Improved Inference for Doubly Robust Estimators of Heterogeneous Treatment Effects

Joseph Antonelli
(Joint work with Heejun Shin)

June 27th, 2022

UF UNIVERSITY of FLORIDA

Motivation

- Estimating treatment effects from observational data is a common scientific goal
  - Effect of environmental exposures on health outcomes
  - Effect of new policing policies on arrest/crime rates
- Number of questions one might ask
  - Is there an average effect of the policy?
  - Are certain subgroups more impacted by the policy than others?
- We aim to answer both of these questions, but will focus more on the second of these two

Setup and notation

- p is the dimension of the covariate space
- n is the sample size
- $\boldsymbol{X}$ are pre-treatment covariates
- $\boldsymbol{V}$ is a subset of the covariates $\boldsymbol{X}$
- T is a binary treatment of interest
- $Y(t)$ is the potential outcome for a unit under treatment $T = t$
- Interested in conditional average treatment effects (CATE)

$$\tau(\boldsymbol{v}) = E(Y(1) - Y(0)|\boldsymbol{V} = \boldsymbol{v})$$

Identifying assumptions

- We will make standard assumptions to identify the treatment effect
- **Assumption 1 (Consistency):** $Y(T) = Y$
- **Assumption 2 (Positivity):** $0 < P(T = 1|\boldsymbol{X}) < 1$ for all $\boldsymbol{X}$
- **Assumption 3 (No unmeasured confounding):**
  $Y(1), Y(0) \perp\!\!\!\perp T|\boldsymbol{X}$
- Under these assumptions and letting $\boldsymbol{V} = \boldsymbol{X}$, we have that

$$\begin{aligned}\tau(\boldsymbol{x}) &= E(Y(1)|\boldsymbol{X} = \boldsymbol{x}) - E(Y(0)|\boldsymbol{X} = \boldsymbol{x}) \\ &= E(Y|T = 1, \boldsymbol{X} = \boldsymbol{x}) - E(Y|T = 0, \boldsymbol{X} = \boldsymbol{x})\end{aligned}$$

Estimating the CATE

- This seems easy. We can just estimate $m(t, \boldsymbol{x}) = E(Y | T = t, \boldsymbol{X} = \boldsymbol{x})$

$$\widehat{\tau}(\boldsymbol{x}) = \widehat{m}(1, \boldsymbol{x}) - \widehat{m}(0, \boldsymbol{x})$$

- This is simply a regression problem
- Some problems with this approach
  - Relies on correct specification of $m(t, \boldsymbol{x})$
  - Slow rates of convergence if $m(t, \boldsymbol{x})$ is estimated flexibly or $\boldsymbol{X}$ is high-dimensional
  - We have little control over the complexity of $m(1, \boldsymbol{x}) - m(0, \boldsymbol{x})$, which is the parameter of interest

Estimating the CATE

- Doubly robust estimation helps address both issues
- Long history of doubly robust estimators in the causal inference literature (Scharfstein et al., 1999; Bang and Robins, 2005)
- Utilized extensively for high-dimensional or semiparametric causal inference problems (Farrell, 2015; Chernozhukov et al., 2018)
- Also used recently for heterogeneous treatment effects (Semenova and Chernozhukov, 2021)

What are doubly robust estimators

- Propensity score: $e(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x})$
- Outcome regression: $m(t, \boldsymbol{x}) = E(Y|T = t, \boldsymbol{X} = \boldsymbol{x})$
- Using these, we can construct a pseudo-outcome:

$$
\begin{aligned}
Z_i = & \frac{1(T_i = 1)}{e(\boldsymbol{X}_i)}(Y_i - m(1, \boldsymbol{X}_i)) + m(1, \boldsymbol{X}_i) \\
& - \frac{1(T_i = 0)}{1 - e(\boldsymbol{X}_i)}(Y_i - m(0, \boldsymbol{X}_i)) - m(0, \boldsymbol{X}_i)
\end{aligned}
$$

What are doubly robust estimators

- It turns out that if *either* 1) the propensity score, or 2) the outcome regression are correctly specified

$$E(Z|\boldsymbol{V} = \boldsymbol{v}) = \tau(\boldsymbol{v})$$

- This suggests a two-stage estimation strategy for $\tau(\boldsymbol{v})$
  1. Estimate the propensity score and outcome regression, and construct Z_i for $i = 1, \ldots, n$
  2. Regress Z_i against $\boldsymbol{V}_i$

- Estimates from this second model will be estimates of $\tau(\boldsymbol{v})$

Features of DR estimators

- A number of pros of doubly robust estimators
  - Consistency even when one model is misspecified
  - Faster rates of convergence if both models are correctly specified
  - Allows for high-dimensional $\boldsymbol{X}$ or nonparametric models
- There are some drawbacks
  - Inference not necessarily doubly robust
  - Inference becomes challenging with high-dimensional or nonparametric models
    - Asymptotic theory may underestimate uncertainty
    - Bootstrap may not apply

Where we come in

- We aim to propose an estimator with all of the aforementioned desirable properties of DR estimators
- Improve inference in finite samples and under model misspecification
- We will utilize Bayesian methods for the propensity score and outcome regression models
    - Allows for a range of Bayesian nonparametric models
- We are not proposing a fully Bayesian procedure!
    - Our inference is ultimately frequentist
    - Simply trying to use posterior distributions to account for difficult sources of uncertainty

Estimation

- Define $\boldsymbol{D}_i = (Y_i, T_i, \boldsymbol{X}_i, \boldsymbol{V}_i)$
- Let $\boldsymbol{\Psi}$ represent all parameters from the treatment and outcome models
- Let $\Delta(\boldsymbol{D}, \boldsymbol{\Psi})$ represent our estimator of $\tau(\boldsymbol{v})$ at the observed data $\boldsymbol{D}$ and parameter values $\boldsymbol{\Psi}$
- Our point estimate is the posterior mean of this quantity
  - Suppose we have B posterior draws, $\boldsymbol{\Psi}^{(b)}$ for $b = 1, \dots, B$

$$\widehat{\Delta} = E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})] \approx \frac{1}{B} \sum_{b=1}^{B} \Delta(\boldsymbol{D}, \boldsymbol{\Psi}^{(b)})$$

What about inference?

- Normally we can easily perform inference once we have the posterior distribution of all unknown parameters
- That's not the case here
  - Estimator is a function of parameters and observed data
  - Not simply a functional of parameters
- How can we use our posterior distribution of $\boldsymbol{\Psi}$ to construct inference that accounts for all sources of uncertainty?

Inference

- We will target the following

$$\text{Var}_{\boldsymbol{D}}\widehat{\Delta} = \text{Var}_{\boldsymbol{D}} E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})]$$

- Ideally we would draw new values of $\boldsymbol{D}$ and calculate the posterior mean each time
- Can't do this for two reasons
  - Don't know the distribution of $\boldsymbol{D}$
  - Computationally infeasible to estimate posterior distribution each time
- We will approximate this process by combining the nonparametric bootstrap with our one posterior distribution

Inference

- We propose the following variance estimator

$$\widehat{V} = \text{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\} + \text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})]$$

- The first term resembles the target variance
  - Outer moment now with respect to $\boldsymbol{D}^{(m)}$, bootstrap resamples of $\boldsymbol{D}$
  - Inner moment still with respect to $\boldsymbol{D}$
  - Ignores uncertainty stemming from the fact that different data sets would lead to different posterior distributions
- Estimate the first term with the standard bootstrap

Inference

$$\widehat{V} = \text{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\} + \text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})]$$

- The second term accounts for uncertainty due to parameter estimation
- Not clear that this estimate of the variance will be any good
  - Not a standard variance decomposition
- It turns out that this variance estimator has some nice properties

Inference

- If both the propensity score and outcome regression models are correctly specified and contract at $n^{-1/4}$ rates or faster, then

$$\widehat{V} - V = o_p(n^{-1})$$

  where V is the true variance of our estimator

- This shows that our variance estimator is consistent

- But what happens in finite samples or when one of the models is misspecified?

Inference

- Our variance estimator also has the following property

$$E_{\boldsymbol{D}}\{\widehat{V} - V\} \gtrapprox 0$$

- This result holds in scenarios when our consistency result doesn't hold
  - Model misspecification
  - Propensity score and outcome regression contract at rates slower than $n^{-1/4}$
  - Small sample sizes

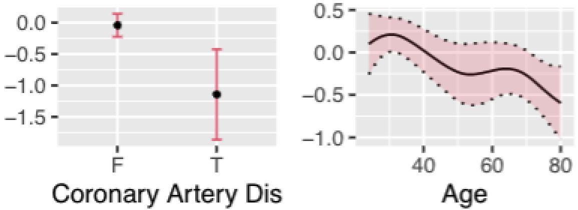
- Conservative variance estimator when conditions of theorem don't hold

Application to environmental exposures

- Now we explore the effects of environmental exposures on various health metrics
  - Dichotomize exposures to being above or below the median level
- We will look at a study from the National Health and Nutrition Survey (NHANES)
- Estimate the effect of diakyl metabolite levels on HDL cholesterol
  - $n = 225$ and $p = 75$, and we will incorporate some high-dimensional techniques

Effect of diakyl metabolite levels on HDL cholesterol

- Here we explore $\tau(V_j)$ for $j = 1, \ldots, p$
  - Trying to find which covariates modify the treatment effect the most
- Below are two covariates that influence the treatment effect
- The negative effects of diakyl are more pronounced in older subjects and those with existing diseases



Conclusions

- We showed that flexible Bayesian methods can be combined with doubly robust estimators
  - Fast convergence rates
  - Improved inferential properties
- Variance estimator is consistent when both models are correctly specified, and conservative otherwise
- Applies to both high-dimensional and nonlinear settings
- Paper available at https://arxiv.org/abs/2111.03594

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal* **21**.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* **189** 1–23.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94** 1096–1120.

Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* **24** 264–289.