Improve Efficiency of Doubly Robust Estimator

when Propensity Score is Misspecified

Liangbo Lyu

University of Michigan.

Molei Liu

Columbia University Mailman School of Public Health

Supplementary Material

S1 Numerical implementation of Algorithm 1

For the RWLS problem in Step 3 of Algorithm 1, inspired by Lemma 1, its solution can be calculated based on:

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{a}}(\widehat{\mathbf{a}}^{\mathsf{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{a}})^{-1}\widehat{\mathbf{a}}^{\mathsf{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{b}} - \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{b}},$$

where $\widehat{\mathbf{a}} = \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}\boldsymbol{X}^{\mathsf{T}}\dot{g}(\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\alpha}})$, $\widehat{\mathbf{b}} = \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}\exp(\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\gamma}})v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) + \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}\boldsymbol{X}^{\mathsf{T}}v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})\widehat{\mathbf{L}}$, and $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}^{\mathsf{T}}v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})$. The other steps in Algorithm 1 can be implemented directly following their descriptions.

The two authors have equal contribution.

S2 Methodological extension

S2.1 Construction with alternative nuisance estimators

As emphasized in Remark 2, there are many potential choices of the estimating equations for the nuisance parameters $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$, e.g., the maximum likelihood estimation. Under the correct specified OR model, the variance of $\widehat{\boldsymbol{\gamma}}$ (determined by its estimating equations) has no influence on the asymptotic properties of the PAD estimator. For $\widehat{\boldsymbol{\alpha}}$ obtained using alternative estimating equations to (2.1), we need to modify the form of $\widehat{V}_{\mu}(\boldsymbol{\beta})$. Meanwhile, we can show that this modification will not change the double robustness and variance-reduction properties of our estimator.

Let $\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{S}'(\boldsymbol{\alpha}) = \mathbf{0}$ be the estimating equation of the outcome model $g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha})$ with the score function $\boldsymbol{S}'(\boldsymbol{\alpha})$ different from $\boldsymbol{S}(\boldsymbol{\alpha}) = \boldsymbol{X}\{Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha})\}$ used in the main text. Suppose $\widehat{\boldsymbol{\alpha}}$ converges to $\bar{\boldsymbol{\alpha}}$, the objective variance function then needs to be changed to

$$\widehat{V}_{\mu}(\boldsymbol{\beta}) = \widehat{\mathbb{E}}_{\mathcal{S}}\{\exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}})+\widehat{\boldsymbol{\Psi}}^{\intercal}\boldsymbol{\beta}\}^2 v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})+2\widehat{\mathbf{L}}^{\intercal}\widehat{\mathbb{E}}_{\mathcal{S}}\{\exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}})+\widehat{\boldsymbol{\Psi}}^{\intercal}\boldsymbol{\beta}\}\{Y-g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\}\boldsymbol{S}'(\widehat{\boldsymbol{\alpha}})$$

correspondingly with

$$\widehat{\mathbf{L}} = \mathbb{E}_{\mathcal{S}}\left\{\left.\frac{\partial \boldsymbol{S}'(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right|_{\widehat{\boldsymbol{\alpha}}}\right\}^{-1}\left\{\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}}) - \widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\right\},$$

and we correspondingly have

$$V_\mu(\boldsymbol{\beta}) = \mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^\intercal\boldsymbol{\beta}\}^2 v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})+2\mathbf{L}^\intercal\mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^\intercal\boldsymbol{\beta}\}\{Y-g(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}\boldsymbol{S}'(\bar{\boldsymbol{\alpha}})$$

the limiting function of $\widehat{V}_\mu(\boldsymbol{\beta})$ with

$$\mathbf{L} = \mathbb{E}_{\mathcal{S}}\left\{\left.\frac{\partial \boldsymbol{S}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right|_{\bar{\boldsymbol{\alpha}}}\right\}^{-1}\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}) - \mathbb{E}_{\mathcal{T}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}$$

When the PS model is correct, we again have $\mathbf{L} = \mathbf{0}$ so the same construction of $\boldsymbol{\Psi}$ ensures double robustness. For correct OR model but wrong PS, the modified $V_\mu(\boldsymbol{\beta})$ can lead to variance reduction following a very similar discussion to Section 2.3.

S2.2 Dual construction to augment OR

In analogy to our PAD estimator, to improve the efficiency our the DR estimator under the correct PS and wrong OR models, we propose the Outcome regression Augmented Doubly robust (OAD) estimator in the following algorithm.

Algorithm 1 Outcome regression Augmented Doubly robust (OAD) estimation

[Step 1] Solve the estimating equations in (2.1) to obtain $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$, and obtain the conditional variance estimator as $\widehat{\boldsymbol{\theta}}$.

[Step 2] Let $\boldsymbol{\Phi} = \boldsymbol{\phi}(\boldsymbol{X})$ with function $\boldsymbol{\phi}(\cdot)$, $\widetilde{g}(\boldsymbol{X}^\intercal\widehat{\alpha}) = g(\boldsymbol{X}^\intercal\widehat{\alpha}) - \widehat{\mathbb{E}}_{\mathcal{T}} g(\boldsymbol{X}^\intercal\widehat{\alpha})$ and

$$\widehat{\boldsymbol{\Psi}} = \boldsymbol{\Phi} - \frac{\widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{\Phi}\widetilde{g}(\boldsymbol{X}^\intercal\widehat{\alpha})}{\widehat{\mathbb{E}}_{\mathcal{T}}\widetilde{g}^2(\boldsymbol{X}^\intercal\widehat{\alpha})}\widetilde{g}(\boldsymbol{X}^\intercal\widehat{\alpha}).$$

[Step 3] Solve the restricted weighted least square (RWLS) problem:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{OAD}} = \mathrm{argmin}_{\boldsymbol{\beta}}\widehat{V}_{\mu,\mathsf{OAD}}(\boldsymbol{\beta}), \quad \text{s.t.} \quad \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\widehat{\boldsymbol{\Psi}}^\intercal\boldsymbol{\beta}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}}) = \mathbf{0}, \tag{S2.1}$$

where

$$\begin{aligned}\widehat{V}_{\mu,\mathsf{OAD}}(\boldsymbol{\beta}) =& n^{-1}\widehat{\mathrm{Var}}_{\mathcal{S}}[\{Y - g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}}) - \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}\}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})] + N^{-1}\widehat{\mathrm{Var}}_{\mathcal{T}}\{g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}}) + \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}\} \\ &+ 2\widehat{\mathbf{L}^*}^\intercal[N^{-1}\widehat{\mathrm{Cov}}_{\mathcal{T}}(\boldsymbol{X}, \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}) + n^{-1}\widehat{\mathrm{Cov}}_{\mathcal{S}}\{\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}}), \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})\}],\end{aligned} \tag{S2.2}$$

and $\widehat{\mathbf{L}^*} = \{\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})\boldsymbol{X}^\intercal\}^{-1}\widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})\boldsymbol{X}$.

[Step 4] Obtain the OAD estimator:

$$\widehat{\mu}_{\mathsf{OAD}} = \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}}) - \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{OAD}}\}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}}) + \widehat{\mathbb{E}}_{\mathcal{T}}\{g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}}) + \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{OAD}}\}.$$

To demonstrate how Algorithm 1 works, we define that

$$\begin{aligned}\widetilde{\mu}_{\mathsf{OAD}} =& \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}}) - \boldsymbol{\Psi}^\intercal\bar{\boldsymbol{\beta}}_{\mathsf{OAD}}\}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}) + \widehat{\mathbb{E}}_{\mathcal{T}}\{g(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}}) + \boldsymbol{\Psi}^\intercal\bar{\boldsymbol{\beta}}_{\mathsf{OAD}}\} \\ &+ \mathbb{E}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})\boldsymbol{X}^\intercal\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})\boldsymbol{X}^\intercal\}^{-1}\{\widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X} - \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})\}.\end{aligned}$$

Then similar to our previous analysis, when the PS model is correct, $\widehat{\mu}_{\mathsf{OAD}}$

is asymptotically equivalent to $\widetilde{\mu}_{\mathsf{OAD}}$, and

$$V_{\mu,\mathsf{OAD}}(\boldsymbol{\beta}) = n^{-1}\operatorname{Var}_{\mathcal{S}}[\{Y - g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}) - \boldsymbol{\Psi}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_{\mathsf{OAD}}\}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}})] + N^{-1}\operatorname{Var}_{\mathcal{T}}\{g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}) + \boldsymbol{\Psi}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_{\mathsf{OAD}}\}$$
$$+ 2\mathbf{L}^{*\mathsf{T}}[N^{-1}\mathrm{Cov}_{\mathcal{T}}(\boldsymbol{X}, \boldsymbol{\Psi}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_{\mathsf{OAD}}) + n^{-1}\mathrm{Cov}_{\mathcal{S}}\{\boldsymbol{X}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}), \boldsymbol{\Psi}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_{\mathsf{OAD}}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}})\}],$$

is the limiting function of $\widehat{V}_{\mu,\mathsf{OAD}}(\boldsymbol{\beta})$ specified in Algorithm 1, where

$$\mathbf{L}^* = \{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\mathsf{T}}\}^{-1}\mathbb{E}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}})\boldsymbol{X}.$$

Similar to the PAD method, when $\boldsymbol{\beta} = \mathbf{0}$, $V_{\mu,\mathsf{OAD}}(\boldsymbol{\beta})$ reduces to the asymptotic variance of the standard DR estimator (with a constant difference invariant with $\boldsymbol{\beta}$). Thus, $\widehat{\mu}_{\mathsf{OAD}}$ has a smaller variance than the standard DR estimator when the PS model is correct and the OR model is wrong, under which we typically have $\bar{\boldsymbol{\beta}} \neq \mathbf{0}$. On the other hand, when the OR is correctly specified, we have $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$, $\mathbf{L}^* = \mathbf{0}$, and thus

$$\frac{\partial V_{\mu,\mathsf{OAD}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\mathbf{0}} = \mathrm{Cov}_{\mathcal{T}}(g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}), \boldsymbol{\Psi}).$$

By definition of $\boldsymbol{\Psi}$, we have $\mathrm{Cov}_{\mathcal{T}}(g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}), \boldsymbol{\Psi}) = \mathbf{0}$, implying that $\bar{\boldsymbol{\beta}}_{\mathsf{OAD}} = 0$ is the (unique) population-level solution of (S2.1). Hence, similar to the analysis in Section 2.2, $\widehat{\mu}_{\mathsf{OAD}}$ preserved the same DR property as $\widehat{\mu}_{\mathsf{DR}}$, i.e., being root-$n$ consistent whenever the PS or the OR model is correctly specified.

S2.3 Ensemble strategy

One might be interested in deriving an estimator more efficient than the standard DR estimator whenever the PS or OR model is misspecified (and the other one is correct). This can be achieved through a natural extension that convexly combines the PAD and OAD estimators with the optimal weights minimizing the asymptotic variance. In specific, we define the ensemble estimator

$$\widehat{\mu}_{\mathsf{ENS}}(c) = c\widehat{\mu}_{\mathsf{PAD}} + (1-c)\widehat{\mu}_{\mathsf{OAD}},$$

where $c \in [0,1]$ is some ensemble weight to be specified. Since both $\widehat{\mu}_{\mathsf{PAD}}$ and $\widehat{\mu}_{\mathsf{OAD}}$ preserve the DR property, $\widehat{\mu}_{\mathsf{ENS}}(c)$ is also an DR estimator of μ_0 with any c. Based on this, we propose to choose the c from $[0,1]$ that minimizes

$$\widehat{\mathrm{aVar}}\left\{\sqrt{n}(\widehat{\mu}_{\mathsf{ENS}}(c) - \mu_0)\right\} = \widehat{\mathrm{aVar}}\left\{\sqrt{n}(c\widehat{\mu}_{\mathsf{PAD}} + (1-c)\widehat{\mu}_{\mathsf{OAD}} - \mu_0)\right\}. \quad \text{(S2.3)}$$

$\widehat{\mathrm{aVar}}\{\sqrt{n}(\widehat{\mu}_{\mathsf{ENS}}(c) - \mu_0)\}$ is a quadratic function of c and can be extracted using the asymptotic covariance matrix of $(\widehat{\mu}_{\mathsf{PAD}}, \widehat{\mu}_{\mathsf{OAD}})$ that can be directly estimated from the asymptotic expansions of $\widehat{\mu}_{\mathsf{PAD}}$ and $\widehat{\mu}_{\mathsf{OAD}}$. Let $\widehat{c}$ be the minimizer of (S2.3), and the ensemble estimator $\widehat{\mu}_{\mathsf{ENS}} = \widehat{\mu}_{\mathsf{ENS}}(\widehat{c})$. It is not hard to see that the asymptotic variance of $\widehat{\mu}_{\mathsf{ENS}}$ is always smaller or equal to those of $\widehat{\mu}_{\mathsf{ENS}}(0) = \widehat{\mu}_{\mathsf{OAD}}$ and $\widehat{\mu}_{\mathsf{ENS}}(1) = \widehat{\mu}_{\mathsf{PAD}}$ as $\widehat{c}$ minimizes (S2.3). Thus, $\widehat{\mu}_{\mathsf{ENS}}$ can realize bias-reduction compared to the standard DR estimator $\widehat{\mu}_{\mathsf{DR}}$, whichever nuisance model is correct and the other one is wrong. This ensemble strategy can also be used to combine our estimator with existing

intrinsic efficient estimators like Shu and Tan (2018).

S2.4 Extension to ATE estimation

Problem setup

In this section, we extend our method to provide an PAD estimator for the average treatment effects (ATE) that attains similar DR and variance reduction properties as that for ATT. Suppose there are independent samples of (Δ_i, Y_i, X_i) for $i = 1, 2, \ldots, n$, where Δ_i is the binary treatment variable. Let $Y(\delta)$ denote the counterfactual outcome under the treatment status $\Delta = \delta \in \{0, 1\}$. We are interested in estimating the ATE parameter $\mu_{\text{ATE}} = \mathbb{E}\{Y(1)\} - \mathbb{E}\{Y(0)\}$, under the standard assumptions (Hernán and Robins, 2010, e.g.) including (i) exchangeability that $\{Y(0), Y(1)\} \perp\!\!\!\perp \Delta \mid \boldsymbol{X}$; (ii) positivity that $\pi(\boldsymbol{X}) = \Pr(\Delta = 1 \mid \boldsymbol{X})$ stays away from 0 and 1 for all $\boldsymbol{X}$; and (iii) consistency that $Y = Y(\delta)$ for all subjects with $\Delta = \delta$.

Following the common setup in Bang and Robins (2005), we specify the PS model $\pi(\boldsymbol{X}) = \text{expit}(\boldsymbol{X}^\intercal \boldsymbol{\gamma})$ with $\text{expit}(a) = e^a/(1 + e^a)$, and the OR models $\mathbb{E}(Y \mid \boldsymbol{X}, \Delta = \delta) = g(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_\delta)$ for $\delta = 0, 1$. Similar to (2.1), we

derive their estimators $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}_\delta$ by solving

$$\widehat{\mathbb{E}}_0 \boldsymbol{X} \exp(\boldsymbol{X}^\intercal \boldsymbol{\gamma}) = \widehat{\mathbb{E}}_1 \boldsymbol{X}, \quad \widehat{\boldsymbol{S}}_\delta(\boldsymbol{\alpha}_\delta) = \widehat{\mathbb{E}}_\delta \boldsymbol{X}\{Y - g(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_\delta)\} = \mathbf{0}, \quad \delta = 0, 1, \tag{S2.4}$$

where $\widehat{\mathbb{E}}_0$ and $\widehat{\mathbb{E}}_1$ represent the empirical mean operator on treatment ($\Delta = 1$) and control ($\Delta = 0$) samples respectively. Then by Bang and Robins (2005), the standard DR estimator of ATE is constructed as

$$\begin{aligned}\widehat{\mu}_{\mathsf{ATE}} =& \widehat{\mathbb{E}}\left[\Delta\{1 + \exp(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\gamma}})\}\{Y - g(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_1)\} + g(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_1)\right] \\ & - \widehat{\mathbb{E}}\left[(1 - \Delta)\{1 + \exp(-\boldsymbol{X}^\intercal \widehat{\boldsymbol{\gamma}})\}\{Y - g(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_0)\} + g(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_0)\right],\end{aligned}$$

where $\widehat{\mathbb{E}}$ is the empirical mean operator on samples $i = 1, 2, \ldots, n$.

PAD estimation of ATE

We propose the PAD estimation of ATE in Algorithm 2 and provide some heuristical justification on it.

Algorithm 2 The PAD estimation of ATE

[Step 1] Solve the estimating equations in (S2.4) to obtain $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\alpha}}_0$ and $\widehat{\boldsymbol{\alpha}}_1$, and derive the conditional variance estimators for Y on treatment and control samples as $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_0$.

[Step 2] For $\delta = 0, 1$, specify $\boldsymbol{\Phi}_\delta = \boldsymbol{\phi}_\delta(\boldsymbol{X})$ of larger dimensionality than $\boldsymbol{X}$ using any basis function $\boldsymbol{\phi}_\delta(\cdot)$, and take $\widehat{\boldsymbol{\Psi}}_\delta = \boldsymbol{\Phi}_\delta - \widehat{\mathbb{E}}[\boldsymbol{\Phi}_\delta v_{\widehat{\boldsymbol{\theta}}_\delta}(\boldsymbol{X})]/\widehat{\mathbb{E}} v_{\widehat{\boldsymbol{\theta}}_\delta}(\boldsymbol{X})$.

[Step 3] Solve the restricted weighted least square (RWLS) problem:

$$\widehat{\boldsymbol{\beta}}_\delta = \text{argmin}_{\boldsymbol{\beta}_\delta} \widehat{V}_{\mu_\delta}(\boldsymbol{\beta}_\delta), \quad \text{s.t.} \quad \widehat{\mathbb{E}}_\delta \boldsymbol{X}^\intercal \dot{g}(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_\delta) \widehat{\boldsymbol{\Psi}}_\delta \boldsymbol{\beta} = \boldsymbol{0}, \tag{S2.5}$$

where the objective function is defined as

$$\widehat{V}_{\mu_\delta}(\boldsymbol{\beta}_\delta) = \widehat{\mathbb{E}} I(\Delta = \delta)[\exp\{(2\delta - 1)\boldsymbol{X}^\intercal \widehat{\boldsymbol{\gamma}}\} + 1 + \widehat{\boldsymbol{\Psi}}_\delta^\intercal \boldsymbol{\beta}_\delta]^2 v_{\widehat{\boldsymbol{\theta}}_\delta}(\boldsymbol{X}) + 2\widehat{\mathbf{L}}_\delta^\intercal \widehat{\mathbb{E}} I(\Delta = \delta) \boldsymbol{X} [\exp\{(2\delta - 1)\boldsymbol{X}^\intercal \widehat{\boldsymbol{\gamma}}\} + 1 + \widehat{\boldsymbol{\Psi}}_\delta^\intercal \boldsymbol{\beta}_\delta] v_{\widehat{\boldsymbol{\theta}}_\delta}(\boldsymbol{X}),$$

and

$$\widehat{\mathbf{L}}_\delta = \left\{ \widehat{\mathbb{E}}_\delta \frac{\partial \widehat{\boldsymbol{S}}_\delta(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^\intercal} \bigg|_{\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}_\delta} \right\}^{-1} \left[\widehat{\mathbb{E}} I(\Delta = \delta) \boldsymbol{X} \dot{g}(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_\delta)[1 + \exp\{(2\delta - 1)\boldsymbol{X}^\intercal \widehat{\boldsymbol{\gamma}}\}] - \widehat{\mathbb{E}} \boldsymbol{X} \dot{g}(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_\delta) \right]$$

[Step 4] Obtain the PAD estimator for $\mu_\delta = \mathbb{E}\{Y(\delta)\}$ as

$$\widehat{\mu}_\delta = \widehat{\mathbb{E}} \left[I(\Delta = \delta)[1 + \exp\{(2\delta - 1)\boldsymbol{X}^\intercal \widehat{\boldsymbol{\gamma}}\} + \widehat{\boldsymbol{\Psi}}_\delta^\intercal \widehat{\boldsymbol{\beta}}_\delta]\{Y - g(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_\delta)\} + g(\boldsymbol{X}^\intercal \widehat{\boldsymbol{\alpha}}_\delta) \right]$$

[Step 5] Obtain the ATE estimator through

$$\widehat{\mu}_{\text{PAD-ATE}} = \widehat{\mu}_1 - \widehat{\mu}_0$$

Suppose that all estimators converge to their limiting values (holding under some reasonable regularity conditions). For $\delta = 0, 1$, let $\boldsymbol{\Psi}_\delta = \boldsymbol{\Phi} -$

$\mathbb{E}\boldsymbol{\Phi} v_{\boldsymbol{\theta}_\delta}(\boldsymbol{X})/\mathbb{E}v_{\boldsymbol{\theta}_\delta}(\boldsymbol{X})$, $\bar{\boldsymbol{\beta}}_\delta$, and $\bar{\boldsymbol{\alpha}}_\delta$ and $\bar{\boldsymbol{\gamma}}$ represent the limit of $\widehat{\boldsymbol{\beta}}_\delta$, $\widehat{\boldsymbol{\alpha}}_\delta$, and $\widehat{\boldsymbol{\gamma}}$ respectively,

$$\mathbf{L}_\delta = \left\{ \mathbb{E}_\delta \frac{\partial \boldsymbol{S}_\delta(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^\intercal}\Big|_{\bar{\boldsymbol{\alpha}}_\delta} \right\}^{-1} \left\{ \mathbb{E}I(\Delta=\delta)\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}}_\delta)[1+\exp\{(2\delta-1)\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}\}] - \mathbb{E}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}}_\delta)\right\},$$

and

$$V_{\mu_\delta}(\boldsymbol{\beta}_\delta) = \mathbb{E}I(\Delta=\delta)[\exp\{(2\delta-1)\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}\}+1+\boldsymbol{\Psi}_\delta^\intercal\boldsymbol{\beta}_\delta]^2 v_{\bar{\boldsymbol{\theta}}_\delta}(\boldsymbol{X}) + 2\mathbf{L}_\delta^\intercal\mathbb{E}I(\Delta=\delta)\boldsymbol{X}[\exp\{(2\delta-1)\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}\}+1+\boldsymbol{\Psi}_\delta^\intercal\boldsymbol{\beta}_\delta]v_{\bar{\boldsymbol{\theta}}_\delta}(\boldsymbol{X})$$

as the limiting function of $\widehat{V}_{\mu_\delta}(\boldsymbol{\beta}_\delta)$. Next, we shall discuss the properties of $\widehat{\mu}_{\mathsf{PAD\text{-}ATE}}$ in two scenarios including (i) correct PS model; and (ii) correct OR and wrong PS models. Similar to the ATT setting, one can show that $\widehat{\mu}_{\mathsf{PAD\text{-}ATE}}$ can simultaneously attain the DR property and variance-reduction compared to the standard DR under wrong PS and correct OR.

Correct PS. When the PS model is correct, we have $\mathbf{L}_\delta = \mathbf{0}$ for both $\delta = 0, 1$ and, thus,

$$\frac{\partial V_{\mu_\delta}(\boldsymbol{\beta}_\delta)}{\partial \boldsymbol{\beta}_\delta}\Big|_{\boldsymbol{\beta}_\delta=\mathbf{0}} = 2\mathbb{E}I(\Delta=\delta)\boldsymbol{\Psi}_\delta[\exp\{(2\delta-1)\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}\}+1]v_{\bar{\boldsymbol{\theta}}_\delta}(\boldsymbol{X}) = 2\mathbb{E}\boldsymbol{\Psi}_\delta v_{\bar{\boldsymbol{\theta}}_\delta}(\boldsymbol{X}) = \mathbf{0}$$

by the definition of $\boldsymbol{\Psi}_\delta$ for $\delta = 0, 1$. This, combined with the strong convexity of $V_{\mu_\delta}(\boldsymbol{\beta}_\delta)$, implies $\bar{\boldsymbol{\beta}}_\delta = \mathbf{0}$ when the PS model is correct. Thus, $\widehat{\mu}_{\mathsf{PAD\text{-}ATE}}$ attains the same doubly robust property as the standard $\widehat{\mu}_{\mathsf{ATE}}$.

Wrong PS and Correct OR. When the OR model is correctly specified and the PS is misspecified, we have that $\widehat{\mu}_{\mathsf{PAD\text{-}ATE}}$ is asymptotically equivalent to $\widetilde{\mu}_{\mathsf{PAD\text{-}ATE}} = \widetilde{\mu}_1 - \widetilde{\mu}_0$, where

$$\begin{aligned}\widetilde{\mu}_\delta =& \widehat{\mathbb{E}} I(\Delta = \delta)[1 + \exp\{(2\delta - 1)\boldsymbol{X}^\intercal \bar{\boldsymbol{\gamma}}\} + \boldsymbol{\Psi}_\delta^\intercal \bar{\boldsymbol{\beta}}_\delta]\{Y - g(\boldsymbol{X}^\intercal \bar{\boldsymbol{\alpha}}_\delta)\} \\ &+ \widehat{\mathbb{E}} g(\boldsymbol{X}^\intercal \bar{\boldsymbol{\alpha}}_\delta) + \mathbf{L}_\delta^\intercal \widehat{\mathbb{E}} I(\Delta = \delta)\boldsymbol{X}\{Y - g(\boldsymbol{X}^\intercal \bar{\boldsymbol{\alpha}}_\delta)\}\end{aligned}$$

for $\delta = 0, 1$. Based on this, have

$$\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD\text{-}ATE}} - \mu_0)\} = V_{\mu_1}(\boldsymbol{\beta}_1) + V_{\mu_0}(\boldsymbol{\beta}_0) + C,$$

which, after dropping the invariant C, is the sum of $V_{\mu_1}(\boldsymbol{\beta}_1)$ and $V_{\mu_0}(\boldsymbol{\beta}_0)$. When the PS model is wrong, $\partial V_{\mu_\delta}(\boldsymbol{\beta}_\delta)/\partial \boldsymbol{\beta}_\delta$ is typically not $\mathbf{0}$ at $\boldsymbol{\beta}_\delta = \mathbf{0}$ so its minimizer $\bar{\boldsymbol{\beta}}_\delta \neq \mathbf{0}$. Thus, similar to the reason in Section 2.3, we have $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD\text{-}ATE}} - \mu_0)\} \leq \mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{ATE}} - \mu_0)\}$, and the strict "$<$" will hold in general when PS is misspecified.

S3 Asymptotic justification

S3.1 Proof of Theorem 1

Proof. Proof of Theorem 1 (i).

When the OR is correctly specified, $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$. Consider $\widetilde{\mu}_{\mathsf{OR}}$ where

$$\widetilde{\mu}_{\mathsf{OR}} = \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal \bar{\boldsymbol{\alpha}})\}\{\exp(\boldsymbol{X}^\intercal \bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^\intercal \bar{\boldsymbol{\beta}}\} + \widehat{\mathbb{E}}_{\mathcal{T}} g(\boldsymbol{X}^\intercal \bar{\boldsymbol{\alpha}})$$

$$+\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})-\mathbb{E}_{\mathcal{T}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\mathbb{E}_{\mathcal{S}}\left\{\frac{\partial \boldsymbol{S}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\intercal}}\bigg|_{\bar{\boldsymbol{\alpha}}}\right\}^{-1}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}.$$

It is obvious that $\mathbb{E}\widetilde{\mu}_{\mathsf{OR}}=\mathbb{E}_{\mathcal{T}}g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})=\mu_0$. Hence, by using central limit theorem, we have that $\widetilde{\mu}_{\mathsf{OR}}-\mu_0=O_p(n^{-1/2})$, $n^{1/2}(\widetilde{\mu}_{\mathsf{OR}}-\mu_0)$ weakly converges to gaussian distribution with mean $\mathbf{0}$. On the other hand, we have that

$$\begin{aligned}
&\widehat{\mu}_{\mathsf{PAD}}-\widetilde{\mu}_{\mathsf{OR}}=\widehat{\mathbb{E}}_{\mathcal{S}}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\intercal}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^{\intercal}(\widehat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})+(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi})^{\intercal}\bar{\boldsymbol{\beta}}\}\\
&-[\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}-\widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)](\widehat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})+o_p(n^{-1/2})\\
&-\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})-\mathbb{E}_{\mathcal{T}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}\mathbb{E}_{\mathcal{S}}\left\{\frac{\partial \boldsymbol{S}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\intercal}}\bigg|_{\boldsymbol{\alpha}_0}\right\}^{-1}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\},
\end{aligned}$$

by using central limit theorem, along with Lemma (1)-(4), we have that

$$\begin{aligned}
&\widehat{\mathbb{E}}_{\mathcal{S}}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\intercal}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^{\intercal}(\widehat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})+(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi})^{\intercal}\bar{\boldsymbol{\beta}}\}\\
&=[\widehat{\mathbb{E}}_{\mathcal{S}}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\intercal}](\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})+[\widehat{\mathbb{E}}_{\mathcal{S}}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}\boldsymbol{\Psi}^{\intercal}](\widehat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})\\
&+[\widehat{\mathbb{E}}_{\mathcal{S}}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}\bar{\boldsymbol{\beta}}^{\intercal}](\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi})=O_p(n^{-1/2})o_p(1)+O_p(n^{-1/2})o_p(1)+O_p(n^{-1/2})o_p(1)\\
&=o_p(n^{-1/2}).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&-[\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}-\widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)](\widehat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})\\
&=[\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})+\boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}-\widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)]\widehat{\mathbb{E}}_{\mathcal{S}}\left\{\frac{\partial \boldsymbol{S}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\intercal}}\bigg|_{\bar{\boldsymbol{\alpha}}}\right\}^{-1}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\\
&=\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})-\mathbb{E}_{\mathcal{T}}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)+O_p(n^{-1/2})\}\\
&\qquad\qquad *\left[\mathbb{E}_{\mathcal{S}}\left\{\frac{\partial \boldsymbol{S}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\intercal}}\bigg|_{\boldsymbol{\alpha}_0}\right\}^{-1}+O_p(n^{-1/2})\right]\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}.
\end{aligned}\tag{S3.6}$$

Hence, we have that

$$
\begin{aligned}
&- [\widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\} - \widehat{\mathbb{E}}_{\mathcal{T}} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)](\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}) \\
&- \{\mathbb{E}_{\mathcal{S}} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0) \exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) - \mathbb{E}_{\mathcal{T}} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\} \mathbb{E}_{\mathcal{S}} \left\{ \frac{\partial \boldsymbol{S}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\intercal}} \bigg|_{\boldsymbol{\alpha}_0} \right\}^{-1} \widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X}\{Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\} \\
&= \widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X}\{Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\} O_p(n^{-1/2}) = o_p(n^{-1/2}).
\end{aligned}
$$

Thus, from previous results, we have that $\widehat{\mu}_{\mathsf{PAD}} - \widetilde{\mu}_{\mathsf{OR}} = o_p(n^{-1/2})$. Together with Slutsky theorem, we futher have that $\widehat{\mu}_{\mathsf{PAD}} - \mu_0 = O_p(n^{-1/2})$ and $n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)$ weakly converges to gaussian distribution with mean $\mathbf{0}$.

When the PS is correctly specified, $\bar{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$, we consider $\widetilde{\mu}_{\mathsf{PS}}$ where

$$
\begin{aligned}
\widetilde{\mu}_{\mathsf{PS}} &= \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\{\exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0) + \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\} + \widehat{\mathbb{E}}_{\mathcal{T}} g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}}) \\
&+ \mathbb{E}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\} \exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) \boldsymbol{X}^{\intercal} \{\mathbb{E}_{\mathcal{S}} \boldsymbol{X} \exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0) \boldsymbol{X}^{\intercal}\}^{-1} \{\widehat{\mathbb{E}}_{\mathcal{T}} \boldsymbol{X} - \widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X} \exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0)\}.
\end{aligned}
$$

Together with the results from Lemma (4), we have that $\mathbb{E}\widetilde{\mu}_{\mathsf{PS}} = \mathbb{E}_{\mathcal{S}} Y \exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) = \mathbb{E}_{\mathcal{T}} Y = \mu_0$. By using the central limit theorem, we have that $\widetilde{\mu}_{\mathsf{PS}} - \mu_0 = O_p(n^{-1/2})$, $n^{1/2}(\widetilde{\mu}_{\mathsf{PS}} - \mu_0)$ weakly converges to gaussian distribution with mean $\mathbf{0}$. On the other hand, we have that

$$
\begin{aligned}
\widehat{\mu}_{\mathsf{PAD}} - \widetilde{\mu}_{\mathsf{PS}} &= \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) \boldsymbol{X}^{\intercal}(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^{\intercal}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})\} \\
&\quad - \mathbb{E}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\} \exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) \boldsymbol{X}^{\intercal} \{\mathbb{E}_{\mathcal{S}} \boldsymbol{X} \exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0) \boldsymbol{X}^{\intercal}\}^{-1} \{\widehat{\mathbb{E}}_{\mathcal{T}} \boldsymbol{X} - \widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X} \exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0)\} \\
&\quad + o_p(n^{-1/2})
\end{aligned}
$$

By using the techniques from (S3.6), we would have

$$\widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\intercal}(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$$

$$- \mathbb{E}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\intercal}\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0)\boldsymbol{X}^{\intercal}\}^{-1}\{\widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X} - \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma}_0)\}$$

$$= o_p(n^{-1/2})$$

And from Lemma A4, we have $\widehat{\boldsymbol{\beta}} = O_p(n^{-1/2})$. Thus, we have $\widehat{\mu}_{\mathsf{PAD}} - \mu_0 = O_p(n^{-1/2})$. On the other hand, it is worth noticing that $\widehat{\boldsymbol{\beta}}$ is the continuous function of $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$, so under central limit theorem and Slutsky theorem, we would have the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. Hence, we further have that $n^{1/2}(\widehat{\mu}_{\mathsf{PS}} - \mu_0)$ weakly converges to gaussian distribution with mean $\mathbf{0}$. □

Proof. Proof of Theorem 1 (ii).

First we denote $\mathbf{U}$ as

$$\mathbf{U} = \mathrm{Var}_{\mathcal{T}}(\mathbb{E}(Y|\boldsymbol{X})) + \mathbf{L}^{\intercal}\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\boldsymbol{X}^{\intercal}\,\mathrm{Var}(Y|\boldsymbol{X})\mathbf{L}$$

When the OR is correctly specified, the asymptotic variance of $\widehat{\mu}_{\mathsf{PAD}}$, $\mathrm{Var}\{n^{-1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)\}$ is

$$\mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}^2 v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2\mathbf{L}^{\intercal}\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + \mathbf{U},$$

and $\bar{\boldsymbol{\beta}}$ contributes to minimizing this variance. When $\bar{\boldsymbol{\beta}} = \mathbf{0}$, the function above is written as

$$\mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\}^2 v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2\mathbf{L}^{\intercal}\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\{\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + \mathbf{U},$$

which is the same as the asymptotic variance of $\widehat{\mu}_{\mathsf{DR}}$, $\mathrm{Var}\{n^{-1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)\}$. Hence, when $\bar{\boldsymbol{\beta}} \neq \mathbf{0}$, $\widehat{\mu}_{\mathsf{PAD}}$ has the smaller asymptotic variance than standard doubly robust estimator $\widehat{\mu}_{\mathsf{DR}}$. □

Proof. Proof of Theorem 1 (iii).

When both the PS and OR is correctly specified, consider $\widetilde{\mu}_B$, where

$$\widetilde{\mu}_B = \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_0)\} \exp(\boldsymbol{X}^\intercal \boldsymbol{\gamma}_0) + \widehat{\mathbb{E}}_{\mathcal{T}} g(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_0).$$

By using central limit theorem, $n^{1/2}(\widetilde{\mu}_B - \mu_0)$ weakly converges to gaussian distribution with mean $\mathbf{0}$. On the other hand, by using Taylor series expansion, we would have $\widehat{\mu}_{\mathsf{PAD}} - \widetilde{\mu}_B = o_p(n^{-1/2})$ and $\widehat{\mu}_{\mathsf{DR}} - \widetilde{\mu}_B = o_p(n^{-1/2})$. Hence, they have the same asymptotic variance. □

S3.2 Technical lemma

Lemma 1. *Define* $\mathbf{a} := \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\intercal \dot{g}(\boldsymbol{X}^\intercal \bar{\boldsymbol{\alpha}})$, $\mathbf{b} := \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \exp(\boldsymbol{X}^\intercal \bar{\boldsymbol{\gamma}}) v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\intercal v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) \mathbf{L}$, *and* $\boldsymbol{\Sigma} := \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{\Psi}^\intercal v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})$, *under Assumption 3, the solution of the RWLS problem in Algorithm 1 is*

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\mathbf{a}(\mathbf{a}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{a})^{-1}\mathbf{a}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{b} - \boldsymbol{\Sigma}^{-1}\mathbf{b}.$$

Proof. First we introduce Lagrange multiplier $\boldsymbol{\lambda}$ and write (2.5) as the Lagrange form:

$$\bar{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}}\, \mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{X}^\intercal \bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^\intercal \boldsymbol{\beta}\}^2 v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})$$

$$+2\mathbf{L}^\top \mathbb{E}_{\mathcal{S}} \boldsymbol{X}\{\exp(\boldsymbol{X}^\top \bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^\top \boldsymbol{\beta}\} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) - \boldsymbol{\lambda}^\top \mathbb{E}_{\mathcal{S}} \boldsymbol{X} \dot{g}(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}) \boldsymbol{\Psi}^\top \boldsymbol{\beta}.$$

Then we have the partial derivative of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$:

$$\mathbb{E}_{\mathcal{S}} \boldsymbol{X} \dot{g}(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}) \boldsymbol{\Psi}^\top \boldsymbol{\beta} = \mathbf{0}, \tag{S3.7}$$

and

$$2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi}\{\exp(\boldsymbol{X}^\top \bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^\top \boldsymbol{\beta}\} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\top v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) \mathbf{L} - \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\top \dot{g}(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}) \boldsymbol{\lambda} = \mathbf{0}. \tag{S3.8}$$

From (S3.8) we have

$$\boldsymbol{\beta} = \{2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{\Psi}^\top v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\}^{-1} \{\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\top \dot{g}(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}) \boldsymbol{\lambda} - 2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \exp(\boldsymbol{X}^\top \bar{\boldsymbol{\gamma}}) v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) - 2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\top v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) \mathbf{L}\},$$

together with (S3.7), we have

$$\mathbb{E}_{\mathcal{S}} \boldsymbol{X} \dot{g}(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}) \boldsymbol{\Psi}^\top \{\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{\Psi}^\top v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\}^{-1}$$
$$* \{\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\top \dot{g}(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}) \boldsymbol{\lambda} - 2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \exp(\boldsymbol{X}^\top \bar{\boldsymbol{\gamma}}) v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) - 2\mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^\top v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) \mathbf{L}\} = \mathbf{0}.$$

this function can be simplified as

$$\mathbf{a}^\top \boldsymbol{\Sigma}^{-1} (\mathbf{a}\boldsymbol{\lambda} - 2\mathbf{b}) = \mathbf{0},$$

and we further have

$$\boldsymbol{\lambda} = 2(\mathbf{a}^\top \boldsymbol{\Sigma}^{-1} \mathbf{a})^{-1} \mathbf{a}^\top \boldsymbol{\Sigma}^{-1} \mathbf{b}.$$

Hence, we have

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1} \mathbf{a} (\mathbf{a}^\top \boldsymbol{\Sigma}^{-1} \mathbf{a})^{-1} \mathbf{a}^\top \boldsymbol{\Sigma}^{-1} \mathbf{b} - \boldsymbol{\Sigma}^{-1} \mathbf{b}.$$

□

Lemma 2. *Under Assumptions 1–4, we have that* $\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi} = O_p(n^{-1/2})$.

Proof. By definition, we would have that

$$\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi} = \frac{\widehat{\mathbb{E}}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})\}}{\widehat{\mathbb{E}}_{\mathcal{T}} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})} - \frac{\mathbb{E}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\}}{\mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})}.$$

Under Assumption 4, we have that

$$\widehat{\mathbb{E}}_{\mathcal{T}} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) - \mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = \widehat{\mathbb{E}}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + \widehat{\mathbb{E}}_{\mathcal{T}} \frac{\partial v_{\boldsymbol{\theta}}(\boldsymbol{X})}{\partial \boldsymbol{\theta}}|_{\widetilde{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - \mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = O_p(n^{-1/2}) \tag{S3.9}$$

for $\widetilde{\boldsymbol{\theta}}$ between $\widehat{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$. By using the same techniques, we have that $\widehat{\mathbb{E}}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})\} - \mathbb{E}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\} = O_p(n^{-1/2})$. And we have

$$\begin{aligned}\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi} =& \frac{\widehat{\mathbb{E}}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})\}\mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) - \mathbb{E}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\}\widehat{\mathbb{E}}_{\mathcal{T}} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})}{\widehat{\mathbb{E}}_{\mathcal{T}} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})\mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})} \\ =& \frac{O_p(n^{-1/2})\mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) - \mathbb{E}_{\mathcal{T}}\{\boldsymbol{\Phi} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\} O_p(n^{-1/2})}{\{\mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + O_p(n^{-1/2})\}\mathbb{E}_{\mathcal{T}} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})} = O_p(n^{-1/2}).\end{aligned}$$

□

Lemma 3. *Under Assumptions 1 and 2, we have that* $\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}} = O_p(n^{-1/2})$ *and* $\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}} = O_p(n^{-1/2})$.

Proof. The estimation of $\boldsymbol{\gamma}$ has been given as

$$\widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X} \exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}}) = \widehat{\mathbb{E}}_{\mathcal{T}} \boldsymbol{X},$$

by applying Taylor series expansion, we have

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\bar{\boldsymbol{\gamma}})+n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\widetilde{\boldsymbol{\gamma}})\boldsymbol{X}_i^{\intercal}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})=N^{-1}\sum_{i=n+1}^{n+N}\boldsymbol{X}_i,$$

where $\widetilde{\boldsymbol{\gamma}}$ is some vector between $\widehat{\boldsymbol{\gamma}}$ and $\bar{\boldsymbol{\gamma}}$. According to (Van der Vaart, 2000, Chapter 5), we have $\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}}=o_p(1)$. Let $\mathbf{J}$ represent matrix $n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\widetilde{\boldsymbol{\gamma}})\boldsymbol{X}_i^{\intercal}$, and we have that

$$\mathbf{J}=n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}_i^{\intercal}+n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}^*)\boldsymbol{X}_i^{\intercal}\boldsymbol{X}_i(\widetilde{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})=\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})\boldsymbol{X}^{\intercal}+o_p(1)$$

for $\boldsymbol{\gamma}^*$ between $\widetilde{\boldsymbol{\gamma}}$ and $\bar{\boldsymbol{\gamma}}$. Hence, by central limit theorem and Slutsky theorem, we have that,

$$\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}}=\mathbf{J}^{-1}\left\{N^{-1}\sum_{i=n+1}^{n+N}\boldsymbol{X}_i-n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\bar{\boldsymbol{\gamma}})\right\}$$

$$=\mathbf{J}^{-1}\left\{N^{-1}\sum_{i=n+1}^{n+N}\boldsymbol{X}_i-\mathbb{E}_{\mathcal{T}}\boldsymbol{X}+\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\gamma}})-n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\exp(\boldsymbol{X}_i^{\intercal}\bar{\boldsymbol{\gamma}})\right\}=O_p(n^{-1/2}).$$

Furthermore, The estimation equation of $\widehat{\alpha}$ is given by

$$\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{S}(\widehat{\boldsymbol{\alpha}})=\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\}=\mathbf{0},$$

by using Taylor series expansion, we have that

$$\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}+\widehat{\mathbb{E}}_{\mathcal{S}}\frac{\partial\boldsymbol{S}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}^{\intercal}}\bigg|_{\widetilde{\alpha}}(\widehat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})=\mathbf{0}$$

for $\widetilde{\alpha}$ between $\widehat{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\alpha}}$, and we have

$$\widehat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}}=-\widehat{\mathbb{E}}_{\mathcal{S}}\left\{\frac{\partial\boldsymbol{S}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}^{\intercal}}\bigg|_{\widetilde{\boldsymbol{\alpha}}}\right\}^{-1}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y-g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}.$$

By using the same techniques as those for obtaining the asymptotic properties of $\widehat{\boldsymbol{\gamma}}$, under Assumptions 1 and 2, we have $\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}} = O_p(n^{-1/2})$. $\square$

Lemma 4. *Under Assumptions 1–4 and by Lemmas 1–3, we can obtain that* $\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}} = O_p(n^{-1/2})$. *In addition, when the PS is correctly specified, we further have* $\bar{\boldsymbol{\beta}} = \mathbf{0}$ *and* $\widehat{\boldsymbol{\beta}} = O_p(n^{-1/2})$.

Proof. By using the same techniques as (S3.9), under Condition 2-4, we first have that

$$\begin{aligned}\widehat{\mathbf{a}} - \mathbf{a} =& \widehat{\mathbb{E}}_{\mathcal{S}} \widehat{\boldsymbol{\Psi}} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal} \widehat{\boldsymbol{\alpha}}) - \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal} \bar{\boldsymbol{\alpha}}) \\ =& \widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal} \bar{\boldsymbol{\alpha}}) - \mathbb{E}_{\mathcal{S}} \boldsymbol{\Psi} \boldsymbol{X}^{\intercal} \dot{g}(\boldsymbol{X}^{\intercal} \bar{\boldsymbol{\alpha}}) + O_p(n^{-1/2}) = O_p(n^{-1/2}).\end{aligned}$$

In addition, we can have that $\widehat{\mathbf{b}} - \mathbf{b} = O_p(n^{-1/2})$ and $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O_p(n^{-1/2})$. Furthermore, we can easily have that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\{\boldsymbol{\Sigma} + O_p(n^{-1/2})\}^{-1} - \boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{\Sigma}^{-1}[\boldsymbol{\Sigma}\{\boldsymbol{\Sigma} + O_p(n^{-1/2})\}^{-1} - \{\boldsymbol{\Sigma} + O_p(n^{-1/2})\}\{\boldsymbol{\Sigma} + O_p(n^{-1/2})\}^{-1}] = O_p(n^{-1/2}),\end{aligned}$$

based on which we can have $(\widehat{\mathbf{a}}^{\intercal}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{a}})^{-1} - (\mathbf{a}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{a})^{-1} = O_p(n^{-1/2})$. Moreover, let $\widehat{\boldsymbol{\Omega}}$ denote $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{a}}(\widehat{\mathbf{a}}^{\intercal}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{a}})^{-1}\widehat{\mathbf{a}}^{\intercal}\widehat{\boldsymbol{\Sigma}}^{-1}$ and $\boldsymbol{\Omega}$ denote $\boldsymbol{\Sigma}^{-1}\mathbf{a}(\mathbf{a}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{a})^{-1}\mathbf{a}^{\intercal}\boldsymbol{\Sigma}^{-1}$, and we can have that $\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} = O_p(n^{-1/2})$. Hence, we have that $\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Omega}}\widehat{\mathbf{b}} - \boldsymbol{\Omega}\mathbf{b} = O_p(n^{-1/2})$.

On the other hand, when the PS is correctly specified, $\mathbf{L} = \mathbf{0}$ and

$\mathbb{E}_{\mathcal{S}}\boldsymbol{\Psi}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}})v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = \mathbb{E}_{\mathcal{T}}\boldsymbol{\Psi}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = \mathbf{0}$, which means

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\Omega}\mathbf{b} = \boldsymbol{\Omega}\{\mathbb{E}_{\mathcal{S}}\boldsymbol{\Psi}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}})v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + \mathbb{E}_{\mathcal{S}}\boldsymbol{\Psi}\boldsymbol{X}^{\mathsf{T}}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\mathbf{L}\} = \boldsymbol{\Omega}\mathbf{0} = \mathbf{0}.$$

And at the same time, we have $\widehat{\boldsymbol{\beta}} = O_p(n^{-1/2})$. □

S4 The mPAD estimator

S4.1 Technical details

To address the range issue of the PS model in aPAD, we propose a multiplicative PAD (mPAD) estimator in Algorithm 3, in which the PS model is taken as $\exp(\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^{\mathsf{T}}\widehat{\boldsymbol{\beta}}')$ with a multiplicative augmentation term $\exp(\widehat{\boldsymbol{\Psi}}'^{\mathsf{T}}\widehat{\boldsymbol{\beta}}')$ added to the original PS model.

Algorithm 3 Multiplicative PAD estimation

[Step 1] Solve the estimating equations in (2.1) to obtain $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$, and obtain the conditional variance estimator as $\widehat{\boldsymbol{\theta}}$.

[Step 2] Specify $\boldsymbol{\Phi} = \boldsymbol{\phi}(\boldsymbol{X})$ of larger dimensionality than $\boldsymbol{X}$ using any basis function $\boldsymbol{\phi}(\cdot)$, and take $\widehat{\boldsymbol{\Psi}}' = \boldsymbol{\Phi} - \widehat{\mathbb{E}}_{\mathcal{T}}[\boldsymbol{\Phi}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})]/\widehat{\mathbb{E}}_{\mathcal{T}}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})$.

[Step 3] Solve the optimization problem:

$$\widehat{\boldsymbol{\beta}}' = \mathrm{argmin}_{\boldsymbol{\beta}}\widehat{V}'_{\mu}(\boldsymbol{\beta}'), \quad \text{s.t.} \quad \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}^\intercal\dot{g}(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})\{\exp(\widehat{\boldsymbol{\Psi}}'^\intercal\boldsymbol{\beta}') - 1\} = \mathbf{0}, \tag{S4.10}$$

where

$$\widehat{V}'_{\mu}(\boldsymbol{\beta}') = \widehat{\mathbb{E}}_{\mathcal{S}}\exp\{2(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^\intercal\boldsymbol{\beta}')\}v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2\widehat{\mathbf{L}}'^\intercal\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^\intercal\boldsymbol{\beta}')v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}), \tag{S4.11}$$

and $\widehat{\mathbf{L}}' = -\widehat{\mathbf{H}}^{-1}\left\{\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^\intercal\widehat{\boldsymbol{\beta}}') - \widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})\right\}$.

[Step 4] Obtain the mPAD estimator through

$$\widehat{\mu}_{\mathsf{mPAD}} = \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^\intercal\widehat{\boldsymbol{\beta}}') + \widehat{\mathbb{E}}_{\mathcal{T}}g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}}).$$

Similar to our discussion in Section 2.3, suppose that all estimators converge to their limiting values. Let $\boldsymbol{\Psi}' = \boldsymbol{\Phi} - \mathbb{E}_{\mathcal{T}}\boldsymbol{\Phi}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})/\mathbb{E}_{\mathcal{T}}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})$,

$$\mathbf{L}' = -\mathbf{H}^{-1}\left\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}} + \boldsymbol{\Psi}'^\intercal\bar{\boldsymbol{\beta}}') - \mathbb{E}_{\mathcal{T}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\right\},$$

$\bar{\boldsymbol{\beta}}'$ be the limits of $\widehat{\boldsymbol{\beta}}'$, and

$$V'_{\mu}(\boldsymbol{\beta}') = \mathbb{E}_{\mathcal{S}}\exp\{2(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}} + \boldsymbol{\Psi}'^\intercal\boldsymbol{\beta}')\}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2\mathbf{L}'^\intercal\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}} + \boldsymbol{\Psi}'^\intercal\boldsymbol{\beta}')v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})$$

be the limiting function of $\widehat{V}'_{\mu}(\boldsymbol{\beta}')$ specified in Algorithm 3. Similarly to

aPAD, we shall comment on two scenarios: (i) correct PS model; and (ii) correct OR and wrong PS, to demonstrate the properties of the proposed mPAD estimator. In specific, similar to aPAD, the mPAD estimator can attain the DR property and a smaller variance than the standard DR under wrong PS. However, one should note that the optimization problem (S4.10) in Algorithm 3 is non-convex and not ensured to converge to the global solution. In addition, as will be discussed in Appendix S4.2, mPAD tends to have worse and less stable finite-sample performance than aPAD due to the presence of squared exponential terms like $\exp\{2(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^\intercal\boldsymbol{\beta}')\}$ in mPAD. Therefore, we still recommend aPAD as the default choice when implementing PAD.

Correct PS model. When the PS model is correct, we have that $\mathbf{L}' = \mathbf{0}$ using the linear constraint in (S4.10) and

$$\frac{\partial V'_\mu(\boldsymbol{\beta}')}{\partial \boldsymbol{\beta}'}\Big|_{\boldsymbol{\beta}'=\mathbf{0}} = 2\mathbb{E}_{\mathcal{S}}\boldsymbol{\Psi}'\exp\{2(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})\}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = 2\mathbb{E}_{\mathcal{T}}\boldsymbol{\Psi}'\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = \mathbf{0}$$

by the definition of $\boldsymbol{\Psi}'$. Thus, $\boldsymbol{\beta}' = \mathbf{0}$ is a (local) minimizer of $V'_\mu(\boldsymbol{\beta}')$, which implies that $\widehat{\boldsymbol{\beta}}'$ converges to $\mathbf{0}$ when PS model is correct. Meanwhile, the PS augmentation does not affect the OR model. Therefore, $\widehat{\mu}_{\mathsf{mPAD}}$ have the same double robustness as $\widehat{\mu}_{\mathsf{DR}}$.

Wrong PS and correct OR. When the OR model is correctly specified and PS is misspecified, $\widehat{\mu}_{\mathsf{mPAD}}$ is asymptotically equivalent with

$$\widetilde{\mu}_{\mathsf{mPAD}} = \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}} + \boldsymbol{\Psi}'^{\mathsf{T}}\bar{\boldsymbol{\beta}}') + \widehat{\mathbb{E}}_{\mathcal{T}}g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}) + \mathbf{L}'^{\mathsf{T}}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y - g(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}})\}.$$

This implies that $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{mPAD}} - \mu_0)\}$ is equal to

$$\mathbb{E}_{\mathcal{S}}\exp^2(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}} + \boldsymbol{\Psi}'^{\mathsf{T}}\boldsymbol{\beta}')v(\boldsymbol{X}) + 2\mathbf{L}'^{\mathsf{T}}\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^{\mathsf{T}}\bar{\boldsymbol{\gamma}} + \boldsymbol{\Psi}'^{\mathsf{T}}\boldsymbol{\beta}')v(\boldsymbol{X}) + C, \tag{S4.12}$$

which, after dropping the invariant C, is equal to $V'_{\mu}(\boldsymbol{\beta}')$. When the PS model is misspecified, $\partial V_{\mu}(\boldsymbol{\beta}')/\partial\boldsymbol{\beta}'$ is typically not $\mathbf{0}$ at $\boldsymbol{\beta}' = \mathbf{0}$ so the minimizer is non-zero. Thus, when the OR model is correct, we have $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{mPAD}} - \mu_0)\} \leq \mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)\}$, and the strict "$<$" will hold in general when PS is wrong.

S4.2 Simulation study

We consider the same settings (L1)–(L3) as introduced in Section 4 and evaluate the performance of mPAD. To solve for the non-convex problem (S4.10), we adopt an iterative Newton-type optimization procedure with a quadratic approximation to the objective function $\widehat{V}'_{\mu}(\boldsymbol{\beta}')$ and a linear approximation to the constraint in (S4.10) at each iteration. Also, we impose an early stop whenever the objective $\widehat{V}'_{\mu}$ will not decrease in the next iteration.

Table 1: The absolute average bias (Bias), standard error (SE), and coverage probability (CP) of the 95% confidence intervals of the mPAD estimators, Relative efficiency (RE) between DR and mPAD, i.e., $\mathrm{Var}(\widehat{\mu}_{\mathsf{DR}})/\mathrm{Var}(\widehat{\mu}_{\mathsf{mPAD}})$ under the settings described in Section 4. All results are based on 1000 repetitions.

	$n = N = 500$				$n = N = 1000$			
Setting	Bias	SE	CP	RE	bias	SE	CP	RE
(L1)	0.001	0.054	0.95	1.09	0.002	0.039	0.94	1.07
(L2)	0.005	0.054	0.96	1.01	0.001	0.041	0.94	0.96
(L3)	0.002	**0.051**	0.94	**1.23**	0.001	**0.035**	0.94	**1.18**

In Table 1, we report mPAD's bias, standard error (SE), and coverage probability (CP) of the 95% CI, and relative efficiency (RE) to the standard DR. Similar to the aPAD estimator studied in Section 4, mPAD performs closely to the standard DR estimator when the PS model is correct, and attains better estimation efficiency than DR when the OR model is correct and PS model is misspecified. Also, mPAD shows similar performance to aPAD when comparing Table 1 with Table 1.

Meanwhile, mainly due to the non-convexity issue and the outlying exponential terms $\exp\{2(\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Psi}}'^{\mathsf{T}}\boldsymbol{\beta}')\}$, mPAD performs poorly in the Gaussian linear settings (G1)–(G3) supposed to be reasonable for the aPAD and standard DR methods. For example, under misspecified OR models,

the mPAD estimator has a bias more than 0.06 in all settings (G1)–(G3) under $N = n = 1000$, which is substantially larger than aPAD.

Bibliography

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Shu, H. and Tan, Z. (2018). Improved estimation of average treatment effects on the treated: Local efficiency, double robustness, and beyond. *arXiv preprint arXiv:1808.01408*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.