widehat{\lambda} := \mathrm{argmin}_{\lambda \in \mathbb{R}^{K-1}} \ \mathbb{P}_n \exp\left(\widehat{g}(X_{S_\cap})^\top \lambda\right),$
4. Compute transfer weights $\widehat{w}(X_{i,S_\cap}) \propto \exp(\widehat{g}(X_{i,S_\cap})^\top \widehat{\lambda})$

Compute estimators and aggregate:

5. Multiply AIPW with transfer weights: $\widehat{\tau}_k^{\mathtt{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k})]$

   Each targets same estimand, so we can use a convex combination!

Main algorithm

1. Use any 'CATE' estimator to obtain $\widehat{\tau}(X_{S_k})$
2. Regress $\widehat{\tau}(X_{S_1}) - \widehat{\tau}(X_{S_k})$ on $X_{S_\cap}$ to obtain $\widehat{g}_k(X_{S_\cap})$

3. Solve the convex problem: $\widehat{\lambda} := \operatorname{argmin}_{\lambda \in \mathbb{R}^{K-1}} \ \mathbb{P}_n \exp\left(\widehat{g}(X_{S_\cap})^\top \lambda\right)$,
4. Compute transfer weights $\widehat{w}(X_{i,S_\cap}) \propto \exp(\widehat{g}(X_{i,S_\cap})^\top \widehat{\lambda})$

5. Multiply AIPW with transfer weights: $\widehat{\tau}_k^{\mathtt{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\,\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k})]$
6. $\widehat{\tau}^{\mathtt{AR}} = \sum \nu_k \widehat{\tau}_k^{\mathtt{R}}$ where $\sum \nu_k = 1$ (works but slow convergence rate)

Main algorithm

1. Use any 'CATE' estimator to obtain $\widehat{\tau}(X_{S_k})$
2. Regress $\widehat{\tau}(X_{S_1}) - \widehat{\tau}(X_{S_k})$ on $X_{S_\cap}$ to obtain $\widehat{g}_k(X_{S_\cap})$

3. Solve the convex problem: $\widehat{\lambda} := \mathrm{argmin}_{\lambda \in \mathbb{R}^{K-1}} \; \mathbb{P}_n \exp\left(\widehat{g}(X_{S_\cap})^\top \lambda\right)$,
4. Compute transfer weights $\widehat{w}(X_{i,S_\cap}) \propto \exp(\widehat{g}(X_{i,S_\cap})^\top \widehat{\lambda})$

5. Multiply AIPW with transfer weights: $\widehat{\tau}_k^{\mathtt{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\,\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k})]$
6. $\widehat{\tau}^{\mathtt{AR}} = \sum \nu_k \widehat{\tau}_k^{\mathtt{R}}$ where $\sum \nu_k = 1$ (works but slow convergence rate)

With additional bias-correction, $\sqrt{n}\left(\widehat{\tau}^{\mathtt{AR}}_{\text{debiased}} - \overline{\tau}^{\mathtt{R}}\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{V}_{\mathtt{AR}})$

Details on debiasing

$\widehat{\tau}_k^{\mathrm{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\,\widehat{\tau}(X_{S_k})]$ has 3 sources of uncertainty in estimation:

$$\underbrace{\text{outcome model,} \quad \text{causal weights balancing trt \& ctrl}}_{\text{use AIPW (doubly robust) estimator}}, \quad \underbrace{\text{transport weights } w^*}_{\text{we need more bias-correction}}$$

Details on debiasing

$\widehat{\tau}_k^{\mathtt{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\,\widehat{\tau}(X_{S_k})]$ has 3 sources of uncertainty in estimation:

$$\underbrace{\text{outcome model, causal weights balancing trt \& ctrl}}_{\text{use AIPW (doubly robust) estimator}}, \quad \underbrace{\text{transport weights } w^*}_{\text{we need more bias-correction}}$$

Key idea:

$$\begin{aligned}
\widehat{\tau}_k^{\mathtt{R}} - \widehat{\tau}_k^{\mathtt{R},*} &= \mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\left(\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k}) - \overline{\tau}^{\mathtt{R}}\right) - w^*(X_{S_\cap})\left(\tau^{\mathtt{AIPW}}(X_{S_k}) - \overline{\tau}^{\mathtt{R}}\right)\right] \\
&= \mathbb{P}_n\left[\left(\widehat{w}(X_{S_\cap}) - w^*(X_{S_\cap})\right)\left(\tau^{\mathtt{AIPW}}(X_{S_k}) - \overline{\tau}^{\mathtt{R}}\right)\right] \\
&\quad + \mathbb{P}_n\left[w^*(X_{S_\cap})\left(\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k}) - \tau^{\mathtt{AIPW}}(X_{S_k})\right)\right] \qquad \text{(small)} \\
&\quad + \mathbb{P}_n\left[\left(\widehat{w}(X_{S_\cap}) - w^*(X_{S_\cap})\right)\left(\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k}) - \tau^{\mathtt{AIPW}}(X_{S_k})\right)\right] \qquad \text{(small)}
\end{aligned}$$

To tackle the first term, we use Taylor expansions.

Details on debiasing

$\widehat{\tau}_k^{\mathtt{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\,\widehat{\tau}(X_{S_k})]$ has 3 sources of uncertainty in estimation:

$\underbrace{\text{outcome model, causal weights balancing trt \& ctrl}}_{\text{use AIPW (doubly robust) estimator}}, \quad \underbrace{\text{transport weights } w^*}_{\text{we need more bias-correction}}$

- Find $\widehat{\nu}$ as follows. $\widehat{\nu}_1 := 1 - \sum_{k=2}^K \widehat{\nu}_k$ and

$$\widehat{\nu}_{2:K} := \left(\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\widehat{g}(X_{S_\cap})\widehat{g}(X_{S_\cap})^\top\right]\right)^{-1}\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\widehat{g}(X_{S_\cap})\left(\widehat{\tau}(X_{S_1}) - \widehat{\tau}_1^{\mathtt{R}}\right)\right].$$

- With $\widehat{\delta}_k(D) := \widehat{\tau}^{\mathtt{AIPW}}(X_{S_1}) - \widehat{\tau}^{\mathtt{AIPW}}(X_{S_k})$, compute a bias-correction term B_n as

$$B_n := \widehat{\lambda}^\top\,\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\left(\widehat{\delta}_{2:K}(D) - \widehat{g}(X_{S_\cap})\right)\left(\sum_{k=1}^K \widehat{\nu}_k\widehat{\mathbb{E}}\left[\widehat{\tau}(X_{S_k}) \mid X_{S_\cap}\right] - \widehat{\tau}_1^{\mathtt{R}}\right)\right].$$

Details on debiasing

$\widehat{\tau}_k^{\mathrm{R}} = \mathbb{P}_n[\widehat{w}(X_{S_\cap})\,\widehat{\tau}(X_{S_k})]$ has 3 sources of uncertainty in estimation:

$$\underbrace{\text{outcome model,}\quad \text{causal weights balancing trt \& ctrl,}}_{\text{use AIPW (doubly robust) estimator}}\quad \underbrace{\text{transport weights } w^*}_{\text{we need more bias-correction}}$$

- Find $\widehat{\nu}$ as follows. $\widehat{\nu}_1 := 1 - \sum_{k=2}^{K} \widehat{\nu}_k$ and

$$\widehat{\nu}_{2:K} := \left(\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\widehat{g}(X_{S_\cap})\widehat{g}(X_{S_\cap})^\top\right]\right)^{-1}\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\widehat{g}(X_{S_\cap})\left(\widehat{\tau}(X_{S_1}) - \widehat{\tau}_1^{\mathrm{R}}\right)\right].$$

- With $\widehat{\delta}_k(D) := \widehat{\tau}^{\mathtt{AIPW}}(X_{S_1}) - \widehat{\tau}^{\mathtt{AIPW}}(X_{S_k})$, compute a bias-correction term B_n as

$$B_n := \widehat{\lambda}^\top\,\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\left(\widehat{\delta}_{2:K}(D) - \widehat{g}(X_{S_\cap})\right)\left(\sum_{k=1}^{K}\widehat{\nu}_k\widehat{\mathbb{E}}\left[\widehat{\tau}(X_{S_k}) \mid X_{S_\cap}\right] - \widehat{\tau}_1^{\mathrm{R}}\right)\right].$$

$$\widehat{\tau}^{\mathtt{AR}}_{\text{debiased}} := \sum_{k=1}^{K}\widehat{\nu}_k\,\mathbb{P}_n\left[\widehat{w}(X_{S_\cap})\,\widehat{\tau}^{\mathtt{AIPW}}(X_{S_k})\right] + B_n$$

Main theoretical result

Assume that the adjustment sets $S_1, \ldots, S_K$ have non-empty intersection $S_\cap$ that satisfy overlap, and at least one adjustment set satisfies ignorability. Then, under regularity assumptions such as $\|\widehat{\mu}_a(X_{S_k}) - \mu_a(X_{S_k})\|_{L^{2+\varepsilon}(\mathbb{P}_n)} = o_P(n^{-1/4})$ $(a = 0, 1)$ and $\|\widehat{e}(X_{S_k}) - e(X_{S_k})\|_{L^{2+\varepsilon}(\mathbb{P}_n)} = o_P(n^{-1/4})$, we have

$$\sqrt{n}\left(\widehat{\tau}^{\mathtt{AR}}_{\text{debiased}} - \overline{\tau}^{\mathtt{R}}\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{V}_{\mathtt{AR}}).$$

Main theoretical result

Assume that the adjustment sets $S_1, \ldots, S_K$ have non-empty intersection $S_\cap$ that satisfy overlap, and at least one adjustment set satisfies ignorability. Then, under regularity assumptions such as $\|\widehat{\mu}_a(X_{S_k}) - \mu_a(X_{S_k})\|_{L^{2+\varepsilon}(\mathbb{P}_n)} = o_P(n^{-1/4})$ $(a = 0, 1)$ and $\|\widehat{e}(X_{S_k}) - e(X_{S_k})\|_{L^{2+\varepsilon}(\mathbb{P}_n)} = o_P(n^{-1/4})$, we have

$$\sqrt{n}\left(\widehat{\tau}^{\mathtt{AR}}_{\text{debiased}} - \overline{\tau}^{\mathrm{R}}\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{V}_{\mathtt{AR}}).$$

$$\mathrm{V}_{\mathtt{AR}} = \mathrm{Var}\left[w^*(X_{S_\cap})\left(\sum_{k=1}^{K} \nu_k^* \tau^{\mathtt{AIPW}}(X_{S_k}) - \overline{\tau}^{\mathrm{R}} + (\delta_{2:K}(D) - g(X_{S_\cap}))^\top \lambda^* f(X_{S_\cap})\right)\right]$$

with $\delta_k = \tau^{\mathtt{AIPW}}(X_{S_1}) - \tau^{\mathtt{AIPW}}(X_{S_k})$, $f(X_{S_\cap}) := \sum_{k=1}^{K} \nu_k^* \mathbb{E}(\tau(X_{S_k}) \mid X_S) - \overline{\tau}^{\mathrm{R}}$, and

$$\tau^{\mathtt{AIPW}}(X) := \tau(X) + \frac{A}{e(X)}(Y - \mu_1(X)) - \frac{1-A}{1-e(X)}(Y - \mu_0(X)),$$

where $\mu_a(x) = \mathbb{E}(Y \mid A = a, X = x)$, and $e(x) = \mathbb{P}(A = 1 \mid X = x)$.

Summary

⚠ In observational studies, multiple adj. sets can seem equally plausible, can yield conflicting inference and we cannot detect which is valid.

Summary

⚠ In observational studies, multiple adj. sets can seem equally plausible, can yield conflicting inference and we cannot detect which is valid.

We provide a single confidence interval which:

- ☑ is specification-robust (valid if at least one of the adj. sets is valid),
- ☑ has width shrinking at fast parametric ($n^{-1/2}$) rate with sample size.

Summary

⚠ In observational studies, multiple adj. sets can seem equally plausible, can yield conflicting inference and we cannot detect which is valid.

We provide a single confidence interval which:

- ☑ is specification-robust (valid if at least one of the adj. sets is valid),
- ☑ has width shrinking at fast parametric ($n^{-1/2}$) rate with sample size.

No free lunch: Our method applies to a target population close to, but different from, the original population (and needs enough heterogeneity).

Preprint available on arXiv

Questions and feedback are welcome!

ghoshadi@stanford.edu