Biometrika (2020), **107**, 3, *pp.* 533–554
Printed in Great Britain

doi: 10.1093/biomet/asaa020
Advance Access publication 3 June 2020

Robust estimation of causal effects via a high-dimensional covariate balancing propensity score

By YANG NING

Department of Statistics and Data Science, Cornell University, 1188 Comstock Hall,
129 Garden Ave., Ithaca, New York 14853, U.S.A.
yn265@cornell.edu

SIDA PENG

Microsoft Research, 14865 NE 36th St., Redmond, Washington 98052, U.S.A.
sidpeng@microsoft.com

and KOSUKE IMAI

Department of Statistics, Harvard University, One Oxford Street, Cambridge,
Massachusetts 02138, U.S.A.
Imai@harvard.edu

Summary

We propose a robust method to estimate the average treatment effects in observational studies when the number of potential confounders is possibly much greater than the sample size. Our method consists of three steps. We first use a class of penalized M-estimators for the propensity score and outcome models. We then calibrate the initial estimate of the propensity score by balancing a carefully selected subset of covariates that are predictive of the outcome. Finally, the estimated propensity score is used to construct the inverse probability weighting estimator. We prove that the proposed estimator, which we call the high-dimensional covariate balancing propensity score, has the sample boundedness property, is root-n consistent, asymptotically normal, and semiparametrically efficient when the propensity score model is correctly specified and the outcome model is linear in covariates. More importantly, we show that our estimator remains root-n consistent and asymptotically normal so long as either the propensity score model or the outcome model is correctly specified. We provide valid confidence intervals in both cases and further extend these results to the case where the outcome model is a generalized linear model. In simulation studies, we find that the proposed methodology often estimates the average treatment effect more accurately than existing methods. We also present an empirical application, in which we estimate the average causal effect of college attendance on adulthood political participation. An open-source software package is available for implementing the proposed methodology.

Some key words: Causal inference; Double robustness; Model misspecification; Post-regularization inference; Semiparametric efficiency.

1. Introduction

The propensity score of Rosenbaum & Rubin (1983) plays a central role in the estimation of causal effects in observational studies. In particular, matching and weighting methods based on propensity score have become part of the applied researcher's toolkit across many scientific disciplines. One important challenge, which is becoming increasingly common as the amount of available data grows, is the question of how to incorporate a large number of potential confounders. For example, Schneeweiss et al. (2009) consider a total of several thousand candidate confounders obtained from the health care claims data.

In this paper we propose a robust method to estimate the average treatment effect and the average treatment effect for the treated in observational studies when the number of potential confounders is possibly much greater than the sample size. Under the standard assumption of strong ignorability, we propose to estimate the propensity score by balancing covariates in high-dimensional settings. The proposed method consists of several steps. We first obtain an initial estimator of the propensity score by maximizing a penalized generalized quasilikelihood, which depends on a user-specified weight function. Next, we apply the weighted least squares method to fit the outcome model. We show that the two weight functions critically determine the performance of the proposed estimator under model misspecification. Third, we refine the initial estimate of the propensity score by balancing a carefully selected set of observed covariates that are predictive of the outcome. Finally, the estimated propensity score is used to construct the inverse probability weighting estimator of the average treatment effect. The extension to the estimation of the average treatment effect for the treated is given in the Supplementary Material.

We prove that under mild conditions the proposed estimator, which we call the high-dimensional covariate balancing propensity score, has the sample boundedness property, is root-n consistent, asymptotically normal and semiparametrically efficient when the propensity score model is correctly specified and the outcome model is linear in covariates. Although this result holds for a broad class of weight functions used in the initial estimation of propensity score, the choice of weight functions determines the rate of convergence under model misspecification. We carefully choose the weight functions such that the estimator remains root-n consistent and asymptotically normal so long as either the propensity score model or the outcome model is correctly specified. In addition, the proposed estimator comes with honest confidence intervals. Finally, we extend these theoretical results to the case where the outcome model is a generalized linear model in order to allow for nonlinearity. An open-source software R package, CBPS (R Development Core Team 2020), is available for implementing our proposed methodology at `https://CRAN.R-project.org/package=CBPS`.

The proposed methodology does not require the variable selection consistency of either the propensity score model or the outcome model. This property is shared by many recent works on high-dimensional inference on regression models; see Zhang & Zhang (2014), Javanmard & Montanari (2014), Van de Geer et al. (2014), Belloni et al. (2016), Ning & Liu (2017), Ning et al. (2017), Cai & Guo (2017), Neykov et al. (2018), Zhu & Bradic (2018), and Dukes et al. (2019), among many others. The main idea of these works is to correct the bias of the lasso-type estimators by inverting the optimality condition or the score function through the projection onto the nuisance tangent space, yielding honest confidence intervals for prespecified coefficients.

Our proposed methodology builds on three strands of research that have recently emerged in the causal inference literature; see § 3.5 for a detailed discussion. First, a number of researchers have proposed estimating the average treatment effect by optimizing covariate balance between the treatment and control groups (e.g., Hainmueller, 2012; Graham et al., 2012; Imai & Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2016; Fan et al., 2016; Zhao, 2019). The proposed methodology extends the covariate balancing propensity score of Imai & Ratkovic (2014) and Fan et al. (2016) to the high-dimensional setting, in which the number of potential confounders is possibly greater than the sample size.

Second, we contribute to the growing literature on the estimation of the average treatment effect in high-dimensional settings (e.g., Belloni et al., 2014, 2017; Farrell, 2015; Chernozhukov et al., 2018). These methods first estimate the nuisance parameters, e.g., propensity score, typically by the penalized maximum likelihood, and then estimate the average treatment effect by solving the efficient score function. Different from this line of work, we rely on the covariate balancing strategy for estimating the propensity score model and use the Horvitz–Thompson estimator. As elaborated in § 3.5, the robustness of the asymptotic distributions of our estimator to model misspecification is the main advantage over these existing methods. Most recently, Tan (2017, 2018) proposed a penalized calibrated propensity score method. Unlike this method, our approach is based on covariate balancing.

Finally, our method is related to the recently proposed approximate residual balancing method (Athey et al., 2018), which, unlike our methodology, does not require a propensity score model. While the approximate residual balancing method requires the outcome model to be linear in covariates, our method yields a consistent and asymptotically normal estimator even under the misspecification of the outcome model so long as the propensity score is correctly specified. In addition, we extend the proposed method and its asymptotic theory to the case in which the outcome variable follows a generalized linear model. This also overcomes the same limitation of the original covariate balancing estimator (Imai & Ratkovic, 2014; Fan et al., 2016).

Throughout the paper, we use the following notation. For $v = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$, and $1 \leqslant q \leqslant \infty$, we define $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $\|v\|_0 = |\mathrm{supp}(v)|$, where $\mathrm{supp}(v) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $v^{\otimes 2} = vv^{\mathrm{T}}$. If the matrix M is symmetric, then $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ are the minimal and maximal eigenvalues of M. For $S \subseteq \{1, \ldots, d\}$, let $v_S = \{v_j : j \in S\}$ and S^c be the complement of S. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if $C \leqslant a_n/b_n \leqslant C'$ for some $C, C' > 0$. Similarly, we use $a_n \lesssim b_n$ to denote $a_n \leqslant Cb_n$ for some constant $C > 0$. A random variable X is subexponential if there exists some constant $K_1 > 0$ such that $\mathrm{pr}(|X| > t) \leqslant \exp(1 - t/K_1)$ for all $t \geqslant 0$. The subexponential norm of X is defined as $\|X\|_{\psi_1} = \sup_{p\geqslant 1} p^{-1}(E|X|^p)^{1/p}$. A random variable X is sub-Gaussian if there exists some constant $K_2 > 0$ such that $\mathrm{pr}(|X| > t) \leqslant \exp(1 - t^2/K_2^2)$ for all $t \geqslant 0$. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{p\geqslant 1} p^{-1/2}(E|X|^p)^{1/p}$. Denote $a \vee b = \max(a, b)$.

2. Proposed methodology

2.1. *Set-up*

Suppose that we observe a simple random sample of size n from a population of interest. For each unit i, we observe a 3-tuple (T_i, Y_i, X_i) where X_i is a d-dimensional vector of pretreatment covariates, Y_i is an outcome variable and T_i is a binary treatment variable denoting whether the observation receives the treatment $(T_i = 1)$ or not $(T_i = 0)$. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under the treatment and control conditions, respectively. This notation implies the stable unit treatment value assumption (Rubin, 1990). Then, the observed outcome can be written as $Y_i = Y_i(T_i)$. Our goal is to infer the average treatment effect, $\mu^* = E\{Y_i(1) - Y_i(0)\}$.

We focus on the estimation of $\mu_1^* = E\{Y_i(1)\}$, since $\mu_0^* = E\{Y_i(0)\}$ can be estimated in a similar manner. We impose a working parametric model $\pi(X_i^{\mathrm{T}}\beta)$ for the treatment assignment mechanism, which is known as the propensity score $\mathrm{pr}(T_i = 1 \mid X_i)$, where $\pi(\cdot)$ is a known function and β is an unknown d-dimensional vector. In this work, we consider settings where the number of covariates is possibly much greater than the sample size, i.e., $d \gg n$. When the propensity score model is correctly specified, we have

$$\mathrm{pr}(T_i = 1 \mid X_i) = \pi(X_i^{\mathrm{T}}\beta^* + u_i), \tag{1}$$

for some $\beta^* \in \mathbb{R}^d$, where u_i is a small approximation error which accounts for the nonsparse effect due to weak signals in the model. Similarly, for the outcome variable we impose a linear working model. When the working model is correctly specified, we have

$$E\{Y_i(1) \mid X_i\} = K_1(X_i) + r_i, \tag{2}$$

where $K_1(X_i) = \alpha^{*\mathrm{T}} X_i$ for some $\alpha^* \in \mathbb{R}^d$ and similarly r_i is the approximation error. An extension to generalized linear models will be studied in § 4. In general, the propensity score model (1) or the outcome model (2) can be misspecified. We begin by assuming both models (1) and (2) hold. When studying the theoretical properties of our proposed methodology in § 3, however, we will consider the situations in which either model (1) or (2) does not hold.

2.2. *High-dimensional covariate balancing propensity score*

In many applications, it is often reasonable to assume that the propensity score model is sparse or approximately sparse. Under the sparsity assumption, Tibshirani (1996) and Fan & Li (2001) proposed the penalized maximum likelihood estimators for parameter estimation and variable selection. Unfortunately, the penalized maximum likelihood estimators cannot be directly used with the Horvitz–Thompson estimator to infer $\mu_1^* = E\{Y_i(1)\}$ because the estimator may incur a large bias due to shrinkage. Thus, the resulting estimator may have a slower rate of convergence.

To address this problem, we estimate the propensity score by optimizing covariate balance between the treatment and control groups. To this end, we distinguish the following two types of covariate balancing properties.

Definition 1 (Covariate balancing properties). *Let $\hat{\pi} = \pi(X^{\mathrm{T}}\hat{\beta})$ denote an estimator of the propensity score* $\mathrm{pr}(T = 1 \mid X)$ *with $\hat{\beta}$ being an estimator of β^*, which is the true value of β.*

(a) We say $\hat{\pi}$ satisfies the strong covariate balancing property if the following equality holds:

$$\sum_{i=1}^{n} \left(\frac{T_i}{\hat{\pi}_i} - 1 \right) X_i = 0. \tag{3}$$

(b) We say $\hat{\pi}$ satisfies the weak covariate balancing property if the following equality holds:

$$\sum_{i=1}^{n} \left(\frac{T_i}{\hat{\pi}_i} - 1 \right) \alpha^{*\mathrm{T}} X_i = 0, \tag{4}$$

where α^* is defined by $K_1(X_i) = \alpha^{*\mathrm{T}} X_i$ in equation (2).

Although the strong covariate balancing property implies the weak one, the converse does not necessarily hold. Existing covariate balancing propensity score methods aim to achieve the strong covariate balancing property, which balances the mean of every component of X_i (e.g., Imai & Ratkovic, 2014; Fan et al., 2016). However, constructing an estimator $\hat{\pi}$ with the strong covariate balancing property is difficult in high-dimensional settings. When $d > n$, the estimator $\hat{\beta}$ that satisfies equation (3) is not unique and therefore not even well defined.

To overcome this difficulty, we propose to estimate the propensity score such that it approximately satisfies the weak covariate balancing property, i.e., $n^{-1} \sum_{i=1}^{n} (T_i/\hat{\pi}_i - 1)\alpha^{*\mathrm{T}} X_i \approx 0$. We show that the weak covariate balancing property is sufficient to remove the bias from the estimation of the propensity score model. Here, we first introduce the proposed methodology, which we call the high-dimensional covariate balancing propensity score.

Step 1. Define a generalized quasilikelihood function as

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\beta^{\mathrm{T}}X_i}\left\{\frac{T_i}{\pi(u)} - 1\right\} w_1(u)\,\mathrm{d}u,$$

where $w_1(\cdot)$ is a positive weight function. Compute the regularized estimator

$$\hat{\beta} \;=\; \underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}}\{-Q_n(\beta) + \lambda\|\beta\|_1\}, \tag{5}$$

where $\lambda > 0$ is a tuning parameter.

Step 2. Define a weighted least square loss function using the treatment group as

$$L_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} T_i w_2(\hat{\beta}^{\mathrm{T}}X_i)(Y_i - \alpha^{\mathrm{T}}X_i)^2,$$

where $w_2(\cdot)$ is another positive weight function. Compute the regularized estimator

$$\tilde{\alpha} \;=\; \underset{\alpha\in\mathbb{R}^d}{\operatorname{argmin}}\{L_n(\alpha) + \lambda'\|\alpha\|_1\}, \tag{6}$$

where $\lambda' > 0$ is a tuning parameter.

Step 3. Let $\tilde{S} = \{j : |\tilde{\alpha}_j| > 0\}$ denote the support of $\tilde{\alpha}$ and $X_{\tilde{S}}$ denote the corresponding subset of X. We calibrate the initial estimator $\hat{\beta}_{\tilde{S}}$ to balance $X_{\tilde{S}}$. Specifically, we solve

$$\tilde{\gamma} \;=\; \underset{\gamma\in\mathbb{R}^{|\tilde{S}|}}{\operatorname{argmin}}\,\|g_n(\gamma)\|_2^2 \quad \text{where } g_n(\gamma) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(\gamma^{\mathrm{T}}X_{i\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}}X_{i\tilde{S}^c})} - 1\right\}X_{i\tilde{S}}. \tag{7}$$

We then set $\tilde{\beta} = (\tilde{\gamma}, \hat{\beta}_{\tilde{S}^c})$ and $\tilde{\pi}_i = \pi(\tilde{\beta}^{\mathrm{T}}X_i)$.

Step 4. Estimate $\mu_1^* = E\{Y_i(1)\}$ by the Horvitz–Thompson estimator $\hat{\mu}_1 = n^{-1}\sum_{i=1}^{n} T_iY_i/\tilde{\pi}_i$.

In Step 1, we obtain an initial estimate of the propensity score via the penalized M-estimation approach. We refer to the function $Q_n(\beta)$ as the generalized quasilikelihood function, as its construction is similar to the quasilikelihood function for generalized linear models (Wedderburn, 1974). To understand how the generalized quasilikelihood function is motivated, we compute the corresponding quasi-score function,

$$\frac{\partial Q_n(\beta)}{\partial\beta} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(\beta^{\mathrm{T}}X_i)} - 1\right\} w_1(\beta^{\mathrm{T}}X_i)X_i. \tag{8}$$

Since (8) is an unbiased estimating function for β, $Q_n(\beta)$ serves as a legitimate quasilikelihood function that integrates the quasi-score function (8). The quasilikelihood function $Q_n(\beta)$ depends on the choice of the weight $w_1(u)$. In particular, we consider the following two examples.

(a) If $w_1(u) = \pi(u)$, (8) is identical to the score function for the logistic regression and thus $Q_n(\beta)$ reduces to the standard quasi-likelihood function for the treatment variable.

(b) If $w_1(u) = 1$, the quasi-score function (8) leads to the strong covariate balancing equation (3). Consequently, we call $Q_n(\beta)$ with $w_1(u) = 1$ as the covariate balancing loss function.

Thus, in Step 1 we allow a broad class of initial estimators $\hat{\beta}$, including the penalized (quasi-)maximum likelihood estimator and many other penalized M-estimators corresponding to different $w_1(u)$. By computing the Hessian matrix of $Q_n(\beta)$, we find that (5) can be a nonconvex optimization problem depending on the choice of $w_1(u)$. The nonconvexity may pose computational challenges. For instance, the gradient descent algorithm can be trapped at a local solution which is far from the global maximizer. To avoid the computational issue, we mainly focus on the concave quasilikelihood function $Q_n(\beta)$. It is easy to verify that $Q_n(\beta)$ with $w_1(u) = \pi(u)$ in case (a) and $w_1(u) = 1$ in case (b) are both concave.

In Step 2 we fit the outcome model using a class of penalized weighted least square estimators. We allow the weight $w_2(\hat{\beta}^{\mathrm{T}} X_i)$ to depend on X_i and also the initial estimator $\hat{\beta}$ from Step 1. For instance, we have the following examples.

(a′) If $w_2(u) = 1$, $L_n(\alpha)$ is the classical least square loss function in the treatment group.
(b′) If $w_2(u) = 1/\pi(u)$, $L_n(\alpha)$ is known as the inverse propensity score weighted least square loss.
(c′) If $w_2(u) = \pi'(u)/\pi^2(u)$, $L_n(\alpha)$ remains a valid loss function for estimating α. It is shown in § 3.4 that this loss function plays an important role when studying the robustness of the proposed estimator to misspecified outcome models. In the following, we call this loss function the propensity score adjusted least square loss.

Step 3 removes the bias induced by the penalized estimators used in Steps 1 and 2. We calibrate the estimated propensity score by balancing a subset of covariates $X_{\tilde{S}}$, which represent the variables selected for the outcome model. The optimization problem (7) implies that the estimated propensity score $\tilde{\pi}_i$ achieves the strong covariate balancing property only for these covariates $X_{\tilde{S}}$ but not for the other covariates $X_{\tilde{S}^c}$. Thus, unlike the original covariate balancing methodology, the proposed method does not achieve the strong covariate balancing property. Interestingly, however, the method does approximately satisfy the weak covariate balancing property if α^* can be approximated well by $\tilde{\alpha}$. Specifically, we have

$$\sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i}-1\right)\alpha^{*\mathrm{T}}X_i \approx \sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i}-1\right)\tilde{\alpha}^{\mathrm{T}}X_i = \sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i}-1\right)\tilde{\alpha}_{\tilde{S}}^{\mathrm{T}}X_{i\tilde{S}} = 0, \qquad (9)$$

where the first equality follows from $\tilde{\alpha}_{\tilde{S}^c} = 0$ and the second equality holds due to (7).

In Step 4, we estimate μ_1^* using the Horvitz–Thompson estimator. The proposed estimator can be rewritten as the Horvitz–Thompson estimator with the normalized weight, and also the augmented inverse probability weighting estimator. A detailed discussion is deferred to the Supplementary Material.

3. Theoretical properties of the proposed estimator

3.1. *Assumptions*

We now study the theoretical properties of the proposed estimator. We begin by presenting the required assumptions.

Assumption 1 (*Unconfoundedness*). The treatment assignment is unconfounded, i.e., $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i$.

Assumption 2 (*Overlap*). There exists a constant $c_0 > 0$ such that $\pi_i^* \geqslant c_0$ for $1 \leqslant i \leqslant n$, where $\pi_i^* = \pi(X_i^{\mathrm{T}}\beta^*)$

Assumption 1 implies that there are no unmeasured confounders, while Assumption 2 requires that all samples have a positive probability of receiving the treatment. Together, these represent the standard strong ignorability condition common to propensity score methods (Rosenbaum & Rubin, 1983). To estimate the treatment effect, one also needs to identify $E\{Y(0)\}$, which requires a similar overlap assumption $\pi_i^* \leqslant 1 - c_1$ for some constant $c_1 > 0$.

Assumption 3 (*Sub-Gaussian condition*). Assume that $\varepsilon_1 = Y(1) - \alpha^{*\mathrm{T}}X$ and X_j satisfy $\|\varepsilon_1\|_{\psi_2} \leqslant C_\varepsilon$ and $\|X_j\|_{\psi_2} \leqslant C_X$ for any $1 \leqslant j \leqslant d$, where C_X and C_ε are positive constants.

Assumption 3 controls the tail behaviour of the error ε_1 and the covariate X_j, which facilitates the use of many existing concentration inequalities in high-dimensional statistics. Similar sub-Gaussian conditions are imposed by Athey et al. (2018) in their Theorem 5. Belloni et al. (2017) and Farrell (2015) relaxed the sub-Gaussian condition on ε_1 to the bounded qth moment for some $q > 4$ under a slightly stronger sparsity assumption than our sparsity assumption below.

Assumption 4 (*Sparsity*). Assume that $(s_1 \vee s_2)\log(d \vee n)/n^{1/2} = o(1)$ as $s_1, s_2, d, n \to \infty$, where $s_1 = \|\beta^*\|_0$ and $s_2 = \|\alpha^*\|_0$. Recall that $a \vee b = \max(a, b)$. The approximation errors in the propensity score model (1) and the outcome model (2) satisfy $\sum_{i=1}^n r_i^2 = O(s_2)$, $\sum_{i=1}^n u_i^2 = O(s_1)$, and $\sum_{i=1}^n r_i u_i = o(n^{1/2})$.

Assumption 4 requires that both the propensity score and outcome models are sparse and the approximation errors are sufficiently small. Since we consider the high-dimensional case with $d \gg n$, the sparsity assumption plays an important role in the regularized M-estimation of the propensity score and outcome models. In particular, if $s_1 \asymp s_2 \asymp n^\kappa$ for some $\kappa < 1/2$, then the condition reduces to $d = o\{\exp(n^{1/2-\kappa})\}$. This condition is similar to that in Belloni et al. (2014, 2017) and Farrell (2015), where they imposed a slightly stronger condition with $\log(d \vee n)$ replaced by $\{\log(d \vee n)\}^q$ for some $q > 1$.

Assumption 5 (*Eigenvalue condition*)> Denote $\Sigma = E(X^{\otimes 2})$. There exists a constant $C > 0$ such that $C \leqslant \lambda_{\min}(\Sigma_{SS}) \leqslant \lambda_{\max}(\Sigma_{SS}) \leqslant 1/C$ for any $S \subset \{1, \ldots, d\}$ with $|S| \lesssim (s_1 \vee s_2)\log n$.

When the dimension d is fixed, this assumption simply requires that the design matrix has full column rank, which is a standard regularity condition for regression problems. For high dimensions, Assumption 5 implies the well-known sparse eigenvalue condition introduced by Bickel et al. (2009); see Lemma 1 in Belloni & Chernozhukov (2013). The same sparse eigenvalue condition is imposed by Belloni et al. (2014). We also refer to Assumption 5 of Athey et al. (2018) and § 6.2 of Farrell (2015) for a similar restricted eigenvalue condition. Since our assumption only applies to any $S \times S$ submatrix of Σ, it is weaker than the $C \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant 1/C$ imposed by Van de Geer et al. (2014), Ning & Liu (2017) and Cai & Guo (2017) for high-dimensional inference on lasso estimators.

Assumption 6 (*Propensity score and weight functions*). Assume that $Q_n(\beta)$ is a concave function. Let C, C' denote positive constants, which may take different values for each of the conditions below.

(i) The propensity score model $\pi(u)$ satisfies $C \leqslant (\pi_i^*)' \leqslant 1/C$, and there exist constants $r > 0$ and $C' > 0$ such that the Lipschitz condition holds locally, i.e., $|\pi'(u) - \pi'(v)| \leqslant C'|u - v|$ for any $u, v \in [X_i^{\mathrm{T}}\beta^* - r, X_i^{\mathrm{T}}\beta^* + r]$ and $1 \leqslant i \leqslant n$.
(ii) The weight $w_1(u)$ satisfies $C \leqslant w_{1i}^* \leqslant 1/C$, $(w_{1i}^*)' \leqslant 1/C$ and the local Lipschitz condition $|w_1'(u) - w_1'(v)| \leqslant C'|u - v|$ for any $u, v \in [X_i^{\mathrm{T}}\beta^* - r, X_i^{\mathrm{T}}\beta^* + r]$ and $1 \leqslant i \leqslant n$, where $w_{1i}^* = w_1(X_i^{\mathrm{T}}\beta^*)$.
(iii) The weight $w_2(u)$ satisfies $C \leqslant w_{2i}^* \leqslant 1/C$ and $(w_{2i}^*)' \leqslant 1/C$ for $1 \leqslant i \leqslant n$, where $w_{2i}^* = w_2(X_i^{\mathrm{T}}\beta^*)$. Assume $w_2'(u)$ is continuous.

Assumption 6 imposes mild regularity conditions on the propensity score function and weight functions. In part (i), we assume $\pi(u)$ is differentiable and its derivative is bounded and Lipschitz around $X_i^{\mathrm{T}}\beta^*$. Under the overlap assumption $c_1 \leqslant \pi_i^* \leqslant 1 - c_1$, part (i) holds for the logistic regression without any further conditions. In parts (ii) and (iii), we assume mild conditions on the magnitude and smoothness of $w_1(u)$ and $w_2(u)$. Again, if $\pi(u)$ is the logistic function and the overlap assumption holds, all examples of $w_1(u)$ and $w_2(u)$ discussed in § 2.2 satisfy the regularity conditions in parts (ii) and (iii). Thus, Assumption 6 automatically holds for the logistic propensity score model under the overlap assumption.

3.2. *Asymptotic distribution under correct model specification*

We now derive the theoretical results for the proposed estimator $\hat{\mu}_1$ when both the propensity score model (1) and the outcome model (2) are correctly specified. Recall that our estimator $\hat{\mu}_1$ depends on the choice of the two weight functions, i.e., $w_1(u)$ in Step 1 and $w_2(u)$ in Step 2. In the following, we establish the asymptotic normality and semiparametric efficiency of $\hat{\mu}_1$ for any weight functions $w_1(u)$ and $w_2(u)$.

Theorem 1 (Asymptotic normality and semiparametric efficiency). *Suppose that both the propensity score model* (1) *and the outcome model* (2) *are correctly specified and Assumptions 1–6 hold. If we take $\lambda \asymp \lambda' \asymp \{\log(d \vee n)/n\}^{1/2}$, then the estimator $\hat{\mu}_1$ with any weight functions $w_1(u)$ and $w_2(u)$ satisfies*

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^*}\{Y_i(1) - \alpha^{*\mathrm{T}}X_i\} + \alpha^{*\mathrm{T}}X_i - \mu_1^*\right] + O_p\left\{\frac{(s_1 \vee s_2)\log(d \vee n)}{n}\right\}$$

as $d, n \to \infty$. Let V be the semiparametric asymptotic variance bound, i.e.,

$$V = E\left\{\frac{1}{\pi^*}E(\varepsilon_1^2 \mid X) + (\alpha^{*\mathrm{T}}X - \mu_1^*)^2\right\}.$$

Assume that $E(\varepsilon_1^2 \mid X) \geqslant c$ for some constant $c > 0$ and $E(\alpha^{*\mathrm{T}}X)^4 = O(s_2^2)$. Then, $n^{1/2}(\hat{\mu}_1 - \mu_1^*)/V^{1/2} \to_d N(0, 1)$.

The theorem shows that $\hat{\mu}_1 - \mu_1^*$ is asymptotically equivalent to the average of the efficient score functions, and hence $\hat{\mu}_1$ is locally efficient under the correct model specification. In addition, the asymptotic distribution of $\hat{\mu}_1$ does not depend on the choice of the weight functions $w_1(u)$ and $w_2(u)$, provided that they satisfy Assumption 6. The intuition is that, as long as the weak covariate balancing property is approximately attained, the choice of the weight functions in the first two steps is less important.

To prove the asymptotic normality of $\hat{\mu}_1$, we further require that the variance of the noise cannot tend to 0, i.e., $E(\varepsilon_1^2 \mid X) \geqslant c > 0$. This guarantees the nondegeneracy of the asymptotic variance V. We also assume $E(\alpha^{*\mathrm{T}}X)^4 = O(s_2^2)$ in order to verify the Lyaponov condition for the central limit theorem. This is a mild technical condition. For instance, if X is a sub-Gaussian vector and $\|\alpha^*\|_2 = O(s_2^{1/2})$, then $\|\alpha^{*\mathrm{T}}X\|_{\psi_2} \leqslant \|\alpha^*\|_2\|X\|_{\psi_2} = O(s_2^{1/2})$. This further implies the desired condition $E(\alpha^{*\mathrm{T}}X)^4 = O(s_2^2)$ by the definition of the sub-Gaussian norm.

The asymptotic variance V depends on the true data-generating process, which is allowed to change with d and n. For this reason, we consider the limiting distribution of the standardized statistic $n^{1/2}(\hat{\mu}_1 - \mu_1^*)/V^{1/2}$ as $n, d \to \infty$. Hahn (1998) proved that V is the semiparametric asymptotic variance bound, when both the propensity score and outcome models are treated as nuisance. He proposed a nonparametric inverse probability weighting estimator for fixed d that attains this semiparametric efficiency bound. We show that, when the high-dimensional models (1) and (2) are both correct, the estimator $\hat{\mu}_1$ attains the same bound and is locally efficient.

Our variance bound V is different from the oracle efficiency bound, which is the semiparametric variance bound with the known support of the propensity score and outcome models (Hahn, 2004). Since the support of both models is unknown and the variable selection consistency does not hold under our assumptions, the estimation of the support set leads to additional uncertainty. This explains why our method cannot attain the oracle efficiency bound. We refer to § 5.3 of Farrell (2015) for further discussion on this point.

Remark 1 (*Sample splitting*). Chernozhukov et al. (2018) proposed a double machine learning method based on the sample splitting technique to reduce the bias and allow for more flexible estimators of nuisance functions; see also Zheng & van der Laan (2011). Newey & Robins (2018) showed that the sample splitting technique can improve the convergence rate of the remainder terms for semiparametric models. In the Supplementary Material we propose a modified algorithm based on sample splitting by replacing Assumption 4 with a weaker assumption: $(s_1 s_2)^{1/2} \log(d \vee n)/n^{1/2} = o(1)$. Ignoring the logarithmic factors of d and n, while Assumption 4 requires $s_1 = o(n^{1/2})$ and $s_2 = o(n^{1/2})$, the sample splitting method only requires the weaker condition $s_1 s_2 = o(n)$, which may still hold if one model is dense, e.g., $n^{1/2} \ll s_1 \ll n$, and the other model is sufficiently sparse, e.g., $s_2 \ll n^{1/2}$. However, the sample splitting method incurs further computational cost and may not be stable when the sample size is small.

Remark 2 (*Sample boundedness*). The proposed method guarantees that $\hat{\mu}_1$ lies in the range of $\{Y_i : T_i = 1, i = 1, \ldots, n\}$. This sample boundedness property (Robins et al., 2007) holds because, by construction, the covariate balancing equation satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i} - 1\right) = 0 \tag{10}$$

so long as an intercept is included in $X_{i\tilde{S}}$. Equation (10) implies that the estimated propensity score $\tilde{\pi}_i$ must be greater than or equal to $1/n$ for any treated observation. In contrast, the estimated propensity score π_i^* for the ith observation via the penalized maximum likelihood estimation can become close to 0, leading to extremely large weights for some observations and unstable causal effect estimates. To see why the sample boundedness property holds, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\tilde{\pi}_i} \geqslant \frac{\min_{i:T_i=1} Y_i}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i} = \min_{i:T_i=1} Y_i,$$

where the last equality holds by (10). Similarly, we can show that $\hat{\mu}_1 \leqslant \max_{i:T_i=1} Y_i$.

Finally, to construct a confidence interval for μ_1^*, we estimate V by

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\tilde{\pi}_i^2}(Y_i - \tilde{\alpha}^{\mathrm{T}}X_i)^2 + (\tilde{\alpha}^{\mathrm{T}}X_i - \hat{\mu}_1)^2\right\}. \tag{11}$$

The following corollary shows that $\hat{V}$ is a consistent estimator of V and therefore we obtain valid confidence intervals for μ_1^*.

COROLLARY 1 (HONEST CONFIDENCE INTERVALS). *Suppose that the assumptions in Theorem 1 hold. Then*

$$|\hat{V} - V| = O_p\left[(s_1 \vee s_2)\left\{\frac{\log(d \vee n)}{n}\right\}^{1/2}\right].$$

Given $0 < \eta \leqslant 1$, *define the* $(1-\eta)$*-confidence interval as* $\mathcal{I} = (\hat{\mu}_1 - z_{1-\eta/2}(\hat{V}/n)^{1/2},\ \hat{\mu}_1 + z_{1-\eta/2}(\hat{V}/n)^{1/2})$, *where* $z_{1-\eta/2}$ *is the* $(1-\eta/2)$ *quantile of the standard normal distribution. Then*

$$\left|\mathrm{pr}(\mu_1^* \in \mathcal{I}) - (1-\eta)\right| = o(1). \tag{12}$$

Indeed, this confidence interval $\mathcal{I}$ is honest in the sense that (12) holds uniformly over all probability distributions that satisfy Assumptions 1–6. In addition, the proof of Corollary 1 holds even if the error ε_1 is heteroskedastic, i.e., $E(\varepsilon_1^2 \mid X)$ depends on the value of X.

In the Supplementary Material we further extend these theoretical results to the estimation of the average treatment effect for the treated.

3.3. *Asymptotic distribution under misspecified propensity score models*

We next investigate the robustness of the proposed estimator to the misspecification of the propensity score model. In this subsection we assume that the true propensity score $\pi^* = \mathrm{pr}(T = 1 \mid X)$ does not belong to the assumed parametric class $\{\pi(X^{\mathrm{T}}\beta) : \beta \in \mathbb{R}^d\}$. To study the limiting behaviour of our estimator $\hat{\mu}_1$ in this setting, we first define the estimand of $\hat{\beta}$ in Step 1. Given the generalized quasilikelihood function $Q_n(\beta)$, the estimand of $\hat{\beta}$ in (5) is defined as

$$\beta^o = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmax}}\, E\left[\int_0^{\beta^{\mathrm{T}}X_i}\left\{\frac{T_i}{\pi(u)} - 1\right\} w_1(u)\, \mathrm{d}u\right].$$

The estimand β^o implicitly depends on the choice of the weight function $w_1(u)$, and when the model is correctly specified, β^o reduces to β^*. We assume that the estimand β^o is sparse, which is a technical assumption required to study misspecified models in high-dimensional settings (Bühlmann & van de Geer, 2015). This assumption ensures that the d-dimensional vector β^o can be well estimated with a fast rate. In practice, this assumption implies that it is sufficient to only adjust for a small number of covariates in order to attain the covariate balance. By choosing the weight $w_1(u) = 1$, β^o satisfies

$$E\left\{\frac{T_i}{\pi(X_i^{\mathrm{T}}\beta^o)} - 1\right\}X_i = 0. \tag{13}$$

Thus, β^o can be viewed as a natural estimand by solving the strong covariate balancing equation (3). This property plays an important role under the misspecified propensity score model. To see this, we rewrite $\hat{\mu}_1 - \mu_1^*$ as

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^o}\{Y_i(1) - \alpha^{*\mathrm{T}}X_i\} + \alpha^{*\mathrm{T}}X_i - \mu_1^*\right] + I_1 + I_2, \tag{14}$$

where $\pi_i^o = \pi(X_i^{\mathrm{T}}\beta^o)$, and the remainder terms I_1 and I_2 are defined as

$$I_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i} - \frac{T_i}{\pi_i^o}\right)\{Y_i(1) - \alpha^{*\mathrm{T}}X_i\}, \qquad I_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i} - 1\right)\alpha^{*\mathrm{T}}X_i.$$

The remainder term I_1 is sufficiently small by exploiting the rate of convergence of the estimated propensity score and the correct specification of the outcome model (2). The remainder term I_2 is identical to the left-hand side of (9). However, in order to rigorously justify (9), we have to require the unbiasedness of the covariate balancing equation (13) when the propensity score model is misspecified. As a side note, if the estimated propensity score satisfies the covariate balancing equation (3) or (4), we immediately have $I_2 = 0$. However, if other estimators of the propensity score model, such as the maximum likelihood estimator, are used, the corresponding estimand is typically different from β^o in (13). By taking the expectation of I_2, we can show that the remainder term I_2 is asymptotically biased. This shows the importance of (13) under misspecified propensity score models. As a comparison, for the augmented inverse probability weighting estimator, the expansion (14) holds with the same I_1 term, and the second remainder term reduces to $I_2' = n^{-1}\sum_{i=1}^{n}(T_i/\tilde{\pi}_i - 1)(\alpha^* - \hat{\alpha})^{\mathrm{T}}X_i$. The method in Belloni et al. (2017) and Farrell (2015) is to estimate β^o by penalized maximum likelihood estimators. However, when the propensity score model is misspecified, the convergence rate of I_2' is slower than root-n; see the Supplementary Material for a detailed analysis.

We now establish the asymptotic properties of $\hat{\mu}_1$ under misspecified propensity score models.

Proposition 1 (Consistency and asymptotic normality under misspecified propensity score models). *Suppose that the outcome model* (2) *is correctly specified, but the propensity score model* (1) *is misspecified. Assumptions 1–6 hold with β^* replaced by β^o. If we take $\lambda \asymp \lambda' \asymp \{\log(d \vee n)/n\}^{1/2}$, then the estimator $\hat{\mu}_1$ with $w_1(u) = 1$ and any weight $w_2(u)$ satisfies*

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^o}\{Y_i(1) - \alpha^{*\mathrm{T}}X_i\} + \alpha^{*\mathrm{T}}X_i - \mu_1^*\right] + O_p\left\{\frac{(s_1 \vee s_2)\log(d \vee n)}{n}\right\},$$

where $\pi_i^o = \pi(X_i^{\mathrm{T}}\beta^o)$. Assume that $E(\varepsilon_1^2 \mid X) \geqslant c$ for some constant $c > 0$ and $E(\alpha^{*\mathrm{T}}X)^4 = O(s_2^2)$. This implies $n^{1/2}(\hat{\mu}_1 - \mu_1^*)/V_{\text{mis-ps}}^{1/2} \to_d N(0,1)$ where

$$V_{\text{mis-ps}} = E\left\{\frac{\pi^*}{(\pi^o)^2}E(\varepsilon_1^2 \mid X) + (\alpha^{*\mathrm{T}}X - \mu_1^*)^2\right\}.$$

If the propensity score model is correctly specified, i.e., $\pi_i^o = \pi_i^*$, then the asymptotic variance $V_{\text{mis-ps}}$ in Proposition 1 reduces to the asymptotic variance V in Theorem 1. To construct the confidence interval for μ_1^*, we need to estimate $V_{\text{mis-ps}}$. By inspecting the proof of Corollary 1, we can show that the estimator $\hat{V}$ defined in (11) is still consistent for $V_{\text{mis-ps}}$, even if the propensity score model is misspecified. Thus, the confidence interval shown in Corollary 1 is valid whether or not the propensity score model is misspecified.

3.4. *Asymptotic distribution under misspecified outcome models*

In this subsection we study the robustness of the proposed estimator to the misspecification of the outcome model. Assume that the propensity score model is correct, but the true conditional mean function $E\{Y_i(1) \mid X_i\}$ is nonlinear in X_i, i.e., there does not exist α^* such that $E\{Y_i(1) \mid X_i\} = \alpha^{*\mathrm{T}}X_i$. Similar to § 3.3, we first define the estimand of $\tilde{\alpha}$ in Step 2 as

$$\alpha^o = \operatorname*{argmin}_{\alpha\in\mathbb{R}^d} E\left\{T_i w_2(\beta^{*\mathrm{T}}X_i)(Y_i - \alpha^{\mathrm{T}}X_i)^2\right\},$$

which in turn depends on the weight function $w_2(\cdot)$. Similarly, we assume α^o is sparse with $s_2 = \|\alpha^o\|_0$. We establish the asymptotic properties of $\hat{\mu}_1$ under misspecified outcome models.

PROPOSITION 2 (CONSISTENCY AND ASYMPTOTIC NORMALITY UNDER MISSPECIFIED OUTCOME MODELS). *Suppose that the propensity score model (1) is correctly specified, but the outcome model (2) is misspecified. Assumptions 1–6 hold with α^* replaced by α^o. If we take $\lambda \asymp \lambda' \asymp \{\log(d \vee n)/n\}^{1/2}$, then the estimator $\hat{\mu}_1$ with any weight functions $w_1(u)$ and $w_2(u) = \pi'(u)/\pi^2(u)$ satisfies*

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^*}\{Y_i(1) - \alpha^{o\mathrm{T}}X_i\} + \alpha^{o\mathrm{T}}X_i - \mu_1^*\right] + O_p\left\{\frac{(s_1 \vee s_2)\log(d \vee n)}{n}\right\}.$$

Assume that $E(\varepsilon_1^{o2} \mid X) \geqslant c$ for some constant $c > 0$ and $E(\alpha^{o\mathrm{T}}X)^4 = O(s_2^2)$, where $\varepsilon_1^o = Y(1) - \alpha^{o\mathrm{T}}X$. This implies that $n^{1/2}(\hat{\mu}_1 - \mu_1^*)/V_{\text{mis-o}}^{1/2} \rightarrow_d N(0,1)$, where

$$V_{\text{mis-o}} = E\left\{\frac{1}{\pi^*}E(\varepsilon^{o2} \mid X) + (\alpha^{o\mathrm{T}}X - \mu_1^*)^2\right\}.$$

The results in this proposition parallel those in Proposition 1. Specifically, when the outcome model is misspecified, if we choose $w_2(u) = \pi'(u)/\pi^2(u)$, or equivalently the propensity score adjusted least square loss $L_n(\alpha)$ in example (c$'$) of § 2.2, the desired properties such as root-n consistency and asymptotic normality are attained. Similarly, the form of $w_2(u)$ is motivated by the remainder term in the asymptotic expansion of $\hat{\mu}_1$ under the misspecified outcome model. Finally, the estimator $\hat{V}$ in (11) is consistent for the asymptotic variance $V_{\text{mis-o}}$.

Remark 3 (*Double robustness and honest confidence intervals*). Propositions 1 and 2 together imply that our estimator $\hat{\mu}_1$ is root-n consistent and asymptotically normal provided either the propensity score model or outcome model is correctly specified. This estimator does not require knowing which of the two models is correct. Since $\hat{V}$ is always consistent, the same confidence interval defined in Corollary 1 is valid as long as one of the two models is correctly specified. Thus, we recommend the use of this estimator and the associated confidence interval in practice. Finally, we comment that when $w_1(u) = 1$ and $w_2(u) = \pi'(u)/\pi^2(u)$, the gradient of $Q_n(\beta)$ and $L_n(\alpha)$ is related to the estimating equations proposed by Robins et al. (2007); see the Supplementary Material for details.

3.5. *Comparison with related work*

In this subsection we compare our method with several related papers. Belloni et al. (2017) and Farrell (2015) showed that the augmented inverse probability weighting estimator is asymptotically normal and efficient in high dimensions when both the propensity score and outcome models are correctly specified. Their assumptions and main results are parallel to those of our Theorem 1. However, the sample boundedness property, see Remark 2, does not hold for the augmented inverse probability weighting estimator in general.

When either the propensity score model or the outcome model is misspecified, Propositions 1 and 2 provide a complete characterization of the asymptotic behaviour of our estimator. In the same context, Farrell (2015) proved that the augmented inverse probability weighting estimator is consistent, but his Theorem 2 does not derive an explicit convergence rate. We show in the Supplementary Material that the augmented inverse probability weighting estimator $\bar{\mu}_1$ satisfies $\bar{\mu}_1 - \mu_1^* = O_p[\{(s_1 \vee s_2)\log(d \vee n)/n\}^{1/2}]$, which is slower than $n^{-1/2}$, and thus the confidence intervals for the treatment effect are not available under model misspecification. In contrast, our estimator is root-n consistent, which leads to honest confidence intervals as shown in § 3.3 and § 3.4. This robustness of the asymptotic distributions to model specification is the main advantage over the augmented inverse probability weighting estimators (Farrell, 2015; Belloni et al., 2017) and the double selection estimator (Belloni et al., 2014).

The approximate residual balancing method proposed by Athey et al. (2018) does not require the propensity score model to be sparse or even well formulated. Thus, their method is robust to the misspecification of the propensity score model so long as the outcome model is correct. In contrast, our method requires both models to be sparse. The advantage of our framework is that it tolerates the misspecified outcome model, so long as the propensity score model is correctly specified. Furthermore, the linearity assumption of the outcome model plays an important role in Athey et al. (2018) and their method is not readily applicable if the outcome model is nonlinear. In contrast, our method is robust to the misspecification of the outcome model and also can be generalized to nonlinear outcome models. As an illustration of its generalizability, we consider the extension of the proposed methodology to generalized linear models in § 4.

In addition, when the propensity score model is correctly specified and is indeed sparse, the estimation of the propensity score can help scientists better understand the treatment assignment mechanism (e.g., Rubin, 2008). Recall that $\tilde{\pi}_i = \pi(\tilde{\beta}^{\mathrm{T}} X_i)$. As a byproduct of Theorem 1, our estimated propensity score is uniformly consistent,

$$\max_{1\leqslant i\leqslant n} |\tilde{\pi}_i - \pi_i^*| = O_p\left\{\frac{(s_1 \vee s_2)\log(d \vee n)(\log n)^{1/2}}{n^{1/2}}\right\}.$$

Thus, the estimated propensity score $\tilde{\pi}_i$ is an accurate approximation to the unknown treatment assignment mechanism. Hirshberg & Wager (2019) recently showed that under suitable regularity conditions the approximate residual balancing method is also semiparametrically efficient and the balancing weights converge to the inverse propensity score but with a slower rate.

Most recently and independently of this paper, Tan (2017, 2018) proposed a penalized calibrated propensity score method and studied its robustness to model misspecification. Our work is closely related to Tan (2017), which can be seen as equivalent to directly plugging the initial estimator $\hat{\beta}$ into the Horvitz–Thompson estimator with $w_1(u) = 1$. However, this method does not balance the covariates as we do in Step 3. Corollary 3 of Tan (2017) implies that the estimator has the rate of convergence $O_p\{(s_1 \log d/n)^{1/2}\}$, which is slower than that of our estimator. In our proof, one can treat $\sum_{i=1}^n \left(T_i/\hat{\pi}_i - 1\right)\alpha^{*\mathrm{T}} X_i$ as the bias of the Horvitz–Thompson estimator, which is eliminated by the covariate balancing step, whereas this term remains in Tan (2017). In the follow-up paper, Tan (2018) removed this bias by constructing an augmented inverse probability weighting estimator. However, our result is more general. First, we show that there exists a large class of estimators that are asymptotically normal under possible model misspecification. Second, our theory holds for generalized linear models, whereas Tan's (2018) method is not applicable if the propensity score model is misspecified.

4. Covariate balancing for generalized linear models

4.1. *Method*

We now extend our method to the setting in which the outcome follows a generalized linear model. The validity of many existing methods critically relies on the assumption that the outcome model is linear in the covariates X_i, or in some transformations, e.g., spline basis, of X_i. Thus, generalizing the covariate balancing approach to nonlinear models is an important extension.

Assume that the working model for $Y_i(1)$ given X_i belongs to the exponential family,

$$p(y \mid X) = h(y, \phi) \exp\left[\frac{1}{a(\phi)}\{y\alpha^{*\mathrm{T}}X - b(\alpha^{*\mathrm{T}}X)\}\right], \tag{15}$$

where $h(\cdot,\cdot)$, $a(\cdot)$ and $b(\cdot)$ are known functions, ϕ is the dispersion parameter and α^* is a d-dimensional vector of unknown regression parameters. For simplicity, we assume that the dispersion parameter ϕ is known. Given this set-up, we propose the following modification of our methodology described in § 2.2.

Step 1. Fit the outcome model via the penalized maximum likelihood method within the treatment group,

$$\hat\alpha = \underset{\alpha\in\mathbb{R}^d}{\operatorname{argmin}}\left[-\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{a(\phi)}\{Y_i\alpha^{\mathrm{T}}X_i - b(\alpha^{\mathrm{T}}X_i)\} + \lambda_0\|\alpha\|_1\right],$$

where $\lambda_0 > 0$ is a tuning parameter.

Step 2. This step is identical to Step 1 in § 2.2, where the weight function $w_1(u)$ is replaced by $w_1(\hat\alpha^{\mathrm{T}}X_i, u)$. This defines the initial estimator $\hat\beta$.

Step 3. Re-estimate the outcome model via the penalized weighted maximum likelihood method within the treatment group,

$$\tilde\alpha = \underset{\alpha\in\mathbb{R}^d}{\operatorname{argmin}}\left[-\frac{1}{n}\sum_{i=1}^{n}\frac{T_i w_2(\hat\beta^{\mathrm{T}}X_i)}{a(\phi)}\{Y_i\alpha^{\mathrm{T}}X_i - b(\alpha^{\mathrm{T}}X_i)\} + \lambda'\|\alpha\|_1\right],$$

where $\lambda' > 0$ is a tuning parameter and $w_2(\cdot)$ is the weight function similar to Step 2 in § 2.2.

Step 4. Define $\tilde S = \{j : |\tilde\alpha_j| > 0\}$ and $f(X) = b''(\tilde\alpha^{\mathrm{T}}X)X_{\tilde S}$. Compute

$$\tilde\gamma = \underset{\gamma\in\mathbb{R}^{|\tilde S|}}{\operatorname{argmin}}\|g_n(\gamma)\|_2^2, \quad \text{where } g_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(\gamma^{\mathrm{T}}\bar X_{i\tilde S} + \hat\beta_{\tilde S^c}^{\mathrm{T}}X_{i\tilde S^c})} - 1\right\}f(X_i).$$

Set $\tilde\beta = (\tilde\gamma, \hat\beta_{\tilde S^c})$ and $\tilde\pi_i = \pi(\tilde\beta^{\mathrm{T}}X_i)$.

Step 5. Estimate μ_1^* by $\hat\mu_1 = n^{-1}\sum_{i=1}^n T_iY_i/\tilde\pi_i - n^{-1}\sum_{i=1}^n(T_i/\tilde\pi_i - 1)b'(\tilde\alpha^{\mathrm{T}}X_i)$.

The current algorithm differs from that in § 2.2 in the following three aspects. First, Step 1 yields an initial estimator of α, which is then incorporated into the weight function $w_1(\hat{\alpha}^{\mathrm{T}}X_i, u)$. As shown later in Theorem 2, the choice of the weight function $w_1(\hat{\alpha}^{\mathrm{T}}X_i, u) = b''(\hat{\alpha}^{\mathrm{T}}X_i)$ leads to a valid asymptotic distribution of $\hat{\mu}_1$ under misspecified propensity score models. Thus, we need this extra step to obtain an initial estimator of α.

Second, Step 4 balances the weighted covariates $f(X) = b''(\tilde{\alpha}^{\mathrm{T}}X)X_{\tilde{S}}$ instead of $X_{i\tilde{S}}$ as done in (7). The reason is that to achieve a similar weak covariate balancing property, one must balance a vector of functions $f(X)$ such that $b'(\alpha^{*\mathrm{T}}X) \in \operatorname{span}\{f(X)\}$, where $\operatorname{span}\{f(X)\}$ represents the linear space generated by the basis functions $f(X)$. Let S denote the support set for α, i.e., $S = \{j : |\alpha_j^*| > 0\}$. Since $b'(\alpha^{*\mathrm{T}}X)$ is unknown in practice, we approximate $b'(\alpha^{*\mathrm{T}}X) = b'(\alpha_S^{*\mathrm{T}}X_S)$ by a local linear estimator $b'(\tilde{\alpha}^{\mathrm{T}}X) + b''(\tilde{\alpha}^{\mathrm{T}}X)(\alpha^* - \tilde{\alpha})_S X_S$. Furthermore, if we replace S by an estimator $\tilde{S}$, this leads to the weighted covariates $f(X) = b''(\tilde{\alpha}^{\mathrm{T}}X)X_{\tilde{S}}$ in Step 4. Unfortunately, balancing $f(X)$ alone does not attain the approximate weak covariate balancing property, because the leading term $b'(\tilde{\alpha}^{\mathrm{T}}X)$ in the local linear approximation has not been considered. It is possible to add $b'(\tilde{\alpha}^{\mathrm{T}}X)$ into the covariate balancing equation, which leads to $f(X) = \{b'(\tilde{\alpha}^{\mathrm{T}}X), b''(\tilde{\alpha}^{\mathrm{T}}X)X_{\tilde{S}}\}$. However, doing so leads to additional technical assumptions on the eigenvalues of $f(X)$. To avoid such assumptions, we only attain partial covariate balancing in Step 4 by choosing $f(X) = b''(\tilde{\alpha}^{\mathrm{T}}X)X_{\tilde{S}}$.

Third, Step 5 applies the augmented inverse probability weighting estimator rather than the Horvitz–Thompson estimator used in § 2.2. The additional term, i.e., $n^{-1}\sum_{i=1}^{n}(T_i/\tilde{\pi}_i - 1)b'(\tilde{\alpha}^{\mathrm{T}}X_i)$, comes from the bias due to the imbalance of $b'(\tilde{\alpha}^{\mathrm{T}}X)$. The augmented inverse probability weighting estimator agrees with the Horvitz–Thompson estimator when the outcome model is linear. In this case we have $b''(u) = 1$ and $b'(u) = u$, and balancing $b''(\tilde{\alpha}^{\mathrm{T}}X)X_{\tilde{S}} = X_{\tilde{S}}$ is sufficient to remove the imbalance effect of $b'(\tilde{\alpha}^{\mathrm{T}}X) = \tilde{\alpha}^{\mathrm{T}}X$. Thus, as expected, the current algorithm reduces to the one in § 2.2 under the linear outcome model. In addition, if the outcome model is the Poisson regression, we can also apply the Horvitz–Thompson estimator because under this model $b''(u) = b'(u) = \exp(u)$ and therefore balancing $b''(\tilde{\alpha}^{\mathrm{T}}X)X_{\tilde{S}}$ is sufficient, provided the intercept term is included.

4.2. *Theoretical results*

Assumption 7 (*Subexponential condition*). Assume that $\varepsilon_1 = Y(1) - b'(\alpha^{*\mathrm{T}}X)$ satisfies $\|\varepsilon_1\|_{\psi_1} \leqslant C$ and $\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant d}|X_{ij}| \leqslant C_n$, where C is a positive constant and we allow C_n to increase with n.

Assumption 8 (*Sparsity*). Let us denote $s_1 = \|\beta^*\|_0$ and $s_2 = \|\alpha^*\|_0$. Assume that $C_n^2(s_1 \vee s_2)\log(d \vee n)/n^{1/2} = o(1)$, where C_n is defined in Assumption 7.

Assumption 9 (*Propensity score, outcome model and weight functions*). Assume that $Q_n(\beta)$ is a concave function. Let C, C' denote positive constants, which may take a different value for each of the conditions below.

(i) The same condition (i) in Assumption 6 holds for the propensity score model $\pi(u)$.

(ii) In the outcome model, $b(u)$ is third-order continuously differentiable and $|X_i^{\mathrm{T}}\alpha^*| \leqslant C'$.

(iii) The weight function $w_1(u, v)$ satisfies the following conditions in a small neighbourhood of $u^* = X_i^{\mathrm{T}}\alpha^*$ and $v^* = X_i^{\mathrm{T}}\beta^*$: $C \leqslant w_1(u, v) \leqslant 1/C$, $0 \leqslant w_1'(u, v) \leqslant 1/C$, and the Lipschitz condition in u, $|w_1(u, v^*) - w_1(u', v^*)| \leqslant C'|u - u'|$, where $u \in [u^* - r, u^* + r]$, $v \in [v^* - r, v^* + r]$ for some small constant $r > 0$ and $w_1'(u, v) = \partial w_1(u, v)/\partial v$.

(iv) The same condition (iii) in Assumption 6 holds for the weight $w_2(u)$.

Unlike the sub-Gaussian condition in Assumption 3, we allow the error ε_1 to be subexponential in Assumption 7. This extension is necessary because many examples of generalized linear models, e.g., exponential regression and Poisson regression, satisfy the subexponential condition but not the sub-Gaussian condition. We also allow C_n to possibly grow with n. Specifically, when X_{ij} is uniformly bounded, C_n is a positive constant. When X_{ij} is sub-Gaussian, then $\max_{1\leqslant i\leqslant n,\, 1\leqslant j\leqslant d} |X_{ij}| = O_p(\{\log(nd)\}^{1/2})$. Assumption 8 requires a similar sparsity condition, and allows C_n to increase with n. For simplicity, we focus on the exact sparse case with no approximation errors in the propensity score and outcome models. Parts (i) and (iv) of Assumption 9 are identical to Assumption 6. Part (ii) is a mild condition, stating that the regression effect in the outcome model is bounded. The third-order differentiability of $b(u)$ holds for most generalized linear models. Part (iii) is a technical condition. To analyse the estimator $\hat{\beta}$ in Step 2, we need to control $w_1(u, v)$ and $w_1'(u, v)$ in a small neighbourhood of the true values. This condition holds for two important examples, $w_1(u, v) = \pi(v)$ and $w_1(u, v) = b''(u)$. The former corresponds to example (a) in § 2.2, and the latter represents the generalization of example (b) to generalized linear models.

To study the performance of our estimator under misspecified models, as in § 3.3 and § 3.4 we define the least false parameters as

$$\beta^o = \operatorname*{argmax}_{\beta\in\mathbb{R}^d} E\left[\int_0^{\beta^{\mathrm{T}}X_i} \left\{\frac{T_i}{\pi(u)} - 1\right\} w_1(X_i^{\mathrm{T}}\alpha^*, u)\, \mathrm{d}u\right],$$

$$\alpha^o = \operatorname*{argmin}_{\alpha\in\mathbb{R}^d} E\left[\frac{T_i w_2(\beta^{*\mathrm{T}}X_i)}{a(\phi)}\{Y_i\alpha^{\mathrm{T}}X_i - b(\alpha^{\mathrm{T}}X_i)\}\right].$$

The following theorem establishes the asymptotic normality of $\hat{\mu}_1$ when the outcome variable follows a generalized linear model. When analysing the theoretical properties under model misspecification, we replace α^* and β^* in all assumptions with α^o and β^o.

THEOREM 2 (ASYMPTOTIC PROPERTIES UNDER GENERALIZED LINEAR MODELS). *Suppose that Assumptions 1, 2, 5, 7, 8 and 9 hold, and that the tuning parameters satisfy $\lambda_0 \asymp \lambda \asymp \lambda' \asymp \{\log(d \vee n)/n\}^{1/2}$.*

(i) Assume that both the propensity score model (1) and the outcome model (15) are correctly specified. Then the estimator $\hat{\mu}_1$ with any weight functions $w_1(u, v)$ and $w_2(u)$ satisfies

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^n \left[\frac{T_i}{\pi_i^*}\{Y_i(1) - b'(\alpha^{*\mathrm{T}}X_i)\} + b'(\alpha^{*\mathrm{T}}X_i) - \mu_1^*\right] + o_p(n^{-1/2}),$$

and $\hat{\mu}_1$ achieves the same semiparametric efficiency bound.

(ii) Assume that the outcome model (15) is correctly specified, but the propensity score model (1) is misspecified. If we choose $w_1(u, v) = b''(u)$, then for any $w_2(u)$ we have

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^n \left[\frac{T_i}{\pi(X_i^{\mathrm{T}}\beta^o)}\{Y_i(1) - b'(\alpha^{*\mathrm{T}}X_i)\} + b'(\alpha^{*\mathrm{T}}X_i) - \mu_1^*\right] + o_p(n^{-1/2}).$$

(iii) Suppose that the propensity score model (1) is correctly specified, but the outcome model (15) is misspecified. If we set $w_2(u) = \pi'(u)/\pi^2(u)$, then for any $w_1(u, v)$ we have

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^*}\{Y_i(1) - b'(\alpha^{o\mathrm{T}}X_i)\} + b'(\alpha^{o\mathrm{T}}X_i) - \mu_1^*\right] + o_p(n^{-1/2}).$$

Part (i) of Theorem 2 is the extension of Theorem 1 to generalized linear models. Under the correct model specification, the asymptotic normality of $\hat{\mu}_1$ holds for any weight functions $w_1(u, v)$ and $w_2(u)$ that satisfy Assumption 9. This result agrees with the theory of augmented inverse probability weighting estimators in Belloni et al. (2017) and Farrell (2015). Unlike the existing work, parts (ii) and (iii) provide novel results on the asymptotic normality of $\hat{\mu}_1$ when either the propensity score model or the outcome model is misspecified. Similar to Propositions 1 and 2, these results hold only if particular forms of $w_1(u, v)$ and $w_2(u)$ are chosen to remove the bias from model misspecification. In particular, we use the weight $w_1(u, v) = b''(u)$ in part (ii), which requires knowledge of α^* in the outcome model. This explains why Step 1 is needed. Since part (ii) holds for any weight function $w_2(u)$, the estimator $\hat{\mu}_1$ remains asymptotically normal even if we skip Step 3 and replace $\tilde{\alpha}$ in Step 4 with $\hat{\alpha}$ in Step 1. Similarly, part (iii) holds for any weight function $w_1(u, v)$. Thus, if we set $w_2(u) = \pi'(u)/\pi^2(u)$ and $w_1(u, v) = \pi(v)$, we may skip Step 1 of our algorithm and the same result in part (iii) still applies.

Similar to Remark 3, when $w_1(u, v) = b''(u)$ and $w_2(u) = \pi'(u)/\pi^2(u)$ the proposed estimator $\hat{\mu}_1$ is asymptotically normal provided that either the propensity score model or the outcome model is correctly specified. This estimator does not require knowing which of the two models is correct, and therefore is recommended for practical use.

5. Simulation studies

In this section we conduct simulation studies to evaluate the finite-sample performance of the proposed methodology. We consider the following data-generating processes. First, we generate the d-dimensional covariate $X_i \sim N(0, \Sigma)$, where $\Sigma_{jk} = \rho^{|j-k|}$ with $\rho = 1/2$. We generate the binary treatment T_i using the logistic regression model of the form $\pi(X_i) = 1 - 1/\{1 + \exp(-X_{i1} + X_{i2}/2 - X_{i3}/4 - X_{i4}/10 - X_{i5}/10 + X_{i6}/10)\}$. For the potential outcomes, we consider both linear and logistic regression models as specified later. The observed outcome is $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$.

The simulation is repeated 200 times under each setting. Throughout the simulation studies, whenever possible we compare our method to the approximate residual balancing method (Athey et al., 2018), the regularized augmented inverse probability weighting method (Farrell, 2015; Belloni et al., 2017) and the double selection estimator (Belloni et al., 2014). For the sake of comparison, we use the lasso penalty in both our method and the regularized augmented inverse probability weighting method, and all tuning parameters are determined by five-fold cross-validation. The weight functions in our method are chosen according to Remark 3. For the approximate residual balancing method we use the default values of the tuning parameters in the R package `balanceHD`. The double selection method is implemented using the R package `hdm`.

We first consider the linear regression models for the potential outcomes,

$$\begin{aligned} Y_i(1) &= 2 + 0.137(X_{i5} + X_{i6} + X_{i7} + X_{i8}) + \varepsilon_{1i}, \\ Y_i(0) &= 1 + 0.291(X_{i5} + X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10}) + \varepsilon_{0i}, \end{aligned}$$

where ε_{1i} and ε_{0i} are independent standard normal random variables. Under this setting, we consider the following four scenarios. In the first scenario, (A), we assume that the propensity score

Table 1. *Bias, standard error, standardized root-mean-squared error, coverage probability of 95% confidence intervals and length of 95% confidence intervals for the estimation of the average treatment effect. Four methods are compared: the proposed method, approximate residual balancing, regularized augmented inverse probability weighting estimator and double selection*

$n = 500$		$d = 1000$				$d = 2000$		
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(A) Both models are correct								
Bias	−0.0026	−0.0017	−0.0498	−0.0910	−0.0595	−0.0580	−0.1200	−0.0397
Std err	0.0936	0.1074	0.0926	0.0979	0.1061	0.1155	0.1011	0.1279
RMSE	0.0936	0.1074	0.1052	0.1337	0.1216	0.1292	0.1569	0.1334
Coverage	0.965	0.930	0.915	0.890	0.910	0.910	0.855	0.945
CI length	0.3867	0.4231	0.3775	0.4294	0.3862	0.4359	0.3731	0.5034
(B) Propensity score model is misspecified								
Bias	−0.0120	−0.0303	−0.1078	−0.0782	−0.0446	−0.0685	−0.1234	−0.0357
Std err	0.0984	0.1153	0.0963	0.1034	0.0924	0.1041	0.0921	0.1214
RMSE	0.0991	0.1193	0.1446	0.1296	0.1025	0.1246	0.1540	0.1265
Coverage	0.965	0.945	0.815	0.905	0.930	0.910	0.740	0.940
CI length	0.3864	0.4431	0.3732	0.4227	0.3839	0.4382	0.3702	0.5023
(C) Outcome model is misspecified								
Bias	−0.0034	−0.0321	−0.0562	−0.0991	−0.0317	−0.0572	−0.1215	−0.0443
Std err	0.0917	0.0982	0.0914	0.1023	0.0944	0.0992	0.0921	0.1026
RMSE	0.0917	0.1033	0.1072	0.1424	0.0995	0.1145	0.1525	0.1118
Coverage	0.960	0.960	0.905	0.845	0.950	0.955	0.770	0.945
CI length	0.3874	0.4292	0.3815	0.4327	0.3890	0.4403	0.3728	0.4261
(D) Both models are misspecified								
Bias	−0.0547	−0.1201	−0.1873	−0.1005	−0.0243	−0.0599	−0.1393	−0.0518
Std err	0.1106	0.1038	0.0903	0.0950	0.0969	0.1060	0.0921	0.0965
RMSE	0.1234	0.1588	0.2079	0.1383	0.0999	0.1218	0.1670	0.1095
Coverage	0.890	0.815	0.775	0.875	0.940	0.940	0.720	0.950
CI length	0.3994	0.4586	0.3790	0.4333	0.3948	0.4545	0.3781	0.4334

HD-CBPS, the proposed method; RB, approximate residual balancing; AIPW, regularized augmented inverse probability weighting estimator; D-SELECT, double selection; Std err, standard error; RMSE, standardized root-mean-squared error; Coverage, coverage probability of 95% confidence intervals; CI length, length of 95% confidence intervals.

and outcome models are correctly specified. In the second scenario, (B) the outcome models are correctly specified but the propensity score model is misspecified. We use the transformed variables $X_{\text{mis}} = \{\exp(X_1/2), X_2/\{1+\exp(X_1)\}+10, (X_1X_3/25+0.6)^3, (X_2+X_4+20)^2, X_6, \exp(X_6+X_7), X_9^2, X_7^3-20, X_9, \ldots, X_d\}$ to generate the treatment but the original variables X to generate the outcome variables. In the third scenario, (C) the propensity score model is correctly specified but the outcome models are misspecified. We use the same transformed variables X_{mis} to generate the outcomes, but the original variables X to generate the treatment. Finally, in scenario (D) both the outcome and propensity score models are misspecified using the transformed covariates. This model misspecification follows the work of Kang & Schafer (2007), who evaluated the empirical performance of the augmented inverse probability weighting estimator in low-dimensional settings.

Table 1 shows the bias, standard error, standardized root-mean-squared error $\{E(\hat{\mu}-\mu)^2\}^{1/2}/\mu$, coverage probability of 95% confidence intervals and their length for the estimation of the average treatment effect under the four scenarios. We focus on the high-dimensional setting with $d = 1000, 2000$ and sample size $n = 500$. Additional simulation studies under a variety of different settings appear in the Supplementary Material. Table 1 shows that the proposed method tends to have a smaller standardized root-mean-squared error in most scenarios. More importantly, as seen in scenarios (B) and (C), the fact that the proposed method has an accurate coverage probability under model misspecification provides empirical support for the robustness property established in Propositions 1 and 2. In contrast, the augmented inverse probability weighting estimator has a significant bias under scenarios (B) and (C). As a result, its coverage probability is too small in most cases. The approximate residual balancing and double selection methods perform reasonably well under model misspecification. But their confidence intervals tend to be wider than those of the proposed method.

We also consider the simulation with logistic outcome models. When the outcome variable is binary, the approximate residual balancing method is not directly applicable. Thus, we only compare our method with the regularized augmented inverse probability weighting and the double selection method. The simulation results illustrate the same conclusion. Due to space limitations, we defer the details to the Supplementary Material.

In summary, the proposed estimator tends to have a smaller mean-squared error, is more robust to model misspecification, and exhibits accurate coverage probability in finite samples. Our results are consistent with the empirical findings of Imai & Ratkovic (2014) and Fan et al. (2016) that covariate balancing tends to outperform the augmented inverse probability weighting estimator in low-dimensional settings. Our simulation studies imply that the same conclusion appears to hold in high-dimensional settings.

6. Empirical illustration

For an empirical illustration we consider a dataset based on the Political Socialization Panel Study, which was originally analysed by Kam & Palmer (2008). One purpose of this study was to understand the effect of higher education on political participation. The dataset consists of 1051 randomly selected high school seniors in the class of 1965. The information about each sample was collected via in-person interviews in the first wave of the study, which we treat as pretreatment covariates. The second wave of the study, conducted in 1973, collected the outcome variable, political participation and the dichotomous treatment variable, college attendance.

For the purpose of comparison, we follow the original study and use 81 pretreatment covariates, including gender, race, club participation and academic performance. Since many of the covariates are categorical variables with more than two levels, we create an indicator variable that represents each level. Therefore, a total of 204 pretreatment variables are used in the propensity score and outcome models. The outcome variable represents an index of adult political participation, which is equal to the sum of eight acts including the turnout in the 1972 presidential election, attending campaign rallies, making a donation to a campaign, and displaying a campaign button and bumper sticker. Since this variable takes an integer value ranging from zero to eight, we use the binomial logistic regression for the outcome model. The propensity score model is assumed to be the logistic regression. We then estimate both the average treatment effect and average treatment effect for the treated college attendance on political participation. The number of treated observations is 675.

We apply seven methods: the proposed methodology, the regularized augmented inverse probability weighting method (Farrell, 2015), the covariate balancing propensity score method (Imai & Ratkovic, 2014), the augmented inverse probability weighting method without regularization (Robins et al., 1994), the inverse probability weighting estimator with regularized logistic regression, the approximate residual balancing method and the double selection method. The estimation procedures for the first two methods are identical to those described in the simulation studies. For the covariate balancing propensity score methodology, it is designed for the linear outcome model, which does not provide an ideal balance of pre-treatment variables. In addition, we use the bootstrap method to approximate the standard error of the estimator based on the covariate balancing propensity score and inverse probability weighting methods.

Table 2. *The estimated average effects of college attendance on political participation. The estimates based on the proposed methodology are compared with those of the covariate balancing propensity score estimator, the regularized augmented inverse probability weighting estimator, the augmented inverse probability weighting estimator without regularization, the inverse probability weighting estimator with the regularized logistic regression, approximate residual balancing and double selection. Standard errors appear in parentheses*

	HD-CBPS	CBPS	AIPW	AIPW-NR	IPW	RB	D-SELECT
Overall (ATE)	0.8293 (0.1247)	1.0163 (0.2380)	0.8796 (0.1043)	0.4904 (0.6009)	1.0666 (0.1588)	0.6706 (0.1643)	0.7859 (0.2682)
Overall (ATT)	0.8439 (0.1420)	1.1232 (0.3094)	0.9790 (0.1617)	0.4761 (1.2614)	1.0023 (0.1272)	0.7348 (0.2237)	1.780 (0.2529)
Whites (ATE)	0.8445 (0.1279)	NA	0.8977 (0.1089)	0.1205 (9.4522)	1.1371 (0.1548)	0.6971 (0.1786)	0.8004 (0.2298)

ATE and ATT, the estimated average effects; HD-CBPS, the estimates based on the proposed methodology; CBPS, the covariate balancing propensity score estimator; AIPW, the regularized augmented inverse probability weighting estimator; AIPW-NR, the augmented inverse probability weighting estimator without regularization; IPW, the inverse probability weighting estimator with the regularized logistic regression; RB, approximate residual balancing; D-SELECT, double selection.

The results are shown in Table 2. All methods except the augmented inverse probability weighting estimator without regularization imply that the overall average treatment effect of college education on political participation is positive and statistically significant. The average treatment effect estimates and their associated standard errors based on the regularized methods are similar to each other. These estimates are, however, smaller than that of the covariate balancing propensity score method. Importantly, our estimator has much smaller standard errors than the covariate balancing propensity score method.

There are at least two reasons for this difference in standard errors. First, as shown in § 4, the proposed methodology uses a different covariate balancing equation than the original covariate balancing propensity score method when the outcome model is nonlinear, achieving the semiparametric efficiency bound. Second, the original covariate balancing propensity score methodology tends to be unstable when balancing a large number of covariates, 204 in this case. Thus, the proposed method improves the existing covariate balancing methods when the outcome model belongs to the class of generalized linear models and the number of covariates is large.

We also apply the same methods to the subsample of whites separately. Among a total of 1051 respondents there are 966 white respondents. In this case, the covariate balancing propensity score method does not converge, and therefore the estimate is unavailable. The results appear in the last row of Table 2. Again, the estimates of the regularized methods are similar, as are the standard errors. However, the augmented inverse probability weighting method without regularization yields a large variance mainly because the maximum likelihood estimate of propensity score tends to be unstable when the number of covariates is large.

7. Discussion

There are several future directions that are worthy of further investigation. First, it is important to extend these high-dimensional causal inference methods to nonbinary treatment regimes, including continuous treatment, individualized treatment rules and dynamic treatment regimes (Imai & Ratkovic, 2015; Fong et al., 2018; Zhao et al., 2019). Second, we plan to further study the effect of tuning parameters on statistical inference. In numerical experiments, the tuning parameters are chosen by cross-validation, which leads to reasonable finite sample results. Based on the sensitivity analysis, the results appear to be stable with respect to a small perturbation of the tuning parameters. However, one challenge is to formally justify the validity of the inference based on the cross-validated estimators. The recent work by Chetverikov et al. (2020) provides some promising results along this line. Further theoretical development is needed to address this important problem. Finally, it is of importance to consider how to relax the sparsity assumptions. One of the most recent works in this area by Bradic et al. (2019) proposes an estimator that is rate/sparsity doubly robust: the estimator remains root-n consistent even if either the propensity score model or the outcome model is nonsparse so long as the other model is sufficiently sparse; see also Smucler et al. (2019). This contrasts with the double robustness of our estimator, which addresses possible model misspecification under the sparsity assumption.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation. The authors are grateful to the editor, an associate editor and three reviewers for helpful comments, and to Jamie Robins for insightful discussions. Imai is also affiliated with the Department of Government at Harvard University.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theoretical results, additional simulation results and further technical details.

References

Athey, S., Imbens, G. W. & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Statist. Soc.* B **80**, 597–623.

Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**, 521–47.

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85**, 233–98.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Studies* **81**, 608–50.

Belloni, A., Chernozhukov, V. & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *J. Bus. Econ. Statist.* **34**, 606–19.

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

Bradic, J., Wager, S. & Zhu, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv:*1905.00744.

Bühlmann, P. & van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Statist.* **9**, 1449–73.

Cai, T. T. & Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45**, 615–46.

Chan, K. C. G., Yam, S. C. P. & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc.* B **78**, 673–700.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Economet. J.* **21**, 733–50.

Chetverikov, D., Liao, Z. & Chernozhukov, V. (2020). On cross-validated lasso. *arXiv:*1605.02214v6.

Dukes, O., Avagyan, V. & Vansteelandt, S. (2019). High-dimensional doubly robust tests for regression parameters. *arXiv:*1805.06714v3.

Fan, J., Imai, K., Liu, H., Ning, Y. & Yang, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical report, Princeton University.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Economet.* **189**, 1–23.

Fong, C., Hazlett, C. & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Ann. Appl. Statist.* **12**, 156–77.

Graham, B. S., de Xavier Pinto, C. C. & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Studies* **79**, 1053–79.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–31.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *Rev. Econ. Statist.* **86**, 73–6.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**, 25–46.

Hirshberg, D. A. & Wager, S. (2019). Augmented minimax linear estimation. *arXiv:*1712.00038v5.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc.* B **76**, 243–63.

Imai, K. & Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *J. Am. Statist. Assoc.* **110**, 1013–23.

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–2909.

Kam, C. D. & Palmer, C. L. (2008). Reconsidering the effects of education on political participation. *J. Polit.* **70**, 612–31.

Kang, J. D. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523–39.

Newey, W. K. & Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv:*1801.09138.

Neykov, M., Ning, Y., Liu, J. S. & Liu, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.* **33**, 427–43.

Ning, Y. & Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high-dimensional models. *Ann. Statist.* **45**, 158–95.

Ning, Y., Zhao, T. & Liu, H. (2017). A likelihood ratio framework for high-dimensional semiparametric regression. *Ann. Statist.* **45**, 2299–327.

R Development Core Team (2020). R:A Language and Environment for Statistical Computing.Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statist. Sci.* **22**, 544–59.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

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. *Statist. Sci.* **5**, 472–80.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Statist.* **2**, 808–40.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H. & Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**, 512.

Smucler, E., Rotnitzky, A. & Robins, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv:*1904.03737v3.

Tan, Z. (2017). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *arXiv:*1710.08074.

Tan, Z. (2018). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv:*1801.09817.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–47.

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–42.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47**, 965–93.

Zhao, Y.-Q., Laber, E. B., Ning, Y., Saha, S. & Sands, B. E. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *J. Mach. Learn. Res.* **20**, 1–23.

Zheng, W. & van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pp. 459–74. New York: Springer.

Zhu, Y. & Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Am. Statist. Assoc.* **113**, 1583–600.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Assoc.* **110**, 910–22.

[*Received on* 19 *December* 2018. *Editorial decision on* 17 *October* 2019]