Statistica Sinica Preprint No: SS-2023-0029

Title	Improve Efficiency of Doubly Robust Estimator when Propensity Score is Misspecified
Manuscript ID	SS-2023-0029
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0029
Complete List of Authors	Liangbo Lyu and Molei Liu
Corresponding Authors	Molei Liu
E-mails	moleiliu95@gmail.com

Notice: Accepted version subject to English editing.

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

Abstract:

Doubly robust (DR) estimation is a crucial technique in causal inference and missing data problems. We propose a novel **P**ropensity score **A**ugmented **D**oubly robust (PAD) estimator to enhance the commonly used DR estimator for average treatment effect on the treated (ATT), or equivalently, the mean of the outcome under covariate shift. Our proposed estimator attains a lower asymptotic variance than the conventional DR estimator when the propensity score (PS) model is misspecified and the outcome regression (OR) model is correct while maintaining the double robustness property that it is valid when either the PS or OR model is correct. These are realized by introducing some properly calibrated adjustment covariates to linearly augment the PS model and solving a restricted weighted least square (RWLS) problem to minimize the variance of the augmented estimator. Both the asymptotic

The two authors have equal contribution.

analysis and simulation studies demonstrate that PAD can significantly reduce the estimation variance compared to the standard DR estimator when the PS model is wrong and the OR is correct, and maintain close performance to DR when the PS model is correct. We further applied our method to study the effects of eligibility for 401(k) plan on the improvement of net total financial assets using data from the Survey of Income and Program Participation of 1991. *Key words and phrases:*

Causal inference; Covariate shift correction; Double robustness; Intrinsic efficiency; Outcome regression; Propensity score.

1. Introduction

1.1 Background

Doubly robust (DR) estimation has attracted extensive interest in the literature on semiparametric theory and causal inference and is frequently used in biomedical science, economics, and policy science studies. It incorporates two nuisance models, a propensity score (PS) model, and an outcome regression (OR) model to characterize distributions of the exposure and outcome against the adjustment covariates respectively, and draws valid inferences when either one of them is correctly specified. It has been well-established that when both the PS and OR models are correct, the DR estimator is semiparametric efficient and its asymptotic variance does not really depend on the estimating equations for the nuisance models (Tsiatis, 2006, e.g.). Nevertheless, there still remains an intriguing question on how to improve the asymptotic efficiency of the DR estimator when one nuisance model is misspecified. For the scenario with correct PS and wrong OR models, there is a track of work (Cao et al., 2009; Tan, 2010, e.g.) proposing the so-called intrinsic efficient estimator that will be reviewed in Section 1.3. This type of estimator preserves the double robustness property and achieves improved efficiency over the standard DR estimator when the PS model is correct and the OR is wrong. Interestingly, we notice that the dual problem of this, i.e., improving the (intrinsic) efficiency of the DR estimator under wrong PS and correct OR, is supposed to be equally important but has not been handled yet due to certain technical reasons that will be discussed later. Aimed in this paper, filling this methodological blank can effectively complement the existing tools for DR and semiparametric inference.

1.2 Problem Setup

To make our idea easier to understand, we focus on a specific missing data problem: transfer estimation of the outcome's mean in the presence of covariate shift (Huang et al., 2007, e.g.). This is also equivalent to estimating the average treatment effect on the treated (ATT) (Hahn, 2004, e.g.) in the context of causal inference and matching-adjusted indirect comparison frequently conducted in biomedical studies (Signorovitch et al., 2010). Our method could be generalized to other settings such as estimating the average treatment effect (ATE) and transfer learning of regression models (Liu et al., 2020).

Suppose there are n labeled samples with observed outcome Y and covariates $\boldsymbol{X} \in \mathbb{R}^d$, and N unlabeled samples only observed on $\boldsymbol{X}$. Let $\Delta = 1$ indicate that the sample is labeled and $\Delta = 0$ otherwise. The labeled observations $(Y_i, \boldsymbol{X}_i)$ are collected from a source population $\mathcal{S}$ with $\Delta_i = 1$ for $i = 1, 2, \ldots, n$. Assume $(Y_i, \boldsymbol{X}_i) \sim p_{\mathcal{S}}(\boldsymbol{x})q(y|\boldsymbol{x})$ for $i = 1, 2, \ldots, n$ where $p_{\mathcal{S}}(\boldsymbol{x})$ and $q(y|\boldsymbol{x})$ represent the density of $\boldsymbol{X}$ on $\mathcal{S}$ and the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ respectively. Meanwhile, there are unlabeled samples from a target population $\mathcal{T}$ indicated by $\Delta_i = 0$ and only observed on covariates $\boldsymbol{X}_i$ for $i = n+1, \ldots, N+n$. Assume that on $\mathcal{T}$, $(Y_i, \boldsymbol{X}_i) \sim p_{\mathcal{T}}(\boldsymbol{x})q(y|\boldsymbol{x})$ with $p_{\mathcal{T}}(\boldsymbol{x})$ representing the density of $\boldsymbol{X}$ on $\mathcal{T}$ and the distribution of $Y \mid \boldsymbol{X}$ remaining to be the same as that on $\mathcal{S}$. Our goal is to estimate $\mu_0 = \mathbb{E}_{\mathcal{T}} Y$, the marginal mean of Y on $\mathcal{T}$. In the absence of observed Y on the target samples, two simple strategies to estimate μ_0 are introduced below.

(PS) Define the propensity score (PS) or density ratio between the two populations as $r_0(\boldsymbol{x}) = p_{\mathcal{T}}(\boldsymbol{x})/p_{\mathcal{S}}(\boldsymbol{x})$. Estimate $r_0(\boldsymbol{x})$ with some $\widehat{r}(\boldsymbol{x})$ and average the observed Y_i weighted by $\widehat{r}(\boldsymbol{X}_i)$ over $i = 1, 2, \ldots, n$ from $\mathcal{S}$.

(OR) Define the outcome regression (OR) or imputation model for Y as $m_0(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$. Estimate $m_0(\boldsymbol{x})$ with some $\widehat{m}(\boldsymbol{x})$ obtained using the labeled samples and average $\widehat{m}(\boldsymbol{X}_i)$ over $i = n+1, \ldots, n+N$ from $\mathcal{T}$.

Both the PS and OR strategies are built upon the assumption that the distribution of $Y \mid \boldsymbol{X}$ is the same between $\mathcal{S}$ and $\mathcal{T}$ so the knowledge of Y on $\mathcal{S}$ is transferable to $\mathcal{T}$. This is in the same spirit as the *no unmeasured confounding* assumption in the context of causal inference.

1.3 Related literature

Our work is based on the doubly robust (DR) inference framework that has been frequently studied and applied in the past years (Robins et al., 1994; Bang and Robins, 2005; Kang and Schafer, 2007; Tan, 2010; Vermeulen and Vansteelandt, 2015, e.g.). It combines the PS and OR models introduced in Section 1.2 to construct an estimator that is valid when at least one of the two nuisance models are correct and, thus, regarded as a more robust statistical inference procedure than the simple PS and OR strategies. Early work in DR inference (Bang and Robins, 2005; Kang and Schafer, 2007, e.g.) mainly used *working* low-dimensional parametric regression to construct the PS and OR models. Recent progress has been made to accommodate the use of high-dimensional regression or complex machine learning methods in estimating the nuisance models (Chernozhukov et al., 2018; Tan, 2020, e.g.), which is less prone to model misspecification. We focus the scope of this paper on the low-dimensional parametric setting that is technically less involved but more user-friendly and less sensitive to over-fitting in practice. It is also possible and valuable to generalize our work to the settings of high-dimensional parametric (Tan, 2020; Dukes and Vansteelandt, 2020, e.g.) or semi-non-parametric (Liu et al., 2020) nuisance models, in which model misspecification is still an important concern.

There has risen great interest in studying and improving the asymptotic efficiency of the DR estimator. One track of literature studied the local efficiency of the DR estimator, i.e., if it is semiparametric efficient when both the PS and OR models are known or correctly specified. While it was shown that the standard DR estimator for the ATE (Robins et al., 1994) achieves such local efficiency (Hahn, 1998; Tsiatis, 2006). This result cannot be directly applied to the ATT estimator because unlike ATE, the PS model of ATT is informative (or non-ancillary) (Hahn, 1998, 2004). Shu and Tan (2018) further studied this subtle issue and proposed locally efficient DR estimators for ATT based on its influence function.

Meanwhile, another track of literature focuses on improving the efficiency of the DR estimator in the presence of correct PS and potentially wrong OR models and, thus, is more relevant to our work that also aims at automatic variance reduction under model misspecification. A class of *intrinsic efficient*

DR estimator has been proposed for the efficient estimation of ATE (Cao et al., 2009; Tan, 2010), ATT (Shu and Tan, 2018), casual regression model (Rotnitzky et al., 2012), longitudinal data (Han, 2016), individual treatment rule (Pan and Zhao, 2021), etc. This type of estimator is (i) valid when either nuisance model is correct; (ii) equivalent with the standard DR estimator when both models are correct; and (iii) of the minimum variance under correct PS and wrong OR, among all the DR estimators with the same parametric specification of the OR model, and, consequently, more efficient than the standard DR estimator.

Although the correct PS and wrong OR setting has been frequently studied, there is still a paucity of solutions to its dual problem, i.e., enhancing the DR estimator under the wrong PS and correct OR, which is the goal of this work. We also notice some early work like Kang and Schafer (2007) and Cao et al. (2009) arguing that the simple OR strategy is an ideal choice when one knows the PS model is wrong since it is free of PS weighting that may decrease the effective sample size. However, since there are no perfect ways to examine model correctness without any additional assumptions, this strategy cannot be as robust as the DR estimator to OR's misspecification.

It was shown that including more prognostic covariates or auxiliary basis when fitting the PS model can help to reduce the variance of the ATE estimator

(Hahn, 2004; Tsiatis, 2006). Motivated by this, Cheng et al. (2020) proposed a double-index PS estimator for ATE that smooths the treatment over the parametric PS and OR models to achieve the DR property as well as variance reduction under correct PS and wrong OR. Nevertheless, methods in this track of work are ensured to reduce the variance of the ATE estimator only when the PS model is correct. In contrast, our work aims at variance reduction under wrong PS and correct OR. Also, our idea can be used for ATT, ATE (see Supplement S2.4), and other casual parameters while this existing strategy only works for ATE but not ATT.

We also notice a large body of work in statistical learning and causal inference that aims at leveraging some auxiliary data or information to boost the asymptotic efficiency of certain estimators using the idea of augmentation. For example, Kawakita and Kanamori (2013), Chakrabortty et al. (2018), Azriel et al. (2021), and Gronsbell et al. (2022) proposed different semi-supervised learning methods that improve estimation efficiency leveraging large unlabeled data drawn from the same distribution as the labeled samples. Methods like Chen and Chen (2000) and Yang and Ding (2019) utilized external data with error-prone outcomes or covariates to construct control variate for variance reduction. These methods, as well as other examples, rely on some auxiliary data to construct estimators that always converge to zero and are asymptotically correlated with the target estimator. These zero estimators are then used to augment the target estimator properly for variance reduction. Our work also adapts the high-level idea of augmentation. But different from these methods, ours does not leverage any auxiliary samples or knowledge and additionally cares about the need of prioritizing validity (double robustness) over statistical power. Consequently, the asymptotic behavior of our augmented estimator actually varies according to the correctness of the nuisance models and is more technically involved in to study.

1.4 Our contribution

To estimate μ_0 introduced in Section 1.2 efficiently, we propose a novel **P**ropensity score **A**ugmented **D**oubly robust (PAD) estimation method that enhances the standard DR estimator of μ_0 through linear additive or multiplicative augmentation of the PS model. The augmentation functions and their linear coefficients are carefully constructed such that the augmentation term always reduces the variance of the DR estimator if the PS is wrong and the OR is correct while it automatically converges to zero if the PS is correct, in order to avoid bias and ensure double robustness. Also, when both models are correct, our PAD estimator becomes asymptotically equivalent to the standard DR estimator.

To our best knowledge, the proposed estimator is the first one to simultaneously have the DR property and a smaller variance than the standard DR estimator under wrong PS and correct OR models. Thus, our work serves as an important complement to existing DR inference approaches, especially to the intrinsically efficient DR estimators proposed to work for the setting with correct PS and wrong OR (Cao et al., 2009; Tan, 2010, e.g.). In this work, our idea is primarily studied in the covariate shift setup with specific nuisance model estimators, aiming at variance reduction under wrong PS and correct OR. All these can be extended to different or more general scenarios. Thus, we also propose in Supplement S2 the extension of our method to address general nuisance estimators, variance reduction under wrong OR and correct PS, and the efficient estimation of the ATE.

2. Method

2.1 Doubly robust estimator

As a prerequisite of our proposal, we first introduce the standard DR estimator for μ_0 under the setup described in Section 1.2, which has been studied for years (Hahn, 1998, 2004; Shu and Tan, 2018, e.g.). Following a common strategy (Bang and Robins, 2005; Shu and Tan, 2018; Liu et al., 2020, e.g.), we form the PS and OR models as $r(\boldsymbol{x}) = \exp(\boldsymbol{x}^\intercal \boldsymbol{\gamma})$ and $m(\boldsymbol{x}) = g(\boldsymbol{x}^\intercal \boldsymbol{\alpha})$ where $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are model coefficients and $g(\cdot)$ is a known and differentiable link function.

We say that the PS (or OR) model is correct if there exists $\boldsymbol{\gamma}_0$ (or $\boldsymbol{\alpha}_0$) such that the true $r_0(\boldsymbol{x}) = \exp(\boldsymbol{x}^\intercal\boldsymbol{\gamma}_0)$ (or $m_0(\boldsymbol{x}) = g(\boldsymbol{x}^\intercal\boldsymbol{\alpha}_0)$). Denote the empirical mean operator on $\mathcal{S}$ and $\mathcal{T}$ as $\widehat{\mathbb{E}}_{\mathcal{S}}$ and $\widehat{\mathbb{E}}_{\mathcal{T}}$ such that

$$\widehat{\mathbb{E}}_{\mathcal{S}}a(\boldsymbol{X}, Y) = n^{-1}\sum_{i=1}^{n} a(\boldsymbol{X}_i, Y_i), \quad \widehat{\mathbb{E}}_{\mathcal{T}}a(\boldsymbol{X}, Y) = N^{-1}\sum_{i=n+1}^{n+N} a(\boldsymbol{X}_i, Y_i)$$

for any function $a(\cdot)$. Suppose the two nuisance estimators $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$ are obtained respectively by solving the estimating equations:

$$\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\exp(\boldsymbol{X}^\intercal\boldsymbol{\gamma}) = \widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}, \quad \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y - g(\boldsymbol{X}^\intercal\boldsymbol{\alpha})\} = \mathbf{0}. \tag{2.1}$$

The estimating equations for $\boldsymbol{\gamma}$ in (2.1) is usually referred as covariate balancing (Imai and Ratkovic, 2014; Zhao and Percival, 2017), and those for $\boldsymbol{\alpha}$ correspond to the ordinary least square regression when $g(a) = a$ and the logistic regression when Y is binary and $g(a) = \operatorname{expit}(a) = e^a/(1 + e^a)$. Note that one can use alternative estimation procedures to obtain $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, e.g., running a logistic regression on Δ against $\boldsymbol{X}$ to estimate $\boldsymbol{\gamma}$, and our proposed method could naturally be adapted to different choices; see Remark 2.

Based on $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$, the PS and OR estimators introduced in Section 1.2 can be specified as $\widehat{\mu}_{\mathsf{PS}} = \widehat{\mathbb{E}}_{\mathcal{S}}Y\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}})$ and $\widehat{\mu}_{\mathsf{OR}} = \widehat{\mathbb{E}}_{\mathcal{T}}g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})$ respectively. Then the standard DR estimator is constructed by augmenting one of them with another nuisance model:

$$\widehat{\mu}_{\mathsf{DR}} = \widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}})\}\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}}) + \widehat{\mathbb{E}}_{\mathcal{T}}g(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\alpha}}). \tag{2.2}$$

When the PS model is correct and $\widehat{\boldsymbol{\gamma}}$ converges to $\boldsymbol{\gamma}_0$, $\widehat{\mathbb{E}}_{\mathcal{T}} g(\boldsymbol{X}^\top \widehat{\boldsymbol{\alpha}}) - \widehat{\mathbb{E}}_{\mathcal{S}} g(\boldsymbol{X}^\top \widehat{\boldsymbol{\alpha}}) \exp(\boldsymbol{X}^\top \widehat{\boldsymbol{\gamma}})$ converges to zero and the remainder term $\widehat{\mathbb{E}}_{\mathcal{S}} Y \exp(\boldsymbol{X}^\top \widehat{\boldsymbol{\gamma}})$ is exactly the PS estimator converging to μ_0. Similarly, when OR is correct, we can show that $\widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\top \widehat{\boldsymbol{\alpha}})\} \exp(\boldsymbol{X}^\top \widehat{\boldsymbol{\gamma}})$ converges to zero and $\widehat{\mathbb{E}}_{\mathcal{T}} g(\boldsymbol{X}^\top \widehat{\boldsymbol{\alpha}})$ converges to μ_0. Thus $\widehat{\mu}_{\text{DR}}$ is doubly robust in the sense that it is consistent when either the PS or OR model is correctly and consistently estimated.

2.2 Expansion of DR estimator under correct OR model

To help the readers understand our method more intuitively, we now heuristically derive and analyze the asymptotic expansion of $\widehat{\mu}_{\text{DR}}$ when the OR model is correctly specified. Suppose that $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$ converge to some $\bar{\boldsymbol{\gamma}}$ and $\bar{\boldsymbol{\alpha}}$ defined as the solutions to the population-level estimating equations $\mathbb{E}_{\mathcal{S}} \boldsymbol{X} \exp(\boldsymbol{X}^\top \boldsymbol{\gamma}) = \mathbb{E}_{\mathcal{T}} \boldsymbol{X}$ and $\mathbb{E}_{\mathcal{S}} \boldsymbol{X}\{Y - g(\boldsymbol{X}^\top \boldsymbol{\alpha})\} = \mathbf{0}$, respectively. Let $\widehat{r}(\boldsymbol{x}) = \exp(\boldsymbol{X}^\top \widehat{\boldsymbol{\gamma}})$, $\bar{r}(\boldsymbol{x}) = \exp(\boldsymbol{X}^\top \bar{\boldsymbol{\gamma}})$, and $\boldsymbol{S}(\boldsymbol{\alpha}) = \boldsymbol{S}(Y, \boldsymbol{X}, \boldsymbol{\alpha}) = \boldsymbol{X}\{Y - g(\boldsymbol{X}^\top \boldsymbol{\alpha})\}$. Suppose that the OR model is correct, i.e., $m_0(\boldsymbol{x}) = g(\boldsymbol{X}^\top \boldsymbol{\alpha}_0)$ and $\boldsymbol{\alpha}_0 = \bar{\boldsymbol{\alpha}}$, and $n^{1/2}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$ is asymptotically normal with mean zero following the standard M-estimation theory (Van der Vaart, 2000). Then we have

$$\widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^\top \widehat{\boldsymbol{\alpha}})\}\{\widehat{r}(\boldsymbol{X}) - \bar{r}(\boldsymbol{X})\} = o_p(n^{-1/2})$$

due to Neyman orthogonality (Neyman, 1959), which, as will be strictly proved in Section 3, implies that $\widehat{\mu}_{\mathsf{DR}}$ defined in (2.2) is asymptotically equivalent with

$$\begin{aligned}\widetilde{\mu}_{\mathsf{DR}} =&\widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\bar{r}(\boldsymbol{X}) + \widehat{\mathbb{E}}_{\mathcal{T}}g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}}) \\ &+ \left[\widehat{\mathbb{E}}_{\mathcal{S}}\{g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}}) - g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\}\bar{r}(\boldsymbol{X}) + \widehat{\mathbb{E}}_{\mathcal{T}}\{g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}}) - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\right] \\ \approx&\widehat{\mathbb{E}}_{\mathcal{S}}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\}\bar{r}(\boldsymbol{X}) + \widehat{\mathbb{E}}_{\mathcal{T}}g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}}) + \mathbf{L}^{\intercal}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{Y - g(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\},\end{aligned}$$

where $\mathbf{L} = -\bar{\mathbf{H}}^{-1}\left\{\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\bar{r}(\boldsymbol{X}) - \mathbb{E}_{\mathcal{T}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})\right\}$, $\bar{\mathbf{H}} = \mathbb{E}_{\mathcal{S}}\boldsymbol{X}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\bar{\boldsymbol{\alpha}})$, and $\dot{g}(a)$ is the derivative of $g(a)$. To derive the above result, we use the standard asymptotic expansion of $\widehat{\boldsymbol{\alpha}}$ given by our technical lemma in Supplement S3.2, and the symbol "$\approx$" indicates that the difference between the two lines is up to $o_p(n^{-1/2})$ and, thus, asymptotically negligible. So when OR is correct, the asymptotic variance of $n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)$ is equal to that of $n^{1/2}(\widetilde{\mu}_{\mathsf{DR}} - \mu_0)$, which can be expressed as

$$\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)\} = \mathbb{E}_{\mathcal{S}}\{\bar{r}(\boldsymbol{X})\}^2 v(\boldsymbol{X}) + 2\mathbf{L}^{\intercal}\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\bar{r}(\boldsymbol{X})v(\boldsymbol{X}) + C, \quad (2.3)$$

where $v(\boldsymbol{x}) = \mathrm{Var}(Y \mid \boldsymbol{X})$ and C is some positive constant free of $\bar{r}(\cdot)$ and, thus, needs not to be considered in the following derivation. Note that when the PS model also is correct, i.e., $\bar{r}(\cdot) = r_0(\cdot)$, we further have $\mathbf{L} = \mathbf{0}$.

Empirically, term $\mathbf{L}$ in (2.3) can be estimated by

$$\widehat{\mathbf{L}} = -\widehat{\mathbf{H}}^{-1}\left\{\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}}) - \widehat{\mathbb{E}}_{\mathcal{T}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\right\}, \quad (2.4)$$

where $\widehat{\mathbf{H}} = \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\boldsymbol{X}^{\intercal}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})$. Estimation of $v(\boldsymbol{x})$ relies on our *working* assumption on the form of $\mathrm{Var}(Y \mid \boldsymbol{X})$. For example, one may assume $Y = m_0(\boldsymbol{X})+\epsilon$ where $\epsilon \sim \mathrm{N}(0,\sigma^2)$ so $v(\boldsymbol{x})$ is invariant of $\boldsymbol{x}$ and can be simply imputed with the moment estimator of σ^2. Also, for the common Poisson model $Y \sim \mathrm{Poisson}\{\exp(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}$ and logistic model $Y \sim \mathrm{Bernoulli}\{\mathrm{expit}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}$, one can naturally estimate $v(\boldsymbol{x})$ by $\exp(\boldsymbol{x}^{\intercal}\widehat{\boldsymbol{\alpha}})$ and $\mathrm{expit}(\boldsymbol{x}^{\intercal}\widehat{\boldsymbol{\alpha}})\{1-\mathrm{expit}(\boldsymbol{x}^{\intercal}\widehat{\boldsymbol{\alpha}})\}$ respectively. To preserve generality, we introduce a *working* model $v_{\boldsymbol{\theta}}(\boldsymbol{x})$ for $v(\boldsymbol{x})$ with some nuisance parameter $\boldsymbol{\theta}$ to be estimated as $\widehat{\boldsymbol{\theta}}$ that could be partially or fully determined by $\widehat{\boldsymbol{\alpha}}$. Suppose that $\widehat{\boldsymbol{\theta}}$ converges to some $\bar{\boldsymbol{\theta}}$. As will be shown in Section 3, violation of this conditional variance model, i.e., $v(\boldsymbol{x}) \neq v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})$ does not impact the double robustness of our proposed estimator but only affects its efficiency gain when PS is wrong and OR is correct.

2.3 PAD estimator

Now we formally introduce the propensity score augmented doubly robust (PAD) estimator. Our central idea is to augment the PS model $\bar{r}(\boldsymbol{X}) = \exp(\boldsymbol{X}^{\intercal}\bar{\gamma})$ as $\bar{r}_{\mathsf{aug}}(\boldsymbol{X};\boldsymbol{\beta}) = \exp(\boldsymbol{X}^{\intercal}\bar{\gamma}) + \boldsymbol{\Psi}^{\intercal}\boldsymbol{\beta}$ and use $\bar{r}_{\mathsf{aug}}(\cdot)$ to replace $\bar{r}(\cdot)$ in the DR estimator. Here $\boldsymbol{\Psi}$ is some properly constructed basis function of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ is some loading coefficient vector to be estimated. We first describe the empirical construction procedures for PAD in Algorithm 1 and then discuss the implementation and intuition of the key steps in this algorithm.

Algorithm 1 Propensity score Augmented Doubly robust (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 $\widehat{\boldsymbol{\Psi}} = \boldsymbol{\Phi} - \widehat{\mathbb{E}}_{\mathcal{T}}[\boldsymbol{\Phi} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})]/\widehat{\mathbb{E}}_{\mathcal{T}} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})$.

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

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \widehat{V}_{\mu}(\boldsymbol{\beta}), \quad \text{s.t.} \quad \widehat{\mathbb{E}}_{\mathcal{S}} \boldsymbol{X} \dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}}) \widehat{\boldsymbol{\Psi}}^{\intercal} \boldsymbol{\beta} = \mathbf{0}, \tag{2.5}$$

with the objective function defined as

$$\widehat{V}_{\mu}(\boldsymbol{\beta}) = \widehat{\mathbb{E}}_{\mathcal{S}}\{\exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}}) + \widehat{\boldsymbol{\Psi}}^{\intercal}\boldsymbol{\beta}\}^{2} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2\widehat{\mathbf{L}}^{\intercal}\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\{\exp(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\gamma}}) + \widehat{\boldsymbol{\Psi}}^{\intercal}\boldsymbol{\beta}\} v_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}), \tag{2.6}$$

and $\widehat{\mathbf{L}}$ defined in equation (2.4).

[Step 4] Obtain the PAD estimator through

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

Beside estimating $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$ that are also needed by the standard DR method, our method only requires to solve one more restricted weighted least square (RWLS) problem (2.5). This RWLS can be shown to have a unique and explicit-form solution under Assumption 3 consisting of common and mild regularity conditions on $\boldsymbol{X}$ and $\boldsymbol{\Psi}$. Since $\boldsymbol{\Psi}$ has a larger dimension than $\boldsymbol{X}$ and is non-singular, the feasible set of the linear constraint in (2.5) is non-empty, and $\widehat{V}_\mu(\boldsymbol{\beta})$ is strongly convex, which, combined together, implies that (2.5) has a unique solution; See our technical lemma in Supplement S3.2 for more details. In Supplement, we provide the specific form of this solution. Therefore, numerical implementation of Algorithm 1 is stable, straightforward, and not time-consuming. This is an advantage of our method over existing intrinsic efficient DR methods like Rotnitzky et al. (2012) that require solving non-convex optimization problems prone to the local minima issue.

For heuristic analysis, suppose that all estimators used in (2.6) converge to their limiting values. Then let $\boldsymbol{\Psi} = \boldsymbol{\Phi} - \mathbb{E}_{\mathcal{T}}[\boldsymbol{\Phi} v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})]/\mathbb{E}_{\mathcal{T}} v_{\boldsymbol{\theta}}(\boldsymbol{X})$ be the limits of $\widehat{\boldsymbol{\Psi}}$, $\bar{\boldsymbol{\beta}}$ the limits of $\widehat{\boldsymbol{\beta}}$, with its specific form given by our technical lemma in Supplement S3.2, and

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

the limiting function of $\widehat{V}_\mu(\boldsymbol{\beta})$ specified in Algorithm 1. We shall consider two scenarios separately to demonstrate that our proposed PAD estimator not only maintains double robustness property but also has a lower asymptotic variance than $\widehat{\mu}_{\mathsf{DR}}$ when the OR model is correctly specified and PS is wrong. Rigorous justification for these results will be provided in Section 3.

Correct PS model. When the PS model is correct, we easily have $\mathbf{L} = \mathbf{0}$ as stated in Section 2.2 so $V_\mu(\boldsymbol{\beta}) = \mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}}) + \boldsymbol{\Psi}^\intercal\boldsymbol{\beta}\}^2 v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})$, and

$$\frac{\partial V_\mu(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=0} = 2\mathbb{E}_{\mathcal{S}}\boldsymbol{\Psi}\exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = 2\mathbb{E}_{\mathcal{T}}\boldsymbol{\Psi}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}).$$

By definition of $\boldsymbol{\Psi}$, we have $\mathbb{E}_{\mathcal{T}}\boldsymbol{\Psi}v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X}) = \mathbf{0}$, as ensured by the mean shift of $\boldsymbol{\Phi}$ in Step 2 of Algorithm 1. Thus, $\boldsymbol{\beta} = \mathbf{0}$ minimizes $V_\mu(\boldsymbol{\beta})$ and consequently, is the solution of the population-level version of the RWLS problem (2.5) since the linear constraints in (2.5) is trivially satisfied by $\boldsymbol{\beta} = \mathbf{0}$. Also, as discussed above, this solution is unique by the strong convexity of $V_\mu(\boldsymbol{\beta})$. These imply that as long as the PS model is correct, $\widehat{\boldsymbol{\beta}}$ converges to $\mathbf{0}$ so the augmented PS estimator $\exp(\boldsymbol{X}^\intercal\widehat{\boldsymbol{\gamma}}) + \widehat{\boldsymbol{\Psi}}^\intercal\widehat{\boldsymbol{\beta}}$ converges to the correct PS model, which ensures $\widehat{\mu}_{\mathsf{PAD}}$ to converge to the true μ_0. Meanwhile, it is clear that the augmentation of PS does not change the OR model at all. Therefore, $\widehat{\mu}_{\mathsf{PAD}}$ preserves 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.

Correct OR and wrong PS. Note that $\widehat{\mu}_{\mathsf{PAD}} = \widehat{\mu}_{\mathsf{DR}} + \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}}\{Y - g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\}$ and when the OR model is correct,

$$\begin{aligned}\widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}}\{Y - g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\} =& \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}}\{Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\} + \widehat{\mathbb{E}}_{\mathcal{S}}\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}}\{g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0) - g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\}\\ \approx& \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\{Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\} + \widehat{\mathbb{E}}_{\mathcal{S}}(\boldsymbol{\alpha}_0 - \widehat{\boldsymbol{\alpha}})^{\intercal}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}},\end{aligned} \tag{2.7}$$

in which we use the orthogonality between $\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}} - \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}$ and $Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)$ on the first term, as well as expansion on $g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0) - g(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})$ in the second term of the first line, to derive the "≈" relation shown in the second line. Here, "≈" in (2.7) again means that the difference between the first and second line is up to $o_p(n^{-1/2})$ and, thus, becomes asymptotically negligible.

In addition, due to the moment constraint in (2.5), $\widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{X}\dot{g}(\boldsymbol{X}^{\intercal}\widehat{\boldsymbol{\alpha}})\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\boldsymbol{\beta}}$ converges to $\mathbf{0}$. So the second term in the second line of (2.7) is also negligible and $\widehat{\mu}_{\mathsf{PAD}} \approx \widehat{\mu}_{\mathsf{DR}} + \widehat{\mathbb{E}}_{\mathcal{S}}\boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\{Y - g(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha}_0)\}$. Combining this with equation (2.3) as well as the asymptotic equivalence between $\widehat{\mu}_{\mathsf{DR}}$ and $\widetilde{\mu}_{\mathsf{DR}}$ discussed in Section 2.2, we have

$$\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)\} = \mathbb{E}_{\mathcal{S}}\{\bar{r}(\boldsymbol{X}) + \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}^2 v(\boldsymbol{X}) + 2\mathbf{L}^{\intercal}\mathbb{E}_{\mathcal{S}}\boldsymbol{X}\{\bar{r}(\boldsymbol{X}) + \boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\beta}}\}v(\boldsymbol{X}) + C, \tag{2.8}$$

which, after dropping the invariant C, is equal to $V_{\mu}(\bar{\boldsymbol{\beta}})$, the limiting value of the minimized objective function $\widehat{V}_{\mu}(\widehat{\boldsymbol{\beta}})$ in the RWLS problem (2.5). Note that $\boldsymbol{\beta} = \mathbf{0}$ is always feasible to the linear constraint in (2.5) and if we simply replace

$\bar{\boldsymbol{\beta}}$ with $\mathbf{0}$ in the right-hand side of (2.8), it reduces to the asymptotic variance of $n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)$ derived in (2.3). Meanwhile, when the PS model is wrong, $\partial V_\mu(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is typically not $\mathbf{0}$ at $\boldsymbol{\beta} = \mathbf{0}$ so the population-level minimizer $\bar{\boldsymbol{\beta}} \neq 0$. Thus, $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)\} \leq \mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)\}$ when the OR model is correct and the strict "<" will hold in general when the PS model is wrong.

Remark 1. The construction of Step 3 in Algorithm 1 seems complicated because the nuisance $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$ could influence $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)\}$, which needs to be accounted for. When both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$ are given or not varying with the data, the empirical variance of $\sqrt{n}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)$ can be reduced to

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

which leads to a simplification on the form of $\widehat{V}_\mu(\boldsymbol{\beta})$ by dropping the second term in (2.6). In this case, the linear constraint in (2.5) can also be removed when solving for $\boldsymbol{\beta}$. This simplified form corresponds to the variance of the influence function $\{r(\boldsymbol{X})+\boldsymbol{\Psi}^\intercal\boldsymbol{\beta}\}\{Y-m(\boldsymbol{X})\}$ on $\mathcal{S}$ when the imputation model is correct.

In our construction, the second term in (2.6) is used to account for the influence of $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0$. The constraint in (2.5) is unavoidable because without it, the form of $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)\}$ will not allow one to simultaneously achieve (i) validity when PS is correct; (ii) variance reduction when PS is wrong and OR is correct.

Remark 2. In Step I of Algorithm 1, we specify $\widehat{\gamma}$ and $\widehat{\boldsymbol{\alpha}}$ as the solutions to the estimating equations in (2.1). In general, Step I can accommodate other choices of the estimating equations or procedures for the nuisance model parameters, e.g., the maximum likelihood estimation of $\widehat{\gamma}$ and $\widehat{\boldsymbol{\alpha}}$. For $\widehat{\gamma}$, it has no asymptotic influence on $\mathrm{aVar}\{n^{1/2}(\widehat{\mu}_{\mathsf{PAD}}-\mu_0)\}$ when the OR model is correct, so any change of its estimating equations in Step I will not affect constructions of the following steps in Algorithm 1. In response to an alternative $\widehat{\boldsymbol{\alpha}}$, we only need to modify the form of $\widehat{V}_{\mu}(\boldsymbol{\beta})$ in Step III, more specifically, the second term in (2.6) according to its asymptotic expansion. Under this modification, (2.5) is still a RWLS with a unique solution of an explicit form. Details are presented in Supplement S2.

2.4 The multiplicative PAD estimation

Since the augmentation term $\widehat{\boldsymbol{\Psi}}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}$ in Algorithm 1 is additive to the PS model, we also refer the method introduced in Section 2.3 as additive PAD (aPAD). Though this aPAD method can attain desirable asymptotic and numerical results as shown in next sections, its augmented PS function $\exp(\boldsymbol{X}^{\mathsf{T}}\widehat{\gamma}) + \widehat{\boldsymbol{\Psi}}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}$ is not ensured to be positive, which presents a problem of interpretability. In response to this range issue, we introduce and study a multiplicative PAD (mPAD) estimator in Supplement S4, in which the PS model is taken as $\exp(\boldsymbol{X}^{\mathsf{T}}\widehat{\gamma} + \widehat{\boldsymbol{\Psi}}'^{\mathsf{T}}\widehat{\boldsymbol{\beta}}')$ with a multiplicative augmentation term $\exp(\widehat{\boldsymbol{\Psi}}'^{\mathsf{T}}\widehat{\boldsymbol{\beta}}')$ specified and derived following a similar idea as the aPAD method.

Again similar to aPAD, we discuss and numerically investigate the performance of mPAD in two scenarios: (i) correct PS model; and (ii) correct OR and wrong PS, to demonstrate its properties. We demonstrate that the mPAD estimator could attain the doubly robustness property as well as variance-reduction compared to the standard DR under wrong PS. However, we note that the mPAD's version of the optimization problem (2.5) is no longer an RWLS and even non-convex. This presents an additional challenge in terms of optimization and obtaining the global solution for the coefficients $\widehat{\boldsymbol{\beta}}'$. In addition, as shown in Supplement S4, mPAD tends to have less stable finite-sample performance than aPAD in some cases, due to the presence of outlying squared exponential terms in its construction. Therefore, if there is no strong willing to specify a strictly positive PS model, we still recommend aPAD as the default choice over mPAD when implementing our method.

3. Asymptotic analysis

In this section, we rigorously present the asymptotic properties of the proposed PAD estimator and compare PAD with the standard DR estimator. We first introduce some mild and common regularity assumptions. Without loss of generality, we assume that $n/N = O(1)$ so the desirable parametric rate of the

DR estimators will be $O(n^{-1/2})$.

Assumption 1. The supports of $\boldsymbol{X}$ and $\boldsymbol{\Phi}$ are compact and $\mathbb{E}Y^4 < \infty$.

Assumption 2. The link function $g(\cdot)$ is differentiable with derivative $\dot{g}(\cdot)$ and there exists a constant L such that $|\dot{g}(x_1) - \dot{g}(x_2)| < L|x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$.

Assumption 3. The dimension of $\boldsymbol{\Psi}$ is larger than that of $\boldsymbol{X}$. Matrices $\mathbb{E}_{\mathcal{S}}\{\boldsymbol{\Psi}\boldsymbol{\Psi}^\intercal v_{\bar{\boldsymbol{\theta}}}(\boldsymbol{X})\}$, $\mathbb{E}_{\mathcal{S}}\{\boldsymbol{X}\boldsymbol{X}^\intercal \exp(\boldsymbol{X}^\intercal\bar{\boldsymbol{\gamma}})\}$, $\mathbb{E}_{\mathcal{S}}\{\boldsymbol{X}\boldsymbol{X}^\intercal\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}$ and $\mathbb{E}_{\mathcal{S}}\{\boldsymbol{\Psi}\boldsymbol{X}^\intercal\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}$ have all their eigenvalues bounded and staying away from zero.

Assumption 4. The conditional variance function $v_{\boldsymbol{\theta}}(\boldsymbol{x})$ is differentiable on $\boldsymbol{\theta}$ with a bounded partial derivative $\partial_{\boldsymbol{\theta}} v_{\boldsymbol{\theta}}(\boldsymbol{x})$. The estimator $\widehat{\boldsymbol{\theta}}$ converges to some $\bar{\boldsymbol{\theta}}$ in probability and satisfies that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})$ is asymptotic normal with mean zero.

Remark 3. Assumptions 1–3 are all mild, standard, and commonly used to justify the asymptotic properties of M-estimation (Van der Vaart, 2000). Note that in Assumption 3, we take $\boldsymbol{\Psi}$ to have larger dimension than $\boldsymbol{X}$ and make regularity conditions on $\mathbb{E}_{\mathcal{S}}\{\boldsymbol{X}\boldsymbol{X}^\intercal\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}$ and $\mathbb{E}_{\mathcal{S}}\{\boldsymbol{\Psi}\boldsymbol{X}^\intercal\dot{g}(\boldsymbol{X}^\intercal\bar{\boldsymbol{\alpha}})\}$. These are to ensure that $\widehat{\boldsymbol{\beta}}$ is not zero and properly converges to $\bar{\boldsymbol{\beta}}$. Assumption 4 constrains the way of specifying $v_{\boldsymbol{\theta}}(\boldsymbol{x})$ and estimating $\boldsymbol{\theta}$. Under Assumptions 1–3, this assumption is satisfied when either $\boldsymbol{\theta}$ is fully determined by $\boldsymbol{\alpha}$, e.g., in a Poisson or logistic model for Y against $\boldsymbol{X}$, or when $\boldsymbol{\theta}$ is estimated by additionally fitting some parametric model of $\mathrm{Var}(Y \mid \boldsymbol{X})$ against $\boldsymbol{X}$.

Now we present the main results about the robustness and efficiency of our proposed PAD estimator in Theorem 1 with its proof given in Supplement S3. Some important heuristics of this theorem has already been discussed in Section 2.3.

Theorem 1. *Under Assumptions 1–4, it holds that*

(i) **Double robustness**. *When either the PS or the OR model is correctly specified, i.e.,* $r_0(\boldsymbol{x}) = \exp(\boldsymbol{x}^\intercal \boldsymbol{\gamma}_0)$ *for some* $\boldsymbol{\gamma}_0$ *or* $m_0(\boldsymbol{x}) = g(\boldsymbol{x}^\intercal \boldsymbol{\alpha}_0)$ *for some* $\boldsymbol{\alpha}_0$, $\widehat{\mu}_{\mathsf{PAD}} \xrightarrow{p} \mu_0$ *and* $n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)$ *weakly converges to some normal distribution with mean zero.*

(ii) **Variance reduction under wrong PS**. *When the OR model is correct while the PS model may be misspecified, the asymptotic variance of* $n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)$ *is always not larger than that of* $n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)$*. Further when* $\bar{\boldsymbol{\beta}} \neq 0$ *(the explicit form of* $\bar{\boldsymbol{\beta}}$ *is given by our technical lemmas in Supplement S3.2,* $n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)$ *has a strictly smaller asymptotic variance than* $n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)$.

(iii) **Equivalence under correct PS and OR**. *When both the PS and OR models are correct,* $n^{1/2}(\widehat{\mu}_{\mathsf{PAD}} - \mu_0)$ *and* $n^{1/2}(\widehat{\mu}_{\mathsf{DR}} - \mu_0)$ *are asymptotically*

equivalent and have the same asymptotic variance.

Remark 4. In Theorem 1, we considered fixed dimensional and regular $\boldsymbol{\Psi}$ resulting in $n^{-1/2}$-consistent $\widehat{\boldsymbol{\beta}}$; see our technical lemmas in Supplement S3.2. Adding more features in $\boldsymbol{\Psi}$ could potentially lead to more variance-reduction. It is tentative to extend our method and theory for the augmentation high dimensional $\boldsymbol{\Psi}$, e.g., sieve methods (Newey, 1997, e.g.) and high-dimensional sparse regression (Tibshirani, 1996). In the latter one, one should impose ℓ_1-penalty on $\boldsymbol{\beta}$ in the RWLS (2.5) and could choose $\boldsymbol{\Psi}$ with higher dimensionality than n. In both cases, $\widehat{\boldsymbol{\beta}}$ will have lower convergence rates than $n^{-1/2}$. Nevertheless, Theorem 1 may still hold for them under proper smoothness or sparsity conditions. For sieve, one could adopt under-smoothing to achieve the $n^{-1/2}$-consistency and normality of $\widehat{\mu}_{\mathsf{PAD}}$. For high-dimensional sparse regression, when the OR model is correct, the excessive impact of $\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$ could be automatically removed leveraging the orthogonality between $\boldsymbol{\Psi}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$ and $Y - g(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_0)$; when PS is correct, since the true $\bar{\boldsymbol{\beta}} = \mathbf{0}$, the lasso shrinkage would not inflate the error of $\widehat{\boldsymbol{\beta}}$ beyond $n^{-1/2}$.

4. Simulation study

We conduct simulation studies to evaluate our proposed estimator in Section 2.3 and compare it with the standard DR estimator. In our studies, we generate covariates $\boldsymbol{X} = (X_1, X_2, X_3)^\intercal$ from $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{ij}) \in \mathbb{R}^{3\times 3}$ and $\sigma_{ij} = 0.3^{|i-j|}$. For generation of the population assignment Δ and outcome Y, we consider six settings, namely:

(G1) **Gassuian Y, Correct PS, Correct OR.** $\Pr(\Delta = 1 \mid \boldsymbol{X})\} = \mathrm{expit}(X_1 - 2X_2 + X_3)$ and $Y = 0.5X_1 + 0.5X_2 + X_3 + \epsilon$ where $\epsilon \mid \boldsymbol{X} \sim \mathrm{N}(0, 1)$.

(G2) **Gassuian Y, Correct PS, Wrong OR.** $\Pr(\Delta = 1 \mid \boldsymbol{X}) = \mathrm{expit}(X_1 - 2X_2 + X_3)$ and $Y = 0.5X_1 + 0.5X_2 + \sin(X_2 + 0.5X_3) + \epsilon$.

(G3) **Gassuian Y, Wrong PS, Correct OR.** $\Pr(\Delta = 1 \mid \boldsymbol{X}) = \mathrm{expit}(4 + X_1 + X_2 + X_3 - 1.5|X_1| - 1.5|X_2| - |X_3|)$ and $Y = 0.5X_1 + 0.5X_2 + X_3 + \epsilon$.

(L1) **Binary Y, Correct PS, Correct OR.** $\Pr(\Delta = 1 \mid \boldsymbol{X}) = \mathrm{expit}(X_1 - 2X_2 + X_3)$ and $\Pr(Y = 1 \mid \boldsymbol{X}) = \mathrm{expit}(0.5X_1 + 0.5X_2 + X_3)$.

(L2) **Binary Y, Correct PS, Wrong OR.** $\Pr(\Delta = 1 \mid \boldsymbol{X}) = \mathrm{expit}(X_1 - 2X_2 + X_3)$ and $\Pr(Y = 1 \mid X)\} = \mathrm{expit}(0.5X_1 + 0.5X_2 + \sin(X_2 + 0.5X_3))$

(L3) **Binary Y, Wrong PS, Correct OR.** $\Pr(\Delta = 1 \mid \boldsymbol{X}) = \mathrm{expit}(4 + X_1 + X_2 + X_3 - 1.5|X_1| - 1.5|X_2| - |X_3|)$ and $\Pr(Y = 1 \mid \boldsymbol{X}) = \mathrm{expit}(0.5X_1 + 0.5X_2 + X_3)$.

In Settings (G1)–(G3), Y is a gaussian variable and we fit linear models for $Y \sim \boldsymbol{X}$ with $v_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1$. While in Settings (L1)–(L3), we fit logistic models for the binary Y against $\boldsymbol{X}$ with $v_{\boldsymbol{\theta}}(\boldsymbol{x}) = \text{expit}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha})\{1 - \text{expit}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha})\}$. We consider different scenarios about the correctness of the PS and OR models to examine the robustness and efficiency of PAD. Bootstrap is used for estimating the asymptotic variance and constructing the confidence interval (CI). For effective variance reduction on PAD when PS is wrong, i.e. under Settings (G3) and (L3), we include in the augmentation covariates $\boldsymbol{\Phi}$ a decent amount of $\boldsymbol{X}$'s basis functions including X_j, $|X_j|$, $\exp(-X_j)$, $\exp(-X_{j_1} - X_{j_2})$, and $\exp(-X_1 - X_2 - X_3)$ for all j and $j_1 \neq j_2 \in \{1, 2, 3\}$. We set $N = n = 500$ or $N = n = 1000$ separately and generate 1000 realizations for each setting.

Table 1 reports the absolute average bias (Bias), standard error (SE), and coverage probability (CP) of the 95% CI of the DR and PAD estimators. When at least one nuisance models are correct, DR and PAD attain very close bias, which is much smaller compared to their SE and, thus, grants their CPs to be close to the nominal level. This indicates that PAD achieves the double robustness property just like the standard DR estimator under finite samples. To compare PAD and DR in terms of their estimation variance and efficiency, we present in Table 2 their relative efficiency (RE) defined as $\text{Var}(\widehat{\mu}_{\mathsf{DR}})/\,\text{Var}(\widehat{\mu}_{\mathsf{PAD}})$. Under Settings (G1), (G2), (L1), and (L2) where the PS model is correct, the two estimators show nearly identical variance, with their REs located between 1 ± 0.04. Under Settings (G3) and (L3) with misspecified PS and correct OR models, our proposed PAD estimator shows 20% to 40% smaller variance than the standard DR estimator. All these results demonstrate that conclusions in Theorem 1 also apply well for finite samples. In specific, PAD performs very closely to the standard DR when the PS model is correct and is potentially better than DR in the presence of wrong PS models.

Table 1: The absolute average bias (Bias), standard error (SE), and coverage probability (CP) of the 95% confidence intervals of the DR and PAD estimators under the settings described in Section 4. All results are produced based on 1000 repetitions.

		$n = N = 500$			$n = N = 1000$		
Setting	Method	Bias	SE	CP	bias	SE	CP
(G1)	DR	0.006	0.145	0.94	0.005	0.106	0.92
	PAD	0.005	0.142	0.93	0.004	0.105	0.92
(G2)	DR	0.007	0.152	0.92	0.008	0.111	0.92
	PAD	0.005	0.149	0.92	0.007	0.112	0.92
(G3)	DR	0.010	**0.162**	0.93	0.001	**0.121**	0.92
	PAD	0.005	**0.136**	0.93	0.001	**0.105**	0.93
(L1)	DR	0.000	0.055	0.92	0.001	0.040	0.92
	PAD	0.001	0.054	0.93	0.001	0.040	0.93
(L2)	DR	0.001	0.054	0.92	0.004	0.040	0.92
	PAD	0.001	0.053	0.92	0.004	0.040	0.92
(L3)	DR	0.005	**0.057**	0.93	0.003	**0.038**	0.92
	PAD	0.005	**0.052**	0.93	0.002	**0.035**	0.93

Specification of the augmentation basis $\mathbf{\Phi}$ can have a substantial influence on the performance of PAD. We investigate and compare the relative efficiency (RE) to the standard DR estimator of the PAD estimators derived under different choices on the basis $\mathbf{\Phi}$. Beside $\mathbf{\Phi}^{(1)} = \mathbf{\Phi}$ used beforehand including X_j, $|X_j|$, $\exp(-X_j)$, $\exp(-X_{j_1} - X_{j_2})$, and $\exp(-X_1 - X_2 - X_3)$ for all j and $j_1 \neq j_2 \in \{1,2,3\}$, we also successively define $\mathbf{\Phi}^{(2)}$ by excluding the term $\exp(-X_1 - X_2 - X_3)$ in $\mathbf{\Phi}^{(1)}$; $\mathbf{\Phi}^{(3)}$ by excluding $\exp(-X_{j_1} - X_{j_2})$ for all pairs (j_1, j_2) in $\mathbf{\Phi}^{(2)}$; and $\mathbf{\Phi}^{(4)}$ by excluding all $\exp(-X_j)$ terms in $\mathbf{\Phi}^{(3)}$. We then generate data in the settings (G1)–(G3) and (L1)–(L3) with $N = n = 500$ and evaluate REs to the standard DR of the PAD estimators with $\mathbf{\Phi}^{(1)}$, ..., $\mathbf{\Phi}^{(4)}$, denoted as $\text{PAD}_{\mathbf{\Phi}^{(1)}}$,...,$\text{PAD}_{\mathbf{\Phi}^{(4)}}$.

The results are presented in Table 2. Under Settings (G1), (G2), (L1), (L2) where the PS model is correct, all the PAD estimators show nearly identical variance to the standard DR, with their REs locating around 1. Under (G3) and (L3) with wrong PS and correct OR, the PAD estimators attain smaller variances than the standard DR estimator. $\text{PAD}_{\mathbf{\Phi}^{(1)}}$ with the highest dimensional basis functions shows 42% smaller variance than DR in (G3) and 20% in (L3). Along with the form-reduction of the basis functions from $\text{PAD}_{\mathbf{\Phi}^{(1)}}$ to $\text{PAD}_{\mathbf{\Phi}^{(4)}}$, the resulted RE to DR decreases and becomes gradually closer to 1. For example, after removing $\exp(-X_1 - X_2 - X_3)$ in $\mathbf{\Phi}^{(1)}$, the RE of

PAD under (G3) decrease from 1.42 to 1.16. These results demonstrate the importance of properly specifying the augmentation basis $\mathbf{\Phi}$.

Table 2: Relative efficiency (RE) between DR and PAD, i.e., $\mathrm{Var}(\widehat{\mu}_{\mathsf{DR}})/\,\mathrm{Var}(\widehat{\mu}_{\mathsf{PAD}})$, under the settings described in Section 4.

Method	(G1)	(G2)	(G3)	(L1)	(L2)	(L3)
$\mathrm{PAD}_{\mathbf{\Phi}^{(1)}}$	1.04	1.04	**1.42**	1.04	1.04	**1.20**
$\mathrm{PAD}_{\mathbf{\Phi}^{(2)}}$	1.03	1.03	**1.16**	1.01	1.02	**1.17**
$\mathrm{PAD}_{\mathbf{\Phi}^{(3)}}$	1.00	1.01	**1.08**	1.01	0.98	**1.09**
$\mathrm{PAD}_{\mathbf{\Phi}^{(4)}}$	0.99	1.00	**1.02**	1.00	0.98	**1.01**

Interestingly, in our simulation, (moderately) adding more terms to $\mathbf{\Phi}$ does not cause bias inflation but can contribute to variance-reduction under wrong PS. This is probably due to the linear constraint in (2.5) that imposes certain linear projections of $\boldsymbol{\beta}$ to be zero. Therefore, it is appealing to extend our approach to accommodate high-dimensional or nonparametric estimation of the augmentation term as discussed in Remark 4. In addition, exponential terms like $\exp(-X_j)$, $\exp(-X_{j_1} - X_{j_2})$, and $\exp(-X_1 - X_2 - X_3)$ seem to be more effective in variance reduction because they are more likely to be correlated with the original PS function that also has an exponential form.

5. Real example

The effects of the 401(k) program have been investigated for a long time (Abadie, 2003; Chernozhukov et al., 2018, e.g.). Different from other plans like Individual Retirement Accounts (IRAs), eligibility for 401(k) is completely decided by employers. Therefore, unobserved personal preferences for savings may make little difference in 401(k) eligibility. However, there may be some other confounders affecting the causal studies of 401(k), such as job choice, income, and age. To address this problem, (Abadie, 2003) and (Chernozhukov et al., 2018) proposed to adjust for certain covariates related to job choice so that 401(k) eligibility can be regarded exogenous.

Whether 401(k) eligibility contributes to the improvement of people's net total financial assets is an important topic studied in existing literature like Abadie (2003) and Chernozhukov et al. (2018). However, whether 401(k) can improve the financial assets of those actually not eligible for 401(k) is still an open and interesting problem. To investigate this problem, we analyze the data from the Survey of Income and Program Participation of 1991. The data set consists of $n + N = 9275$ observations. The outcome of our interests, Y is defined as the indication of having positive net total financial assets. There are 9 adjustment covariates in $\boldsymbol{X}$, including age, income, family size, years of education, benefit pension status, marriage, two-earner household status, individual participation in IRA plan, and home ownership status. The source (treated) samples $\mathcal{S}$ with $\Delta = 1$ are taken as those eligible for 401(k) and the target (untreated) samples $\mathcal{T}$ are those without 401(k) eligibility. We applied PAD and standard DR to estimate μ, the effect of 401(k) eligibility on improving the positive rate of net total financial assets among people without 401(k) eligibility. The PS model is specified as $\exp(\boldsymbol{X}^{\intercal}\boldsymbol{\gamma})$ and the OR model is $\text{expit}(\boldsymbol{X}^{\intercal}\boldsymbol{\alpha})$. In our method, the augmentation covariates vector $\boldsymbol{\Phi}$ consists of $\boldsymbol{X}$, $\exp(-0.3X_j)$, $|X_j|$, and X_j^2 for all X_j's that are not binary. We again use bootstrap to estimate SEs and construct CIs.

In Table 3, we report the point estimation, their estimated standard errors (ESE), and 95% CIs for the treatment effect μ, obtained using the standard DR and our proposed PAD methods (including aPAD and mPAD). All methods indicate that 401(k) eligibility has a significant effect on improving the rate of having positive net total financial assets among people who are actually not eligible for 401(k). The estimated treatment effect is 0.169 (95% CI: $0.142, 0.196$) by the standard DR, 0.150 (95% CI: $0.126, 0.175$) by aPAD and 0.156 (95% CI: $0.132, 0.180$) by mPAD. Moreover, the ESEs of our methods are smaller than that of the standard DR, with their estimated RE, i.e., $\text{Var}(\widehat{\mu}_{\mathsf{DR}})/\,\text{Var}(\widehat{\mu}_{\mathsf{PAD}})$ being around 1.25 and $\text{Var}(\widehat{\mu}_{\mathsf{DR}})/\,\text{Var}(\widehat{\mu}_{\mathsf{mPAD}})$ being around 1.30, which means our proposed can characterize the treatment effect μ more precisely.

Table 3: The point estimation (PE), its estimated standard error (ESE), and 95% confidence interval (CI) for μ, the effect of 401(k) eligibility on improving the positive rate of net total financial assets among people without 401(k) eligibility, derived using the standard DR, aPAD, and mPAD methods.

Method	PE	ESE	CI
DR	0.169	0.0140	$(0.142, 0.196)$
aPAD	0.150	0.0125	$(0.126, 0.175)$
mPAD	0.156	0.0123	$(0.132, 0.180)$

6. Discussion

In analogy to our PS model augmentation strategy, we also propose an OR model augmentation strategy (OAD) that augments the OR model with some bases of $\boldsymbol{X}$ satisfying certain moment conditions like $\boldsymbol{\Psi}$ in Algorithm 1. Description and discussion of this method are presented in Supplement S2. Similar to Theorem 1, we are able to show that this OAD estimator is doubly robust, of a smaller variance than the standard DR estimator when the PS model is correct but the OR model is wrong, and equivalent with DR when both nuisance models are correct. Just like PAD, this OAD method is easy to implement and only requires convex optimization. We notice that some existing methods in intrinsic efficient DR estimation like Rotnitzky et al. (2012) and Gronsbell et al. (2022) rely on non-convex training to construct the OR model when it is not linear. This OAD strategy could mitigate this practical problem and still achieves the purpose of variance reduction in the presence of misspecified OR models.

Furthermore, we introduce in Suppplement S2 a natural and effective ensemble method that convexly combines the PAD and OAD estimators (or potentially other existing estimators), using the optimal allocation weights. The proposed ensemble estimator is doubly robust and more efficient than the standard DR estimator whenever the PS or OR model is wrong and the other one is correct. For ease of demonstration, we focus on covariate shift correction, or equivalently the problem of ATT estimation in this paper. Our proposed PAD estimation can be potentially generalized to address other causal or missing data problems like ATE estimation (Bang and Robins, 2005, e.g.), casual model estimation Rotnitzky et al. (2012), transfer learning of a regression model Liu et al. (2020), etc. In Supplement S2, we present the detailed implementation procedure for the PAD estimation of the ATE, and heuristically justify its validity and effectiveness. At last, as discussed in Remark 4 and Section 4, finding the optimal choice or enlarging the dimensionality of the augmentation basis function $\mathbf{\Phi}$ in our framework is a crucial yet open problem warranting future research.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Azriel, D., Brown, L. D., Sklar, M., Berk, R., Buja, A., and Zhao, L. (2021). Semi-supervised linear regression. *Journal of the American Statistical Association*, pages 1–14.

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

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Chakrabortty, A., Cai, T., et al. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572.

Chen, Y.-H. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):449–460.

Cheng, D., Chakrabortty, A., Ananthakrishnan, A. N., and Cai, T. (2020). Estimating average treatment effects with a double-index propensity score. *Biometrics*, 76(3):767–777.

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.

Dukes, O. and Vansteelandt, S. (2020). Inference on treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*.

Gronsbell, J., Liu, M., Tian, L., and Cai, T. (2022). Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(4):1353–1391.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics*, 86(1):73–76.

Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103(3):683–700.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Kawakita, M. and Kanamori, T. (2013). Semi-supervised learning with density-ratio estimation. *Machine learning*, 91(2):189–209.

Liu, M., Zhang, Y., Liao, K. P., and Cai, T. (2020). Augmented transfer regression learning with semi-non-parametric nuisance models. *arXiv*.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234.

Pan, Y. and Zhao, Y.-Q. (2021). Improved doubly robust estimation in learning optimal individualized treatment rules. *Journal of the American Statistical Association*, 116(533):283–294.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

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*.

Signorovitch, J. E., Wu, E. Q., Yu, A. P., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R., and Mulani, P. M. (2010). Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*, 28(10):935–945.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

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

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Yang, S. and Ding, P. (2019). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*.

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1).