Efficient Generalization and Transportation

By Zhenghao Zeng, Edward H. Kennedy, Lisa M. Bodnar, Ashley I. Naimi

Keith Barnatchez

May 17th, 2023

Outline

1. Background
2. Intuition behind estimators based on influence functions
3. The paper

Roadmap

1. Background
2. Intuition behind estimators based on influence functions
3. The paper

Background

In public health, increasingly common to integrate data from various sources

- Data sharing in hospital networks
- Multi-site cohort studies
- Small study + U.S. Census data

Lots of challenges pop up in these settings...

- Covariate shift
- Potentially different sources of biases
- Different sets of covariates available at each site

Background

Common issue in causal inference:

- The population we're interested in making inferences on is not the same population we're able to conduct our study on, e.g.
  - Clinical trials
  - Observational studies with hard-to-measure variables
  - Training data vs testing data

Background

Common issue in causal inference:

- The population we're interested in making inferences on is not the same population we're able to conduct our study on, e.g.
  - Clinical trials
  - Observational studies with hard-to-measure variables
  - Training data vs testing data

- A consequence is that treatment effects we calculate with our study sample may not "generalize" to the population we want to give the treatment to

Background

Common issue in causal inference:

- The population we're interested in making inferences on is not the same population we're able to conduct our study on, e.g.
  - Clinical trials
  - Observational studies with hard-to-measure variables
  - Training data vs testing data
- A consequence is that treatment effects we calculate with our study sample may not "generalize" to the population we want to give the treatment to
- Can't always conduct the exact study we want; how do we still answer the main question of interest under sub-optimal study design?
  - And how do we do this in a way that maximizes *efficiency*?

Goals for this JC

1. Give some intuition for “doubly-robust” / influence-function based estimators
    - Heuristics to explain what role the influence function plays in causal inference
    - Discuss some of the nice properties these types of estimators have
2. Connect back to multi-study problems via the concepts of generalizability and transportability
3. Discuss Zeng et al. (2023)

Roadmap

1. Background
2. Intuition behind estimators based on influence functions
3. The paper

Estimating treatment effects

Focus on a simple example. Suppose we observe

$$(Y_i, A_i, \boldsymbol{X}_i) \sim \mathbb{P}$$

and we want to estimate the average treatment effect

$$\tau \stackrel{\text{def}}{=} \mathbb{E}(Y(1) - Y(0))$$

Estimating treatment effects

Focus on a simple example. Suppose we observe

$$(Y_i, A_i, \boldsymbol{X}_i) \sim \mathbb{P}$$

and we want to estimate the average treatment effect

$$\tau \stackrel{\text{def}}{=} \mathbb{E}(Y(1) - Y(0))$$

If the standard causal assumptions hold, then

$$\tau = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}(Y|A=1,\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}\mathbb{E}(Y|A=0,\boldsymbol{X})$$

A natural "plug-in" estimator is then

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mathbb{E}}(Y_i|A_i=1,\boldsymbol{X}_i) - \hat{\mathbb{E}}(Y_i|A_i=0,\boldsymbol{X}_i)]$$

Two approaches

Two common ways to proceed:

- **Parametric approach** Assume

$$(Y_i, A_i, \boldsymbol{X}_i) \sim \mathbb{P}_\theta, \quad \theta \in \mathbb{R}^p$$

- **Non-parametric approach**: Assume

$$(Y_i, A_i, \boldsymbol{X}_i) \sim \mathbb{P}_\eta, \quad \eta \text{ is infinite dimensional}$$

  - E.g. $\mathbb{P}$ may be indexed by possible outcome/treatment models (with no restrictions on functional forms) instead of parameter values

Non-parametric approach

If taking the non-parametric approach, then we're assuming

$$(Y_i, A_i, \boldsymbol{X}_i) \sim \mathbb{P}_\eta, \quad \eta \text{ is infinite dimensional}$$

The treatment effect τ also depends on η

- To be explicit, we'll write $\tau = \tau(\eta)$

Again, instead of $\mathbb{P}_\eta$ being indexed by parameter values, it's indexed by functions (outcome model, treatment model etc) with no a priori restrictions

Non-parametric approach

If taking the non-parametric approach, then we're assuming

$$(Y_i, A_i, \boldsymbol{X}_i) \sim \mathbb{P}_\eta, \quad \eta \text{ is infinite dimensional}$$

The treatment effect τ also depends on η

- To be explicit, we'll write $\tau = \tau(\eta)$

Again, instead of $\mathbb{P}_\eta$ being indexed by parameter values, it's indexed by functions (outcome model, treatment model etc) with no a priori restrictions

To estimate $\tau(\eta)$, one can

1. Use flexible ML models to estimate $\mathbb{E}(Y|A = a, \boldsymbol{X})$
2. Compute $\tau(\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mathbb{E}}(Y_i|A_i = 1, \boldsymbol{X}_i) - \hat{\mathbb{E}}(Y_i|A_i = 0, \boldsymbol{X}_i)]$

Bias of the plug-in estimator

A natural question: how does this estimator behave as the sample size grows?

- Similar to how we could use Taylor expansions to make sense of MLE, turns out there's an analogue here

Bias of the plug-in estimator

A natural question: how does this estimator behave as the sample size grows?

- Similar to how we could use Taylor expansions to make sense of MLE, turns out there's an analogue here

Heuristics – hope the following Taylor-type expansion exists:

$$\tau(\eta) = \tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta}) + O(||\eta - \hat{\eta}||^2)$$

Bias of the plug-in estimator

A natural question: how does this estimator behave as the sample size grows?

- Similar to how we could use Taylor expansions to make sense of MLE, turns out there's an analogue here

Heuristics – hope the following Taylor-type expansion exists:

$$\tau(\eta) = \tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta}) + O(||\eta - \hat{\eta}||^2)$$

- The second order terms vanish to 0 quickly
- The first order term will only vanish to 0 at the rate that our ML models estimate $\eta \rightarrow$ in general *much* slower than $\sqrt{n}$
  - i.e. $O(||\eta - \hat{\eta}||)$

Correcting first-order bias

$$\tau(\eta) = \tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta}) + O(||\eta - \hat{\eta}||^2)$$

The first order term is abstract

- What does it even mean to compute a derivative w.r.t. η here?

Correcting first-order bias

$$\tau(\eta) = \tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta}) + O(||\eta - \hat{\eta}||^2)$$

The first order term is abstract

- What does it even mean to compute a derivative w.r.t. η here?

Take a leap of faith and suppose we *are* able to compute/estimate it

- Notice

$$\tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta})$$

will approach $\tau(\eta)$ at faster rates than the initial estimator

Correcting first-order bias

$$\tau(\eta) = \tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta}) + O(||\eta - \hat{\eta}||^2)$$

The first order term is abstract

- What does it even mean to compute a derivative w.r.t. η here?

Take a leap of faith and suppose we *are* able to compute/estimate it

- Notice

  $$\tau(\hat{\eta}) + \tau'(\hat{\eta})(\eta - \hat{\eta})$$

  will approach $\tau(\eta)$ at faster rates than the initial estimator

The first order term is closely related to what's often called the *efficient influence function*

- The intuition of IF-based estimators is to "add back" first-order bias of an initial plug-in estimator that would otherwise vanish to 0 slowly

(Efficient) Influence function

More specifically, it turns out for estimating the ATE that

$$\tau'(\hat{\eta})(\eta - \hat{\eta}) = \mathbb{E}_{\mathbb{P}_\eta}[\mathsf{EIF}_{\hat{\eta}}(Y, A, \boldsymbol{X})]$$

(Efficient) Influence function

More specifically, it turns out for estimating the ATE that

$$\tau'(\hat{\eta})(\eta - \hat{\eta}) = \mathbb{E}_{\mathbb{P}_\eta}[\mathsf{EIF}_{\hat{\eta}}(Y, A, \boldsymbol{X})]$$

- We can represent this derivative as the expectation of something called the "efficient influence function" (EIF)
  - Where the expectation is taken over the truth $\mathbb{P}_\eta$, not our estimate $\mathbb{P}_{\hat{\eta}}$
  - Under the truth the EIF has mean zero: $\mathbb{E}_{\mathbb{P}_\eta}[\mathsf{EIF}_\eta(Y, A, \boldsymbol{X})]$

(Efficient) Influence function

More specifically, it turns out for estimating the ATE that

$$\tau'(\hat{\eta})(\eta - \hat{\eta}) = \mathbb{E}_{\mathbb{P}_\eta}[\mathrm{EIF}_{\hat{\eta}}(Y, A, \boldsymbol{X})]$$

- We can represent this derivative as the expectation of something called the "efficient influence function" (EIF)
  - Where the expectation is taken over the truth $\mathbb{P}_\eta$, not our estimate $\mathbb{P}_{\hat{\eta}}$
  - Under the truth the EIF has mean zero: $\mathbb{E}_{\mathbb{P}_\eta}[\mathrm{EIF}_\eta(Y, A, \boldsymbol{X})]$

Won't go into the specifics of when this representation holds/how to calculate the EIF. Taking the above for granted, **if** we can find an expression for the EIF, then

$$\mathbb{E}_{\mathbb{P}_\eta}[\mathrm{EIF}_{\hat{\eta}}(Y, A, \boldsymbol{X})] \text{ can be estimated by } \frac{1}{n}\sum_{i=1}^{n}\mathrm{EIF}_{\hat{\eta}}(Y_i, A_i, \boldsymbol{X}_i)$$

Bias correction

Specifically, the one-step bias-correction estimator is

$$\tau(\hat{\eta}) + \frac{1}{n} \sum_{i=1}^{n} \mathrm{EIF}_{\hat{\eta}}(Y_i, A_i, \boldsymbol{X}_i)$$

Making sense of the influence function

The Taylor-type expansion, heuristically, is decribing how the estimator changes if we slightly perturb the distribution $P_{\hat{\eta}}$ towards P_{η}

Consider a mixture of our estimated distribution and the truth:

$$P_{\varepsilon} = (1 - \varepsilon)\mathbb{P}_{\eta} + \varepsilon\mathbb{P}_{\hat{\eta}}$$

Making sense of the influence function

The Taylor-type expansion, heuristically, is decribing how the estimator changes if we slightly perturb the distribution $P_{\hat{\eta}}$ towards P_{η}

Consider a mixture of our estimated distribution and the truth:

$$P_{\varepsilon} = (1 - \varepsilon)\mathbb{P}_{\eta} + \varepsilon\mathbb{P}_{\hat{\eta}}$$

- As we move ε from 1 to 0, can think of our estimate as getting closer to the truth
- In turn, can trace out a curve of estimates as a function of ε

Making sense of the influence function

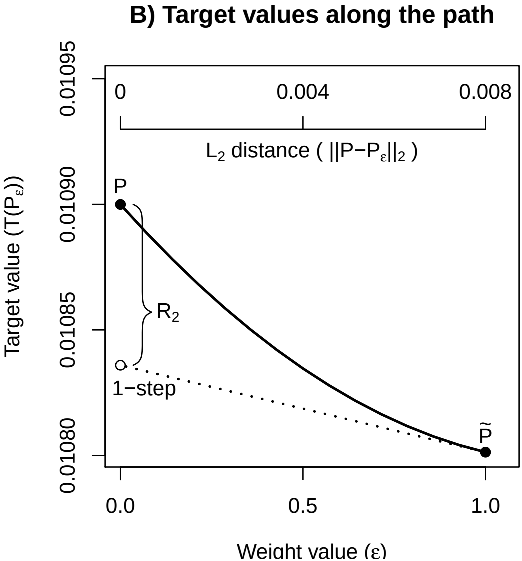
The Taylor-type expansion, heuristically, is decribing how the estimator changes if we slightly perturb the distribution $P_{\hat{\eta}}$ towards P_{η}

Consider a mixture of our estimated distribution and the truth:

$$P_{\varepsilon} = (1-\varepsilon)\mathbb{P}_{\eta} + \varepsilon\mathbb{P}_{\hat{\eta}}$$

- As we move ε from 1 to 0, can think of our estimate as getting closer to the truth
- In turn, can trace out a curve of estimates as a function of ε
- The issue: only observe the state where $\varepsilon = 1$
- It turns out the efficient influence function gives the slope of this curve at $\varepsilon = 1$

Visual Intuition



Nice properties of EIF estimators

A couple key benefits of EIF-based estimation:

- $\sqrt{n}$ rates of convergence when the nuisance estimators themselves converge much slower
- Asymptotic normality $\rightarrow$ straightforward way to compute CIs

Nice properties of EIF estimators

A couple key benefits of EIF-based estimation:

- $\sqrt{n}$ rates of convergence when the nuisance estimators themselves converge much slower
- Asymptotic normality $\rightarrow$ straightforward way to compute CIs

EIF-based estimators have other nice properties:

- **Variance lower bound**: Achieve smallest possible asymptotic variance among all non-parametric estimators

Nice properties of EIF estimators

A couple key benefits of EIF-based estimation:

- $\sqrt{n}$ rates of convergence when the nuisance estimators themselves converge much slower
- Asymptotic normality $\rightarrow$ straightforward way to compute CIs

EIF-based estimators have other nice properties:

- **Variance lower bound**: Achieve smallest possible asymptotic variance among all non-parametric estimators
- (Sometimes) **Double-robustness**: Depending on the quantity being estimated, the EIF may include more than 1 nuisance model
  - e.g. for the average treatment effect, the EIF includes an outcome model + propensity score model
  - In these scenarios, the math usually works out so that the estimator itself is consistent as long as one of the two nuisance models is consistent

AIPW

The popular AIPW estimator is based on the EIF for the average treatment effect:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\underbrace{\hat{\mu}_1(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_i)}_{\text{plug-in}} + \underbrace{A_i\frac{Y_i - \hat{\mu}_1(\boldsymbol{X}_i)}{\hat{\pi}(\boldsymbol{X}_i)} - (1 - A_i)\frac{Y_i - \hat{\mu}_0(\boldsymbol{X}_i)}{1 - \hat{\pi}(\boldsymbol{X}_i)}}_{\text{bias correction}}\right]$$

where

- $\hat{\mu}_a(\boldsymbol{X}) = \hat{\mathbb{E}}(Y|A = a, \boldsymbol{X})$ (outcome model)
- $\hat{\pi}(\boldsymbol{X}) = \hat{\mathbb{P}}(A = 1|\boldsymbol{X})$ (propensity score model)

A Point of Confusion

Notice that double-robustness was not a motivating factor at all in the EIF-based framework

- This property is nice, and the easiest advantageous feature of these estimators to explain
- As a result, EIF-based estimators are often referred to as "doubly-robust" estimators for short

A Point of Confusion

Notice that double-robustness was not a motivating factor at all in the EIF-based framework

- This property is nice, and the easiest advantageous feature of these estimators to explain
- As a result, EIF-based estimators are often referred to as "doubly-robust" estimators for short

In general, estimators that model the outcome + propensity score jointly are referred to as "doubly-robust," but they may not share the same advantages as EIF estimators for more complicated problems

- The problem covered in Zeng et al. (2023) is one example

Roadmap

1. Background
2. Intuition behind estimators based on influence functions
3. The paper

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
    - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X \mathbb{E}(Y|A=1,X) - \mathbb{E}_X \mathbb{E}(Y|A=0,X)$

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X\mathbb{E}(Y|A=1,X) - \mathbb{E}_X\mathbb{E}(Y|A=0,X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator

General Flow of a "Doubly-Robust" Estimator Paper

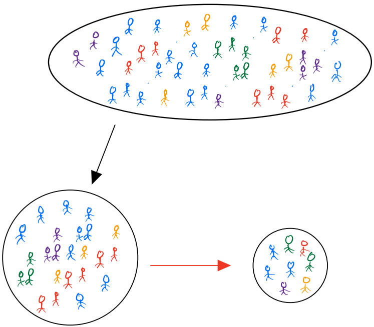
Tend to follow a pattern:

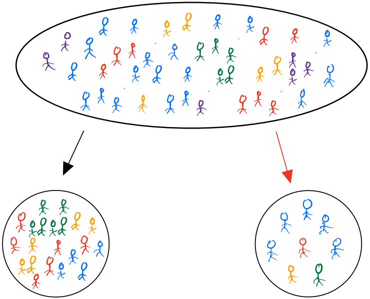
1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X\mathbb{E}(Y|A=1,X) - \mathbb{E}_X\mathbb{E}(Y|A=0,X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator
4. Prove useful stuff about the estimator (double/triple/multiple robustness, minimax rates, ...), simulation studies, etc
5. Data example

The paper

Focuses on the problems of *generalizability* and *transportability*

- In both settings, there is a source population and a target population
- **Generalizability**: The source data is a subset of the target
- **Transportability**: The source and target datasets are disjoint





The paper

Motivation: Growing interest in multi-source studies/combining information from different sources in causal inference

- But existing methods mostly don't make use of the efficiency theory motivating estimators like AIPW

Idea: Develop influence-function-based/"doubly-robust" estimators for generalizability/transportability problems

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X\mathbb{E}(Y|A=1,X) - \mathbb{E}_X\mathbb{E}(Y|A=0,X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator
4. Prove useful stuff about the estimator (double/triple/multiple robustness, minimax rates, ...), simulation studies, etc
5. Data example

Notation

- Y: outcome
- A: binary treatment indicator
- $\boldsymbol{X}$: measured confounding variables
- S: binary indicator of whether observation is part of the source population
- $Y(1), Y(0)$: *potential outcomes* under treatment/no treatment

The problem of interest

Authors focus on two settings:

1. Estimating ATE in the *generalizability* case:

$$\tau_{\text{gen}} = \mathbb{E}(Y(1) - Y(0))$$

2. Estimating ATE in the *transportability* case:

$$\tau_{\text{transp}} = \mathbb{E}(Y(1) - Y(0) | S = 0)$$

We'll just focus on their results for the transportability case

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X \mathbb{E}(Y|A=1, X) - \mathbb{E}_X \mathbb{E}(Y|A=0, X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator
4. Prove useful stuff about the estimator (double/triple/multiple robustness, minimax rates, ...), simulation studies, etc
5. Data example

Identification

Need the 3 "core" causal assumptions

1. Consistency: $Y = A \times Y(a) + (1 - A) \times Y(0)$
2. No unmeasured confounding in the source: $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}, S = 1$
3. Positivty in the source: $0 < \mathbb{P}(A = 1|\boldsymbol{X} = x, S = 1) < 1, \ \forall x$

Also need

1. Source exchangeability: $Y(a) \perp\!\!\!\perp S|\boldsymbol{X}$
   - All effect modifiers are measured
2. Source positivity: $0 < \mathbb{P}(S = 1|\boldsymbol{X} = x) < 1$
   - Anyone has a positive probability of being in the source populatin

Identification

If there is *no* covariate mismatch...

$$\begin{aligned}
\mathbb{E}(Y^a|S=0) &= \mathbb{E}_{\boldsymbol{X}|S=0}\mathbb{E}(Y^a|\boldsymbol{X}, S=0) && \text{(Law of total exp)} \\
&= \mathbb{E}_{\boldsymbol{X}|S=0}\mathbb{E}(Y^a|\boldsymbol{X}, S=1) && (Y^a \perp\!\!\!\perp S|\boldsymbol{X}) \\
&= \mathbb{E}_{\boldsymbol{X}|S=0}\mathbb{E}(Y^a|\boldsymbol{X}, S=1, A=a) && (Y^a \perp\!\!\!\perp A|\boldsymbol{X}) \\
&= \mathbb{E}_{\boldsymbol{X}|S=0}\mathbb{E}(Y|\boldsymbol{X}, S=1, A=a) && \text{(Consistency)}
\end{aligned}$$

Intuitively, we

- Fit the outcome regression model in the source data (where we have data to do this)
- Average the outcome regression estimates across the dist. of $\boldsymbol{X}$ in the *target* population

General Flow of a “Doubly-Robust” Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X \mathbb{E}(Y|A=1,X) - \mathbb{E}_X \mathbb{E}(Y|A=0,X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator
4. Prove useful stuff about the estimator (double/triple/multiple robustness, minimax rates, ...), simulation studies, etc
5. Data example

Plug-in Estimator

The identification condition implies the following plug-in estimator

$$\hat{\tau}_{\text{plug}} = \frac{1}{|\text{Target}|} \sum_{i \in \text{Target}} [\hat{\mathbb{E}}(Y_i | \boldsymbol{X}_i, A_i = 1, S = 1) - \hat{\mathbb{E}}(Y_i | \boldsymbol{X}_i, A_i = 0, S = 1)]$$

where $\hat{\mathbb{E}}(Y_i | \boldsymbol{X}_i, A_i = a, S = 1)$ is non-parametrically estimated in the source data

The problem: In general this will converge slower than $\sqrt{n}$ and be sub-optimal in general

Proposed Estimator

Idea: Find the efficient influence function for

$$\mathbb{E}_{\boldsymbol{X}|S=0}\mathbb{E}(Y|\boldsymbol{X}, S=1, A=a),$$

estimate it, and add it on to the plug-in

Proposed Estimator: Influence Function

Efficient influence function is found to be:

$$\text{IF}(Y, A, \boldsymbol{X}) = \frac{I(A = a, S = 1)(1 - \rho(\boldsymbol{X}))(Y - \mu_a(\boldsymbol{X}))}{\rho(\boldsymbol{X})\pi_a(\boldsymbol{X})} + I(S = 0)\mu_a(\boldsymbol{X})$$

So the resulting estimator is

$$\hat{\psi}_a = \frac{1}{\hat{\mathbb{P}}(S = 1)} \sum_{i=1}^{n} \left(\frac{I(A_i = a, S_i = 0)(1 - \hat{\rho}(\boldsymbol{X}_i))(Y - \hat{\mu}_a(\boldsymbol{X}_i))}{\hat{\rho}(\boldsymbol{X}_i)\hat{\pi}_a(\boldsymbol{X}_i)} + I(S_i = 0)\hat{\mu}_a(\boldsymbol{X}) \right)$$

where $\hat{\tau} = \hat{\psi}_1 - \hat{\psi}_0$ and $\rho(\boldsymbol{X}) = \mathbb{P}(S = 1|\boldsymbol{X})$

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

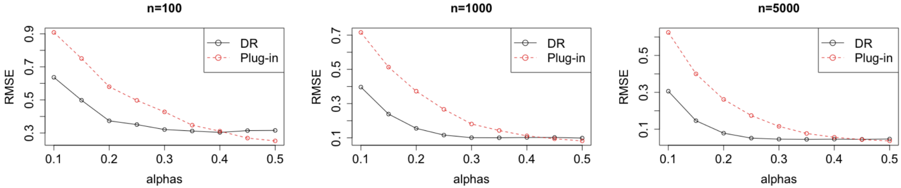
1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X \mathbb{E}(Y|A=1,X) - \mathbb{E}_X \mathbb{E}(Y|A=0,X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator
4. Prove useful stuff about the estimator (double/triple/multiple robustness, minimax rates, ...), simulation studies, etc
5. Data example

Simulation study

Main idea:

- Compare "doubly-robust" (EIF-based) estimator to plug-in estimator in terms of RMSE
- Alter the data-generating process so that convergence rate of nuisance estimators is $O(n^{-\alpha})$
  - Have control over α
  - $\alpha = 1/2$ is standard parametric setting
  - $\alpha < 1/2$ implies slow convergence of nuisance estimators

Simulation study



(a) n=100 (b) n=1000 (c) n=5000

Figure 1: RMSE V.S. α

General Flow of a "Doubly-Robust" Estimator Paper

Tend to follow a pattern:

1. Define the problem of interest (e.g. estimate the ATE in an obs. study)
2. Find an identification expression for the problem of interest
   - E.g. in an observational study, the ATE can be identified as $\mathbb{E}_X\mathbb{E}(Y|A=1,X) - \mathbb{E}_X\mathbb{E}(Y|A=0,X)$
3. Derive the corresponding efficient influence function
   - This also gives you your implied D-R estimator
4. Prove useful stuff about the estimator (double/triple/multiple robustness, minimax rates, ...), simulation studies, etc
5. Data example

Data example

Background: Maternal malnutrition is strongly linked with adverse pregnancy and birth outcomes

- Nurition is one of the few known modifiable risk factors for adverse pregnancy outcomes

Focus: Estimate the average causal effects of fruit + vegetable intake on adverse pregnancy outcomes in *the whole U.S. population*

Data: Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b)

- Non-representative sample of the U.S. population
- E.g. 23% of participants college educated, vs. 10% for the U.S. pregnant population

Methods: data

High-level: Estimate treatment effects in the nuMoM2b study, and transport back to the U.S. population

- To this end, authors use representative sample from National Survey of Family Growth
- Contains demographic data for 9553 women in the U.S.

nuMoM2b data:

- Outcomes: preterm birth, SGA birth, gestational diabetes, preeclampsia
- Confounders: maternal age, race, education, pre-pregnancy BMI, smoking status, marital status, health insurance status, working status, other dietary components (e.g. grains, fish)
- Treatments: Discretize consumption of fruits/vegetables
  - $A = 1$ if fruit/vegetable consumption is higher than the 80th percentile

Methods: data

National Survey of Family Growth:

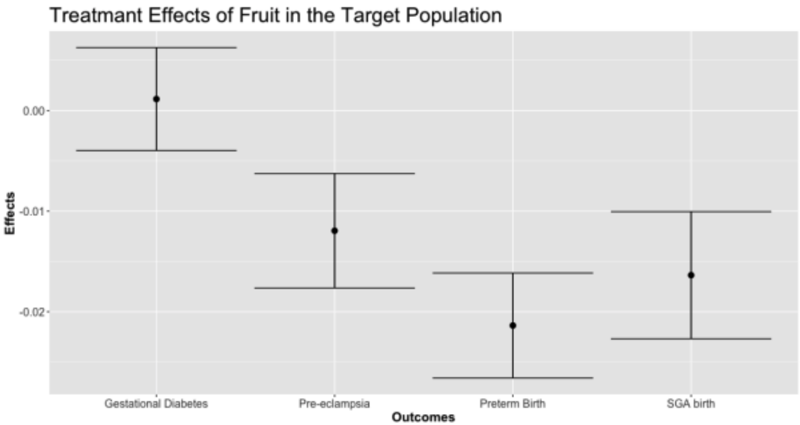
- Representative U.S. sample
- Contains info on *most* of the covariates found in nuMoM2b data, but not all
- Authors use a *slightly* different version of the estimator we've discussed to account for this
    - Basically, assume all of the *effect modifiers* are measured in both datasets, and all effect modifiers + confounders available in nuMoM2b dataset

Methods: model

Implement the EIF estimator, using ensemble learners (SuperLearner) to fit all of the nuisance models $(\hat{\mu}(\cdot), \hat{\pi}(\cdot), \hat{\rho}(\cdot))$

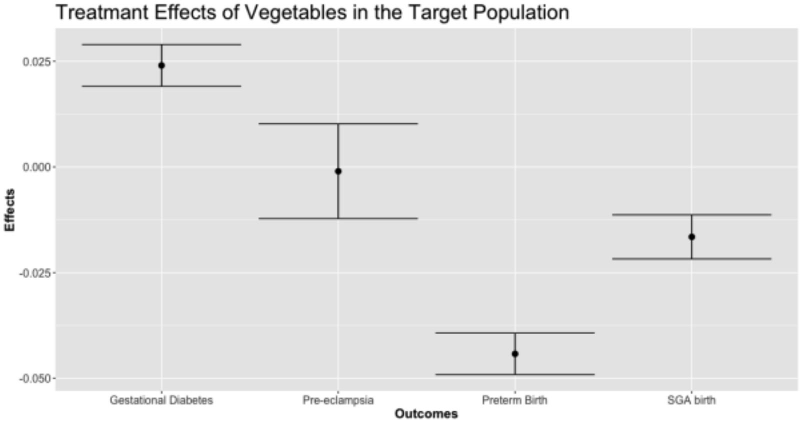
- 8 total treatment effect estimates
- 4 outcomes, 2 treatments of interest

Results: fruit



(a) Fruit

Results: vegetables



(b) Vegetables

Discussion

References

Zeng, Z., Kennedy, E. H., Bodnar, L. M., and Naimi, A. I. (2023). Efficient generalization and transportation. *arXiv preprint arXiv:2302.00092*.