Variance estimation for the weighted average treatment effect estimators

Yi Liu

Department of Statistics, North Carolina State University

yliu297@ncsu.edu

NC STATE UNIVERSITY

Introduction

It has been widely noted in causal inference literature that the sources of variability for estimating WATE come from not only estimating the estimand itself, but also from estimating the nuisance functions, i.e., propensity scores and outcomes.

Developing valid, flexible and computing-time-efficient variance estimation methods is an important focus in nonparametric resampling field. The goal of this work is to leverage the power for such methods into the causal inference setting. We consider estimators of a general class of causal estimands, the weighted average treatment effect (WATE).

Main take-away of this poster:

• Methods review and limitation comparisons.

• Preliminary results by the proposed methods.

Setup

• **Notation**

Treatment: $Z \in \{0,1\}$; covariate/ancillary variable/confounder: X; observed outcome: Y; potential outcomes: $Y(0), Y(1)$; propensity score: $e(x) = P(Z=1|X=x)$.

• **Assumptions**

– (SUTVA) There is only one version of the potential outcome. The potential outcome of a subject is unrelated to others' treatment status.

– (Consistency) $Y = ZY(1) + (1-Z)Y(0)$.

– (Unconfoundness) $(Y(0), Y(1)) \perp\!\!\!\perp Z|X$.

– (Positivity) $0 < e(X) < 1$ almost everywhere.

• **Interested causal estimands—WATE**

$$\tau_g = \frac{\mathbb{E}[g(X)\tau(X)]}{\mathbb{E}[g(X)]} = \frac{\int \tau(x)f(x)g(x)d\mu(x)}{\int g(x)f(x)d\mu(x)}$$

with a base measure $\mu(\cdot)$, $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X=x]$, and $g(x)$ is a *selection function* which specifies a target population. We consider, in our work, it is always a function of $e(x)$. See figure below for an illustration of different selection functions we may use.
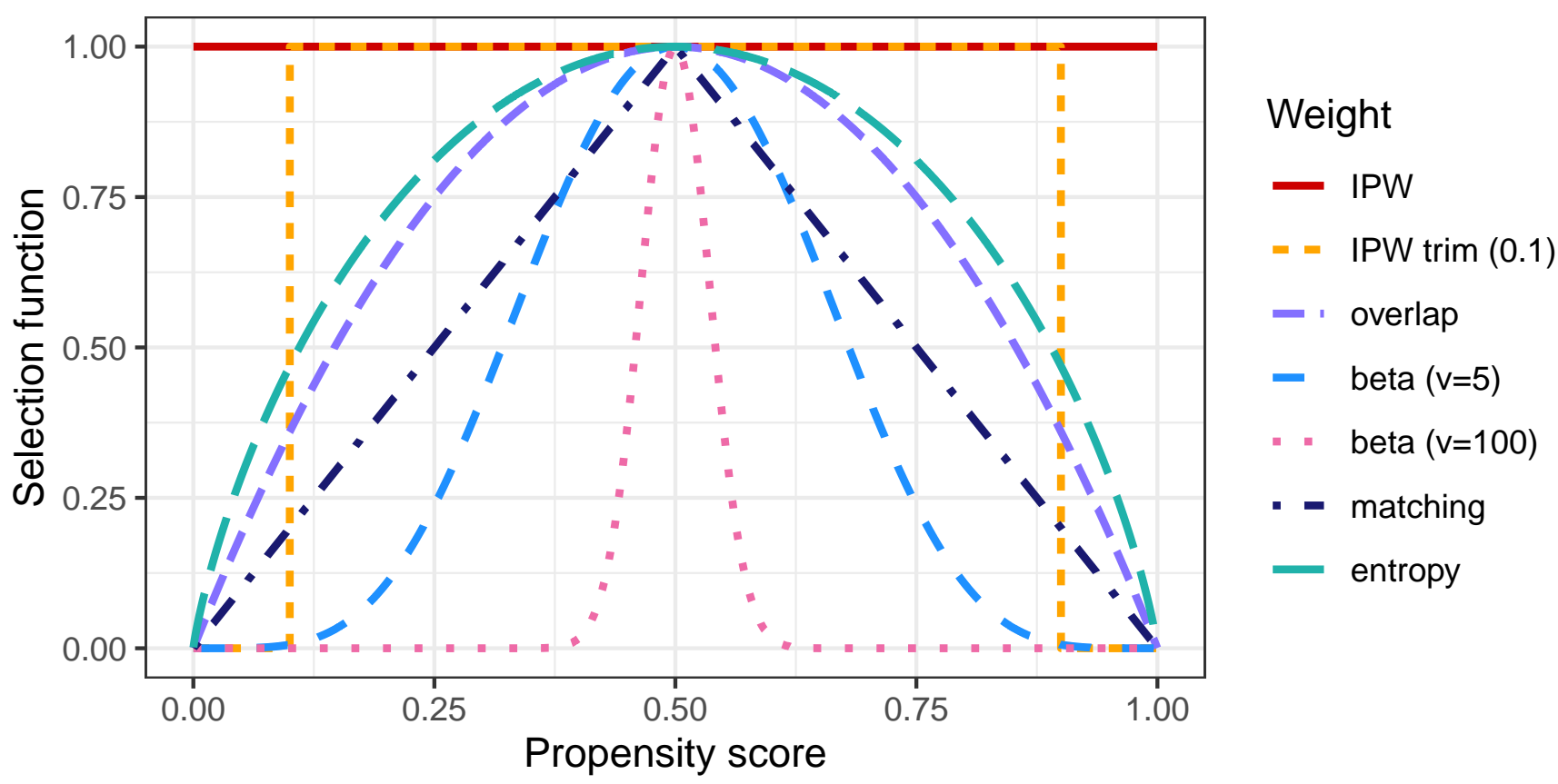
• **Simple weighting estimator (Hájek)**

$$\widehat{\tau}_g^{\mathrm{H}} = \frac{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)Y_i}{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)} - \frac{\sum_{i=1}^N (1-Z_i)\widehat{w}_0(X_i)Y_i}{\sum_{i=1}^N (1-Z_i)\widehat{w}_0(X_i)},$$

where $w_z(x), z=0,1$ is the balancing weight with $(w_0(x), w_1(x)) \propto \left(\frac{g(x)}{1-e(x)}, \frac{g(x)}{e(x)}\right)$.

• **Achieving semiparametric efficiecy via augmented estimators**

$$\widehat{\tau}_g^{\mathrm{aug}} = \widehat{\tau}_g^{\mathrm{H}} + \frac{\sum_{i=1}^N g(X_i)\{\widehat{m}_1(X_i) - \widehat{m}_0(X_i)\}}{\sum_{i=1}^N g(X_i)}$$
$$- \frac{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)\widehat{m}_1(X_i)}{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)} - \frac{\sum_{i=1}^N (1-Z_i)\widehat{w}_0(X_i)\widehat{m}_0(X_i)}{\sum_{i=1}^N (1-Z_i)\widehat{w}_0(X_i)}$$

where $m_z(X) = \mathbb{E}[Y(z)|X], z=0,1$, is the outcome regression model for $Y(z)$.

• **Large sample properties and variance bound**

$\widehat{\tau}_g^{\mathrm{H}}$ and $\widehat{\tau}_g^{\mathrm{aug}}$ are consistent, regular and asymptotic linear (RAL) estimators, and $\widehat{\tau}_g^{\mathrm{aug}}$ is efficient when both the propensity score and outcome models are correctly specified. The efficient influence function (EIF) of τ_g is (shown by [3])

$$F_g(X,Y,Z) = \frac{g(X)}{\mathbb{E}[g(X)]}\{\psi(X,Z,Y) + \tau(X) - \tau_g\}, \quad (1)$$

where

$$\psi(X,Z,Y) = \frac{Z}{e(X)}\{Y - m_1(X)\} - \frac{1-Z}{1-e(X)}\{Y - m_0(X)\}.$$

An issue with this EIF: it is based on assuming the propensity score is known. Technically speaking, the propensity score is not included in the paramertic submodel of the semiparametric model, and thus the nuisance tangent space is not large enough to quantify the model uncertainty. In practice, we need to estimate the propensity via some postulated models.

Another EIF is by Crump et al. [2], which includes the propensity score as a part of the parametric model, given by

$$\phi_g(X,Z,Y) = F_g(X,Z,Y) + \frac{\partial g(X)}{\partial e(X)}\{\tau(X) - \tau\}\{Z - e(X)\}. \quad (2)$$

When the propensity score is unknown, the variance bound is given by $\mathbb{E}[\phi_g \phi_g']$.

Variance Estimations

• **Bootstrap**

The first procedure is the most standard way. We resample the data (X,Y,Z) with replacement by R replications first, then calculate propensity score weights, fitted outcomes, and then WATE estimates with each bootstrap sample, and use the empirical variance of these estimates as the variance estimate.

Limitation: *time-consuming, especially dealing with massive data, [5] and random violation of positivity may occur. [4]*

• **Post-weighting bootstrap**

in this second procedure, we obtain the propensity score weights and fit the outcome models first in the original sample. Then we resample $(X,Y,Z,\widehat{g},\widehat{w}_1,\widehat{w}_0,\widehat{m}_1)$ together, where $\widehat{g},\widehat{w}_1,\widehat{w}_0$ and $\widehat{m}_1$ are calculated using the original sample. There is an analogy by Austin and Small [1] in matching methods.

Limitation: *No justification but people were using it in practice.*

• **Sandwich variance estimation**

We obtain an unbiased estimating equation

$$\sum_{i=1}^N \Psi_\theta(X_i,Z_i,Y_i) = \sum_{i=1}^N \begin{bmatrix} \psi_\beta(X_i,Z_i) \\ Z_i \psi_{\alpha_1}(X_i,Y_i) \\ (1-Z_i)\psi_{\alpha_0}(X_i,Y_i) \\ g(X_i)\{m_1(X_i) - \tau_{1g}^m\} \\ g(X_i)\{m_0(X_i) - \tau_{0g}^m\} \\ Z_i w_1(X_i)(Y_i - m_1(X_i) - \mu_{1g}) \\ (1-Z_i)w_0(X_i)(Y_i - m_0(X_i) - \mu_{0g}) \end{bmatrix} = 0$$

where $\theta = (\beta', \alpha_1', \alpha_0', \tau_{1g}^m, \tau_{0g}^m, \mu_{1g}, \mu_{0g})'$. Using M-theory, the asymptotic variance of $\widehat{\theta}$ is $A(\theta)^{-1}B(\theta)\{A(\theta)'\}^{-1}$ under regularity conditions, where $A(\theta) = -\mathbb{E}[\partial \Psi_\theta/\partial \theta]$ and $B(\theta) = \mathbb{E}[\Psi_\theta \Psi_\theta']$. Then $\widehat{Var}(\widehat{\tau}_g) = c'\widehat{Var}(\widehat{\theta})c$, for some constant vector c.

Limitation: *The formula depends on model you postulate to the nuisance functions, thus is not as flexible as nonparametric methods. Regularity conditions it relies on may be violated. [4]*

• **Wild bootstrap**

We propose to consider the wild bootstrap for variance estimation. Consider $\widehat{\phi}_g(Z,X,Y)$ an estimator of $\phi_g(Z,X,Y)$ in (2), where we replace τ_g by its related estimator $\widehat{\tau}_g^{\mathrm{H}}$ or $\widehat{\tau}_g^{\mathrm{aug}}$. We generate random vectors $\xi = (\xi_1,\ldots,\xi_N)$ of IID random variables—with mean 0 (or 1), variance 1, and $\mathbb{E}(\xi_i^3) < \infty$, where ξ does not depend on the observed data. Calculate a perturbed estimator $\widehat{\tau}_g^* = \widehat{\tau}_g + \frac{1}{N}\sum_{i=1}^N \xi \widehat{\phi}_g(Z_i,X_i,Y_i)$. Under some conditions,

$$\sqrt{N}(\widehat{\tau}_g^* - \tau_g) = \frac{1}{\sqrt{N}}\sum_{i=1}^N (1+\xi)\widehat{\phi}_g(Z_i,X_i,Y_i) + o_p(1)$$

$$\sqrt{N}(\widehat{\tau}_g^* - \tau_g) \xrightarrow{d} N(0,\Sigma_g^*), \Sigma_g^* = \mathbb{E}[(1+\xi)^2 \phi_g \phi_g'], \text{ as } N \to \infty.$$

Using the RAL property of τ_g^{H} and τ_g^{aug}, the above approximation can be used for variance estimation under repeated sampling on specific perturbed distribution for multiple replications.

Limitation: *It may not work uniformly well. We are investigating the algorithm using extensive empirical studies.*

Simulation findings

We generate data from a random sample of $N = 5000$, in each $M = 1000$ Monte Carlo replication. We consider $X_4 \sim Bern(0.5), X_3 \sim Bern(0.6X_4 + 0.2)$,

$$(X_1,X_2)' \sim \mathcal{N}\left(\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 4 & 0.25 \\ 0.25 & 4 \end{bmatrix}\right),$$

and their quadratic and interaction terms $X_5 = X_1^2, X_6 = X_1X_2, X_7 = X_2^2$. Next, we generate the treatment $Z \sim Bern(\{1 + \exp(-X'\beta)\}^{-1})$ where β is the regression coefficients and $X = (1, X_1, \ldots, X_7)'$ is the design matrix, where $1 = (1,\ldots,1)'$. We choose the following $\beta = (\beta_0, \gamma(0.3, 0.4, 0.4, 0.4, 0.1, -0.1, -0.1))$ with varying $(\beta_0, \gamma) \in \{(0.5, 0.25), (0.5, 0.25)\}$ to specify different propensity score models, and thus to simulate two cases of overlap (good and poor).

We simulate a heterogeneous treatment effect via potential outcomes. The true $Y(0)$ is a linear model given by $Y(0) = 0.25 + X_1 + 0.75X_2 + X_3 - 0.25X_4 + 0.5(X_1 + X_2)^2 + \varepsilon$ and then $Y(1) = Y(0) + \delta(X)$ with $\varepsilon \sim \mathcal{N}(0,4)$, where $\delta(X) = 4 + 2(X_1 + X_2)^2 + X_1X_3$.

We present the results of wild bootstrap variance estimation (the coverage rate) as follows.

Estimand	Estimator			
	Hájke	Augmented	Hájke	Augmented
	Good overlap		Poor overlap	
ATE	0.92	0.94	0.08	0.94
ATE (trim)	0.92	0.94	0.81	0.93
ATT	0.94	0.95	0.69	0.93
ATO	0.96	0.96	0.98	0.99
ATM	0.99	0.99	0.99	1.00
ATEN	0.95	0.96	0.96	0.99

The coverage rates of Hájek estimators of ATE and ATT are bad mainly because under poor overlap, there are extreme weights by inverse of propensity scores, thus there is substantial bias in the point estimates under finite sample.

Concluding Remarks

Summary

• Wild bootstrap is an efficient, flexible, and time-saving method for variance estimation for WATE estimators, when we correctly involve the uncertainty associated with estimating nuisance functions and the estimand itself.

Ongoing work

• Extend the simulation study to compare other methods.

• Other nonparametric methods: BLB, SDB, etc. [5]

References

[1] Peter C Austin and Dylan S Small. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319, 2014.

[2] Richard K Crump, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

[3] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[4] Roland A Matsouaka, Yi Liu, and Yunji Zhou. Variance estimation for the average treatment effects on the treated and on the controls. *Statistical Methods in Medical Research*, 32(2):389–403, 2023.

[5] Srijan Sengupta, Stanislav Volgushev, and Xiaofeng Shao. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111(515):1222–1232, 2016.