Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Optimal Covariate Balancing Conditions in Propensity Score Estimation

Jianqing Fan, Kosuke Imai, Inbeom Lee, Han Liu, Yang Ning & Xiaolin Yang

To cite this article: Jianqing Fan, Kosuke Imai, Inbeom Lee, Han Liu, Yang Ning & Xiaolin Yang (2021): Optimal Covariate Balancing Conditions in Propensity Score Estimation, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2021.2002159

To link to this article: https://doi.org/10.1080/07350015.2021.2002159

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2021, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/07350015.2021.2002159

Optimal Covariate Balancing Conditions in Propensity Score Estimation

Jianqing Fan[a], Kosuke Imai[b], Inbeom Lee[c], Han Liu[d], Yang Ning[c], and Xiaolin Yang[a]

[a]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ; [b]Department of Government and Department of Statistics, Harvard University, Cambridge, MA; [c]Department of Statistics and Data Science, Cornell University, Ithaca, NY; [d]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL; [e]Amazon, Seattle, WA

ABSTRACT

Inverse probability of treatment weighting (IPTW) is a popular method for estimating the average treatment effect (ATE). However, empirical studies show that the IPTW estimators can be sensitive to the misspecification of the propensity score model. To address this problem, researchers have proposed to estimate propensity score by directly optimizing the balance of pretreatment covariates. While these methods appear to empirically perform well, little is known about how the choice of balancing conditions affects their theoretical properties. To fill this gap, we first characterize the asymptotic bias and efficiency of the IPTW estimator based on the covariate balancing propensity score (CBPS) methodology under local model misspecification. Based on this analysis, we show how to optimally choose the covariate balancing functions and propose an optimal CBPS-based IPTW estimator. This estimator is doubly robust; it is consistent for the ATE if either the propensity score model or the outcome model is correct. In addition, the proposed estimator is locally semiparametric efficient when both models are correctly specified. To further relax the parametric assumptions, we extend our method by using a sieve estimation approach. We show that the resulting estimator is globally efficient under a set of much weaker assumptions and has a smaller asymptotic bias than the existing estimators. Finally, we evaluate the finite sample performance of the proposed estimators via simulation and empirical studies. An open-source software package is available for implementing the proposed methods.

ARTICLE HISTORY
Received June 2020
Accepted November 2021

KEYWORDS
Average treatment effect; Causal inference; Double robustness; Model misspecification; Semiparametric efficiency; Sieve estimation

1. Introduction

Suppose that we have a random sample of n units from a population of interest. For each unit i, we observe $(T_i, Y_i, \boldsymbol{X}_i)$, where $\boldsymbol{X}_i \in \mathbb{R}^d$ is a d-dimensional vector of pretreatment covariates, T_i is a binary treatment variable, and Y_i is an outcome variable. In particular, T_i takes 1 if unit i receives the treatment and is equal to 0 if unit i belongs to the control group. The observed outcome can be written as $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$, where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under the treatment and control conditions, respectively. This notation implicitly requires the stable unit treatment value assumption (Rubin 1990). In addition, throughout this article, we assume the strong ignorability of the treatment assignment (Rosenbaum and Rubin 1983),

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i \quad \text{and} \quad 0 < \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) < 1. \tag{1.1}$$

Next, we assume that the conditional mean functions of potential outcomes exist and denote them by,

$$\begin{aligned} \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) &= K(\boldsymbol{X}_i) \quad \text{and} \\ \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i) &= K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i), \end{aligned} \tag{1.2}$$

for some functions $K(\cdot)$ and $L(\cdot)$, which represent the conditional mean of the potential outcome under the control condition and the conditional average treatment effect, respectively. Under this setting, we are interested in estimating the average treatment effect (ATE),

$$\mu = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(L(\boldsymbol{X}_i)). \tag{1.3}$$

The propensity score is defined as the conditional probability of treatment assignment (Rosenbaum and Rubin 1983),

$$\pi(\boldsymbol{X}_i) = \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i). \tag{1.4}$$

In practice, since $\boldsymbol{X}_i$ can be high dimensional, the propensity score is usually parameterized by a model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$ where $\boldsymbol{\beta}$ is a q-dimensional vector of parameters. A popular choice is the logistic regression model, that is, $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i) = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})/\{1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})\}$. Once the parameter $\boldsymbol{\beta}$ is estimated (for example, by the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$), the Horvitz-Thompson estimator (Horvitz and Thompson 1952), which is based on the inverse probability of treatment weighting (IPTW), can be used to obtain an estimate of the ATE,

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} - \frac{(1 - T_i)Y_i}{1 - \pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}\right). \tag{1.5}$$

However, it has been shown that the IPTW estimator with the known propensity score does not attain the semiparametric

CONTACT Yang Ning yn265@cornell.edu Cornell University, Department of Statistical Science, Ithaca, NY 14853.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

efficiency bound (Hahn 1998). A variety of efficient ATE estimators have been proposed (see, e.g., Robins, Rotnitzky, and Zhao 1994; Bang and Robins 2005; Tan 2006; Qin and Zhang 2007; Robins et al. 2007; Cao, Tsiatis, and Davidian 2009; Tan 2010; van der Laan 2010; Rotnitzky et al. 2012; Han and Wang 2013; Vermeulen and Vansteelandt 2015, among many others). Despite the popularity of these methods, researchers have found that in practice the estimators can be sensitive to the misspecification of the propensity score model and the outcome model (e.g., Kang and Schafer 2007). To overcome this problem, several researchers have recently considered the estimation of the propensity score by optimizing covariate balance rather than maximizing the accuracy of predicting treatment assignment (e.g., Hainmueller 2012; Graham, Pinto, and Egel 2012; Imai and Ratkovic 2014; Chan, Yam, and Zhang 2016; Zubizarreta 2015; Zhao and Percival 2017; Zhao 2019). Recently, Ai, et al. (2021) proposed a weighted estimation framework by maximizing the entropy. In this article, we focus on the Covariate Balancing Propensity Score (CBPS) methodology (Imai and Ratkovic 2014). In spite of its simplicity, several scholars independently found that the CBPS performs well in practice (e.g., Wyss et al. 2014; Frölich, Huber, and Wiesenfarth 2015). The method can also be extended for the analysis of longitudinal data (Imai and Ratkovic 2015), general treatment regimes (Fong, Hazlett, and Imai 2018a) and high-dimensional propensity score (Ning, Peng, and Imai 2018). In this article, we conduct a theoretical investigation of the CBPS. Given the similarity between the CBPS and some other methods, our theoretical analysis may also provide new insights for understanding other covariate balancing methods.

The CBPS method estimates the parameters of the propensity score model, $\boldsymbol{\beta}$, by solving the following m-dimensional estimating equation,

$$\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i) \;=\; 0 \quad \text{where}$$
$$\boldsymbol{g}_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i) = \left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)\mathbf{f}(\boldsymbol{X}_i), \qquad (1.6)$$

for some covariate balancing function $\mathbf{f}(\cdot) : \mathbb{R}^d \to \mathbb{R}^m$ when the number of equations m is equal to the number of parameters q. Imai and Ratkovic (2014) point out that the common practice of fitting a logistic model is equivalent to balancing the score function with $\mathbf{f}(\boldsymbol{X}_i) = \pi'_{\boldsymbol{\beta}}(\boldsymbol{X}_i) = \partial\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)/\partial\boldsymbol{\beta}$. They find that choosing $\mathbf{f}(\boldsymbol{X}_i) = \boldsymbol{X}_i$, which balances the first moment between the treatment and control groups, significantly reduces the bias of the estimated ATE. Some researchers also include higher moments and/or interactions, e.g., $\mathbf{f}(\boldsymbol{X}_i) = (\boldsymbol{X}_i\ \boldsymbol{X}_i^2)$, in their applications. This guarantees that the treatment and control groups have an identical sample mean of $\mathbf{f}(\boldsymbol{X}_i)$ after weighting by the estimated propensity score.

When $m > q$, then $\widehat{\boldsymbol{\beta}}$ can be estimated by optimizing the covariate balance by the generalized method of moments (GMM) method (Hansen 1982):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\Theta}{\operatorname{argmin}}\ \bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})^{\top}\,\widehat{\mathbf{W}}\,\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}), \qquad (1.7)$$

where Θ is the parameter space for $\boldsymbol{\beta}$ in $\mathbb{R}^q$ and $\widehat{\mathbf{W}}$ is an $(m \times m)$ positive definite weighting matrix, which we assume in this article does not depend on $\boldsymbol{\beta}$. Alternatively, the empirical likelihood method can be used (Owen 2001). Once the estimate of $\boldsymbol{\beta}$ is obtained, we can estimate the ATE using the IPTW estimator in (1.5).

The main idea of the CBPS and other related methods is to directly optimize the balance of covariates between the treatment and control groups so that even when the propensity score model is misspecified we still obtain a reasonable balance of the covariates between the treatment and control groups. However, one open question remains in this literature: How shall we choose the covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$? In particular, if the propensity score model is misspecified, this problem becomes even more important.

This article makes two main contributions. First, we conduct a thorough theoretical study of the CBPS-based IPTW estimator with an arbitrary covariate balancing function $\mathbf{f}(\cdot)$. We characterize the asymptotic bias and efficiency of this estimator under locally misspecified propensity score models. Based on these findings, we show how to optimally choose the covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ for the CBPS methodology (Section 2).

However, the optimal choice of $\mathbf{f}(\boldsymbol{X}_i)$ requires some initial estimators for the unknown propensity score model and the outcome models. This limits the application of the CBPS method with the optimal $\mathbf{f}(\boldsymbol{X}_i)$ in practice. Our second contribution is to overcome this problem by developing an optimal CBPS method that does not require an initial estimator. We show that the IPTW estimator based on the optimal CBPS (oCBPS) method retains the double robustness property. The proposed estimator is semiparametrically efficient when both the propensity score and outcome models are correctly specified. More importantly, we show that the rate of convergence of the proposed oCBPS estimator is faster than the augmented inverse probability weighted (AIPW) estimator (Robins, Rotnitzky, and Zhao 1994) under locally misspecified models (Section 3).

To relax the parametric assumptions on the propensity score model and the outcome model, we further extend the proposed oCBPS method to the nonparametric settings, by using a sieve estimation approach (Newey 1997; Chen 2007). In Section 4, we establish the semiparametric efficiency result for the IPTW estimator under the nonparametric setting. Compared to the existing nonparametric propensity score methods (e.g., Hirano, Imbens, and Ridder 2003; Chan, Yam, and Zhang 2016), our theoretical results require weaker smoothness assumptions. For instance, the theories in Hirano, Imbens, and Ridder (2003), Imbens, Newey, and Ridder (2007), and Chan, Yam, and Zhang (2016) require $s/d > 7$, $s/d > 9$ and $s/d > 13$, respectively, where s is the smoothness parameter of the corresponding function class and $d = \dim(\boldsymbol{X}_i)$. In comparison, we only require $s/d > 3/4$, which is significantly weaker than the existing conditions. To prove this result, we exploit the matrix Bernstein's concentration inequalities (Tropp 2015) and a Bernstein-type concentration inequality for U-statistics (Arcones 1995). Moreover, we show that our estimator has smaller asymptotic bias than the usual nonparametric method (e.g., Hirano, Imbens, and Ridder 2003). Therefore, the asymptotic normality result is expected to be more accurate in practice (Section 4). The proof of the theoretical results are deferred to the supplementary material.

An open-source R software package CBPS is available for implementing the proposed estimators (Fong, Ratkovic, and Imai 2018b). In Section 5, we conduct simulation studies to evaluate the performance of the proposed methodology and show that the oCBPS methodology indeed performs better than the standard CBPS methodology in a variety of settings. Finally, we conduct an empirical study using a canonical application in labor economics. We show that the oCBPS method is able to yield estimates closer to the experimental benchmark when compared to the standard CBPS method.

2. CBPS Under Locally Misspecified Propensity Score Models

Our theoretical investigation starts by examining the consequences of model misspecification for the CBPS-based IPTW estimator. While researchers can avoid gross model misspecification through careful model fitting, in practice it is often difficult to nail down the exact specification. The prominent simulation study of Kang and Schafer (2007), for example, is designed to illustrate this phenomenon. We therefore consider the consequences of local misspecification of propensity score model in the general framework of Copas and Eguchi (2005). In particular, we assume that the true propensity score $\pi(\boldsymbol{X}_i)$ is related to the working model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$ through the exponential tilt for some $\boldsymbol{\beta}^*$,

$$\pi(\boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\exp(\xi\, u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)), \tag{2.1}$$

where $u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)$ is a function determining the direction of misspecification and $\xi \in \mathbb{R}$ represents the magnitude of misspecification. We assume $\xi = o(1)$ as $n \to \infty$ so that the true propensity score $\pi(\boldsymbol{X}_i)$ is in a local neighborhood of the working model $\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$.

Intuitively, we can interpret $\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$ as an approximation of the true propensity score $\pi(\boldsymbol{X}_i)$. The main advantage of this exponential tilt approach is that $\pi(\boldsymbol{X})$ is always nonnegative. Although it does not guarantee $\pi(\boldsymbol{X}) \leq 1$, with $\xi = o(1)$ and Assumption B.1 in the supplementary material (i.e., $|u(\boldsymbol{X};\boldsymbol{\beta}^*)| \leq C$ almost surely for some constant $C > 0$), we can show that $\pi(\boldsymbol{X}) \leq 1$ holds with probability tending to 1. Finally, we note that under suitable regularity conditions, Model (2.1) can be approximated by $\pi(\boldsymbol{X}) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}) + \xi\bar{u}(\boldsymbol{X};\boldsymbol{\beta}^*) + O_p(\xi^2)$, for some $\bar{u}(\boldsymbol{X};\boldsymbol{\beta}^*)$. This provides an asymptotically equivalent specification of the locally misspecified model. To keep our presentation focused, in this section we assume Model (2.1) holds.

In the following, we will establish the asymptotic normality of the CBPS-based IPTW estimator in (1.5) under this local model misspecification framework.

To derive the asymptotic bias and variance, let us define some necessary quantities,

$$B = \left\{\mathbb{E}\left[\frac{u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)\{K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\}}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right] + \boldsymbol{H}_y^{*}(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\mathbb{E}\left(\frac{u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)\mathbf{f}(\boldsymbol{X}_i)}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right)\right\}, \tag{2.2}$$

where $K(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i)$ are defined in (1.2), $\mathbf{W}^*$ is the limiting value of $\widehat{\mathbf{W}}$ in (1.7), and

$$\boldsymbol{H}_y^{*} = -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))}\cdot\frac{\partial\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}}\right),$$
$$\boldsymbol{H}_{\mathbf{f}}^{*} = -\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))}\left(\frac{\partial\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}}\right)^{\top}\right).$$

Furthermore, denote $\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i) = \frac{T_iY_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}$,

$$\bar{\boldsymbol{H}}^* = (1, \boldsymbol{H}_y^{*\top}) \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top \\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix}, \tag{2.3}$$

where

$$\begin{aligned}
\Sigma_\mu &= \operatorname{var}\big(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i)\big) \\
&= \mathbb{E}\left(\frac{Y_i(1)^2}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} + \frac{Y_i(0)^2}{1-\pi^*_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - (\mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)))^2\right), \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= (\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\operatorname{var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i))\mathbf{W}^* \\
&\quad \boldsymbol{H}_{\mathbf{f}}^{*}(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}, \\
\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} &= -(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\operatorname{cov}(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i), \\
&\quad \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)),
\end{aligned}$$

in which $\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)$ is defined in (1.6). Under the model in Equation (1.2), we have

$$\begin{aligned}
&\operatorname{var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)) = \mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))}\right), \\
&\operatorname{cov}(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i), \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)) \\
&\quad = \mathbb{E}\left[\frac{\{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)\}\mathbf{f}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))}\right].
\end{aligned}$$

The following theorem establishes the asymptotic normality of the CBPS-based IPTW estimator under the local misspecification of the propensity score model.

Theorem 2.1 (Asymptotic Distribution under Local Misspecification of the Propensity Score Model). If the propensity score model is locally misspecified as in (2.1) with $\xi = n^{-1/2}$ and Assumption B.1 in the supplementary material holds, the estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5), where $\widehat{\boldsymbol{\beta}}$ is obtained by GMM (1.7), has the following asymptotic distribution

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(B,\ \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*), \tag{2.4}$$

where B is the asymptotic bias given in Equation (2.2) and the asymptotic variance $\bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*$ is obtained from (2.3).

The theorem shows that the first order asymptotic bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is given by B under local model misspecification. In particular, this bias term implicitly depends on the covariate balancing function $\mathbf{f}(\cdot)$. Thus, we consider how to choose $\mathbf{f}(\cdot)$ such that the first order bias $|B|$ is minimized. While at the first glance the expression of B appears to be mathematically intractable, the next corollary shows that any $\mathbf{f}(\boldsymbol{X})$ satisfying (2.5) can eliminate the first order bias, $B = 0$.

Corollary 2.1. Suppose that the covariate balancing function $\mathbf{f}(\boldsymbol{X})$ satisfies the following condition: there exits some $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that

$$\boldsymbol{\alpha}^\top \mathbf{f}(\boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i). \tag{2.5}$$

In addition, assume that the dimension of $\mathbf{f}(\boldsymbol{X}_i)$ is equal to the number of parameters, that is, $m = q$. Then, under the conditions in Theorem 2.1, the asymptotic bias of the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is 0, i.e., $B = 0$.

Intuitively, the above result can be viewed as a "local" version of robustness of IPTW with respect to the misspecification of the propensity score model. The form of $\mathbf{f}(\boldsymbol{X}_i)$ in (2.5) implies that when balancing covariates, for any given unit we should give a greater weight to the determinants of the mean potential outcome that is less likely to be realized. For example, if a unit is less likely to be treated, then it is more important to balance the covariates that influence the mean potential outcome under the treatment condition. In the following, we focus on the asymptotic variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in Theorem 2.1. Interestingly, we can show that the same choice of $\mathbf{f}(\boldsymbol{X}_i)$ in (2.5) minimizes the asymptotic variance.

Corollary 2.2. Under the same conditions in Corollary 2.1, the asymptotic variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is minimized by any covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ which satisfies (2.5). In this case, the CBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ attains the semiparametric asymptotic variance bound in Theorem 1 of Hahn (1998), that is,

$$V_{\text{opt}} = \mathbb{E}\left[\frac{\text{var}(Y_i(1) \mid \boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)} + \frac{\text{var}(Y_i(0) \mid \boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)} + \{L(\boldsymbol{X}_i) - \mu\}^2\right]. \tag{2.6}$$

Based on Theorem 2.1, we can define the asymptotic mean squared error (AMSE) of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ as $AMSE = B^2 + \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*$. Corollaries 2.1 and 2.2 together imply that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ with $\mathbf{f}(\boldsymbol{X})$ satisfying (2.5) attains the minimum AMSE over all possible covariate balancing estimators. Thus, we refer to (2.5) as the optimality condition for the covariate balancing function. We note that there may exist many choices of $\mathbf{f}(\boldsymbol{X})$ which satisfy (2.5). For instance, we can choose $f_1(\boldsymbol{X}) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$ and $f_2, ..., f_m$ in an arbitrary way, as long as the estimating equation $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = 0$ is not degenerate. In this case, to implement $f_1(\boldsymbol{X})$, we need to further estimate $\boldsymbol{\beta}^*$ by some initial estimator, for example, the maximum likelihood estimator, and estimate the conditional mean $\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i)$ and $\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$ by some parametric/nonparametric models. While Corollaries 2.1 and 2.2 hold with this choice of $\mathbf{f}(\boldsymbol{X})$, the empirical performance of the resulting estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is often unstable due to the estimation error of the initial estimators. To overcome this problem, we will next construct the optimal CBPS estimator that does not require any initial estimator.

3. The Optimal CBPS Methodology

Recall that the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X})$ is given by (2.5). Plugging $\mathbf{f}(\boldsymbol{X})$ into the estimating function $\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ in (1.6), we obtain that

$$\begin{aligned}\boldsymbol{\alpha}^\top \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right) \\ &\quad \times \left[\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))\right] \\ &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right) K(\boldsymbol{X}_i) \\ &\quad + \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - 1\right) L(\boldsymbol{X}_i). \end{aligned} \tag{3.1}$$

In other words, the optimality condition (2.5) holds if and only if some linear combination of estimating function $\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ satisfies (3.1). Motivated by this observation, we construct the following set of estimating functions,

$$\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = \begin{pmatrix} \bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) \\ \bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) \end{pmatrix}, \tag{3.2}$$

where $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ with

$$\begin{aligned}\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)\boldsymbol{h}_1(\boldsymbol{X}_i), \\ \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - 1\right)\boldsymbol{h}_2(\boldsymbol{X}_i), \end{aligned} \tag{3.3}$$

for some pre-specified functions $\boldsymbol{h}_1(\cdot) : \mathbb{R}^d \to \mathbb{R}^{m_1}$ and $\boldsymbol{h}_2(\cdot) : \mathbb{R}^d \to \mathbb{R}^{m_2}$ with $m_1 + m_2 = m$. It is easy to see that if the functions $K(\cdot)$ and $L(\cdot)$ lie in the linear space spanned by the functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ respectively, then there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that (3.1) holds for $(\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i), \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i))$, further implying that the optimality condition (2.5) is met.

As discussed in Section 2, the choice of the optimal covariate balancing function is not unique. Unlike the one mentioned after Corollary 2.2, the estimating function in (3.2) does not require any initial estimators for $\boldsymbol{\beta}$ or the conditional mean models, and is more convenient for implementation. Given the estimating functions in (3.2), we can estimate $\boldsymbol{\beta}$ by the GMM estimator $\widehat{\boldsymbol{\beta}}$ in (1.7). We call this method as the optimal CBPS method (oCBPS). Similarly, the ATE is estimated by the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5). The implementation of the proposed oCBPS method (e.g., the choice of $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$) will be discussed in later sections.

It is worthwhile to note that $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ has the following interpretation. The first set of functions $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ is the same as the existing covariate balancing moment function in (1.6), which balances the covariates $\boldsymbol{h}_1(\boldsymbol{X}_i)$ between the treatment and control groups. However, unlike the original CBPS method, we introduce another set of functions $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ which matches the weighted covariates $\boldsymbol{h}_2(\boldsymbol{X}_i)$ in the treatment group to the unweighted covariates $\boldsymbol{h}_2(\boldsymbol{X}_i)$ in the control group, because $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = 0$ can be rewritten as

$$\sum_{T_i=1} \frac{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\boldsymbol{h}_2(\boldsymbol{X}_i) = \sum_{T_i=0} \boldsymbol{h}_2(\boldsymbol{X}_i).$$

As seen in the derivation of (3.1), the auxiliary estimating function $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ is required in order to satisfy the optimality condition.

3.1. Theoretical Properties

We now derive the theoretical properties of the IPTW estimator (1.5) based on the proposed oCBPS method. In particular, we will show that the proposed estimator is doubly robust and locally efficient. The following set of assumptions are imposed for the establishment of double robustness.

Assumption 3.1. The following regularity conditions are assumed.

1. There exists a positive definite matrix $\mathbf{W}^*$ such that $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}^*$.
2. For any $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3), the minimizer $\boldsymbol{\beta}^o = \text{argmin}_{\boldsymbol{\beta}\in\Theta}\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}))^\top \mathbf{W}^*\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}))$ is unique.
3. $\boldsymbol{\beta}^o \in \text{int}(\Theta)$, where Θ is a compact set.
4. $\pi_{\boldsymbol{\beta}}(\boldsymbol{X})$ is continuous in $\boldsymbol{\beta}$.
5. There exists a constant $0 < c_0 < 1/2$ such that with probability tending to one, $c_0 \le \pi_{\boldsymbol{\beta}}(\boldsymbol{X}) \le 1 - c_0$, for any $\boldsymbol{\beta} \in \text{int}(\Theta)$.
6. $\mathbb{E}|Y(1)|^2 < \infty$ and $\mathbb{E}|Y(0)|^2 < \infty$.
7. For any $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3) and $\mathbf{W}^*$ in part 1, $\mathbf{G}^* := \mathbb{E}(\partial \boldsymbol{g}(\boldsymbol{\beta}^o)/\partial\boldsymbol{\beta})$ exists where $\boldsymbol{g}(\boldsymbol{\beta}) = (\boldsymbol{g}_{1\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})^\top, \boldsymbol{g}_{2\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})^\top)^\top$ and there is a q-dimensional function $C(\boldsymbol{X})$ and a small constant $r > 0$ such that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial\pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial\beta_k| \le C_k(\boldsymbol{X})$ for $1 \le k \le q$, and $\mathbb{E}(|h_{1j}(\boldsymbol{X})|C_k(\boldsymbol{X})) < \infty$ for $1 \le j \le m_1$, $1 \le k \le q$ and $\mathbb{E}(|h_{2j}(\boldsymbol{X})|C_k(\boldsymbol{X})) < \infty$ for $1 \le j \le m_2$, $1 \le k \le q$, where $\mathbb{B}_r(\boldsymbol{\beta}^o)$ is a ball in $\mathbb{R}^q$ with radius r and center $\boldsymbol{\beta}^o$.

Conditions 1–4 of Assumption 3.1 are the standard conditions for consistency of the GMM estimator (Newey and McFadden 1994). We allow the propensity score model to be misspecified: we use the notation $\boldsymbol{\beta}^o$ in Condition 2 to distinguish it from $\boldsymbol{\beta}^*$ used in the previous section. Condition 5 is the positivity assumption commonly used in the causal inference literature (Robins, Rotnitzky, and Zhao 1994, 1995). Conditions 6 and 7 are technical conditions that enable us to apply the dominated convergence theorem. Note that, $\sup_{\boldsymbol{\beta}\in\mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial\pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial\beta_k| \le C_k(\boldsymbol{X})$ in Condition 7 is a local condition because it only requires the existence of an envelop function $C_k(\boldsymbol{X})$ around a small neighborhood of $\boldsymbol{\beta}^o$.

We now establish the double robustness of the proposed estimator under Assumption 3.1.

Theorem 3.1 (Double Robustness). Under Assumption 3.1, the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\beta}}$ is doubly robust. That is, $\widehat{\mu}_{\widehat{\beta}} \xrightarrow{p} \mu$ if at least one of the following two conditions holds:

1. The propensity score model is correctly specified, that is, $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$;
2. The functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3) and $\mathbf{W}^*$ in Assumption 3.1 satisfy the following condition. There exist some vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^q$ such that $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \mathbf{M}_1\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \mathbf{M}_2\boldsymbol{h}_2(\boldsymbol{X}_i)$, where $\mathbf{M}_1 \in \mathbb{R}^{q\times m_1}$ and $\mathbf{M}_2 \in \mathbb{R}^{q\times m_2}$ are the partitions of $\mathbf{G}^{*\top}\mathbf{W}^* = (\mathbf{M}_1, \mathbf{M}_2)$.

Next, we establish the asymptotic normality of the proposed estimator if either the propensity score model (Condition 1 in Theorem 3.1) or the outcome model is correctly specified (Condition 2 in Theorem 3.1) . For this result, we require an additional set of regularity conditions.

Assumption 3.2. The following regularity conditions are assumed.

1. For any $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3) and $\mathbf{W}^*$ in Assumption 3.1, $\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*$ and $\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i,\boldsymbol{X}_i)\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i,\boldsymbol{X}_i)^\top)$ are nonsingular.
2. The function $C_k(\boldsymbol{X})$ defined in Condition 7 of Assumption 3.1 satisfies $\mathbb{E}(|Y(0)|C_k(\boldsymbol{X})) < \infty$ and $\mathbb{E}(|Y(1)|C_k(\boldsymbol{X})) < \infty$ for $1 \le k \le q$.

Condition 1 of Assumption 3.2 ensures the non-singularity of the asymptotic variance matrix and Condition 2 is a mild technical condition required for the dominated convergence theorem.

Theorem 3.2 (Asymptotic Normality). Suppose that Assumptions 3.1 and 3.2 hold.

1. If Condition 1 of Theorem 3.1 holds, then the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\beta}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\beta}} - \mu) \xrightarrow{d} N\left(0,\ \bar{\mathbf{H}}^{*\top}\boldsymbol{\Sigma}\bar{\mathbf{H}}^*\right), \tag{3.4}$$

where $\bar{\mathbf{H}}^* = (\mathbf{1}, \mathbf{H}^{*\top})^\top$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\boldsymbol{\Omega}\mathbf{W}^*\mathbf{G}^*(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}$ and

$$\mathbf{H}^* = -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))} \cdot \frac{\partial\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}}\right),$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top \\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix}, \text{ with}$$

$$\Sigma_\mu = \mathbb{E}\left(\frac{Y_i^2(1)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} + \frac{Y_i^2(0)}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right) - \mu^2. \tag{3.5}$$

In addition, $\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}$ is given by

$$\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\left\{\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_i^o)L(\boldsymbol{X}_i)}{(1-\pi_i^o)\pi_i^o}\boldsymbol{h}_1^\top(\boldsymbol{X}_i)\right), \mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_i^o)L(\boldsymbol{X}_i)}{\pi_i^o}\boldsymbol{h}_2^\top(\boldsymbol{X}_i)\right)\right\}^\top.$$

2. If Condition 2 of Theorem 3.1 holds, then the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\beta}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\beta}} - \mu) \xrightarrow{d} N\left(0,\ \widetilde{\mathbf{H}}^{*\top}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{H}}^*\right), \tag{3.6}$$

where $\widetilde{\mathbf{H}}^* = (1, \check{\mathbf{H}}^{*\top})^\top$,

$$\check{\mathbf{H}}^* = -\mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)}{(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right\}\frac{\partial\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}^o}\right],$$

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \widetilde{\Sigma}_\mu & \widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}}^\top \\ \widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix} \text{ with}$$

$$\widetilde{\Sigma}_\mu = \mathbb{E}\left(\frac{\pi(\boldsymbol{X}_i)Y_i^2(1)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1-\pi(\boldsymbol{X}_i))Y_i^2(0)}{(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right) - \mu^2.$$

In addition, $\widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}}$ is given by

$$\widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\boldsymbol{S},$$

where $\boldsymbol{S} = (\boldsymbol{S}_1^\top, \boldsymbol{S}_2^\top)^\top$ and

$$\boldsymbol{S}_1 = \mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i)-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)\mu)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1-\pi(\boldsymbol{X}_i))(K(\boldsymbol{X}_i)+(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))\mu)}{(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right\}\boldsymbol{h}_1(\boldsymbol{X}_i)\right],$$

$$\boldsymbol{S}_2 = \mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)[(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i))(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)\mu]}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)+(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))\mu}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right\}\boldsymbol{h}_2(\boldsymbol{X}_i)\right].$$

3. If both Conditions 1 and 2 of Theorem 3.1 hold and $\mathbf{W}^* = \boldsymbol{\Omega}^{-1}$, then the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, V),$$

where

$$V = \Sigma_\mu - (\boldsymbol{\alpha}_1^\top\mathbf{M}_1, \boldsymbol{\alpha}_2^\top\mathbf{M}_2)\mathbf{G}^*(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\begin{pmatrix}\mathbf{M}_1^\top\boldsymbol{\alpha}_1\\ \mathbf{M}_2^\top\boldsymbol{\alpha}_2\end{pmatrix} \tag{3.7}$$

and Σ_μ is defined in (3.5).

The asymptotic variance V in (3.7) contains two terms. The first term Σ_μ represents the variance of each summand in the estimator defined in equation (1.5) with $\widehat{\boldsymbol{\beta}}$ replaced by $\boldsymbol{\beta}^o$. The second term can be interpreted as the effect of estimating $\boldsymbol{\beta}$ via covariate balance conditions. Since this second term is nonnegative, the proposed estimator is more efficient than the standard IPTW estimator with the true propensity score model, that is, $V \leq \Sigma_\mu$. In particular, Henmi and Eguchi (2004) offered a theoretical analysis of such efficiency gain due to the estimation of nuisance parameters under a general estimating equation framework.

Since the choice of $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ can be arbitrary, it might be tempting to incorporate more covariate balancing conditions into $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. However, the following corollary shows that under Conditions 1 and 2 of Theorem 3.1 one cannot improve the efficiency of the proposed estimator by increasing the number of functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ or equivalently, the dimensionality of covariate balance conditions, that is, $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$.

Corollary 3.1. Define $\bar{\boldsymbol{h}}_1(\boldsymbol{X}) = (\boldsymbol{h}_1^\top(\boldsymbol{X}), \boldsymbol{a}_1^\top(\boldsymbol{X}))^\top$ and $\bar{\boldsymbol{h}}_2(\boldsymbol{X}) = (\boldsymbol{h}_2^\top(\boldsymbol{X}), \boldsymbol{a}_2^\top(\boldsymbol{X}))^\top$, where $\boldsymbol{a}_1(\cdot)$ and $\boldsymbol{a}_2(\cdot)$ are some additional covariate balancing functions. Similarly, let $\bar{\boldsymbol{g}}_1(\boldsymbol{X})$ and $\bar{\boldsymbol{g}}_2(\boldsymbol{X})$ denote the corresponding estimating equations defined by $\bar{\boldsymbol{h}}_1(\boldsymbol{X})$ and $\bar{\boldsymbol{h}}_2(\boldsymbol{X})$. The resulting oCBPS-based IPTW estimator is denoted by $\bar{\mu}_{\widehat{\boldsymbol{\beta}}}$ where $\widehat{\boldsymbol{\beta}}$ is in (1.7) and its asymptotic variance is denoted by $\bar{V}$. Under Conditions 1 and 2 of Theorem 3.1, we have $V \leq \bar{V}$, where V is defined in (3.7).

The above corollary shows a potential tradeoff between robustness and efficiency when choosing $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Recall that Condition 2 of Theorem 3.1 implies $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top\mathbf{M}_1\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top\mathbf{M}_2\boldsymbol{h}_2(\boldsymbol{X}_i)$. Therefore, we can make the proposed estimator more robust by incorporating more basis functions into $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$, such that this condition is more likely to hold. However, Corollary 3.1 shows that doing so may inflate the variance of the proposed estimator.

In the following, we focus on the efficiency of the estimator. Using the notations in this section, we can rewrite the semiparametric asymptotic variance bound V_{opt} in (2.6) as

$$V_{\text{opt}} = \Sigma_\mu - (\boldsymbol{\alpha}_1^\top\mathbf{M}_1, \boldsymbol{\alpha}_2^\top\mathbf{M}_2)\boldsymbol{\Omega}\begin{pmatrix}\mathbf{M}_1^\top\boldsymbol{\alpha}_1\\ \mathbf{M}_2^\top\boldsymbol{\alpha}_2\end{pmatrix}. \tag{3.8}$$

Comparing this expression with (3.7), we see that the proposed estimator is semiparametrically efficient if $\mathbf{G}^*$ is a square matrix (i.e., $m = q$) and invertible. This important result is summarized as the following corollary.

Corollary 3.2. Assume $m = q$ and $\mathbf{G}^*$ is invertible. Under Assumption 3.1, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5) is doubly robust in the sense that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \mu$ if either of the following conditions holds:

1. The propensity score model is correctly specified. That is $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$.
2. There exist some vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^q$ such that $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top\boldsymbol{h}_2(\boldsymbol{X}_i)$.

In addition, under Assumption 3.2, if both conditions hold, then the proposed estimator has the asymptotic variance given in (3.8). Thus, our estimator is a locally semiparametric efficient estimator in the sense of Robins, Rotnitzky, and Zhao (1994).

Unlike Theorem 3.1, Case 2 of the above corollary does not involve the matrices M_1 and M_2. As a result, $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ can be interpreted as the basis of $K(\cdot)$ and $L(\cdot)$, respectively.

The corollary shows that the proposed oCBPS method has two advantages over the original CBPS method (Imai and Ratkovic 2014) with balancing first and second moments of $\boldsymbol{X}_i$ and/or the score function of the propensity score model. First, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is robust to model misspecification, whereas the original CBPS estimator does not have that property. Second, the proposed oCBPS estimator can be more efficient than the original CBPS estimator.

Recall that as shown by Hahn (1998), the semiparametric variance bound V_{opt} is derived in a semiparametric setting without imposing specific parametric models for the propensity score or outcome variables. Corollary 3.2 shows that our estimator attains this bound if both conditions hold. Since $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ can be interpreted as the basis of $K(\cdot)$ and $L(\cdot)$, we can improve the robustness of the estimator without sacrificing the efficiency by increasing the number of functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Meanwhile, this also makes the propensity score model more flexible, since we need to increase the number of parameters $\boldsymbol{\beta}$ to match $m = q$ as required in Corollary 3.2. This observation further motivates us to consider a sieve estimation approach to improve the oCBPS method, as shown in Section 4.

Remark 3.1 (Implementation of the oCBPS method). Based on Corollary 3.2, $\boldsymbol{h}_1(\cdot)$ serves as the basis functions for the baseline conditional mean function $K(\cdot)$, while $\boldsymbol{h}_2(\cdot)$ represents the basis functions for the conditional average treatment effect function $L(\cdot)$. Thus, in practice, researchers can choose a set of basis functions for the baseline conditional mean function and the

conditional average treatment effect function when determining the specification for $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Once these functions are selected, they can over-parameterize the propensity score model by including some higher order terms or interactions such that $m = q$ holds. The resulting oCBPS-based IPTW estimator may reduce bias under model misspecification and attain high efficiency.

Remark 3.2. We also extend the oCBPS method to the estimation of the average treatment effect for the treated (ATT). Given the space limitation, we defer the details to the supplementary material.

3.2. Comparison With Related Estimators

Next, we compare the proposed estimator with some related estimators from the literature. We begin with the following standard AIPW estimator of Robins, Rotnitzky, and Zhao (1994),

$$\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - (T_i - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)) \left(\frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha}) + L(\boldsymbol{X}_i,\boldsymbol{\gamma})}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} + \frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha})}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)\right\},$$

where $K(\boldsymbol{X}_i,\boldsymbol{\alpha})$ and $L(\boldsymbol{X}_i,\boldsymbol{\gamma})$ are the conditional mean models indexed by finite dimensional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Assume the linear outcome models: $K(\boldsymbol{X}_i,\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i,\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T\boldsymbol{h}_2(\boldsymbol{X}_i)$. It is interesting to note that our IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in Corollary 3.2 can be rewritten as the AIPW estimator $\widehat{\mu}^{\text{AIPW}}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ (for any $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$), since we have,

$$\frac{1}{n}\sum_{i=1}^{n}(T_i - \pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i))\left(\frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha}) + L(\boldsymbol{X}_i,\boldsymbol{\gamma})}{\pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} + \frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha})}{1-\pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}\right) = 0,$$

by the definition of the covariate balancing estimating equations in (3.2).

It is well known that the AIPW estimator is consistent provided that either the propensity score model or the outcome model is correctly specified. Since both the AIPW estimator and our estimator are doubly robust and locally efficient, in the following we conduct a theoretical investigation of these two estimators under the scenario that *both* propensity score and outcome models are misspecified. Indeed, this scenario corresponds to the simulation settings used in the influential study of Kang and Schafer (2007).

To make the comparison mathematically tractable, we focus on the case that both of these two models are locally misspecified. Similar to Section 2, we assume that the true treatment assignment satisfies, $\pi(\boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\exp(\xi\, u(\boldsymbol{X}_i;\boldsymbol{\beta}^*))$ in (2.1), while the true regression functions $K(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i)$ in (1.2) satisfy

$$K(\boldsymbol{X}_i) = \boldsymbol{\alpha}^{*\top}\boldsymbol{h}_1(\boldsymbol{X}_i) + \delta r_1(\boldsymbol{X}_i), \quad L(\boldsymbol{X}_i) = \boldsymbol{\gamma}^{*\top}\boldsymbol{h}_2(\boldsymbol{X}_i) + \delta r_2(\boldsymbol{X}_i), \tag{3.9}$$

where $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ can be viewed as the approximate true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the functions $r_1(\boldsymbol{X}_i)$ and $r_2(\boldsymbol{X}_i)$ determine the direction of misspecification, and $\delta \in \mathbb{R}$ represents the magnitude of misspecification.

Assume further that the models are locally misspecified, that is, $\xi, \delta = o(1)$. Under regularity conditions similar to Section 2, we can show that the proposed estimator satisfies,

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi(\boldsymbol{X}_i)}\{Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\} - \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\{Y_i(0) - K(\boldsymbol{X}_i)\} + L(\boldsymbol{X}_i) - \mu\right] + O_p(\xi^2\delta + \delta n^{-1/2} + \xi n^{-1/2}), \tag{3.10}$$

whereas the AIPW estimator satisfies,

$$\widehat{\mu}^{\text{AIPW}}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}} - \mu = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi(\boldsymbol{X}_i)}\{Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\} - \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\{Y_i(0) - K(\boldsymbol{X}_i)\} + L(\boldsymbol{X}_i) - \mu\right] + O_p(\xi\delta + \delta n^{-1/2} + \xi n^{-1/2}), \tag{3.11}$$

where $\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\gamma}}$ are the corresponding maximum likelihood and least-square estimators. The derivation of (3.10) and (3.11) is shown in the supplementary material.

The leading terms in the asymptotic expansions of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu$ and $\widehat{\mu}^{\text{AIPW}}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}} - \mu$ are identical and are known as the efficient influence function for μ. However, the remainder terms in (3.10) and (3.11) may have different order. Consider the following two scenarios. First, if $\xi\delta \gg n^{-1/2}$, then we have $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = O_p(\xi^2\delta + n^{-1/2})$ and $\widehat{\mu}^{\text{AIPW}}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}} - \mu = O_p(\xi\delta)$. Thus, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ converges in probability to the ATE at a faster rate than $\widehat{\mu}^{\text{AIPW}}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}}$. Second, if $\xi\delta = o(n^{-1/2})$, the two estimators have the same limiting distribution, that is, $\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N(0, V_{\text{opt}})$, where $\widehat{\mu}$ can be either $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ or $\widehat{\mu}^{\text{AIPW}}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}}$. However, the rates of convergence of the Gaussian approximation determined by the remainder terms in (3.10) and (3.11) are different. For instance, assume that $\xi = \delta = n^{-(1/4+\epsilon)}$ for some small positive $\epsilon < 1/4$. We observe that the remainder term in (3.10) is of order $O_p(n^{-(3/4+\epsilon)})$ and is smaller in magnitude than the corresponding term in (3.11), which is of order $O_p(n^{-(1/2+2\epsilon)})$. As a result, the proposed estimator converges in distribution to $N(0, V_{\text{opt}})$ at a faster rate than the AIPW estimator. The above analysis justifies the theoretical advantage of the proposed oCBPS estimator over the standard AIPW estimator.

Furthermore, the proposed estimator is related to the class of bias-reduced doubly robust estimators (Vermeulen and Vansteelandt 2015), see also Robins et al. (2007). To see this, we consider the derivative of $\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ with respect to the nuisance parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}$. In particular, under the linear outcome models, it is easily shown that $\partial\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\alpha} = \bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})$ and $\partial\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\gamma} = \bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})$, where $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})$ are our covariate balancing functions in (3.2). This provides an alternative justification for the proposed method: the oCBPS estimator $\widehat{\boldsymbol{\beta}}$, which satisfies $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}) = 0$ and $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}) = 0$, removes the local effect of the estimated nuisance parameters, that is, $\partial\widehat{\mu}^{\text{AIPW}}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\alpha} = 0$ and $\partial\widehat{\mu}^{\text{AIPW}}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\gamma} = 0$. This property would not hold if we replace $\widehat{\boldsymbol{\beta}}$ by the maximum likelihood estimator or other convenient estimators of $\boldsymbol{\beta}$. Vermeulen and

Vansteelandt (2015) defined the class of bias-reduced doubly robust estimator as $\widehat{\mu}^{\text{AIPW}}_{\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\gamma}}}$, where $(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\gamma}})$ are the estimators corresponding to the estimating equations $\partial\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\alpha} = 0, \partial\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\gamma} = 0, \partial\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\beta} = 0$. The first two sets of estimating equations are identical to the covariate balancing estimating equations in (3.2), whereas the last set of estimating equations $\partial\widehat{\mu}^{\text{AIPW}}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\beta} = 0$ (leading to the estimators $\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\gamma}}$) is unnecessary in our setting because $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} = \widehat{\mu}^{\text{AIPW}}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ does not rely on how $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are estimated. As expected, all the theoretical properties of the bias-reduced doubly robust estimator in Section 3 of Vermeulen and Vansteelandt (2015) hold for our estimator.

Recently, a variety of empirical likelihood based estimators are proposed to match the moment of covariates in treatment and control groups (e.g., Tan 2006, 2010; Hainmueller 2012; Graham, Pinto, and Egel 2012; Han and Wang 2013; Chan, Yam, and Zhang 2016; Zubizarreta 2015; Zhao and Percival 2017). Usually, these methods aim to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$ (or $\mathbb{E}(Y_i(1) \mid T_i = 0)$ and $\mathbb{E}(Y_i(0) \mid T_i = 1)$) separately and combine then to estimate the ATE. Our approach directly estimates the propensity score and ATE by jointly solving the potentially over-identified estimating functions (3.2). In addition, our asymptotic results and the discussion rely on the GMM theory for over-identified estimating functions which is different from these methods. Another recent article by Zhao (2019) studied the robustness of a general class of loss function based covariate balancing methods. When the goal is to estimate the ATE, his score function reduces to our first set of estimating functions $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})$ in (3.2). In this case, his estimator is robust to the misspecification of the propensity score model under the constant treatment effect model, that is, $L(\boldsymbol{X}) = \tau^*$ for some constant τ^*. In comparison, our methodology and theoretical results cover a broader case that allows for heterogeneous treatment effects.

4. Nonparametric oCBPS Methodology

In this section, we extend our theoretical results of the oCBPS methodology to nonparametric estimation. As seen in Corollary 3.2, the proposed estimator is efficient if both the propensity score $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i)$ and the conditional mean functions $K(\cdot)$ and $L(\cdot)$ are correctly specified. To avoid model misspecification, we can choose a large number of basis functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$, such that the conditional mean functions $K(\cdot)$ and $L(\cdot)$ satisfy the condition 2 in Corollary 3.2.

However, the parametric assumption for the propensity score model $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$ imposed in Corollary 3.2 may be too restrictive. Once the propensity score model is misspecified, the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is inefficient and could even become inconsistent. To relax the strong parametric assumptions imposed in the previous sections, we propose a flexible nonparametric approach for modeling the propensity score and the conditional mean functions. The main advantage of this nonparametric approach is that, the resulting oCBPS-based IPTW estimator is semiparametrically efficient under a much broader class of propensity score models and the conditional mean models than those of Corollary 3.2.

Specifically, we assume $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = J(\psi^*(\boldsymbol{X}_i))$, where $J(\cdot)$ is a known monotonic link function (e.g., $J(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$), and $\psi^*(\cdot)$ is an unknown smooth function. One practical way to estimate $\psi^*(\cdot)$ is to approximate it by the linear combination of κ basis functions, where κ is allowed to grow with n. This approach is known as the sieve estimation (Andrews 1991; Newey 1997). In detail, let $\boldsymbol{B}(\boldsymbol{x}) = \{b_1(\boldsymbol{x}), ..., b_\kappa(\boldsymbol{x})\}$ denote a collection of κ basis functions, whose mathematical requirement is given in Assumption E.1 in the supplementary material. Intuitively, we would like to approximate $\psi^*(\boldsymbol{x})$ by $\boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})$, for some coefficient $\boldsymbol{\beta}^* \in \mathbb{R}^\kappa$.

To estimate $\boldsymbol{\beta}^*$, similar to the parametric case, we define $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}) = \sum_{i=1}^n \boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{T}_i,\boldsymbol{X}_i)/n$, where $\boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{T}_i,\boldsymbol{X}_i) = (\boldsymbol{g}_{1\boldsymbol{\beta}}^\top(\boldsymbol{T}_i,\boldsymbol{X}_i), \boldsymbol{g}_{2\boldsymbol{\beta}}^\top(\boldsymbol{T}_i,\boldsymbol{X}_i))^\top$ with,

$$\boldsymbol{g}_{1\boldsymbol{\beta}}(\boldsymbol{T}_i,\boldsymbol{X}_i) = \left(\frac{T_i}{J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - \frac{1-T_i}{1-J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))}\right)\boldsymbol{h}_1(\boldsymbol{X}_i),$$
$$\boldsymbol{g}_{2\boldsymbol{\beta}}(\boldsymbol{T}_i,\boldsymbol{X}_i) = \left(\frac{T_i}{J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - 1\right)\boldsymbol{h}_2(\boldsymbol{X}_i).$$

Recall that $\boldsymbol{h}_1(\boldsymbol{X}) \in \mathbb{R}^{m_1}$ and $\boldsymbol{h}_2(\boldsymbol{X}) \in \mathbb{R}^{m_2}$ are interpreted as the basis functions for $K(\boldsymbol{X})$ and $L(\boldsymbol{X})$. Let $m_1 + m_2 = m$ and $\boldsymbol{h}(\boldsymbol{X}) = (\boldsymbol{h}_1(\boldsymbol{X})^\top, \boldsymbol{h}_2(\boldsymbol{X})^\top)^\top$. Here, we assume $m = \kappa$, so that the number of equations in $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})$ is identical to the dimension of the parameter $\boldsymbol{\beta}$. Then define $\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\Theta} ||\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})||_2^2$, where Θ is the parameter space for $\boldsymbol{\beta}$ and $||\boldsymbol{v}||_2$ represents the L_2 norm of the vector $\boldsymbol{v}$. The resulting IPTW estimator is,

$$\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} = \frac{1}{n}\sum_{i=1}^n\left(\frac{T_iY_i}{J(\widetilde{\boldsymbol{\beta}}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - \frac{(1-T_i)Y_i}{1-J(\widetilde{\boldsymbol{\beta}}^\top\boldsymbol{B}(\boldsymbol{X}_i))}\right).$$

To establish the large sample properties of $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$, we require a few regularity conditions. Due to the space constraint, we defer the regularity conditions to the supplementary material. The following theorem establishes the asymptotic normality and semiparametric efficiency of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$.

Theorem 4.1 (Efficiency under nonparametric models). Assume that Assumption E.1 in the supplementary material holds, and there exist $r_b, r_h > 1/2$, $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^{*\top}, \boldsymbol{\alpha}_2^{*\top})^\top \in \mathbb{R}^\kappa$, such that the propensity score model satisfies

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\psi^*(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})| = O(\kappa^{-r_b}), \tag{4.1}$$

and the outcome models $K(\cdot)$ and $L(\cdot)$ satisfy

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |K(\boldsymbol{x}) - \boldsymbol{\alpha}_1^{*\top}\boldsymbol{h}_1(\boldsymbol{x})| = O(\kappa^{-r_h}),$$
$$\sup_{\boldsymbol{x}\in\mathcal{X}} |L(\boldsymbol{x}) - \boldsymbol{\alpha}_2^{*\top}\boldsymbol{h}_2(\boldsymbol{x})| = O(\kappa^{-r_h}). \tag{4.2}$$

Assume $\kappa = o(n^{1/3})$ and $n^{\frac{1}{2(r_b+r_h)}} = o(\kappa)$. Then

$$n^{1/2}(\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, V_{\text{opt}}),$$

where V_{opt} is the asymptotic variance bound in (2.6). Thus, $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ is semiparametrically efficient.

This theorem can be viewed as a nonparametric version of Corollary 3.2. It shows that one can construct a globally efficient estimator of the treatment effect without imposing strong parametric assumptions on the propensity score model and the outcome model. Since the estimator is asymptotically equivalent to the sample average of the efficient influence function, it is also adaptive in the sense of Bickel et al. (1998).

In the following, we comment on the technical assumptions of Theorem 4.1. We assume $\psi^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) can be uniformly approximated by the basis functions $\boldsymbol{B}(\boldsymbol{x})$ and $\boldsymbol{h}_1(\boldsymbol{x})$ (also $\boldsymbol{h}_2(\boldsymbol{x})$) in (4.1) and (4.2), respectively. It is well known that the uniform rate of convergence is related to the smoothness of the functions $\psi^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) and the dimension of $\boldsymbol{X}$. For instance, if the function class $\mathcal{M}$ for $\psi^*(\boldsymbol{x})$ and $\mathcal{H}$ for $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) correspond to the Hölder class with smoothness parameter s on the domain $\mathcal{X} = [0,1]^d$, under the assumption that $m_1 \asymp m_2 \asymp \kappa$, (4.1) and (4.2) hold for the spline basis and wavelet basis with $r_b = r_h = s/d$; see Newey (1997); Chen (2007) for details. In the same setting, Hirano, Imbens, and Ridder (2003) considered a nonparametric IPTW estimator, which is globally efficient under the condition $s/d > 7$. Imbens, Newey, and Ridder (2007) established the asymptotic equivalence between a regression based estimator and Hirano, Imbens, and Ridder (2003) estimator under $s/d > 9$. Recently, Chan, Yam, and Zhang (2016) proposed a sieve based calibration estimator under the condition $s/d > 13$. Compared to these existing results, our theorem needs a much weaker condition, that is, $s/d > 3/4$. We refer to the supplementary material for further technical discussion of our nonparametric estimator.

5. Simulation and Empirical Studies

5.1. Simulation Studies

In this section, we conduct a set of simulation studies to examine the performance of the proposed methodology. We consider the following linear model for the potential outcomes,

$$Y_i(1) = 200 + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i,$$
$$Y_i(0) = 200 + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i.$$

where $\varepsilon_i \sim N(0,1)$, independent of $\boldsymbol{X}_i$, and consider the following true propensity score model

$$\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(-\beta_1 x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})}{1 + \exp(-\beta_1 x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})}, \quad (5.1)$$

where β_1 varies from 0 to 1. When implementing the proposed methodology, we set $\boldsymbol{h}_1(\boldsymbol{x}_i) = (1, x_{i2}, x_{i3}, x_{i4})$ and $\boldsymbol{h}_2(\boldsymbol{x}_i) = x_{i1}$ so that the number of equations is equal to the number of parameters to be estimated. Covariate X_{i1} is generated independently from $N(3,2)$ and X_{i2}, X_{i3}, and X_{i4} are generated from $N(0,1)$. Each set of results is based on 500 Monte Carlo simulations.

We examine the performance of the IPTW estimator when the propensity score model is fitted using maximum likelihood (GLM), the standard CBPS with balancing the first moment (CBPS), and the proposed optimal CBPS (oCBPS) as well as the case where the true propensity score (True), that is, $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, is used for the IPTW estimator. In addition, we include the IPTW estimator when the propensity score model is estimated by logistic series (Hirano, Imbens, and Ridder 2003). Since a fully nonparametric logistic series approach is impractical to implement due to the curse of dimensionality, instead we consider a generalized additive model (GAM) and apply the logistic series approach to each of the covariate separately. Finally, we also include the targeted maximum likelihood estimator, a doubly robust estimator (DR, Benkeser et al. 2017), using the R package drtmle.

In the first set of simulations, we use the correctly specified propensity score and outcome models. Table 1 shows the standard deviation, bias, root mean square error (RMSE), and the coverage probability of the constructed 95% confidence intervals of these estimators when the sample size is $n = 300$ and $n = 1000$. The confidence intervals are constructed using estimates of the asymptotic variances of the estimators. The exact formulas can be found in the supplementary material. We find that CBPS and oCBPS substantially outperform True, GLM, and GAM in terms of efficiency, and in most cases outperform DR as well. In addition, oCBPS is more efficient than CBPS in all the cases as well. The efficiency improvement is consistent with Corollary 3.2. The coverage probabilities of True, GLM, CBPS and oCBPS are close to the nominal level. However, GAM yields much lower coverage probability, partly because the estimates of the propensity score from logistic series are unstable. The pattern becomes more evident as β_1 increases, corresponding to the setting that the propensity score can be close to 0 or 1. Similarly, the coverage probability of DR also deteriorates as β_1 increases.

We further evaluate our method by considering different cases of misspecification for the outcome and propensity score models. We begin with the case where the outcome model is linear like before but the propensity score is misspecified. While we use the model given in Equation (5.1) when estimating the propensity score, the actual treatment is generated according to the following different model,

$$\mathbb{P}(T_i = 1 \mid \boldsymbol{X} = \boldsymbol{x}_i) = \frac{\exp(-\beta_1 x^*_{i1} + 0.5x^*_{i2} - 0.25x^*_{i3} - 0.1x^*_{i4})}{1 + \exp(-\beta_1 x^*_{i1} + 0.5x^*_{i2} - 0.25x^*_{i3} - 0.1x^*_{i4})},$$

with $x^*_{i1} = \exp(x_{i1}/3)$, $x^*_{i2} = x_{i2}/\{1 + \exp(x_{i1})\} + 10$, $x^*_{i3} = x_{i1}x_{i3}/25 + 0.6$, and $x^*_{i4} = x_{i1} + x_{i4} + 20$ where β_1 again varies from 0 to 1. In other words, the model misspecification is introduced using nonlinear transformations. Table 2 shows the results for this case. As expected from the double robustness property shown in Theorem 3.1, we find that the bias for the oCBPS becomes significantly smaller than all the other estimators. The oCBPS also dominates the other estimators in terms of efficiency and maintains the desired coverage probability.

We also consider the case when the propensity score is locally misspecified with Equation (2.1). In the case, we use (5.1) as the working model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$, set $\xi = n^{-1/2}$ as in Theorem 2.1 and choose the function $u(\boldsymbol{X}_i; \boldsymbol{\beta}) = X^2_{i1}$ as the direction of misspecification. We compute the true propensity score from the model (2.1) and use it to generate the treatment variables. We note that sometimes the true propensity score may exceed

Table 1. The bias, standard deviation, root mean squared error (RMSE), and the coverage probability of the constructed 95% C.I. of the IPTW estimator with known propensity score (True), the IPTW estimator when the propensity score is fitted using the maximum likelihood (GLM), the IPTW estimator when the propensity score is fitted using the generalized additive model (GAM), the targeted maximum likelihood estimator (DR), the standard CBPS estimator balancing the first moment (CBPS), and the proposed optimal CBPS estimator (oCBPS) under the scenario that both the outcome model and the propensity score model are correctly specified.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	−0.43	−0.01	1.15	−5.19	0.00	0.09	−2.43	9.99
	GLM	−0.18	−0.86	0.15	−4.32	−0.04	0.02	0.32	11.15
	GAM	−0.74	−4.60	−15.55	−35.38	−0.19	−1.16	−2.85	−6.86
	DR	0.08	−1.04	−3.41	−8.32	0.18	−0.56	−2.14	−4.50
	CBPS	−0.05	−0.09	0.54	−0.27	0.04	0.04	0.20	0.45
	oCBPS	−0.04	0.03	0.07	0.06	0.04	0.06	0.16	0.08
	True	29.52	39.46	72.56	138.33	15.73	22.36	38.18	88.33
	GLM	4.45	12.31	63.35	144.25	2.21	5.49	22.93	114.45
Std	GAM	4.31	14.91	43.08	100.16	2.06	5.22	21.27	51.96
Dev	DR	2.39	2.57	4.25	8.06	1.20	1.29	1.76	3.32
	CBPS	2.39	2.35	2.66	15.94	1.24	1.26	1.27	1.45
	oCBPS	2.26	2.16	2.27	2.39	1.20	1.20	1.18	1.22
RMSE	True	29.52	39.46	72.57	138.43	15.73	22.36	38.26	88.89
	GLM	4.46	12.34	63.35	144.32	2.21	5.49	22.93	114.99
	GAM	4.37	15.60	45.81	106.23	2.07	5.35	21.46	52.41
	DR	2.39	2.77	5.45	11.58	1.21	1.41	2.77	5.59
	CBPS	2.39	2.35	2.72	15.94	1.24	1.26	1.29	1.52
	oCBPS	2.26	2.16	2.27	2.39	1.20	1.20	1.19	1.23
	True	0.936	0.938	0.922	0.948	0.962	0.942	0.926	0.948
Coverage	GLM	0.946	0.946	0.946	0.946	0.944	0.954	0.954	0.958
Probability	GAM	0.704	0.310	0.090	0.028	0.754	0.382	0.108	0.048
(of the	DR	0.928	0.876	0.576	0.278	0.960	0.906	0.562	0.268
95% C.I.)	CBPS	0.944	0.944	0.944	0.944	0.960	0.958	0.958	0.968
	oCBPS	0.950	0.964	0.962	0.982	0.956	0.954	0.962	0.966

NOTE: We vary the value of β_1 in the data-generating model (5.1).

Table 2. Correct outcome model with a misspecified propensity score model.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	0.00	2.13	0.08	4.79	−1.28	−0.36	1.83	3.62
	GLM	0.41	−6.67	−18.84	−32.15	0.19	−6.33	−19.21	−32.96
	GAM	15.61	3.11	−7.16	−20.76	4.07	0.28	−4.98	−14.11
	DR	−0.29	−0.68	−1.89	−3.60	−0.21	−0.39	−1.23	−2.75
	CBPS	0.84	−0.05	−2.06	−2.44	0.06	−0.79	−2.74	−3.28
	oCBPS	−0.20	−0.02	−0.13	0.07	−0.04	0.03	0.01	−0.05
	True	45.43	36.03	39.77	77.26	26.32	19.36	39.15	88.45
	GLM	11.23	12.66	15.73	26.82	2.17	5.32	8.61	10.92
Std	GAM	19.91	9.40	8.81	16.18	4.29	2.87	4.14	8.52
Dev	DR	3.35	2.57	2.52	3.16	1.42	1.27	1.28	1.57
	CBPS	3.21	2.74	3.18	3.61	1.25	1.41	1.74	2.04
	oCBPS	2.26	2.30	2.28	2.34	1.24	1.26	1.24	1.29
RMSE	True	45.43	36.10	39.77	77.40	26.36	19.36	39.20	88.52
	GLM	11.24	14.31	24.55	41.86	2.18	8.27	21.05	34.72
	GAM	25.30	9.90	11.35	26.32	5.91	2.89	6.47	16.48
	DR	3.37	2.65	3.15	4.79	1.44	1.33	1.78	3.16
	CBPS	3.32	2.74	3.79	4.36	1.26	1.62	3.24	3.86
	oCBPS	2.27	2.30	2.29	2.34	1.24	1.26	1.24	1.29
	True	0.952	0.936	0.964	0.972	0.946	0.950	0.960	0.988
Coverage	GLM	0.964	0.898	0.740	0.834	0.948	0.714	0.300	0.346
probability	GAM	0.236	0.434	0.286	0.066	0.356	0.648	0.178	0.042
(of the	DR	0.882	0.904	0.822	0.596	0.908	0.938	0.788	0.392
95% C.I.)	CBPS	0.956	0.978	0.924	0.914	0.944	0.928	0.742	0.654
	oCBPS	0.946	0.944	0.952	0.944	0.950	0.950	0.954	0.954

1. In this case we simply replace its value with 0.95. The results are given in Table 3. oCBPS dominates all the other estimators in terms of bias, standard deviation and root mean square error, but CBPS and DR are also noticeably better than True, GLM, and GAM.

We next examine the cases where the outcome model is misspecified. We do this by generating potential outcomes from the following quadratic model

$$\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 200 + 27.4x_{i1}^2 + 13.7x_{i2}^2 + 13.7x_{i3}^2 + 13.7x_{i4}^2,$$

$$\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 200 + 13.7x_{i2}^2 + 13.7x_{i3}^2 + 13.7x_{i4}^2,$$

whereas the propensity score model is the same as the one in (5.1) with β_1 varying from 0 to 0.4. Table 4 shows the results

Table 3. Correctly specified outcome with a locally misspecified propensity score model.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	−1.96	0.69	0.80	4.87	0.04	0.87	−0.42	3.07
	GLM	−16.73	8.43	5.85	19.96	8.55	0.84	4.65	21.07
	GAM	−8.19	7.68	−4.35	−10.79	4.62	−0.25	−0.63	2.95
	DR	0.43	0.34	−0.83	−3.67	0.38	0.08	−1.39	−3.50
	CBPS	−0.76	−2.15	0.56	1.34	−1.92	−0.34	0.22	0.37
	oCBPS	−0.41	0.05	0.10	0.06	−0.05	0.02	−0.01	−0.02
	True	41.03	33.16	41.86	82.09	20.65	18.39	28.44	59.63
	GLM	67.79	9.55	23.67	72.99	9.43	3.23	13.86	81.20
Std	GAM	46.08	8.92	21.56	52.34	11.06	2.91	11.78	52.31
Dev	DR	3.10	2.51	2.87	5.74	1.37	1.29	1.59	2.60
	CBPS	3.26	2.56	2.44	2.77	1.58	1.28	1.33	1.43
	oCBPS	2.47	2.24	2.25	2.26	1.29	1.22	1.26	1.29
RMSE	True	41.07	33.17	41.87	82.24	20.65	18.41	28.44	59.70
	GLM	69.82	12.74	24.39	75.67	12.73	3.34	14.62	83.89
	GAM	46.80	11.77	21.99	53.44	11.98	2.92	11.80	52.39
	DR	3.13	2.53	2.99	6.81	1.42	1.29	2.11	4.36
	CBPS	3.35	3.34	2.51	3.07	2.49	1.32	1.34	1.48
	oCBPS	2.50	2.24	2.26	2.27	1.29	1.22	1.26	1.29
	True	0.962	0.948	0.962	0.938	0.934	0.946	0.954	0.942
Coverage	GLM	0.804	0.788	0.888	0.916	0.652	0.936	0.918	0.910
probability	GAM	0.132	0.294	0.238	0.076	0.154	0.612	0.144	0.052
(of the	DR	0.856	0.922	0.866	0.556	0.916	0.936	0.736	0.332
95% C.I.)	CBPS	0.912	0.914	0.926	0.958	0.752	0.954	0.954	0.952
	oCBPS	0.916	0.946	0.936	0.954	0.950	0.948	0.958	0.954

Table 4. Misspecified outcome model with correct propensity score model.

		$n = 300$				$n = 1000$			
	β_1	0	0.13	0.27	0.4	0	0.13	0.27	0.4
Bias	True	−4.37	−0.03	−4.24	1.51	0.80	−1.00	2.31	2.67
	GLM	0.38	−0.64	−2.67	−1.33	0.11	−0.44	0.05	0.75
	GAM	−2.03	−5.49	−10.43	−13.66	−0.65	−1.72	−1.95	−3.04
	DR	−2.77	−5.06	−9.92	−14.36	−2.98	−4.98	−7.43	−10.11
	CBPS	0.07	−0.69	−2.59	−3.94	0.05	−0.55	−0.71	−1.63
	oCBPS	−0.56	−0.97	−3.05	−4.37	−0.03	−0.68	−0.84	−1.70
	True	49.87	58.75	74.32	100.35	27.61	33.62	44.75	53.58
	GLM	18.12	24.87	34.83	56.17	9.68	12.37	18.45	31.16
Std	GAM	17.59	23.19	34.72	49.87	9.07	11.36	16.85	26.50
Dev	DR	14.02	14.65	15.58	16.65	7.95	8.26	8.21	8.40
	CBPS	15.51	17.60	18.83	20.66	8.74	9.47	10.64	12.05
	oCBPS	14.74	16.15	17.13	18.55	8.44	9.03	9.68	10.87
RMSE	True	50.06	58.75	74.45	100.36	27.62	33.64	44.81	53.60
	GLM	18.13	24.88	34.93	56.18	9.68	12.37	18.45	31.17
	GAM	17.71	23.83	36.25	51.71	9.09	11.49	16.96	26.67
	DR	14.29	15.50	18.47	21.99	8.49	9.65	11.07	13.15
	CBPS	15.51	17.62	19.01	21.03	8.74	9.49	10.66	12.16
	oCBPS	14.75	16.18	17.40	19.06	8.44	9.06	9.72	11.00
	True	0.948	0.954	0.946	0.920	0.938	0.950	0.910	0.922
Coverage	GLM	0.896	0.852	0.870	0.868	0.908	0.862	0.816	0.802
probability	GAM	0.912	0.832	0.676	0.476	0.932	0.846	0.690	0.516
(of the	DR	0.930	0.910	0.838	0.716	0.924	0.874	0.794	0.688
95% C.I.)	CBPS	0.920	0.870	0.790	0.676	0.914	0.862	0.776	0.668
	oCBPS	0.950	0.930	0.908	0.904	0.954	0.920	0.902	0.862

when the outcome model is misspecified but the propensity score model is correct. We find that the magnitude of bias is similar across all estimators with the exception of GAM and DR, which seem to have a significantly larger bias. The DR dominates in terms of standard deviation, but oCBPS closely follows. In terms of the root mean square error, oCBPS is on par with DR.

Finally, when both the outcome and propensity score models are misspecified, we observe that DR and oCBPS dominate all other estimators with respect to all three criteria. In particular, oCBPS performs much better than CBPS in all scenarios. The results are organized in Table 5.

In summary, the proposed oCBPS method outperforms the CBPS method with respect to root mean square error (RMSE) under all five scenarios we examined. In addition, the oCBPS method often yields better or at least comparable results relative to all the other estimators.

Table 5. Misspecified outcome with misspecified propensity score models.

		$n = 300$				$n = 1000$			
	β_1	0	0.13	0.27	0.4	0	0.13	0.27	0.4
Bias	True	0.54	−1.74	1.71	−3.56	−2.66	−2.52	−2.06	−0.36
	GLM	2.94	−1.70	−8.47	−20.25	−0.18	−2.07	−8.89	−18.79
	GAM	20.74	12.05	3.42	−8.06	4.95	2.35	−1.03	−5.01
	DR	9.16	6.66	4.52	0.46	6.55	4.91	2.80	0.36
	CBPS	9.57	4.10	0.37	−7.62	0.46	−0.81	−4.94	−11.18
	oCBPS	2.51	−0.24	−1.62	−4.82	0.04	−0.61	−2.29	−4.54
	True	59.12	55.64	54.16	58.35	34.79	31.31	28.41	31.62
	GLM	25.00	19.44	22.49	26.01	9.67	9.79	11.17	12.44
Std	GAM	30.85	23.01	19.46	21.72	10.23	9.53	9.19	9.25
Dev	DR	15.18	15.14	13.71	13.60	7.86	7.85	7.69	7.70
	CBPS	26.74	18.65	19.74	18.92	9.16	9.11	9.36	9.63
	oCBPS	16.28	15.38	15.08	14.42	8.93	8.60	8.32	8.27
RMSE	True	59.12	55.66	54.19	58.45	34.89	31.42	28.48	31.62
	GLM	25.18	19.51	24.03	32.96	9.67	10.00	14.28	22.53
	GAM	37.17	25.97	19.76	23.17	11.37	9.81	9.25	10.52
	DR	17.73	16.54	14.43	13.60	10.24	9.26	8.19	7.71
	CBPS	28.40	19.10	19.75	20.40	9.17	9.15	10.59	14.76
	oCBPS	16.47	15.38	15.17	15.20	8.93	8.62	8.63	9.43
	True	0.952	0.940	0.936	0.952	0.936	0.940	0.952	0.916
Coverage	GLM	0.854	0.902	0.866	0.788	0.890	0.878	0.772	0.540
probability	GAM	0.714	0.810	0.860	0.832	0.868	0.902	0.916	0.834
(of the	DR	0.878	0.920	0.934	0.946	0.876	0.906	0.936	0.946
95% C.I.)	CBPS	0.866	0.892	0.890	0.866	0.894	0.888	0.852	0.670
	oCBPS	0.940	0.964	0.926	0.934	0.944	0.942	0.926	0.894

5.2. An Empirical Application

We next apply the oCBPS methodology to a well-known study where the experimental benchmark estimate is available. Specifically, LaLonde (1986) conducted a study, in which after the randomized evaluation study was implemented, the experimental control group is replaced with a set of untreated individuals taken from the Panel Study of Income Dynamics. This created an artificial observational study with 297 treated observations and 2490 control observations. Ever since the original study, this dataset has been used for evaluating whether a new statistical methodology can recover the experimental benchmark estimate (see, e.g., Dehejia and Wahba 1999; Smith and Todd 2005). In the original CBPS article, Imai and Ratkovic (2014) use this dataset to show that the propensity score matching estimator based on the CBPS method outperforms the matching estimator based on the standard logistic regression. In the following, we evaluate whether the proposed oCBPS method can further improve the CBPS methodology.

We begin by replicating the original results of Imai and Ratkovic (2014) and then compare those results with those of the proposed oCBPS methodology. To do this, we focus on the estimation of the average treatment effect for the treated (ATT). The response of interest is earnings in 1978 and the treatment variable is whether or not the individual participates the job training program. The original randomized experiment yields the ATT estimate $886, which is used as a benchmark for the later comparison. Imai and Ratkovic (2014) consider the propensity score estimation based on the standard logistic regression (GLM), the just-identified CBPS with moment balance condition only (CBPS1) and the over-identified CBPS with score equation and moment balance condition (CBPS2). Based on each set of these estimated propensity scores, we estimate the ATT using the 1-to-1 nearest neighbor matching with replacement. The estimates of standard errors are based on the results in Abadie and Imbens (2006). We then add the estimated propensity score based on the proposed oCBPS methodology. Since the quantity of interest is the ATT, we use a slightly modified oCBPS estimator described in the supplementary material.

We follow the propensity score model specifications examined in Imai and Ratkovic (2014). The covariates we adjust include age, education, race (white, black, or Hispanic), marriage status, high school degree, earnings in 1974 and earnings in 1975 as pretreatment variables. We consider three different specification of balance conditions: the first moment of covariates (Linear), the first and second moment of covariates (Quadratic), and the Quadratic specification with some interactions selected by Smith and Todd (2005). We compare the performance of each methodology across these three specifications.

The results are shown in Table 6. We find that although the standard error is relatively large as in any evaluation study based on the LaLonde data, the proposed oCBPS method yields much smaller bias than GLM and CBPS1 under all three specifications. The oCBPS also improves the CBPS2 under the linear and Smith and Todd's specifications of covariates. We note that the

Table 6. The bias and standard errors (shown in parentheses) of estimates of the average treatment effect for the treated in the LaLonde's Study.

	GLM	CBPS1	CBPS2	oCBPS
Linear	−1190.92	−462.7	−702.33	−306.01
	(1437.02)	(1295.19)	(1240.79)	(1662.02)
Quadratic	−1808.16	−646.54	207.13	−370.03
	(1382.38)	(1284.13)	(1567.33)	(1773.03)
Smith & Todd	−1620.49	−1154.07	−462.24	−383.12
	(1424.57)	(1711.66)	(1404.15)	(1748.87)

NOTE: We use the benchmark $886 as the true value.

standard error of the oCBPS method appears to be larger than the competing methods. This may be due to the fact that the uncertainty of the estimated propensity score is ignored when we calculate the standard error of the matching estimators (i.e., GLM, CBPS1, and CBPS2). In summary, consistent with the theoretical results, the proposed oCBPS method yields more accurate estimates of ATT than the original CBPS estimator or the standard logistic regression. Finally, it is important to note that these results are only suggestive since we do not know whether the assumptions of propensity score methods hold in this study.

6. Conclusion

This article presents a theoretical investigation of the covariate balancing propensity score methodology that others have found work well in practice (e.g., Wyss et al. 2014; Frölich, Huber, and Wiesenfarth 2015). We derive the optimal choice of the covariate balancing function so that the resulting IPTW estimator is first order unbiased under local misspecification of the propensity score model. Furthermore, it turns out that the CBPS-based IPTW estimator with the same covariate balancing function attains the semiparametric efficiency bound.

Given these theoretical insights, we propose an optimal CBPS methodology by carefully choosing the covariate balancing estimating functions. We prove that the proposed oCBPS-based IPTW estimator is doubly robust and locally efficient. More importantly, we show that the rate of convergence of the proposed estimator is faster than the standard AIPW estimator under locally misspecified models. To relax the parametric assumptions and improve the double robustness property, we further extend the oCBPS method to the nonparametric setting. We show that the proposed estimator can achieve the semiparametric efficiency bound without imposing parametric assumptions on the propensity score and outcome models. The theoretical results require weaker technical conditions than existing methods and the estimator has smaller asymptotic bias. Our simulation and empirical studies confirm the theoretical results, demonstrating the advantages of the proposed oCBPS methodology.

In this work, we mainly focus on the theoretical development of the IPTW estimator with the propensity score estimated by the optimal CBPS approach. It is a very interesting research problem to establish the theoretical results for the matching estimators combined with the optimal CBPS approach or some variants. While the asymptotic theory (i.e., consistency and asymptotic normality) for the estimated propensity score via the optimal CBPS approach can be derived from the current results (by the Delta method), the full development is beyond the scope of this work. We leave it for a future study.

Supplementary Material

The supplementary material contains the appendix of this article which collects the proofs and further technical details.

Acknowledgments

An earlier version of this article was entitled, "Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach."

Funding

Supported by NSF grants DMS-1854637, DMS-1712591, DMS-2053832, CAREER Award DMS-1941945, NIH grant R01-GM072611, and Sloan Grant 2020-13946.

References

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267. [12]

Ai, C., Linton, O., Motegi, K., and Zhang, Z. (2021), "A Unified Framework for Efficient Estimation of General Treatment Models," *Quantitative Economics*, 12, 779–816. [2]

Andrews, D. W. (1991), "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 307–345. [8]

Arcones, M. A. (1995), "A Bernstein-Type Inequality for u-Statistics and u-Processes," *Statistics & Probability Letters*, 22, 239–247. [2]

Bang, H., and Robins, J. M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–973. [2]

Benkeser, D., Carone, M., van der Laan, M. J., and Gilbert, P. B. (2017), "Doubly-Robust Nonparametric Inference on the Average Treatment Effect," *Biometrika*, 104, 863–880. [9]

Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag. [9]

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*, asp033. [2]

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016), "Globally Efficient Nonparametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting," *Journal of the Royal Statistical Society*, Series B, 78, 673–700. [2,8,9]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5549–5632. [2,9]

Copas, J., and Eguchi, S. (2005), "Local Model Uncertainty and Incomplete-Data Bias," *Journal of the Royal Statistical Society*, 67, 459–513. [3]

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [12]

Fong, C., Hazlett, C., and Imai, K. (2018a), "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements," *Annals of Applied Statistics*, 12, 156–177. [2]

——— (2018b), "CBPS: R Package for Covariate Balancing Propensity Score." Comprehensive R Archive Network (CRAN), available at: *https://CRAN.R-project.org/package=CBPS*. [3]

Frölich, M., Huber, M. and Wiesenfarth, M. (2015), "The Finite Sample Performance of Semi- and Nonparametric Estimators for Treatment Effects and Policy Evaluation," Tech. rep., IZA discussion paper No. 8756. [2,13]

Graham, B. S., Pinto, C., and Egel, D. (2012), "Inverse Probability Tilting for Moment Condition Models With Missing Data," *Review of Economic Studies*, 79, 1053–1079. [2,8]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [2,4,6]

Hainmueller, J. (2012), "Entropy Balancing for Causal Effects: Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Political Analysis*, 20, 25–46. [2,8]

Han, P., and Wang, L. (2013), "Estimation With Missing Data: Beyond Double Robustness," *Biometrika*, ass087. [2,8]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [2]

Henmi, M., and Eguchi, S. (2004), "A Paradox Concerning Nuisance Parameters and Projected Estimating Functions," *Biometrika*, 91, 929–941. [6]

Hirano, K., Imbens, G., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1307–1338. [2,9]

Horvitz, D., and Thompson, D. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685. [1]

Imai, K., and Ratkovic, M. (2014), "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society*, 76, 243–263. [2,6,12]

——— (2015), "Robust Estimation of Inverse Probability Weights for Marginal Structural Models," *Journal of the American Statistical Association*, 110, 1013–1023. [2]

Imbens, G. W., Newey, W. K., and Ridder, G. (2007), "Mean-Square-Error Calculations for Average Treatment Effects," Technical Report. [2,9]

Kang, J. D. Y., and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data," *Statistical Science*, 22, 574–580. [2,3,7]

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *The American Economic Review*, 76, 604–620. [12]

Newey, W. K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168. [2,8,9]

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245. [5]

Ning, Y., Peng, S., and Imai, K. (2018), "Robust Estimation of Causal Effects Via High-Dimensional Covariate Balancing Propensity Score." arXiv:1812.08683 . [2]

Owen, A. B. (2001), *Empirical Likelihood*, New York: Chapman & Hall/CRC. [2]

Qin, J., and Zhang, B. (2007), "Empirical-Likelihood-Based Inference in Missing Response Problems and Its Application in Observational Studies," *Journal of the Royal Statistical Society*, 69, 101–122. [2]

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable," *Statistical Science*, 22, 544–559. [2,7]

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, 846–866. [2,5,6,7]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association* 90 106–121. [5]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1]

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). "Improved Double-Robust Estimation in Missing Data and Causal Inference Models," *Biometrika*, 99, 439–456. [2]

Rubin, D. B. (1990), "Comments on 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9' by J. Splawa-Neyman Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed," *Statistical Science*. 472–480. [1]

Smith, J. A., and Todd, P. E. (2005), "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, 305–353. [12]

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637. [2,8]

——— (2010), "Bounded, Efficient and Doubly Robust Estimation With Inverse Weighting," *Biometrika*, 97, 661–682. [2,8]

Tropp, J. A. (2015), "An Introduction to Matrix Concentration Inequalities," arXiv:1501.01571. [2]

van der Laan, M. J. (2010), "Targeted Maximum Likelihood Based Causal Inference: Part I," *The International Journal of Biostatistics*, 6, 2. [2]

Vermeulen, K., and Vansteelandt, S. (2015), "Bias-Reduced Doubly Robust Estimation," *Journal of the American Statistical Association*, 110, 1024–1036. [2,7,8]

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Funk, M. J., LoCasale, R., and Stürmer, T. (2014), "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score," *American Journal of Epidemiology* 180, 645–655. [2,13]

Zhao, Q. (2019), "Covariate Balancing Propensity Score by Tailored Loss Functions," *The Annals of Statistics*, 47, 965–993. [2,8]

Zhao, Q., and Percival, D. (2017), " Primal–Dual Covariate Balance and Minimal Double Robustness Via Entropy Balancing," *Journal of Causal Inference*, 5. [2,8]

Zubizarreta, J. R. (2015), "Stable Weights That Balance Covariates for Estimation With Incomplete Outcome Data," *Journal of the American Statistical Association*, 110, 910–922. [2,8]