Doubly Robust Estimators with Weak Overlap

Yukun Ma (Vanderbilt)
Pedro H. C. Sant'Anna (Emory)
Yuya Sasaki (Vanderbilt)
Takuya Ura (UC Davis)

- A doubly robust estimator is a popular choice.
    - ATE, ATT, LATE, ATT in DID, ...
- Weak overlap.
- We often trim observations, but we lose double robustness.
- We apply Sasaki and Ura's trim-and-correct technique.
- The proposed estimator is doubly robust, and at the same time, can achieve a smaller variance.

Related Literature

- Doubly robust estimation.
- Small denominator problem.
    - Crump, Hotz, Imbens, and Mitnik (2009), Khan and Tamer (2010), Yang (2014), Khan and Nekipelov (2015), Chaudhuri and Hill (2016), Rothe (2017), Yang and Ding (2018), Hong, Leung, and Li (2020), Ma and Wang (2020), Heiler and Kazak (2021), and Sasaki and Ura (2022).
    - Sasaki and Ura (2022): trim observations and correct the bias.

Setup

- Observed variables: Y, D, X
  - $(Y(1), Y(0))$: potential outcomes
  - $Y = Y(D)$: outcome variable
  - $D = 1$ if an observation is treated; =0 otherwise
- Unconfoundedness

$$E[Y(0)|D = 1, X] = E[Y(0)|D = 0, X]$$

- Assume $P(X) < 1$ where $P(X) = P(D = 1 \mid X)$
- ATT

$$ATT = E[Y(1) - Y(0) \mid D = 1]$$

Doubly robust moment

- Doubly robust moment for ATT:

$$ATT = \frac{1}{E\left[D\right]} E\left[\left(D - \frac{P(X)}{1-P(X)}(1-D)\right)\left(Y - \nu(X)\right)\right],$$

where $\nu(X) = E[Y|D=0, X]$ and $P(X) = P(D=1 \mid X)$

- Parametric models
  - Outcome function $\nu(X;\gamma_0)$
  - Propensity score $P(X;\gamma_0)$
- Double robustness: If one of the parametric models is correct,

$$ATT = \frac{1}{E\left[D\right]} E\left[\left(D - \frac{P(X;\gamma_0)}{1-P(X;\gamma_0)}(1-D)\right)\left(Y - \nu(X;\gamma_0)\right)\right].$$

Doubly robust moment

- Doubly robust moment for ATT:

$$ATT = \frac{1}{E\left[D\right]} E\left[\left(D - \frac{P(X)}{1-P(X)}(1-D)\right)(Y - \nu(X))\right],$$

where $\nu(X) = E[Y|D=0, X]$ and $P(X) = P(D = 1 \mid X)$

- Trimming:

$$\frac{1}{E\left[D\right]} E\left[(D - \frac{P(X)}{1-P(X)} 1\{1 - P(X) \geq h\}(1-D))(Y - \nu(X))\right]$$

- It can be (asymptotically) biased even if $h \to 0$.
- Doubly robustness is gone.

Sasaki and Ura (2022) for $E\left[B/A\right]$

- Trimmed Moment with $h \to 0$:

$$E\left[\frac{B}{A} 1\{|A| \geq h\}\right]$$

- When $E[B|A = 0] = 0$, then we can approximate the bias by

$$\begin{aligned}
-\text{bias} &= E\left[\frac{B}{A} 1\{|A| < h\}\right] \\
&= E\left[\frac{E\left[B|A\right] - E\left[B|A = 0\right]}{A} 1\{|A| < h\}\right] \\
&\approx E\left[\frac{c_1 A + c_2 A^2 + \cdots + c_k A^k}{A} 1\{|A| < h\}\right] \\
&= E\left[(c_1 + c_2 A + \cdots + c_k A^{k-1}) 1\{|A| < h\}\right]
\end{aligned}$$

- Trimmed-Then-Corrected Moment:

$$E\left[\frac{B}{A}1\{|A|\geq h\}+(c_1+c_2A+\cdots+c_kA^{k-1})1\{|A|<h\}\right]$$

  - We estimate $(c_1,\ldots,c_k)$ from the sieve regression of $E[B\mid A]$.
  - Trimming can yield smaller variance for the estimator
  - Bias correction: $O(h^kE[1\{|A|<h\}])$ from $O(E[1\{|A|<h\}])$.
  - Asymptotic normality for

$$E_n\left[\frac{B}{A}1\{|A|\geq h\}+(\hat{c}_1+\hat{c}_2A+\cdots+\hat{c}_kA^{k-1})1\{|A|<h\}\right]$$

ATT's Double Robust Moment

- Doubly robust moment for ATT:

$$ATT = \frac{1}{E\left[D\right]} E\left[\left(D - \frac{P(X)}{1-P(X)}(1-D)\right)\left(Y - \nu(X)\right)\right],$$

where $\nu(X) = E[Y|D=0,X]$ and $P(X) = P(D=1 \mid X)$.

- Trimmed-Then-Corrected Moment:

$$\frac{1}{E\left[D\right]} E\left[D\left(Y - \nu(X)\right) + \frac{B}{A}1\{|A| \geq h\} + (c_1 + \cdots + c_k A^{k-1})1\{|A| < h\}\right]$$

where $A = 1 - P(X)$ and $B = -P(X)(1-D)\left(Y - \nu(X)\right)$

- Parametric models
  - Outcome function $\nu(X;\gamma_0)$
  - Propensity score $P(X;\gamma_0)$
- $\hat{A} = 1 - P(X;\hat{\gamma})$ and $\hat{B} = -P(X;\hat{\gamma})(1-D)\left(Y - \nu(X;\hat{\gamma})\right)$
- The resulting estimator:

$$\hat{\theta} = \frac{1}{E_n[D]} E_n\left[D\left(Y - \nu(X;\hat{\gamma})\right) + \frac{\hat{B}}{\hat{A}} 1\{|\hat{A}| \geq h\} + (\hat{c}_1 + \cdots + \hat{c}_k \hat{A}^{k-1}) 1\{|\hat{A}| < h\}\right]$$

Asymptotic Result

Theorem: If

1. $P(D = 1) > 0$,
2. $P(D = 1 \mid X) < 1$,
3. Either $\nu(X; \gamma_0)$ or $P(X; \gamma_0)$ (or both) is correctly specified,
4. several regularity conditions (e.g., smoothness, rate for h, moment conditions),

then

$$\frac{\hat{\theta} - ATT}{\sqrt{E[(\varphi - E[\varphi])^2]/n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $n \to \infty$.

Double Robustness

- Trimmed-Then-Corrected Moment:

  $$\frac{1}{E\left[D\right]}E\left[D\left(Y-\nu(X)\right)+\frac{B}{A}1\{|A|\geq h\}+(c_1+\cdots+c_kA^{k-1})1\{|A|<h\}\right]$$

  where $A=1-P(X)$ and $B=-P(X)(1-D)\left(Y-\nu(X)\right)$

- Parametric models
  - Outcome function $\nu(X;\gamma_0)$
  - Propensity score $P(X;\gamma_0)$

- Doubly robustness: If one of the parametric models is correct,

  $$\text{Untrimmed Moment} = ATT.$$

- Case 1: ν is correctly specified (and P can be misspecified).

$$\text{Trimmed-Then-Corrected Moment} = ATT$$

$$\implies \text{Robust!}$$

- Case 2: P is correctly specified (and ν can be misspecified).

$$\text{Trimmed-Then-Corrected Moment} = ATT + \underbrace{O(h^k P(|A| \leq h))}_{o(n^{-1/2}) \text{ for large } k}$$

$$\implies \text{(Asymptotically) Robust!}$$

Simulations

- Let $X = (X_1, X_2, X_3, X_4)'$ be independent student-t random variables with df degrees of freedom.
    - $df = 30$ degrees of freedom (relatively strong overlap)
    - $df = 20$ degrees of freedom (weaker overlap)
    - $df = 10$ degrees of freedom (much weaker overlap)

DGP1: Two models ν and P are correct

$$Y(d) \mid X \sim \mathcal{N}(1 + sum(X), 1)$$

$$P(D = 1 \mid X) = \frac{\exp(sum(X))}{1 + \exp(sum(X))}$$

DGP2: Only outcome model ν is correct

$$Y(d) \mid X \sim \mathcal{N}(1 + sum(X), 1)$$

$$P(D = 1 \mid X) = \frac{\exp(sum(polynomial(X)))}{1 + \exp(sum(polynomial(X)))}$$

DGP3: Only propensity score model P is correct

$$Y(d) \mid X \sim \mathcal{N}(1 + sum(polynomial(X)), 1)$$

$$P(D = 1 \mid X) = \frac{\exp(sum(X))}{1 + \exp(sum(X))}$$

Polynomial $= X_1 + (X_1^2 - X_2^2) + X_3^3 + X_4^3$

- $n = 500$

- 10,000 Monte Carlo iterations

- NEW: our new method with $h = 0.01$ and $k = 3$ (with the sieve dimension $= 3$).

- CON: conventional method based on ATT's doubly robust moment ($=$ NEW with $h = 0.00$)

$df = 30$ degrees of freedom (relatively strong overlap)

	DGP1		DGP2		DGP3	
	CON	NEW	CON	NEW	CON	NEW
BIAS	0.001	-0.000	0.555	0.006	-0.052	-0.099
SD	0.605	0.249	101.525	0.253	0.505	0.319
RMSE	0.605	0.249	101.527	0.253	0.507	0.334
95%	0.916	0.924	0.952	0.943	0.922	0.925

$df = 20$ degrees of freedom (weaker overlap)

	DGP1		DGP2		DGP3	
	CON	NEW	CON	NEW	CON	NEW
BIAS	-0.006	0.001	-0.087	0.000	-0.055	-0.101
SD	0.397	0.241	58.239	0.257	0.661	0.330
RMSE	0.397	0.241	58.239	0.257	0.664	0.345
95%	0.916	0.926	0.954	0.942	0.918	0.920

df = 10 degrees of freedom (much weaker overlap)

	DGP1		DGP2		DGP3	
	CON	NEW	CON	NEW	CON	NEW
BIAS	-0.002	0.001	-3.317	-0.000	-0.070	-0.115
SD	0.556	0.234	268.827	0.257	0.973	0.330
RMSE	0.556	0.234	268.848	0.257	0.975	0.350
95%	0.910	0.922	0.958	0.947	0.915	0.923

Conclusion

Doubly Robust Estimators with Weak Overlap
Yukun Ma, Pedro H. C. Sant'Anna, Yuya Sasaki, Takuya Ura

- Intersection of double robustness and weak overlap
    - Doubly robust after bias correction
    - ATE, ATT, LATE, ATT in DID, ...
- The proposed estimator is doubly robust, and at the same time can achieve a smaller variance.

Exact Statements i

Define

$$\theta_h = \Lambda\left(\alpha_1(h, \gamma_0), \ldots, \alpha_L(h, \gamma_0)\right),$$

where $m_l(a; \gamma) = E[B_l(\gamma) \mid A_l(\gamma) = a]$ and

$$\alpha_l(h, \gamma) = E\left[\frac{B_l(\gamma)}{A_l(\gamma)} 1\{|A_l(\gamma)| \geq h\}\right] + \sum_{\kappa=1}^{k} \frac{E\left[A_l(\gamma)^{\kappa-1} 1\{|A_l(\gamma)| < h\}\right]}{\kappa!} \cdot m_l^{(\kappa)}(0; \gamma).$$

Exact Statements ii

Assumption

For each $l = 1, \ldots, L$ *with* $0 \in \mathrm{support}(A_l(\gamma_0))$*, (i)* $m_l(0; \gamma_0) = 0$*; and (ii)* $m_l(\cdot; \gamma_0)$ *is* $(k+1)$*-times continuously differentiable in a neighborhood of* 0*.*

Assumption

$\Lambda(\cdot)$ *is twice continuously differentiable in a neighborhood of* $(\alpha_1(0, \gamma_0), \ldots, \alpha_L(0, \gamma_0))$*.*

Exact Statements iii

Assumption

There are independent random variables $\phi_1, \ldots, \phi_n$ such that

$$\alpha_l(h, \hat{\gamma}) - \alpha_l(h, \gamma_0) = \frac{\partial}{\partial \gamma'} \alpha_l(h, \gamma)|_{\gamma=\gamma_0} (E_n - E)[\phi] + o_p(n^{-1/2})$$

for each $l = 1, \ldots, L$.

Exact Statements iv

Assumption

For each $l = 1, \dots, L$ *and* $\kappa = 1, \dots, k$,

$$\hat{m}_l^{(\kappa)}(0;\gamma_0) - m_l^{(\kappa)}(0;\gamma_0) - (E_n - E)[\psi_{l,\kappa}(\gamma_0)] = o_p(n^{-1/2}h^{1-\kappa}),$$

where

$$\psi_{l,\kappa}(\gamma) = p_K^{(\kappa)}(0)'E[p_K(A_l(\gamma))p_K(A_l(\gamma))']^{-1}p_K(A_l(\gamma))(B_l(\gamma) - m_l(A_l(\gamma);\gamma)).$$

Exact Statements v

For each $l = 1, \ldots, L$, we can define

$$\begin{aligned}\omega_l(h,\gamma) = {} & \frac{B_l(\gamma)}{A_l(\gamma)} 1\{|A_l(\gamma)| \geq h\} \\ & + \sum_{\kappa=1}^{k} \frac{A_l(\gamma)^{\kappa-1} 1\{|A_l(\gamma)| < h\}}{\kappa!} \cdot m_l^{(\kappa)}(0;\gamma) \\ & + \sum_{\kappa=1}^{k} \frac{E\left[A_l(\gamma)^{\kappa-1} 1\{|A_l(\gamma)| < h\}\right]}{\kappa!} \cdot \psi_{l,\kappa}(\gamma) + \frac{\partial}{\partial \gamma'} \alpha_l(h,\gamma)\phi.\end{aligned}$$

Assumption

For each $l = 1, \ldots, L$, $E[\omega_l(h,\gamma_0)^2] = o(n^{1/2})$.

Exact Statements vi

Assumption

For each $l = 1, \ldots, L$*,*
$\hat{\alpha}_l(h, \hat{\gamma}) - \alpha_l(h, \hat{\gamma}) - \hat{\alpha}_l(h, \gamma_0) + \alpha_l(h, \gamma_0) = o_p(n^{-1/2})$*.*

Assumption

$nh^{2k} = O(1)$ *as* $n \to \infty$*.*

Define

$$\varphi = \sum_{l=1}^{L} \Lambda_l(\alpha_1(0, \gamma_0), \ldots, \alpha_L(0, \gamma_0)))\omega_l(h, \gamma_0).$$

Exact Statements vii

Theorem

Suppose that the above assumptions are satisfied. (i) The estimator $\hat{\theta}$ has the asymptotically linear representation

$$\hat{\theta} - \theta_0 = (E_n - E)[\varphi] + o_p(n^{-1/2}).$$

(ii) If in addition, $E[\varphi^2]$ is bounded away from zero and $\frac{E[(\varphi - E[\varphi])^{2+\delta}]}{n^{\delta/2} E[(\varphi - E[\varphi])^2]^{(2+\delta)/2}} = o(1)$ for some $\delta > 0$, then

$$\frac{\hat{\theta} - \theta_0}{\sqrt{E[(\varphi - E[\varphi])^2]/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$.