Outcome model free causal inference with ultra-high dimensional covariates

July 29, 2020

Dingke Tang

Department of Statistical Sciences, University of Toronto

Dehan Kong

Department of Statistical Sciences, University of Toronto

Wenliang Pan

Department of Statistical Science, School of Mathematics, Sun Yat-Sen University

Linbo Wang

Department of Statistical Sciences, University of Toronto

for the Alzheimer's Disease Neuroimaging Initiative [1]

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Abstract

Causal inference has been increasingly reliant on observational studies with rich covariate information. To build tractable causal models, including the propensity score models, it is imperative to first extract important features from high dimensional data. Unlike the familiar task of variable selection for prediction modeling, our feature selection procedure aims to control for confounding while maintaining efficiency in the resulting causal effect estimate. Previous empirical studies imply that one should aim to include all predictors of the outcome, rather than the treatment, in the propensity score model. In this paper, we formalize this intuition through rigorous proofs, and propose the causal ball screening for selecting these variables from modern ultra-high dimensional data sets. A distinctive feature of our proposal is that we do not require any modeling on the outcome regression, thus providing robustness against misspecification of the functional form or violation of smoothness conditions. Our theoretical analyses show that the proposed procedure enjoys a number of oracle properties including model selection consistency, normality and efficiency. Synthetic and real data analyses show that our proposal performs favorably with existing methods in a range of realistic settings.

Keywords: Average causal effect; Ball covariance; Confounder selection; Propensity score modeling

1 Introduction

Modern observational databases hold great promise for drawing causal conclusions. In these studies, both the treatment and outcome of interest are often associated with some baseline covariates, called confounders. Insufficient adjustment for confounders leads to biased causal effect estimates. In their seminal work, Rosenbaum and Rubin (1983) showed that the propensity score, defined as the probability of assignment to a particular treatment conditioning on baseline covariates, can be used to remove bias due to observed confounders. Traditionally, specification of the propensity score model is typically driven by expert knowledge. This is becoming increasingly difficult in modern data applications, where researchers are often presented with high or even ultra-high dimensional covariates and have to decide on a set of variables to include in the propensity score model.

In response to this challenge, recently there has been a growing interest in developing data-driven procedures for covariate selection in propensity score models. A central aim of these methods is to reduce bias and improve efficiency in the final causal effect estimator. This is in sharp contrast to covariate selection in prediction modeling (e.g. Tibshirani, 1996; Fan and Li, 2001; Fan and Lv, 2008), where the goal is to find a sparse representation of the association structure with good prediction accuracy (Witte and Didelez, 2019). In particular, a good prediction model for the propensity score includes all strong predictors of the treatment. However, empirical evidence implies that inclusion of variables associated with treatment but not the outcome may inflate the variance of the resulting causal effect estimates, while inclusion of variables only related to the outcome may improve efficiency (e.g. Austin et al., 2007; Brookhart et al., 2006; Patrick et al., 2011). These empirical results suggest that a good propensity score model for causal effect estimation should instead model the dependence of the treatment on predictors of the outcome, including confounders and other outcome predictors.

Various procedures have been developed for selecting these variables into a propensity score model. A naive approach is prediction modeling for the outcome (e.g. Tibshirani, 1996), while specifying the treatment as a fixed covariate in the model. When used with a small sample size, it may miss confounders that are weakly associated with the outcome but strongly associated with the treatment (Wilson and Reich, 2014). The omission of such variables leads to bias but reduces standard error; in some scenarios, it even reduces the mean squared error (Brookhart et al., 2006). Alternatively, Zigler and Dominici (2014) proposed a Bayesian model averaging approach based on a PS model and an outcome model conditional on the estimated PS and baseline covariates. Shortreed and Ertefaie (2017) proposed the outcome-adaptive lasso, which penalizes the coefficients of a propensity score model inversely proportional to their coefficients in a separate outcome regression model. Ertefaie et al. (2018) proposed a variable selection method using a penalized objective function based on both a linear outcome and a logistic propensity score model. Under a sparse linear outcome model, Antonelli et al. (2019) proposed a Bayesian approach that uses continuous spike-and-slab priors on the regression coefficients corresponding to the confounders. The validity of these methods all relies on correct specification of the outcome regression model. However, one salient benefit of PS-based causal effect estimation is that it does not necessarily require a parametric model for the outcome (e.g. Zigler and Dominici, 2014). If one is willing to specify a parametric outcome model, then the utility of propensity score modeling is undermined, as causal effect estimates based on a correctly specified outcome model is generally more efficient than a PS model-based estimate, or doubly robust methods that combine outcome and PS modeling (e.g. Robins et al., 1994; Chernozhukov et al., 2018). Furthermore, "doubly robust" methods based on PS model selection in these manners are no longer doubly robust, as covariate selection in the PS model hinges on the outcome model being correct. An additional pitfall of existing methods is that none of them is well-suited for covariate selection from an ultra-high dimensional feature set, for which penalization or Bayesian selection methods face challenges in computational cost and estimation accuracy (Fan and Lv, 2008).

In this paper, we make several important contributions to covariate selection in propensity score models. First, building on existing works of Lunceford and Davidian (2004) and Hahn (2004) and through theoretical justifications, we show that the target adjustment set for propensity score modeling indeed includes the confounders and other outcome predictors, thus confirming the empirical findings in the literature (e.g. Austin et al., 2007; Brookhart et al., 2006; Patrick et al., 2011). Along the way, we introduce a novel definition of confounders and show that it satisfies several desirable properties of confounders previously proposed in the literature (VanderWeele and Shpitser, 2013). Second, we propose the Causal Ball Screening (CBS), a novel covariate selection procedure that combines an assumption-free screening step motivated by the ball covariance (Pan et al., 2019, 2020) with a refined selection step. In contrast to existing approaches, our proposal is outcome model-free: it does not require specifications of the outcome regression model, nor does it involve any smoothness assumptions on the outcome regression. Furthermore, to the best of our knowledge, our method is the first in the causal inference literature that applies to ultra-high dimensional settings. Third, we provide novel theoretical guarantees for our procedure. In addition to model selection consistency and asymptotic normality that are standard in the variable selection literature, we introduce a novel concept termed oracle efficiency. We show that our causal effect estimator is as efficient as if one knew the target adjustment set *a priori* and use that to fit a propensity score model and estimate the causal effect. Notably, unlike the case in a prediction problem, the oracle propensity score model for causal inference is different from the "true" propensity score model, yet asymptotically our procedure is still able to achieve the efficiency under the oracle model. Following existing literature, we develop our results under the propensity score weighting framework, which includes popular procedures such as inverse probability weighting and doubly robust estimation (Horvitz and Thompson, 1952; Hájek, 1971; Robins et al., 1994; Chernozhukov et al., 2018).

The rest of the article is organized as follows. Section 2 introduces background on propensity score weighting and the ball covariance. We present a theoretical analysis of the target adjustment set in Section 3, and introduce our outcome model-free causal selection procedure, namely the CBS, in Section 4. Section 5 provides theoretical justifications for the CBS. Simulations studies described in Section 6 evaluate the finite-sample performance of the proposed method. In Section 7, we apply our method to the Alzheimer's Disease Neuroimaging Initiative study and estimate the causal effect of tau protein level in cerebrospinal fluid on Alzheimer's behavioral score while accounting for ultra high dimensional genetic covariates. We finish with a brief discussion in Section 8.

2 Background

2.1 The propensity score

Following the potential outcome framework, we use D to denote a binary treatment assignment, $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)})$ to denote baseline covariates, and $Y(d)$ to denote the outcome that would have been observed under treatment assignment d for $d = 0, 1$. We make the stable unit treatment value assumption (Rubin, 1980) that is standard in the causal inference literature.

Assumption 1. *(Stable unit treatment value assumption): The potential outcomes for any unit do not vary with the treatments assigned to other units; for each unit, there are no different forms or versions of each treatment level, which leads to different potential outcomes.*

Under Assumption 1, the observed outcome Y satisfies $Y = DY(1) + (1 - D)Y(0)$. Suppose we observe n independent samples from the joint distribution of $(\boldsymbol{X}, D, Y)$, denoted as $(\boldsymbol{X}_i, D_i, Y_i), i = 1, \ldots, n$. We are interested in estimating the average causal effect (ACE) $\Delta_0 = E[Y(1)] - E[Y(0)]$. The ACE can be non-parametrically identified under the following assumptions.

Assumption 2. *(Weak Ignorability): There exists $\boldsymbol{X}^{\mathcal{S}}$ such that $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}}$ for $d = 0, 1$, where $S \subset \{1, \ldots, p\}$.*

Assumption 3. *(Positivity)* $0 < P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}}) < 1, a.e..$

Rosenbaum and Rubin (1983) introduced the notion of propensity score $e(\boldsymbol{X}^{\mathcal{S}}) = P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})$ and show that under assumptions 2 and 3, adjusting for the propensity score is sufficient to remove confounding:

$$D \perp\!\!\!\perp Y(d) \mid e(\boldsymbol{X}^{\mathcal{S}}), d = 0, 1.$$

2.2 Propensity score weighting

Propensity scores are commonly used to construct inverse probability weighting (IPW) estimators for estimating the ACE. Let $\hat{e}_i = \hat{e}(\boldsymbol{X}^{\mathcal{S}})$ be the estimated propensity score. Several leading examples of IPW estimators include the Horvitz-Thompson estimator (Horvitz and Thompson, 1952):

$$\hat{\Delta}_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i Y_i}{\hat{e}_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - D_i) Y_i}{1 - \hat{e}_i}, \tag{1}$$

the Ratio estimator (Hájek, 1971):

$$\hat{\Delta}_{Ratio} = \frac{\sum_{i=1}^{n} D_i Y_i / \hat{e}_i}{\sum_{i=1}^{n} D_i / \hat{e}_i} - \frac{\sum_{i=1}^{n} (1 - D_i) Y_i / (1 - \hat{e}_i)}{\sum_{i=1}^{n} (1 - D_i) / (1 - \hat{e}_i)}, \tag{2}$$

and the classical doubly robust estimator (Robins et al., 1994):

$$\hat{\Delta}_{DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i - (D_i - \hat{e}_i)\hat{b}_1(\boldsymbol{X}_i^{\mathcal{S}})}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)Y_i + (D_i - \hat{e}_i)\hat{b}_0(\boldsymbol{X}_i^{\mathcal{S}})}{1-\hat{e}_i}, \tag{3}$$

where $\hat{b}_d(\boldsymbol{X}_i^{\mathcal{S}})$ is an estimate of $E(Y \mid D = d, \boldsymbol{X}_i^{\mathcal{S}})$ obtained via outcome regression.

2.3 The ball covariance

The ball covariance is a generic measure of dependence in Banach space with many desirable properties (Pan et al., 2020). Importantly, it is completely model-free for data in Euclidean spaces, and its empirical version is easy to compute as a test statistic of independence. Compared with other measures of dependence between two random variables such as the mutual information (Cover and Thomas, 2012) and distance correlation (Feuerverger, 1993; Székely et al., 2007), the ball correlation does not assume finite moments and hence provides robustness for data with a heavy-tailed distribution (Pan et al., 2019).

Specifically, let X,Y be two random variables on separable Banach spaces $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \xi)$, respectively, where ρ and ξ are distance functions in the respective spaces. And let θ, μ, ν be probability measures defined by (X, Y), X, Y, respectively. Denote $\overline{B}_\rho(x_1, x_2)$ a closed ball in space $(\mathcal{X}, \rho)$ centering in x_1 with radius $\rho(x_1, x_2)$, and $\overline{B}_\xi(y_1, y_2)$ a closed ball in space $(\mathcal{Y}, \xi)$ centering in y_1 with radius $\xi(y_1, y_2)$.

Definition 1. *The ball covariance is defined as the square root of $BCov^2(X, Y)$, which is an integral of the Hoeffdings dependence measure on the coordinate of radius over poles:*

$$BCov^2(X, Y) = \int [\theta - \mu \otimes \nu]^2 (\overline{B}_\rho(x_1, x_2) \times \overline{B}_\xi(y_1, y_2))$$
$$\theta(dx_1, dy_1)\theta(dx_2, dy_2),$$

where $\mu \otimes \nu$ is a product measure on $\mathcal{X} \times \mathcal{Y}$.

Let $\delta^X_{ij,k} = I(X_k \in \overline{B}_\rho(X_i, X_j))$, where $I(\cdot)$ is an indicator function. Further define $\delta^X_{ij,kl} = \delta^X_{ij,k}\delta^X_{ij,l}$ and $\xi^X_{ij,klst} = (\delta^X_{ij,kl} + \delta^X_{ij,st} - \delta^X_{ij,ks} - \delta^X_{ij,lt})/2$. We can similarly define $\delta^Y_{ij,k}, \delta^Y_{ij,kl}, \delta^Y_{ij,klst}$ for Y.

Proposition 1. *(Separability property) Let (X_i, Y_i), i = 1, 2, ..., 6 be i.i.d samples from the joint distribution of (X, Y). We have:*

$$BCov^2(X, Y) = E(\xi^X_{12,3456}\xi^Y_{12,3456}).$$

Pan et al. (2020) show that $BCov(X, Y) = 0$ if and only if X is independent of Y. Therefore, the ball covariance can be used to perform independent tests.

We now introduce the empirical version of ball covariance.

Definition 2. *(Empirical ball covariance) The empirical ball covariance $BCov_n$(X, Y) is defined as the square root of:*

$$BCov^2_n(X, Y) = \frac{1}{n^6}\sum_{i,j,k,l,s,t=1}^{n} \xi^X_{ij,klst}\xi^Y_{ij,klst}.$$

3 Target Adjustment Set in Propensity Score Modeling

We now study the best set of variables to include in a propensity score model in order to eliminate bias and reduce variance in the resulting causal effect estimates. We first define confounders and related variables under the framework of a causal directed acyclic graph (DAG) (Pearl, 2009). We refer readers to Richardson and Robins (2013) for a discussion of the connection between the potential outcome and DAG frameworks for causality, and VanderWeele and Shpitser (2013) for alternative definitions of confounders.

To begin with, we introduce some definitions we shall use later. A DAG is a finite directed graph with no directed cycles. If there is a directed edge that starts from node X and goes to node

Y, we say X is a parent of Y, and Y is a child of X. A directed path is a path trace out entirely along arrows tail-to-head. If there is a directed path from X to Y, then X is an ancestor of Y and Y is a descendant of X. Next, we define d-separation (Pearl, 2009).

Definition 3. *(d-separation) A path is blocked by a set of nodes Z if and only if*

1. *The path contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (i.e., B is conditioned on), or*
2. *The path contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.*

If Z blocks every path between two nodes X and Y, then X and Y are d-separated conditional on Z, otherwise X and Y are d-connected conditional on Z.

Assumption 4. *(Causal sufficiency) The causal relationships among $(\boldsymbol{X}, D, Y)$ can be represented by a causal DAG.*

Assumption 5. *(Temporal ordering) Every $X_i \in \boldsymbol{X}$ is a non-descendant of D, which is in turn a non-descendant of Y.*

Let $\boldsymbol{X}^{\mathcal{C}} = \{X^{(j)} \in pa(Y) : D$ and $X^{(j)}$ are d-connected given $pa(Y) \setminus \{X^{(j)}\}\}$, where $pa(Y)$ denotes the parents of Y. In the following we define $\boldsymbol{X}^{\mathcal{C}}$ as confounders, $\boldsymbol{X}^{\mathcal{P}} = pa(Y) \setminus \boldsymbol{X}^{\mathcal{C}}$ as precision variables, $\boldsymbol{X}^{\mathcal{I}} = pa(D) \setminus \boldsymbol{X}^{\mathcal{C}}$ as instrumental variables, and $\boldsymbol{X}^{\mathcal{N}} = \boldsymbol{X} \setminus (\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}} \cup \boldsymbol{X}^{\mathcal{N}})$ as null variables. Figurear 1 provides the simplest causal diagram associated with these definitions. The illustration remains valid if one adds arrows among $(\boldsymbol{X}^{\mathcal{C}}, \boldsymbol{X}^{\mathcal{I}}, \boldsymbol{X}^{\mathcal{P}}, \boldsymbol{X}^{\mathcal{N}})$ in Figure 1, as long as the causal sufficiency assumption is satisfied.

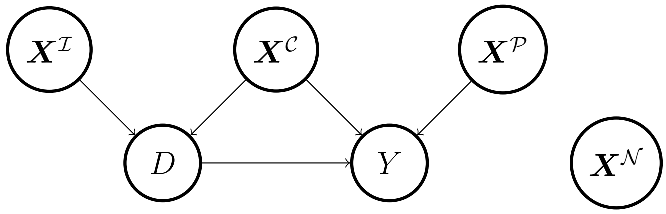

Figure 1: A causal directed acyclic graph illustrating the four types of baseline covariates: confounders $\boldsymbol{X}^{\mathcal{C}}$, precision variables $\boldsymbol{X}^{\mathcal{P}}$, instrumental variables $\boldsymbol{X}^{\mathcal{I}}$ and null variables $\boldsymbol{X}^{\mathcal{N}}$

Remark 1. *The instrumental variable is commonly used to estimate causal effects when the causal sufficiency assumption may be violated. Under the causal sufficiency assumption, our definition here coincides with the definition of instrumental variables in the literature (e.g. Wang and Tchetgen Tchetgen, 2018).*

Proposition 2. *For any $\mathcal{S}$ such that $\boldsymbol{X}^{\mathcal{C}} \subset \boldsymbol{X}^{\mathcal{S}} \subset \boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}} \cup \boldsymbol{X}^{\mathcal{I}}$, assumptions 4 and 5 imply that $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}}, d = 0, 1$.*

Proposition 2 shows that our definition of confounders satisfies Property 1 proposed by VanderWeele and Shpitser (2013). We now introduce the faithfulness assumption that is common in the causal inference literature (e.g. Uhler et al., 2013). Under faithfulness, our definition of confounders also satisfies Property 2A of VanderWeele and Shpitser (2013).

Assumption 6. *(Faithfulness) A distribution P is faithful to a DAG $\mathcal{G}$ if no conditional independence relations other than the ones entailed by the Markov property are present.*

Proposition 3. *Under assumptions 4 – 6, there is no proper subset $\mathcal{C}'$ of $\mathcal{C}$ such that $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{C}'}$. In other words, $\boldsymbol{X}^{\mathcal{C}}$ is a minimally sufficient adjustment set.*

The faithfulness assumption also allows one to show that if one already adjusts for a sufficient adjustment set, then adjusting for precision variables and instrumental variable does not introduce new bias.

Proposition 4. *For any $\mathcal{S}$ such that $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}}, d = 0, 1$, let $\mathcal{S}_P = \mathcal{S} \cup \mathcal{P}', \mathcal{S}_I = \mathcal{S} \cup \mathcal{I}'$, where $\mathcal{P}'$ and $\mathcal{I}'$ is an arbitrary subset of $\mathcal{P}$ and $\mathcal{I}$, respectively. Under assumptions 4 – 6 we have $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_P}, d = 0, 1$ and $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_I}, d = 0, 1$.*

Remark 2. *Adjusting for the null variable may, however, introduce bias. Consider the causal DAG with $D \to Y$ and $D \leftarrow X^{\mathcal{I}} \to X^{collider} \leftarrow X^{\mathcal{P}} \to Y$. Under our definitions, $X^{collider}$ is a null variable and the empty set is a sufficient adjustment set. However, under faithfulness, adjusting for $X^{collider}$ leads to a biased causal effect estimate.*

Previous empirical findings suggest that inclusion of instrumental variables in addition to the confounding variables in the adjustment set may result in efficiency loss (Brookhart et al., 2006; Schisterman et al., 2009; Myers et al., 2011; Patrick et al., 2011) while inclusion of precision variables may bring efficiency gains (Brookhart et al., 2006; Patrick et al., 2011). We now show that when the PS is known, this is indeed the case for the PS weighting estimators discussed in Section 2.2. Together with Proposition 4, these results suggest that in addition to the minimally sufficient adjustment set $\boldsymbol{X}^{\mathcal{C}}$, one should also include the precision variables but not the instrumental variables in the propensity score model.

For any $\mathcal{S}$ such that $\boldsymbol{X}^{\mathcal{C}} \subset \boldsymbol{X}^{\mathcal{S}} \subset \boldsymbol{X}$ and $Y(d) \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{S}}, d = 0, 1$, let $\tilde{\Delta}_{HT}^{\mathcal{S}}$ be the infeasible estimator replacing $\hat{e}_i$ in $\hat{\Delta}_{HT}$ with the true propensity score conditional on $\boldsymbol{X}^{\mathcal{S}}$, namely $P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})$. Let $\mathcal{S}_P = \mathcal{S} \cup \mathcal{P}', \mathcal{S}_I = \mathcal{S} \cup \mathcal{I}'$, where $\mathcal{P}'$ and $\mathcal{I}'$ is an arbitrary subset of $\mathcal{P}$ and $\mathcal{I}$, respectively.

Theorem 1. *(Adjusting for precision variables improves efficiency) Under standard regularity conditions, as $n \to \infty$, we have $\sqrt{n}(\tilde{\Delta}_{HT}^{\mathcal{S}_P} - \Delta_0) \xrightarrow{d} N(0, \sigma^2_{P,HT})$ and $\sqrt{n}(\tilde{\Delta}_{HT}^{\mathcal{S}} - \Delta_0) \xrightarrow{d}$*

$N(0, \sigma^2_{0,HT})$. Furthermore, $\sigma^2_{P,HT} \leq \sigma^2_{0,HT}$. The same conclusion holds for $\tilde{\Delta}_{Ratio}$ and $\tilde{\Delta}_{DR}$.

Theorem 2. *(Adjusting for instrumental variables inflate variance) Under standard regularity conditions, as $n \to \infty$, we have $\sqrt{n}(\tilde{\Delta}^{\mathcal{S}_I}_{HT} - \Delta_0) \xrightarrow{d} N(0, \sigma^2_{I,HT})$, and $\sqrt{n}(\tilde{\Delta}^{\mathcal{S}}_{HT} - \Delta_0) \xrightarrow{d} N(0, \sigma^2_{0,HT})$. Furthermore, $\sigma^2_{I,HT} \geq \sigma^2_{0,HT}$. The same conclusion holds for $\tilde{\Delta}_{Ratio}$ and $\tilde{\Delta}_{DR}$.*

Remark 3. *Lunceford and Davidian (2004, §3.3) proved Theorem 1 in the case where $\mathcal{S} = \mathcal{C}$. Hahn (2004) proved Theorem 2 for $\hat{\Delta}_{DR}$. The remainder of these results appears to be new.*

Remark 4. *In the Supplementary Material we also prove Theorem 1 for the comparison between $\hat{\Delta}^{\mathcal{S}_p}_{HT}$ and $\hat{\Delta}^{\mathcal{S}}_{HT}$. The comparisons between $\hat{\Delta}^{\mathcal{S}_I}_{HT}$ and $\hat{\Delta}^{\mathcal{S}}_{HT}$, and $\hat{\Delta}^{\mathcal{S}_I}_{Ratio}$ and $\hat{\Delta}^{\mathcal{S}}_{Ratio}$, are more challenging since the functional form of the PS model may change as one adds an additional instrumental variable into the PS model.*

4 Causal ball screening

Given theorems 1 and 2, our target adjustment set for propensity score modeling is $\mathcal{A} = \mathcal{C} \cup \mathcal{P}$. In this section, we develop the causal ball screening (CBS), a two-step procedure for covariate selection in a propensity score model. The first step involves a generic sure independence screening procedure to screen out most null and instrumental variables while keeping all the confounding and precision variables. This procedure is based on the conditional ball covariance, a novel concept we introduce based on the ball covariance (Pan et al., 2019, 2020). The second step is a refined selection step that involves a propensity score model but no outcome regression modeling.

4.1 Conditional ball covariance screening

When the candidate feature set is ultra-high dimensional, a common strategy is to use sure independence screening based on marginal (Fan and Lv, 2008) or conditional correlations (Barut et al., 2016). From Figure 1, if the DAG is faithful, then one can read off the following (conditional) independences:

$$\boldsymbol{X}^{\mathcal{C}} \not\perp\!\!\!\perp Y, \qquad \boldsymbol{X}^{\mathcal{P}} \not\perp\!\!\!\perp Y, \qquad \boldsymbol{X}^{\mathcal{I}} \not\perp\!\!\!\perp Y, \qquad \boldsymbol{X}^{\mathcal{N}} \perp\!\!\!\perp Y; \tag{4}$$

$$\boldsymbol{X}^{\mathcal{C}} \not\perp\!\!\!\perp Y \mid D, \;\; \boldsymbol{X}^{\mathcal{P}} \not\perp\!\!\!\perp Y \mid D, \;\; \boldsymbol{X}^{\mathcal{I}} \not\perp\!\!\!\perp Y \mid D, \;\; \boldsymbol{X}^{\mathcal{N}} \perp\!\!\!\perp Y \mid D. \tag{5}$$

On the surface, it seems that independence screening based on (4) or (5) works equally well. In practice, however, note that the dependence between $\boldsymbol{X}^{\mathcal{I}}$ and Y after conditioning on D is induced by collider bias (Pearl, 2009). Previous qualitative analyses (Greenland, 2008; Ding and Miratrix, 2015) and numerical analysis (Liu et al., 2012) show that collider bias tends to be small in many realistic settings. Consequently, we expect that the instrumental variables $\boldsymbol{X}^{\mathcal{I}}$ have weaker dependence with the outcome Y after conditioning on the treatment variable D, and hence are more likely to be screened out.

To perform conditional independence screening based on (5), we first introduce the notion of conditional ball covariance. Let $\omega = P(D = 1)$ be the probability of receiving treatment. Let $X^{(d)}, Y^{(d)}, d = 0, 1$ be random variables such that $(X^{(d)}, Y^{(d)}) \overset{d}{=} (X, Y \mid D = d), d = 0, 1$.

The ball covariance between X and Y given D is defined as the square root of

$$BCov^2(X, Y \mid D) = \omega BCov^2(X^{(1)}, Y^{(1)}) + (1 - \omega) BCov^2(X^{(0)}, Y^{(0)}).$$

Analogously, we can define the sample version of the conditional ball covariance. Let $n_1 = \sum_{i=1}^{n} D_i$ be the number of subject who receive treatment and $n_0 = n - n_1$. Let $\hat{\omega} = n_1/n$ be the empirical estimator of ω. Now we are ready to define the empirical conditional ball covariance.

Definition 4. *The empirical conditional ball covariance $BCov_n(X, Y \mid D)$ is defined as the square root of:*

$$BCov_n^2(X, Y \mid D) = \frac{\hat{\omega}}{n_1^6} \sum_{(i,j,k,l,s,t):D_i,D_j,D_k,D_l,D_s,D_t=1} \xi_{ij,klst}^X \xi_{ij,klst}^Y + \frac{1-\hat{\omega}}{n_0^6} \sum_{(i,j,k,l,s,t):D_i,D_j,D_k,D_l,D_s,D_t=0} \xi_{ij,klst}^X \xi_{ij,klst}^Y.$$

We have the following proposition, which is an extension of Lemma 2.1 in Pan et al. (2019).

Proposition 5. $BCov(X, Y \mid D) = 0 \Leftrightarrow X \perp\!\!\!\perp Y \mid D.$

To perform conditional ball covariance screening, we first calculate the empirical conditional ball covariance between the outcome Y and each baseline covariate $X^{(j)}$, $j = 1, \ldots, p$, and then select baseline covariates into the next step if the corresponding conditional ball covariance is larger than a pre-determined threshold. This is summarized in the first two steps of Algorithm 1.

4.2 Refined selection

We now propose a second refined selection step to further exclude instrumental variables and null variables. Traditionally, in the independence screening literature, refined selection entails a second-step joint penalized regression (e.g. Fan and Lv, 2008) that requires specification of the outcome regression model. When the outcome model cannot be specified, although one may use model-free screening procedures (Li et al., 2012; Pan et al., 2019) for independence screening, refined selection after such a screening step has not been studied before.

To overcome this challenge, we adapt the idea of adaptive lasso (Zou, 2006; Shortreed and Ertefaie, 2017) to our setting. We use $e(\boldsymbol{X})$ and $e(\boldsymbol{X};\boldsymbol{\beta})$ exchangeably to denote the parametric propensity score model. In the following, we use logistic regression as a working model for the propensity score, although all the results can be extended to other parametric models in a straightforward manner. Specifically, we assume

$$\operatorname{logit}\{P(D=1 \mid \boldsymbol{X}^{\mathcal{A}})\} = \operatorname{logit}\{e(\boldsymbol{X}^{\mathcal{A}};\boldsymbol{\beta}^{\mathcal{A}})\} = (\boldsymbol{X}^{\mathcal{A}})^{\mathrm{T}}\boldsymbol{\beta}^{\mathcal{A}},$$

where without loss of generality, we assume the target adjustment set $\mathcal{A} = \mathcal{C} \cup \mathcal{P} = \{1, \ldots, p_0\}$, and $\boldsymbol{\beta}_0^{\mathcal{A}} \in \mathbb{R}^{p_0}$ is the true value for $\boldsymbol{\beta}^{\mathcal{A}}$.

Without loss of generality, we assume the selected set after the screening step $\mathcal{K} = \{1, 2, \ldots, q\}$, and let

$$\hat{\boldsymbol{\beta}}_{\mathcal{K}} = \underset{\boldsymbol{\beta}_{\mathcal{K}}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left[D_i \log\left\{ \frac{1 - e(\boldsymbol{X}_i^{\mathcal{K}};\boldsymbol{\beta}_{\mathcal{K}})}{e(\boldsymbol{X}_i^{\mathcal{K}};\boldsymbol{\beta}_{\mathcal{K}})} \right\} - \log\left\{1 - e(\boldsymbol{X}_i^{\mathcal{K}};\boldsymbol{\beta}_{\mathcal{K}})\right\} \right] + \lambda_n \sum_{j=1}^{q} \frac{1}{\hat{\omega}_j} |\beta_j| \right), \tag{6}$$

where $\hat{\omega}_j$ is a nonparametric estimator of the importance of covariate $X^{(j)}$ in the outcome model and λ_n is a tuning parameter. In practice, $\hat{\omega}_j$ can be obtained based on the inverse of the conditional mutual information (Berrett et al., 2019), the conditional distance correlation (Wang et al., 2015), or the conditional ball covariance introduced in Section 4.1. In the simulations and data analysis, we let $\hat{w}_j = |BCov_n^2(X^{(j)}, Y \mid D)|^{\gamma}$. Following Shortreed and Ertefaie (2017), given a particular value of λ_n, we choose γ so that $\lambda_n n^{\gamma-1} = n^2$.

In our empirical data analyses, we let $\lambda_n = n^a$, where a is a real constant ranging from -15 to 0.49; this range is chosen according to the condition in Theorem 4. We then select the parameter $\lambda_n = n^a$ that minimizes the weighted absolute mean difference (Shortreed and Ertefaie, 2017)

$$wAMD(\lambda_n) = \sum_{j=1}^{q} |\beta_j| \times \left| \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda_n} X_i^{(j)} D_i}{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda_n} D_i} - \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda_n} X_i^{(j)} (1 - D_i)}{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda_n} (1 - D_i)} \right|,$$

where $\hat{\tau}_i^{\lambda_n} = D_i/\hat{e}_i^{\lambda_n} + (1 - D_i)/(1 - \hat{e}_i^{\lambda_n})$ and $\hat{e}_i^{\lambda_n}$ is the estimated propensity score with tuning parameter λ_n. Finally, plug-in estimator $\hat{e}_i = e(\boldsymbol{X}_i^{\mathcal{K}};\hat{\boldsymbol{\beta}}_{\mathcal{K}})$ is used to construct weighting estimators for Δ. These procedures are summarized in Steps 3-5 of Algorithm 1.

Algorithm 1 Causal Ball Screening

1: For $j = 1, \dots, p$, calculate $\hat{\rho}_j = BCov_n^2(X^{(j)}, Y \mid D)$;

2: Select variables such that $\mathcal{K} = \{j : \hat{\rho}_j \geq \tau_n\}$, where τ_n is a preset parameter;

3: Set $\hat{\boldsymbol{\beta}}^{\mathrm{T}} = ((\hat{\boldsymbol{\beta}}^{\mathcal{K}})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})$, where $\hat{\boldsymbol{\beta}}^{\mathcal{K}}$ is obtained via (6);

4: For $i = 1, \dots, n$, calculate $\hat{e}_i = e(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})$;

5: Use (1), (2) and (3) and $\hat{e}_i$ to construct $\hat{\Delta}_{HT}^{CBS}$, $\hat{\Delta}_{Ratio}^{CBS}$, and $\hat{\Delta}_{DR}^{CBS}$ as estimates of Δ_0.

5 Theoretical Properties

In this section, we study the theoretical properties of CBS. We first prove the sure independence screening property, which guarantees that the set selected by the conditional ball covariance screening procedure in Section 4.1 includes all confounders and precision variables with high probability. We first assume the following two conditions.

(A1): There exist constants $c > 0$ and $0 \leq \kappa < 1/2$ such that: $\min_{j \in \boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}}} \rho_j \geq 2cn^{-\kappa}$, where $\rho_j = BCov^2(X^{(j)}, Y \mid D)$;

(A2): $\log(p) = o(n^{1-2\kappa})$.

Condition (A1) specifies the minimum marginal association strength that can be identified by our screening procedure. Condition (A2) allows the dimension of covariates to grow at the exponential order of the sample size.

Theorem 3. *(Sure independence screening property) Under conditions (A1) and (A2), if we let* $\tau_n = cn^{-\kappa}$, *then* $P((\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}}) \subset \mathcal{K}) \to 1$ *as* $n \to \infty$.

We then present theoretical guarantees for the variable selection and estimation step. We first assume that the data-driven weights $\hat{w}_j^{(n)}$ satisfy the following conditions:

(B1): (Convergence inside the target set) For each $j \in \mathcal{A}$, $\hat{w}_j^{(n)} \xrightarrow{p} c_j$, where c_j is a positive constant;

(B2): (Uniform convergence to zero outside of the target set) There exists some constant $s > 0$ such that for all $j \in \mathcal{A}^c$, $\hat{w}_j^{(n)} = O_P(n^{-s})$;

(B3): (Finite moment conditions) For $k = 1, 2$ and $1 \leq j_1, j_2 \leq p$, we have the following moment conditions: $E\{YDX^{(j_1)}/e^k(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\} < \infty$, $E\{YDX^{(j_1)}X^{(j_2)}/e^k(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\} < \infty$, $E[Y(1-D)X^{(j_1)}/\{1-e(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\}^2] < \infty$ and $E[Y(1-D)X^{(j_1)}X^{(j_2)}/\{1-e(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\}^2] < \infty$.

We next introduce some conditions that are useful for studying the properties of our DR estimator.

(B4): (Bounded outcome regression model) For $d = 0, 1$, $E(Y \mid D = d, \boldsymbol{X}) = b_d(\boldsymbol{X}; \boldsymbol{\alpha}_d)$, such that $b_d(\boldsymbol{X}; \boldsymbol{\alpha})$ is continuous with respect to $\boldsymbol{X}$, $\boldsymbol{\alpha}$, and $|b_d(\boldsymbol{X}; \boldsymbol{\alpha})| \leq f_d(\boldsymbol{X})g_d(\boldsymbol{\alpha})$; here f_d, g_d are finite-valued, continuous functions. Furthermore, $\hat{\boldsymbol{\alpha}}_d$ is a consistent estimator for $\boldsymbol{\alpha}_d$ and the plug-in estimator $\hat{b}_d(\boldsymbol{X}_i) = b_d(\boldsymbol{X}_i; \hat{\alpha}_d)$ is used in constructing $\hat{\Delta}_{DR}^{CBS}$;

(B5): (Finite moment conditions) For $1 \leq j_1, j_2 \leq p$, we have the following moment conditions: $E\{f_1(\boldsymbol{X})X^{(j_1)}/e^2(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\} < \infty$, $E\{f_1(\boldsymbol{X})X^{(j_1)}X^{j_2}/e^2(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\} < \infty$, $E[f_0(\boldsymbol{X})X^{(j_1)}/\{1 - e(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\}^2] < \infty$ and $E[f_0(\boldsymbol{X})X^{(j_1)}X^{(j_2)}/\{1 - e(\boldsymbol{X}^{\mathcal{A}}; \boldsymbol{\beta}_0^{\mathcal{A}})\}^2] < \infty$.

Condition (B4) holds under a wide range of parametric models. For example, if the outcome model is linear that $b_d(\boldsymbol{X}; \boldsymbol{\alpha}_d) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}$, then $f_d(\boldsymbol{X}) = \|\boldsymbol{X}\|_1 = \sum_{j=1}^p |X^{(j)}|$, $g_d(\boldsymbol{\alpha}_d) = \|\boldsymbol{\alpha}_d\|_1 = \sum_{j=1}^p |\alpha_j|$ satisfy condition (B4). If the covariates $\boldsymbol{X}$ is bounded so that $|X^{(j)}| \leq M$ for all $1 \leq j \leq p$, then $f_d = 1$, $g_d(\boldsymbol{\alpha}) = \sup_{|X^{(j)}| \leq M, |\pi_j| \leq |\alpha_j|} |b_d(\boldsymbol{X}; \boldsymbol{\pi})|$ satisfy condition (B4).

Theorem 4 summarizes the key theoretical properties of our proposed estimators.

Theorem 4. *Under conditions (B1) and (B2), when $\gamma > 1$ and the regularization parameter λ_n satisfies $\lambda_n/\sqrt{n} \to 0$, $\lambda_n n^{\gamma-1} \to \infty$, we have:*

(a) Consistency in variable selection: $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^{\mathcal{A}^c} = \mathbf{0}) = 1$*;*

(b) Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathcal{A}} - \boldsymbol{\beta}^{\mathcal{A}}) \xrightarrow{d} N(0, I_{11}^{-1})$*, where* $I_{11} = I(\boldsymbol{\beta}_0^{\mathcal{A}})$ *is the Fisher information matrix of* $\boldsymbol{\beta}^{\mathcal{A}}$ *evaluated at* $\beta_0^{\mathcal{A}}$*;*

(c) Oracle efficiency: Let $\hat{\Delta}_{HT}^o$*,* $\hat{\Delta}_{Ratio}^o$ *and* $\hat{\Delta}_{DR}^o$ *be the oracle estimators where one uses the target set* $\mathcal{A}$ *to fit the propensity score model and then plug those estimates into the HT, Ratio or the DR estimator to estimate the average causal effect, respectively.*

(i) *We have* $\sqrt{n}(\hat{\Delta}_{HT}^o - \Delta_0) \xrightarrow{d} N(0, \sigma_{HT}^2)$*. If condition (B3) also holds, then* $\sqrt{n}(\hat{\Delta}_{HT}^{CBS} - \Delta_0) \xrightarrow{d} N(0, \sigma_{HT}^2)$*;*

(ii) *We have* $\sqrt{n}(\hat{\Delta}_{Ratio}^o - \Delta_0) \xrightarrow{d} N(0, \sigma_{Ratio}^2)$*. If condition (B3) also holds, then* $\sqrt{n}(\hat{\Delta}_{Ratio}^{CBS} - \Delta_0) \xrightarrow{d} N(0, \sigma_{Ratio}^2)$*;*

(iii) *We have* $\sqrt{n}(\hat{\Delta}_{DR}^o - \Delta_0) \xrightarrow{d} N(0, \sigma_{DR}^2)$*. If conditions (B3), (B4) and (B5) also hold, then* $\sqrt{n}(\hat{\Delta}_{DR}^{CBS} - \Delta_0) \xrightarrow{d} N(0, \sigma_{DR}^2)$*.*

Remark 5. *Statements (a) and (b) in Theorem 4 are parallel to Theorem 4 in Zou (2006) and Theorem 1 in Shortreed and Ertefaie (2017). The oracle efficiency property in Statement (c) appears to be new.*

Remark 6. *Our CBS method may be combined with other outcome model-based method to deliver doubly robust estimation. In Section 6.1 we illustrate this point by combining our CBS procedure for estimating the propensity score with lasso for estimating the outcome regression, and showing that the resulting estimator is doubly robust in our simulations. Existing methods* *for propensity score model selection, such as the OAL, cannot be used to construct a doubly robust estimator in this fashion as the validity of their propensity score estimates depends on the outcome regression model being correctly specified.*

6 Simulation Studies

In this section, we evaluate the finite-sample performance of the proposed method. We consider four different combinations of sample size n and covariate dimension p: $(n, p) = (200, 100), (200, 1000), (500, 200), (500, 2000)$. We generate D from a Bernoulli distribution with mean 0.5, and $(\boldsymbol{X} \mid D = d)$ from $N(\boldsymbol{\mu}_d, \Sigma)$, where $\Sigma = I_p$, $\boldsymbol{\mu}_1 = (0.4, 0.4, 0, 0, 1, 1, 0, \cdots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, $\boldsymbol{\mu}_0 = \mathbf{0} \in \mathbb{R}^p$. The following Proposition 6 show that in this case, the propensity score model follows a logistic regression model for any marginalization of $\boldsymbol{X}$.

Proposition 6. *Suppose $\boldsymbol{X} \in \mathbb{R}^p$ is a random vector, and $\mathcal{S} \subset \{1, \ldots, p\}$. If $(\boldsymbol{X} \mid D = d) \sim N(\mathbf{u}_d, \boldsymbol{\Sigma})$, then for any $\boldsymbol{X}^{\mathcal{S}} \subset \boldsymbol{X}$, there exist coefficients α_0 and $\boldsymbol{\alpha}$ such that logit$\{P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})\} = \alpha_0 + \boldsymbol{\alpha}^T \boldsymbol{X}^{\mathcal{S}}$.*

In particular, in our simulation setting, logit$\{P(D = 1 \mid \boldsymbol{X})\} = \gamma_0 + \sum_{j=1}^{p} \gamma_j X^{(j)}$, where $\gamma_1 = \gamma_2 = 0.4$, $\gamma_5 = \gamma_6 = 1$, and $\gamma_j = 0, j \notin \{1, 2, 5, 6\}$.

Given D and $\boldsymbol{X}$, we consider four different generating models for the outcome Y:

$$
\begin{aligned}
&\text{Case I (linear model):} && Y = \alpha_1 + \boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{X} + D\eta_1 + \epsilon_1; \\
&\text{Case II (non-linear model):} && Y = \exp(\alpha_2 + D\boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{X} + \epsilon_2); \\
&\text{Case III (non-linear model):} && Y = \exp(\alpha_3 + \boldsymbol{\beta}_3^{\mathrm{T}} \boldsymbol{X} + D\eta_3 + \epsilon_3); \\
&\text{Case IV (non-linear model):} && Y = \alpha_4 + \boldsymbol{\beta}_4^{\mathrm{T}} \boldsymbol{X}^2 + D\eta_4 + \epsilon_4,
\end{aligned}
$$

where $\epsilon_j \sim N(0,1), j = 1,2,3,4$, $\alpha_1 = 0, \alpha_2 = -1, \alpha_3 = -2, \alpha_4 = 0, \boldsymbol{\beta_1} = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = (0.6, 0.6, 0.6, 0.6, 0, \cdots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, $\boldsymbol{\beta}_4 = (1,1,1,1,0,\cdots,0)^{\mathrm{T}} \in \mathbb{R}^p$, and $\eta_1 = \eta_3 = \eta_4 = 2$.

We compare the following methods for estimating the average causal effect:

(i) Regression: We use lasso (Tibshirani, 1996) to select relevant variables $\boldsymbol{X}^{\mathcal{S}}$ from a linear outcome regression model of Y on X and D, with the tunning parameter selected using the Extended Bayesian Information Criteria (Chen and Chen, 2012). This is implemented using `R` code based on the pseudo-code provided in Chen and Chen (2012) and the `R` package glmnet. We then refit a linear outcome regression model on the selected variables $\mathbf{X}^{\mathcal{S}}$: $Y \sim \mathbf{X}^{\mathcal{S}} + D$, and the estimated ACE is the ordinary least squared (OLS) estimate of the coefficient corresponding to D.

(ii) OAL: We use the `R` code documented in Shortreed and Ertefaie (2017) to estimate the propensity score. The ACE is then estimated using the Ratio estimator (2). In cases where $p > n$, as the method in Shortreed and Ertefaie (2017) is not directly applicable, we first apply a sure-independence screening procedure conditioning on D (Barut et al., 2016) and select the top 30 covariates. We then apply the method in Shortreed and Ertefaie (2017) on the selected set.

(iii) CBS: The proposed Algorithm 1, where we select the top 30 covariates in Step 2 of Algorithm 1.

Table 1 reports the average bias in estimating the ACE as well as the mean squared error among 2,000 Monte Carlo runs. When the outcome model is linear, the regression estimator has the smallest MSE in all scenarios. The OAL performs better than the CBS when the sample size is small ($n = 200$) while the CBS performs better when the sample size is moderate ($n = 500$). Note that in this case, the outcome model used by the OAL is correctly specified. As expected,

Table 1: Simulation results based on $2,000$ Monte Carlo runs. We report bias $\times 100$ and MSE $\times 100$ for each estimator. Standard errors of bias and MSE are reported in parentheses. Bold numbers represent the best method in each scenario

(n,p)	Method	Case I (linear)		Case II (non-linear)	
		Bias×100(SE×100)	MSE×100(SE×100)	Bias×100(SE×100)	MSE×100 (SE×100)
(200,100)	Regression	**0.69(0.31)**	**1.95(0.06)**	16.6(2.05)	86.38(12.87)
	OAL	3.6(0.42)	3.62(0.12)	18.82(1.58)	53.19(4.43)
	CBS	6.11(0.44)	4.25(0.15)	**9.76(1.19)**	**29.46(2.39)**
(200,1000)	Regression	**0.07(0.33)**	**2.14(0.07)**	43.96(1.02)	40.3(1.51)
	OAL	6.9(0.39)	3.56(0.11)	22.11(1.56)	53.82(9.86)
	CBS	10.31(0.46)	5.34(0.19)	**15.51(1.23)**	**32.53(4.46)**
(500,200)	Regression	**0.11(0.21)**	**0.89(0.03)**	18.18(0.95)	21.52(1.33)
	OAL	1.11(0.26)	1.36(0.05)	5.76(0.84)	14.27(1.39)
	CBS	1.27(0.24)	1.18(0.04)	**0.7(0.67)**	**8.86(0.52)**
(500,2000)	Regression	**0.22(0.21)**	**0.86(0.03)**	34.17(0.79)	24.09(0.96)
	OAL	3.79(0.23)	1.22(0.04)	12.94(0.93)	18.83(1.79)
	CBS	1.76(0.23)	1.07(0.03)	**4.18(0.68)**	**9.47(0.55)**
(n,p)	Method	Case III (non-linear)		Case IV (non-linear)	
		Bias×100(SE×100)	MSE×100(SE×100)	Bias×100(SE×100)	MSE×100 (SE×100)
(200,100)	Regression	35.62(6.16)	771.14(95.62)	5.2(1.45)	42.58(1.55)
	OAL	55.59(4.35)	409.31(39.52)	7.95(1.35)	37.26(1.27)
	CBS	**16.5(3.12)**	**197.7(15.63)**	**1.23(1.2)**	**28.77(0.93)**
(200,1000)	Regression	163.08(2.95)	439.65(12.2)	34.49(1.28)	44.68(1.46)
	OAL	61.17(4.31)	409.61(72.37)	21.17(1.22)	34.09(1.05)
	CBS	**19.19(2.97)**	**180.31(19.73)**	**6.24(1.19)**	**28.78(0.93)**
(500,200)	Regression	36.7(3.06)	200.08(11.17)	5.17(0.84)	14.24(0.46)
	OAL	20.69(2.37)	116.48(10.22)	5.35(0.86)	15.21(0.5)
	CBS	**0.57(1.82)**	**65.91(4.63)**	**0.53(0.72)**	**10.43(0.32)**
(500,2000)	Regression	118.01(2.44)	258.46(11.74)	17.44(0.79)	15.58(0.51)
	OAL	32.24(2.5)	135.35(12.28)	13.84(0.73)	12.63(0.4)
	CBS	**3.83(1.84)**	**67.47(4.14)**	**2.42(0.72)**	**10.3(0.34)**

when the outcome model is non-linear, CBS significantly outperforms the comparison methods. This confirms the robustness of CBS against outcome model mis-specification.

6.1 Double robustness

We further examine the performance of the doubly robust estimator (3) based on the CBS and OAL. We generate data under Case I, where the true outcome model is linear in D and $\boldsymbol{X}$. We use the method "Regression" described above to obtain $\hat{b}_d(\boldsymbol{X}_i^{\mathcal{S}}), d = 0, 1$, and the OAL or the CBS to obtain $\hat{e}_i$. We also consider settings where the outcome regression or the propensity score model is mis-specified. In these settings, the analyst assumes that the propensity score model is logistic in $\{(X^{(j)} - 0.2)^2; j = 1, \ldots, p\}$, and/or that the outcome model is linear in D and $\{(X^{(j)})^2; j = 1, \ldots, p\}$. We let $(n, p) = (2000, 200)$. Figure 2 compares boxplots of $\hat{\Delta}_{DR}^{CBS}$ and $\hat{\Delta}_{DR}^{OAL}$ in four different scenarios, representing different combinations of correct/incorrect specifications of the outcome regression and propensity score models. One can see that $\hat{\Delta}_{DR}^{OAL}$ is consistent as long as one, but not necessarily more than one of the OR and PS models is correctly specified, and is thus doubly robust. In contrast, $\hat{\Delta}_{DR}^{OAL}$ is only consistent if the outcome regression model is correct.

7 Real data application

In this session, we analyze data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The data use acknowledgement is included in the Supplementary Material. We consider the clinical, genetic, and behavioral measures in the ADNI dataset. The exposure of interest is the tau protein level in cerebrospinal fluid (CSF) observed at Month 12. Tau is a microtubule-associated protein that promotes microtubule polymerization and stabiliza-

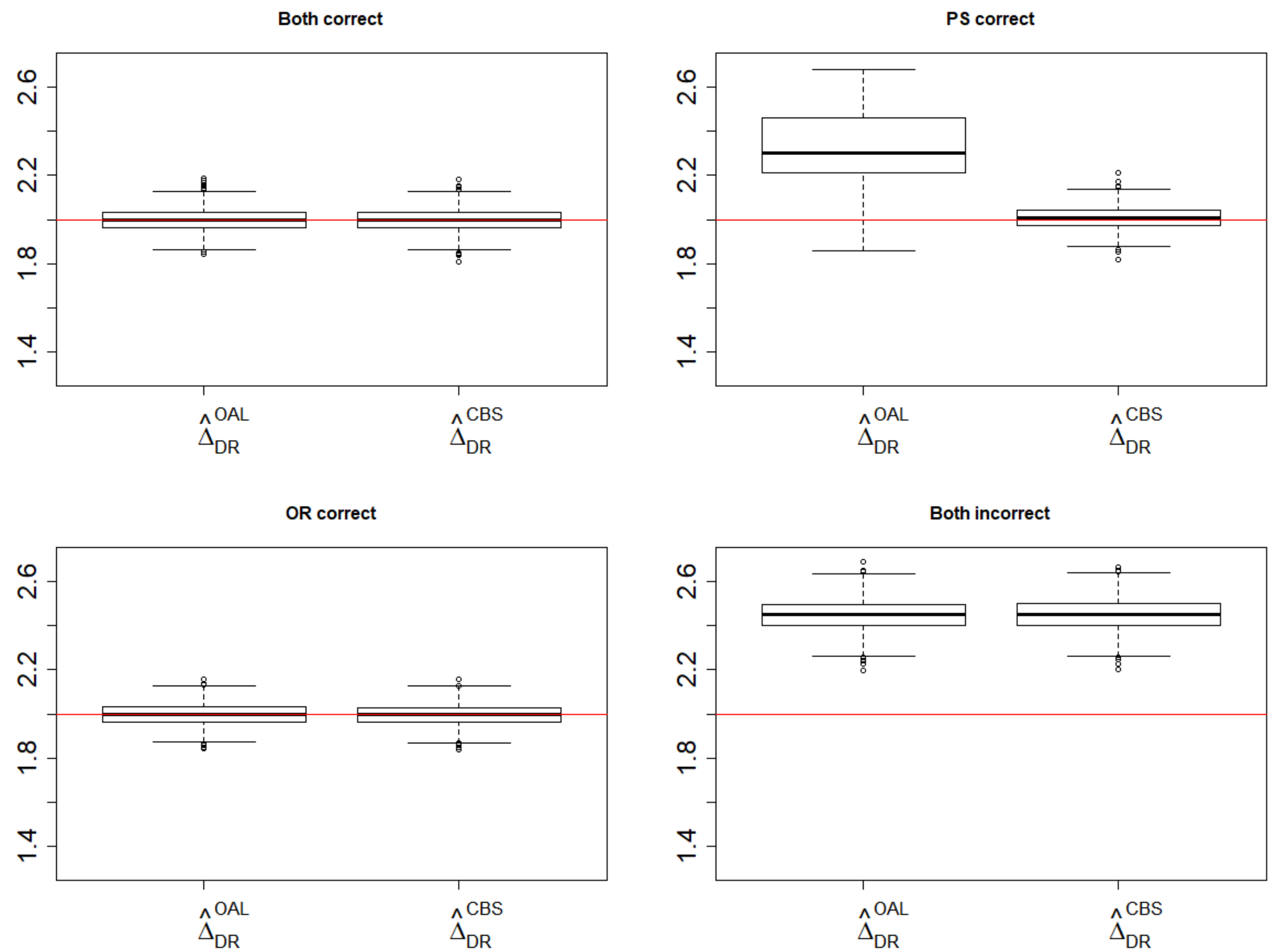

Figure 2: Boxplots of causal effect estimates obtained by $\hat{\Delta}_{DR}^{CBS}$ and $\hat{\Delta}_{DR}^{OAL}$. The red lines correspond to the true causal effect. Results are based on 2,000 Monte Carlo runs

tion (Kametani and Hasegawa, 2018). Studies (Iqbal et al., 2010) have found that tau protein abnormalities initiate the Alzheimer's Disease (AD) cascade and cause neurodegeneration and dementia. Under physiological conditions, tau regulates the assembly and maintenance of the structural stability of microtubules. In the diseased brain, however, tau becomes abnormally hyperphosphorylated, which ultimately causes the microtubules to disassemble, and the free tau molecules aggregate into paired helical filaments (Medeiros et al., 2011). Scientists have found that CSF-tau was markedly increased in Alzheimers disease (Blennow et al., 2015). In this study, we go beyond association and study whether the CSF-tau protein level affects the severity of the Alzheimer's Disease.

We dichotomize the CSF-tau protein level using the cutoff value 350 pg/mL (Tapiola et al., 2009), i.e. $D = 1$ if CSF-tau protein level is over 350 pg/mL, and $D = 0$ otherwise. The severity of the AD is measured by the 11-item Alzheimers Disease Assessment Scale (ADAS-11) cognitive score observed at Month 24, a widely used measure of cognitive behavior ranging from 0 to 70. A higher ADAS-11 score indicates greater severity of AD.

In our analysis, we adjust for clinical covariates including baseline age, gender, and education length, as they have been shown to be the main risk factors for AD (Guerreiro and Bras, 2015; Vina and Lloret, 2010; Zhang et al., 1990). We also consider genetic covariates extracted from whole-genome sequencing data from all of the 22 autosomes. We provide details of how we preprocess the genetic data in the Supplementary Material. After pre-processing, $6,087,205$ bi-allelic markers (including SNPs and indels) were retained in the data analysis.

The data set has 268 subjects with complete information on CSF tau protein data, the Month 24 ADAS-11 score and genetic information. Among these subjects, the average baseline age is 75.4 years old with a standard deviation of 6.80, and the average education length is 15.7 years, with a standard deviation of 3.01; 61.2 percent of them are female.

Table 2: The top 10 SNPs selected by our CBS procedure

Rank	SNP name	Gene	Chromosome number	Related references
1	rs429358	ApoE	19	Cramer et al. (2012)
2	rs56131196	ApoC1	19	Guerreiro et al. (2012)
3	rs4420638	ApoC1	19	Guerreiro et al. (2012)
4	rs12721051	ApoC1	19	Gao et al. (2016); Yashin et al. (2018)
5	rs769449	ApoE	19	Cruchaga et al. (2013)
6	rs10414043	ApoC1	19	Zhou et al. (2014)
7	rs7256200	ApoC1	19	Takei et al. (2009)
8	rs73052335	ApoC1	19	Zhou et al. (2018, 2014)
9	rs111789331	ApoC1	19	Rajabli et al. (2018)
10	rs6857	NECTIN2	19	Kamboh et al. (2012)

We use $\boldsymbol{Z}$ to denotes the clinical covariates including age, gender and education length, and fit a linear regression $Y \sim \boldsymbol{Z}$ to adjust for these clinical covariates. We then apply our CBS procedure using the fitted residuals from the linear regression as the outcome and select the top 30 genetic covariates. The top 10 SNPs, listed in Table 2, are all located on Chromosome 19 and have previously been found to be strongly associated with Alzheimer's.

Since some of the SNPs selected through our first step screening are perfectly correlated, we only keep one SNP among a cluster of SNPs that are perfectly correlated with each other. We further apply our refined selection procedure in Section 4.2 on these selected covariates and three clinical covariates age, gender and education length, where the coefficients corresponding to the clinical covariates are not penalized. Finally, we use the ratio estimator (2) to estimate the average causal effect of CSF-tau protein level on ADAS-11 score. Analysis results suggest that on average, one pg/mL increase in CSF-tau protein level will lead to 5.66 points increase in ADAS-11 score. In other words, increase in CSF-tau protein level is harmful for cognition.

8 Discussion

In this paper we propose a novel variable selection procedure for propensity score modeling, called causal ball screening. In contrast to previous approaches, validity of our procedure does not depend on any assumptions on the outcome regression model. It can moreover be applied to select variables important for causal inference from millions of baseline covariates, as we illustrate in the real data analysis.

To support our methodological developments, we provide a formal theory characterizing the asymptotic behaviors of our proposed estimator. Theorem 4 in this paper is conditional on the set obtained via the screening step. It would be an interesting topic for future research to establish post selection inference for our estimator while accounting for the randomness in the first-step screening procedure. This is a challenging problem, and one possible direction is to adapt state-of-art post selection methods (e.g. Van de Geer et al., 2014; Lee et al., 2016)

We have so far considered causal effect estimation using three classical propensity score weighting estimators, including the Horvitz-Thompson estimator, the ratio estimator and the classical doubly robust estimator. Our developments on propensity score model selection and estimation can also be combined with other techniques for estimating the propensity score weights such as the covariate balancing propensity score (Imai and Ratkovic, 2014) and the subclassification weights (Wang et al., 2016).

References

Antonelli, J., Parmigiani, G., Dominici, F., et al. (2019), "High-dimensional confounding adjustment using continuous spike and slab priors," *Bayesian Analysis*, 14, 825–848.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007), "A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study," *Statistics in Medicine*, 26, 734–753.

Barut, E., Fan, J., and Verhasselt, A. (2016), "Conditional sure independence screening," *Journal of the American Statistical Association*, 111, 1266–1277.

Berrett, T. B., Samworth, R. J., Yuan, M., et al. (2019), "Efficient multivariate entropy estimation via k-nearest neighbour distances," *The Annals of Statistics*, 47, 288–318.

Blennow, K., Dubois, B., Fagan, A. M., Lewczuk, P., de Leon, M. J., and Hampel, H. (2015), "Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease," *Alzheimer's & Dementia*, 11, 58–69.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006), "Variable selection for propensity score models," *American Journal of Epidemiology*, 163, 1149–1156.

Chen, J. and Chen, Z. (2012), "Extended BIC for small-n-large-P sparse GLM," *Statistica Sinica*, 555–574.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Consortium, . G. P. et al. (2012), “An integrated map of genetic variation from 1,092 human genomes,” *Nature*, 491, 56–65.

Cover, T. M. and Thomas, J. A. (2012), *Elements of information theory*, John Wiley & Sons.

Cramer, P. E., Cirrito, J. R., Wesson, D. W., Lee, C. D., Karlo, J. C., Zinn, A. E., Casali, B. T., Restivo, J. L., Goebel, W. D., James, M. J., et al. (2012), “ApoE-directed therapeutics rapidly clear β-amyloid and reverse deficits in AD mouse models,” *Science*, 335, 1503–1506.

Cruchaga, C., Kauwe, J. S., Harari, O., Jin, S. C., Cai, Y., Karch, C. M., Benitez, B. A., Jeng, A. T., Skorupa, T., Carrell, D., et al. (2013), “GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer’s disease,” *Neuron*, 78, 256–268.

Ding, P. and Miratrix, L. W. (2015), “To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias,” *Journal of Causal Inference*, 3, 41–57.

Ertefaie, A., Asgharian, M., and Stephens, D. A. (2018), “Variable selection in causal inference using a simultaneous penalization method,” *Journal of Causal Inference*, 6, 20170010.

Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2008), “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Feuerverger, A. (1993), “A consistent test for bivariate dependence,” *International Statistical Review/Revue Internationale de Statistique*, 419–433.

Gao, L., Cui, Z., Shen, L., and Ji, H.-F. (2016), "Shared genetic etiology between type 2 diabetes and Alzheimer's disease identified by bioinformatics analysis," *Journal of Alzheimer's Disease*, 50, 13–17.

Greenland, S. (2008), "Invited commentary: variable selection versus shrinkage in the control of multiple confounders," *American Journal of Epidemiology*, 167, 523–529.

Guerreiro, R. and Bras, J. (2015), "The age factor in Alzheimers disease," *Genome Medicine*, 7, 106.

Guerreiro, R. J., Gustafson, D. R., and Hardy, J. (2012), "The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE," *Neurobiology of Aging*, 33, 437–456.

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

Hájek, J. (1971), "Comment on An essay on the logical foundations of survey sampling by Basu, D," *Foundations of Statistical Inference*.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

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

Iqbal, K., Liu, F., Gong, C.-X., and Grundke-Iqbal, I. (2010), "Tau in Alzheimer disease and related tauopathies," *Current Alzheimer Research*, 7, 656–664.

Kamboh, M. I., Barmada, M. M., Demirci, F. Y., Minster, R. L., Carrasquillo, M. M., Pankratz, V. S., Younkin, S. G., Saykin, A. J., Sweet, R. A., Feingold, E., et al. (2012), "Genome-wide association analysis of age-at-onset in Alzheimer's disease," *Molecular Psychiatry*, 17, 1340–1346.

Kametani, F. and Hasegawa, M. (2018), "Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease," *Frontiers in Neuroscience*, 12, 25.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016), "Exact post-selection inference, with application to the lasso," *The Annals of Statistics*, 44, 907–927.

Li, R., Zhong, W., and Zhu, L. (2012), "Feature screening via distance correlation learning," *Journal of the American Statistical Association*, 107, 1129–1139.

Liu, E., Li, M., Wang, W., and Li, Y. (2013), "MaCH-Admix: Genotype Imputation for Admixed Populations," *Genetic Epidemiology*, 37, 25–37.

Liu, W., Brookhart, M. A., Schneeweiss, S., Mi, X., and Setoguchi, S. (2012), "Implications of M bias in epidemiologic studies: a simulation study," *American Journal of Epidemiology*, 176, 938–948.

Lunceford, J. K. and Davidian, M. (2004), "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study," *Statistics in Medicine*, 23, 2937–2960.

Medeiros, R., Baglietto-Vargas, D., and LaFerla, F. M. (2011), "The role of tau in Alzheimer's disease and related disorders," *CNS Neuroscience & Therapeutics*, 17, 514–524.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011), "Effects of adjusting for instrumental variables on bias and precision of effect estimates," *American Journal of Epidemiology*, 174, 1213–1222.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2019), "A generic sure independence screening procedure," *Journal of the American Statistical Association*, 114, 928–937.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2020), "Ball covariance: A generic measure of dependence in banach space," *Journal of the American Statistical Association*, 115, 307–317.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., and Stürmer, T. (2011), "The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration," *Pharmacoepidemiology and Drug Safety*, 20, 551–559.

Pearl, J. (2009), *Causality*, Cambridge university press.

Pearl, J. and Paz, A. (1985), *Graphoids: A graph-based logic for reasoning about relevance relations*, University of California (Los Angeles). Computer Science Department.

Rajabli, F., Feliciano, B. E., Celis, K., Hamilton-Nelson, K. L., Whitehead, P. L., Adams, L. D., Bussies, P. L., Manrique, C. P., Rodriguez, A., Rodriguez, V., et al. (2018), "Ancestral origin of ApoE $\varepsilon 4$ Alzheimer's disease risk in Puerto Rican and African American populations," *PLoS Genetics*, 14, e1007791.

Richardson, T. S. and Robins, J. M. (2013), "Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality," *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper, 128*.

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.

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.

Rubin, D. B. (1980), "Comment," *Journal of the American Statistical Association*, 75, 591–593.

Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009), "Overadjustment bias and unnecessary adjustment in epidemiologic studies," *Epidemiology (Cambridge, Mass.)*, 20, 488.

Shortreed, S. M. and Ertefaie, A. (2017), "Outcome-adaptive lasso: Variable selection for causal inference," *Biometrics*, 73, 1111–1122.

Stefanski, L. A. and Boos, D. D. (2002), "The calculus of M-estimation," *The American Statistician*, 56, 29–38.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007), "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, 35, 2769–2794.

Takei, N., Miyashita, A., Tsukie, T., Arai, H., Asada, T., Imagawa, M., Shoji, M., Higuchi, S., Urakami, K., Kimura, H., et al. (2009), "Genetic association study on in and around the APOE in late-onset Alzheimer's disease in Japanese," *Genomics*, 93, 441–448.

Tapiola, T., Alafuzoff, I., Herukka, S.-K., Parkkinen, L., Hartikainen, P., Soininen, H., and Pirttilä, T. (2009), "Cerebrospinal fluid β-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain," *Archives of Neurology*, 66, 382–389.

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

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013), "Geometry of the faithfulness assumption in causal inference," *The Annals of Statistics*, 436–463.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014), "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, 42, 1166–1202.

VanderWeele, T. J. and Shpitser, I. (2013), "On the definition of a confounder," *Annals of statistics*, 41, 196–220.

Vina, J. and Lloret, A. (2010), "Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-β peptide," *Journal of Alzheimer's Disease*, 20, S527–S533.

Wang, L. and Tchetgen Tchetgen, E. (2018), "Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 531–550.

Wang, L., Zhang, Y., Richardson, T., and Zhou, X.-H. (2016), "Robust Estimation of Propensity Score Weights via Subclassification," *arXiv preprint arXiv:1602.06366*.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015), "Conditional distance correlation," *Journal of the American Statistical Association*, 110, 1726–1734.

Wilson, A. and Reich, B. J. (2014), "Confounder selection via penalized credible regions," *Biometrics*, 70, 852–861.

Witte, J. and Didelez, V. (2019), "Covariate selection strategies for causal inference: Classification and comparison," *Biometrical Journal*, 61, 1270–1289.

Yashin, A. I., Fang, F., Kovtun, M., Wu, D., Duan, M., Arbeev, K., Akushevich, I., Kulminski, A., Culminskaya, I., Zhbannikov, I., et al. (2018), "Hidden heterogeneity in Alzheimer's disease:

insights from genetic association studies and other analyses," *Experimental Gerontology*, 107, 148–160.

Zhang, M., Katzman, R., Salmon, D., Jin, H., Cai, G., Wang, Z., Qu, G., Grant, I., Yu, E., Levy, P., et al. (1990), "The prevalence of dementia and Alzheimer's disease in Shanghai, China: impact of age, gender, and education," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 27, 428–437.

Zhou, Q., Zhao, F., Lv, Z.-p., Zheng, C.-g., Zheng, W.-d., Sun, L., Wang, N.-n., Pang, S., de Andrade, F. M., Fu, M., He, X.-h., Hui, J., Jiang, W.-y., Yang, C.-y., Shi, X.-h., Zhu, X.-q., Pang, G.-f., Yang, Y.-g., Xie, H.-q., Zhang, W.-d., Hu, C.-y., and Yang, Z. (2014), "Association between APOC1 Polymorphism and Alzheimers Disease: A Case-Control Study and Meta-Analysis," *PloS One*, 9, e87017.

Zhou, X., Chen, Y., Mok, K. Y., Zhao, Q., Chen, K., Chen, Y., Hardy, J., Li, Y., Fu, A. K., Guo, Q., et al. (2018), "Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis," *Proceedings of the National Academy of Sciences*, 115, 1697–1706.

Zigler, C. M. and Dominici, F. (2014), "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects," *Journal of the American Statistical Association*, 109, 95–107.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Supplementary Material for "Outcome model free causal inference with ultra-high dimensional covariates"

Abstract

The Supplementary File is organized as follows. Section 1 gives the proofs of Propositions 2–6. The proofs of Theorems 1–4 are given in Section 2. Section 3 contains additional information in the real data application including data usage acknowledgement and details in preprocessing the genetics data.

1 Proofs of Propositions

1.1 Proof of Proposition 2

We first introduce some graphoid axioms (Pearl and Paz, 1985) we will use later:

$$\text{Intersection:} \quad D \perp\!\!\!\perp Y \mid W, Z; D \perp\!\!\!\perp W \mid Y, Z \Rightarrow D \perp\!\!\!\perp Y, W \mid Z, \tag{S1}$$

$$\text{Contraction:} \quad D \perp\!\!\!\perp Y \mid Z; D \perp\!\!\!\perp W \mid Y, Z \Rightarrow D \perp\!\!\!\perp Y, W \mid Z, \tag{S2}$$

$$\text{Weak union:} \quad D \perp\!\!\!\perp X \cup Y \mid Z \Rightarrow D \perp\!\!\!\perp X \mid Z \cup Y, \tag{S3}$$

$$\text{Decomposition:} \quad D \perp\!\!\!\perp X \cup Y \mid Z \Rightarrow D \perp\!\!\!\perp X \mid Z. \tag{S4}$$

We first show that any superset of $pa(Y)$ is sufficient to adjust for confounding:

$$D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{M}}, \tag{S5}$$

where $pa(Y) \subseteq \boldsymbol{X}^{\mathcal{M}}$. We show this by contradiction. Assume D and $Y(d)$ are d-connected given $\boldsymbol{X}^{\mathcal{M}}$. Due to Assumption 2, there is no direct edge between D and $Y(d)$. Furthermore, D

and $Y(d)$ are not ancestral to each other due to Assumptions 4 and 5. Then any path connecting $Y(d)$ and D must be one of the following:

- $Y(d) \leftarrow Q \cdots D$, where Q is a parent of $Y(d)$. Since $Q \in pa(Y) \subset \boldsymbol{X}^{\mathcal{M}}$, this path is blocked by $\boldsymbol{X}^{\mathcal{M}}$;

- $Y(d) \rightarrow Q \cdots D$. This is impossible since $X^{(j)'}s$ are non-descendants of $Y(d)$.

We now show that a precision variable is independent of the treatment conditional on confounders (and other precision variables):

$$D \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}, \tag{S6}$$

where $\boldsymbol{X}^{\tilde{\mathcal{P}}} = pa(Y) \setminus \boldsymbol{X}^{\mathcal{S}}$.

To see this, note that if $j \in \tilde{\mathcal{P}} \subset \mathcal{P}$, by the definition of $\mathcal{P}$ we have D and $X^{(j)}$ are d-separated given $pa(Y) \setminus X^{(j)}$, which implies $D \perp\!\!\!\perp X^{(j)} \mid \{pa(Y) \setminus X^{(j)}\}$. Without loss of generality, assume $\tilde{\mathcal{P}} = \{1, 2, 3, \ldots, d_0\}$. We then have

$$D \perp\!\!\!\perp X^{(1)} \mid [X^{(2)} \cup \{pa(Y) \setminus X^{(1,2)}\}],$$
$$D \perp\!\!\!\perp X^{(2)} \mid [X^{(1)} \cup \{pa(Y) \setminus X^{(1,2)}\}].$$

By the intersection property (S1), we have $D \perp\!\!\!\perp X^{(1,2)} \mid \{pa(Y) \setminus X^{(1,2)}\}$. Repeat this process $d_0 - 1$ time, we then have $D \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}$.

Combining (S5) and (S6), by the contraction property (S2), we can show that adjusting for all the confounders and any precision variables are sufficient to control for confounding:

$$D \perp\!\!\!\perp Y(d) \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}.$$

We now show that an instrument variable set is d-separated from, and hence independent of a precision variable conditional on confounders and other precision variables:

$$\boldsymbol{X}^{\tilde{\mathcal{I}}} \perp\!\!\!\perp X^{(j)} \mid \{pa(Y) \setminus X^{(j)}\},$$

where $\tilde{\mathcal{I}} \subset \mathcal{I}$, $j \in \mathcal{P}$.

We again show by contradiction. Assume there exists $X^{(j)} \in \boldsymbol{X}^{\mathcal{P}}$ such that $X^{(j)}$ and $\boldsymbol{X}^{\tilde{\mathcal{I}}}$ are d-connected given $pa(Y) \setminus X^{(j)}$. By definition $\boldsymbol{X}^{\mathcal{I}} \subset pa(D)$, there is a path $D \leftarrow \boldsymbol{X}^{\tilde{\mathcal{I}}}$. Then D and $X^{(j)}$ are d-connected given $pa(Y) \setminus X^{(j)}$, which is a contradiction to the definition of $\mathcal{P}$.

We now show that a set of instruments is independent of any subset of precision variables conditional on confounders and other precision variables:

$$\boldsymbol{X}^{\tilde{\mathcal{I}}} \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}, \tag{S7}$$

where $\tilde{\mathcal{I}} \subset \mathcal{I}$. Again, without loss of generality, we assume $\tilde{\mathcal{P}} = \{1, 2, 3, \ldots, d_0\}$. We then have:

$$\boldsymbol{X}^{\tilde{\mathcal{I}}} \perp\!\!\!\perp X^{(1)} \mid [X^{(2)} \cup \{pa(Y) \setminus X^{(1,2)}\}],$$
$$\boldsymbol{X}^{\tilde{\mathcal{I}}} \perp\!\!\!\perp X^{(2)} \mid [X^{(1)} \cup \{pa(Y) \setminus X^{(1,2)}\}].$$

By the intersection property (S1), we have $\boldsymbol{X}^{\tilde{\mathcal{I}}} \perp\!\!\!\perp X^{(1,2)} \mid \{pa(Y) \setminus X^{(1,2)}\}$. Repeat this process $d_0 - 1$ time, we have $\boldsymbol{X}^{\tilde{\mathcal{I}}} \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}$.

Finally, we show

$$D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}}. \tag{S8}$$

Using the same argument as in our proof of (S5), we can show that

$$D \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid [\boldsymbol{X}^{\mathcal{I}} \cup \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}]. \tag{S9}$$

This relationship holds as $pa(D) \subset \boldsymbol{X}^{\mathcal{I}} \cup \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}$, which means $\boldsymbol{X}^{\mathcal{I}} \cup \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}$ is a superset of $pa(D)$. By letting $\tilde{\mathcal{I}} = \mathcal{I}$ in (S7), combining (S7), (S9) and the contraction property (S2), we have

$$(\boldsymbol{X}^{\mathcal{I}} \cup D) \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}. \tag{S10}$$

Combining (S10) and the decomposition property (S4), for $\tilde{\mathcal{I}} \subset \mathcal{I}$ we have

$$(\boldsymbol{X}^{\tilde{\mathcal{I}}} \cup D) \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\}.$$

Using the weak union property (S3), we have the following result:

$$D \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid [\{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\} \cup \boldsymbol{X}^{\tilde{\mathcal{I}}}].$$

We set $\tilde{\mathcal{I}} = \mathcal{S} \cap \mathcal{I} \subset \mathcal{I}$. We note that $\{pa(Y) \setminus \boldsymbol{X}^{\tilde{\mathcal{P}}}\} \cup (\boldsymbol{X}^{\mathcal{S}} \cap \boldsymbol{X}^{\mathcal{I}}) = [pa(Y) \cap \{pa(Y) \cap (\boldsymbol{X}^{\mathcal{S}})^c\}^c] \cup (\boldsymbol{X}^{\mathcal{S}} \cap \boldsymbol{X}^{\mathcal{I}}) = \{pa(Y) \cap \boldsymbol{X}^{\mathcal{S}}\} \cup (\boldsymbol{X}^{\mathcal{S}} \cap \boldsymbol{X}^{\mathcal{I}}) = \boldsymbol{X}^{\mathcal{S}}$. The last equality holds because $\mathcal{S} \subset \mathcal{P} \cup \mathcal{I} \cup \mathcal{C}$. So we have

$$D \perp\!\!\!\perp \boldsymbol{X}^{\tilde{\mathcal{P}}} \mid \boldsymbol{X}^{\mathcal{S}}. \tag{S11}$$

We let $\mathcal{M} = pa(Y) \cup \mathcal{S}$ in equation (S5). We note that $\boldsymbol{X}^{\mathcal{S}} \subset \boldsymbol{X}^{\mathcal{M}}$ and $\boldsymbol{X}^{\mathcal{S}} \cup \boldsymbol{X}^{\tilde{\mathcal{P}}} = \boldsymbol{X}^{\mathcal{M}}$. Combining (S11) and (S5), by the contraction property (S2), we have result (S8). $\square$

1.2 Proof of Proposition 3

We are going to prove this Proposition by contradiction. Suppose that there exists a proper set $\mathcal{C}' \subset \mathcal{C}$ such that

$$Y(d) \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{C}'}, \tag{S12}$$

where $d = 0, 1$. Without loss of generality, we assume $\boldsymbol{X}^{\mathcal{C}} \setminus \boldsymbol{X}^{\mathcal{C}'}$ contains at least one element $X^{(1)}$. We first show that for all $\mathcal{E} \subseteq pa(Y) \setminus \mathcal{C}'$, we have

$$\boldsymbol{X}^{\mathcal{E}} \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{C}'}. \tag{S13}$$

We will show this by contradiction. If this is not true, then $\boldsymbol{X}^{\mathcal{E}} \not\perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{C}'}$. So we have $\boldsymbol{X}^{\mathcal{E}}$ and D are d-connected given $\boldsymbol{X}^{\mathcal{C}'}$. Since $\mathcal{E} \subset pa(Y) \setminus \boldsymbol{X}^{\mathcal{C}'} \subset pa(Y)$, there is a direct edge

$\boldsymbol{X}^{\mathcal{E}} \to Y$, which means D and $Y(d)$ are d-connected given $\boldsymbol{X}^{\mathcal{C}'}$. This contradicts (S12) under the faithfulness assumption.

Let $\mathcal{E} = pa(Y) \setminus \boldsymbol{X}^{\mathcal{C}'}$ in (S13) and $\mathcal{M} = pa(Y)$ in (S5). Given (S13), (S5) and the contraction property (S2), we have

$$\{\boldsymbol{X}^{\mathcal{E}} \cup Y(d)\} \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{C}'}. \tag{S14}$$

Combining the decomposition property (S4) and (S14), we have

$$[\{\boldsymbol{X}^{\mathcal{E}} \setminus X^{(1)}\} \cup Y(d)] \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{C}'}.$$

We note that $\boldsymbol{X}^{\mathcal{C}'} \cup \{\boldsymbol{X}^{\mathcal{E}} \setminus X^{(1)}\} = \boldsymbol{X}^{\mathcal{C}'} \cup [\{pa(Y) \setminus \boldsymbol{X}^{\mathcal{C}'}\} \setminus X^{(1)}] = pa(Y) \setminus X^{(1)}$. We then use the weak union property (S3) to obtain

$$Y(d) \perp\!\!\!\perp D \mid \{pa(Y) \setminus X^{(1)}\},$$

where $d = 0, 1$. Under faithfulness, this contradicts with the fact that $X^{(1)} \in \boldsymbol{X}^{\mathcal{C}}$, which suggests $X^{(1)}$ and D are d-connected given $pa(Y) \setminus X^{(1)}$. $\square$

1.3 Proof of Proposition 4

Part A In this part, we will show $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_I}$ for $d = 0, 1$. For $j \in \mathcal{I}' \subset \mathcal{I}$, we first show that an instrument variable is independent of the potential outcome conditional on a sufficient set $\mathcal{S}$:

$$X^{(j)} \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}} \text{ for } d = 0, 1. \tag{S15}$$

We will show this by contradiction. Assume this is not true, then $Y(d)$ and $X^{(j)}$ are dependent conditional on $\boldsymbol{X}^{\mathcal{S}}$. Given the faithfulness assumption 6, we know that $Y(d)$ and $X^{(j)}$ are d-connected conditional on $\boldsymbol{X}^{\mathcal{S}}$. Since $X^{(j)} \in pa(D)$, there is a direct path that $X^{(j)} \leftarrow D$, so there exists a back door path $D \leftarrow X^{(j)} \cdots Y(d)$ conditional on $\boldsymbol{X}^{\mathcal{S}}$, which means D and $Y(d)$ are d-connected conditional on $\boldsymbol{X}^{\mathcal{S}}$. Given the faithfulness assumption 6, we know that $Y(d)$ and D are dependent conditional on $\boldsymbol{X}^{\mathcal{S}}$, which is a contradiction to the condition $Y(d) \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{S}}$.

We now show $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_I}$ for $d = 0, 1$. Given (S15), $Y(d) \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{S}}$, and the faithfulness assumption 6, we know that $Y(d)$ and $(X^{(j)}, D)$ are d-separated given $\boldsymbol{X}^{\mathcal{S}}$. So we have

$$(\boldsymbol{X}^{\mathcal{I}'} \cup D) \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}} \text{ for } d = 0, 1. \tag{S16}$$

Combining (S16) and the weak union property (S3), we have

$$D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_I} \text{ for } d = 0, 1. \qquad \square$$

Part B In this part, we will show $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_P}$ for $d = 0, 1$. We first show that a precision variable is independent of treatment given sufficient set $\mathcal{S}$:

$$X^{(j)} \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{S}}, \tag{S17}$$

where $j \in \mathcal{P}$. Again we show this by contradiction. Assume this is not true, then D and $X^{(j)}$ are dependent given $\boldsymbol{X}^{\mathcal{S}}$. Given the faithfulness assumption 6, we know that D and $X^{(j)}$ are d-connected given $\boldsymbol{X}^{\mathcal{S}}$. Since $X^{(j)} \in pa(Y)$, there is a direct path that $X^{(j)} \leftarrow Y$, so there exists a back door path $Y \leftarrow X^{(j)} \cdots D$ given $\boldsymbol{X}^{\mathcal{S}}$, which means D and $Y(d)$ are d-connected given $\boldsymbol{X}^{\mathcal{S}}$. Given the faithfulness assumption 6, we know that $Y(d)$ and D are dependent given $\boldsymbol{X}^{\mathcal{S}}$, which is a contradiction to the condition.

Next, we show $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_P}$. Given (S17), $Y(d) \perp\!\!\!\perp D \mid \boldsymbol{X}^{\mathcal{S}}$, and the faithful assumption 6, we know that $(Y(d), X^{(j)})$ and D are d-separated given $\boldsymbol{X}^{\mathcal{S}}$. So we have

$$D \perp\!\!\!\perp \{Y(d) \cup X^{\mathcal{P}'}\} \mid \boldsymbol{X}^{\mathcal{S}} \text{ for } d = 0, 1. \tag{S18}$$

Combining (S18) and the weak union property (S3), we have

$$D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}_P} \text{ for } d = 0, 1. \qquad \square$$

1.4 Proof of Proposition 5

Under conditions of Lemma 2.1 in Pan et al. (2019) we have

$$\begin{aligned} & Bcov(X, Y \mid D) = 0 \\ \Longleftrightarrow & Bcov(X^{(d)}, Y^{(d)}) = 0, d = 1, 0. \\ \Longleftrightarrow & X^{(d)} \perp\!\!\!\perp Y^{(d)}, d = 1, 0. \\ \Longleftrightarrow & X \perp\!\!\!\perp Y \mid D. \end{aligned}$$

1.5 Proof of Proposition 6

We assume $(\boldsymbol{X} \mid D = d) \sim N(\mathbf{u_d}, \boldsymbol{\Sigma})$, then $(\boldsymbol{X}^{\mathcal{S}} \mid D = d) \sim N(\tilde{\mathbf{u}}_d, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\mathbf{u}}_d$ is a subvector of $\mathbf{u_d}$ that contains the first p_0 elements of $\mathbf{u_d}$, and $\tilde{\boldsymbol{\Sigma}}$ is a submatrix of $\boldsymbol{\Sigma}$ that contains first p_0 rows and first p_0 columns of $\boldsymbol{\Sigma}$. We can write

$$\frac{P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})}{P(D = 0 \mid \boldsymbol{X}^{\mathcal{S}})} = \frac{P(\boldsymbol{X}^{\mathcal{S}} \mid D = 1)}{P(\boldsymbol{X}^{\mathcal{S}} \mid D = 0)} \frac{P(D = 1)}{P(D = 0)}.$$

Let $P(D = 1)/P(D = 0) = \exp(c)$, where c is some real constant. We have

$$\frac{P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})}{P(D = 0 \mid \boldsymbol{X}^{\mathcal{S}})} = \exp(c) \exp\{(\tilde{\mathbf{u}}_1 - \tilde{\mathbf{u}}_0)^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}^{\mathcal{S}} - \frac{1}{2}(\tilde{\mathbf{u}}_1^T \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{u}}_1 - \tilde{\mathbf{u}}_0^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{u}}_0)\}.$$

Let $\alpha_0 = c - \frac{1}{2}(\tilde{\mathbf{u}}_1^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{u}}_1 - \tilde{\mathbf{u}}_0^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{u}}_0)$, and $\boldsymbol{\alpha}^{\mathrm{T}} = (\tilde{\mathbf{u}}_1 - \tilde{\mathbf{u}}_0)^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}^{-1}$. We have

$$\frac{P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})}{1 - P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})} = \exp(\alpha_0 + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}^{\mathcal{S}}).$$

This finishes the proof. $\square$

2 Proof of Theorems

2.1 Proof of Theorem 1

Without loss of generality, we assume $\mathcal{S} = \{1, 2, \ldots, d\}$ and $\mathcal{S}_P = \{1, 2, \ldots, p_0\}$, i.e. we add precision variables $X^{(j)}$, $j = d+1, \ldots, p_0$ to the set $\mathcal{S}$. We prove the theorem using standard M-estimation theories. For completeness, we first introduce the theories here.

An M-estimator $\hat{\boldsymbol{\theta}}$ satisfies the following estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{Y}_i, \hat{\boldsymbol{\theta}}) = 0.$$

Denote $\boldsymbol{\theta}_0$ the solution of vector function $E\{\boldsymbol{\phi}(\boldsymbol{Y}, \boldsymbol{\theta})\} = 0$. Stefanski and Boos (2002) showed that an M-estimator is asymptotically normally distributed with $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{0, V(\boldsymbol{\theta}_0)\}$, where $V(\boldsymbol{\theta}_0) = A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0) \{A(\boldsymbol{\theta}_0)^{-1}\}^{\mathrm{T}}$ with $A(\boldsymbol{\theta}_0) = E\{-\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{\phi}(\boldsymbol{Y}, \boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$ and $B(\boldsymbol{\theta}_0) = E\{\boldsymbol{\phi}(\boldsymbol{Y}, \boldsymbol{\theta}) \boldsymbol{\phi}(\boldsymbol{Y}, \boldsymbol{\theta})^{\mathrm{T}} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$.

Part A We first prove the results for the ratio estimator (2).

When we use $\boldsymbol{X}^{\mathcal{S}}$ to estimate the average causal effect, we introduce the parameters of interest $\boldsymbol{\theta}^{\mathcal{S}} = (\Delta^{\mathcal{S}}, \{\boldsymbol{\beta}^{\mathcal{S}}\}^{\mathrm{T}}, \lambda^{\mathcal{S}}, \kappa^{\mathcal{S}})^{\mathrm{T}} \in \mathbb{R}^{d+3}$, where $\lambda^{\mathcal{S}} = E\{D/e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}^{\mathcal{S}})\}, \kappa^{\mathcal{S}} = E[(1-D)/\{1 - e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}^{\mathcal{S}})\}]$ are the ratios, $\boldsymbol{\beta}^{\mathcal{S}} \in \mathbb{R}^d$ is the regression coefficients of $\boldsymbol{X}^{\mathcal{S}}$ in the propensity score model, and $\Delta^{\mathcal{S}}$ is the average causal effect, our primary parameter of interest. We define the functions $\boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}}) = (\phi_0, \ldots, \phi_{d+2})^{\mathrm{T}} : \mathbb{R}^{d+3} \to \mathbb{R}^{d+3}$, where

$$\begin{cases} \phi_0 = -\Delta^{\mathcal{S}} + \lambda^{\mathcal{S}} \frac{YD}{e} - \kappa^{\mathcal{S}} \frac{Y(1-D)}{1-e}, \\ \phi_i = (\frac{D}{e} - \frac{1-D}{1-e}) \partial e / \beta_i, \\ \phi_{d+1} = -\lambda^{\mathcal{S}} \frac{D}{e} + 1, \\ \phi_{d+2} = -\kappa^{\mathcal{S}} \frac{1-D}{1-e} + 1, \end{cases}$$

for $1 \le i \le d$, and $e = e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}^{\mathcal{S}})$ is the propensity score. The estimate $\hat{\boldsymbol{\theta}}^{\mathcal{S}} = (\hat{\Delta}^{\mathcal{S}}, (\hat{\boldsymbol{\beta}}^{\mathcal{S}})^{\mathrm{T}}, \hat{\lambda}^{\mathcal{S}}, \hat{\kappa}^{\mathcal{S}})^{\mathrm{T}}$ satisfies $\sum_{i=1}^{n} \boldsymbol{\phi}(Y_i, D_i, \boldsymbol{X}_i^{\mathcal{S}}; \hat{\boldsymbol{\theta}}^{\mathcal{S}}) = \mathbf{0}$, and the first element in $\hat{\boldsymbol{\theta}}^{\mathcal{S}}$ is our ratio estimator (2).

We first calculate the asymptotic variance of $\hat{\Delta}^{\mathcal{S}}$, denoted by $V^{\mathcal{S}}(\boldsymbol{\theta}_0^{\mathcal{S}})$. In this case, the true parameter values $\boldsymbol{\theta}_0^{\mathcal{S}}$, defined as $E\{\Phi(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}_0^{\mathcal{S}})\} = \mathbf{0}$, satisfies $\boldsymbol{\theta}_0^{\mathcal{S}} = (\Delta_0, (\boldsymbol{\beta}_0^{\mathcal{S}})^{\mathrm{T}}, 1, 1)^{\mathrm{T}}$, where $\boldsymbol{\beta}_0^{\mathcal{S}}$ is the coefficient of $\boldsymbol{X}^{\mathcal{S}}$ in the true propensity score model that $e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}_0^{\mathcal{S}}) = P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}})$. We have

$$A^{\mathcal{S}} = E\{-\frac{\partial}{(\partial \boldsymbol{\theta}^{\mathcal{S}})^{\mathrm{T}}} \boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}}) \mid_{\boldsymbol{\theta}^{\mathcal{S}} = \boldsymbol{\theta}_0^{\mathcal{S}}}\} = \begin{bmatrix} 1 & (H_{Ratio}^{\mathcal{S}})^{\mathrm{T}} & -\mu_1 & \mu_0 \\ 0 & E^{\mathcal{S},\mathcal{S}} & 0 & 0 \\ 0 & -E\{\frac{1}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\} & 1 & 0 \\ 0 & -E\{\frac{1}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\} & 0 & 1 \end{bmatrix},$$

$$B^{\mathcal{S}} = E\{\boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}}) \boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}})^{\mathrm{T}} \mid_{\boldsymbol{\theta}^{\mathcal{S}} = \boldsymbol{\theta}_0^{\mathcal{S}}}\} =$$

$$\begin{bmatrix} \sigma^2 & (H_{Ratio}^{\mathcal{S}})^{\mathrm{T}} & \Delta_0 - E\{\frac{Y(1)}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\} & \Delta_0 + E\{\frac{Y(0)}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\} \\ H_{Ratio}^{\mathcal{S}} & E^{\mathcal{S},\mathcal{S}} & -E\{\frac{1}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\} & E\{\frac{1}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\} \\ \Delta_0 - E\{\frac{Y(1)}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\} & -E\{\frac{1}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\} & E\{\frac{1}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} - 1\} & -1 \\ \Delta_0 + E\{\frac{Y(0)}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\} & E\{\frac{1}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\} & -1 & E\{\frac{1}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\} - 1 \end{bmatrix},$$

where

$$H_{Ratio}^{\mathcal{S}} = E\left[\left\{\frac{Y(1)}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} + \frac{Y(0)}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\right\} \frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\right], \quad \mu_1 = E\{Y(1)\}, \quad \mu_0 = E\{Y(0)\},$$

$$\sigma^2 = E\left\{\frac{Y(1)^2}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} + \frac{Y(0)^2}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\right\},\ e(\boldsymbol{X}^{\mathcal{S}};\boldsymbol{\beta}_0^{\mathcal{S}}) = P(D=1 \mid \boldsymbol{X}^{\mathcal{S}}),$$

$$E^{\mathcal{S},\mathcal{S}} = E\left[\frac{1}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})\{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})\}}\frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\frac{\partial e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\right].$$

The asymptotic variance of our average treatment effect $\hat{\boldsymbol{\Delta}}^{\mathcal{S}}$ is the entry in the first row and first column of the matrix $V^{\mathcal{S}} = \{A^{\mathcal{S}}\}^{-1}B^{\mathcal{S}}\{(A^{\mathcal{S}})^{-1}\}^{\mathrm{T}}$. We obtain the asymptotic variance

$$\sigma_0^2 = \sigma_{\mathcal{S}}^2 - (H_{Ratio}^{\mathcal{S}})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}H_{Ratio}^{\mathcal{S}},$$

where

$$\sigma_{\mathcal{S}}^2 = E\left[\frac{\{Y(1)-\mu_1\}^2}{e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})} + \frac{\{Y(0)-\mu_0\}^2}{1-e(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}})}\right].$$

Similarly, when we use $\boldsymbol{X}^{\mathcal{S}_P}$ to estimate the average causal effect, the parameters of interest are $\boldsymbol{\theta}^{\mathcal{S}_P} = (\Delta^{\mathcal{S}_P}, (\boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}, (\boldsymbol{\gamma}^{\mathcal{P}'})^{\mathrm{T}}, \lambda^{\mathcal{S}_P}, \kappa^{\mathcal{S}_P})^{\mathrm{T}} \in \mathbb{R}^{p_0+3}$, where $\kappa^{\mathcal{S}_P} = E[(1-D)/\{1-e(\boldsymbol{X}^{\mathcal{S}_P};\boldsymbol{\beta}^{\mathcal{S}},\boldsymbol{\gamma}^{\mathcal{P}'})\}], \lambda^{\mathcal{S}_P} = E\{D/e(\boldsymbol{X}^{\mathcal{S}_P};\boldsymbol{\beta}^{\mathcal{S}},\boldsymbol{\gamma}^{\mathcal{P}'})\}$ are the ratios, $\boldsymbol{\beta}^{\mathcal{S}} \in \mathbb{R}^d$ is the coefficient of $\boldsymbol{X}^{\mathcal{S}}$ in the propensity score model, $\boldsymbol{\gamma}^{\mathcal{P}'} \in \mathbb{R}^{p_0-d}$ is the coefficient of $\boldsymbol{X}^{\mathcal{P}'}$ in the propensity score model, and $\Delta^{\mathcal{S}_P}$ is the average causal effect, the primary parameter of interest. Define $\boldsymbol{\Phi}(Y,D,\boldsymbol{X}^{\mathcal{S}_P};\boldsymbol{\theta}^{\mathcal{S}_P}) = (\phi_0,\phi_1,\ldots,\phi_{p_0+2})^{\mathrm{T}} : \mathbb{R}^{p_0+3} \to \mathbb{R}^{p_0+3}$, where

$$\begin{cases} \phi_0 = -\Delta^{\mathcal{S}_P} + \lambda^{\mathcal{S}_P}\frac{YD}{\tilde{e}} - \kappa^{\mathcal{S}_P}\frac{Y(1-D)}{1-\tilde{e}} \\ \phi_i = (\frac{D}{\tilde{e}} - \frac{1-D}{1-\tilde{e}})\partial\tilde{e}/\beta_i \\ \phi_{d+j} = (\frac{D}{\tilde{e}} - \frac{1-D}{1-\tilde{e}})\partial\tilde{e}/\partial\gamma_j \\ \phi_{p_0+1} = -\lambda^{\mathcal{S}_P}\frac{D}{\tilde{e}} + 1 \\ \phi_{p_0+2} = -\kappa^{\mathcal{S}_P}\frac{1-D}{1-\tilde{e}} + 1, \end{cases}$$

$1 \le i \le d$, $1 \le j \le p_0 - d$, and $\tilde{e} = \tilde{e}(\boldsymbol{X}^{\mathcal{S}_P};\boldsymbol{\beta}^{\mathcal{S}},\boldsymbol{\gamma}^{\mathcal{P}'})$ is the parametric propensity score model. Although the true value of $\boldsymbol{\gamma}^{\mathcal{P}'} \in \mathbb{R}^{p_0-d}$ satisfies $\boldsymbol{\gamma}_0^{P'} = \mathbf{0}$.

In this case, the true value of the parameter, $\boldsymbol{\theta}_0^{\mathcal{S}_P} = (\Delta_0, (\boldsymbol{\beta}_0^{\mathcal{S}})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}, 1, 1)^{\mathrm{T}}$, which is defined as the solution of $E\{\Phi(Y, D, \boldsymbol{X}^{\mathcal{S}_P}; \boldsymbol{\theta}_0^{\mathcal{S}_P})\} = \mathbf{0}$. Similar as the previous derivation, we can calculate $A^{\mathcal{S}_P}$, $B^{\mathcal{S}_P}$, $V^{\mathcal{S}_P}$. We have

$$\sigma_P^2 = \sigma_{\mathcal{S}_P}^2 - (H_{Ratio}^{\mathcal{S}_P})^{\mathrm{T}}\{E^{(\mathcal{S},\mathcal{P}'),(\mathcal{S},\mathcal{P}')}\}^{-1} H_{Ratio}^{\mathcal{S}_P},$$

where

$$\begin{aligned}
H_{Ratio}^{\mathcal{S}_P} =& E\left[\left\{\frac{Y(1)-\mu_1}{\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})} + \frac{Y(0)-\mu_0}{1-\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}\right\}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\right., \\
&\left.\left\{\frac{Y(1)-\mu_1}{\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})} + \frac{Y(0)-\mu_0}{1-\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}\right\}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\gamma}^{\mathcal{P}'}}\right], \\
\sigma_{\mathcal{S}_P}^2 =& E\left[\frac{\{Y(1)-\mu_1\}^2}{\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})} + \frac{\{Y(0)-\mu_0\}^2}{1-\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}\right], \\
E^{(\mathcal{S},\mathcal{P}'),(\mathcal{S},\mathcal{P}')} =& \begin{bmatrix} E\{\mathcal{U}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\} & E\{\mathcal{U}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{(\partial \boldsymbol{\gamma}^{\mathcal{P}'})^{\mathrm{T}}}\} \\ E\{\mathcal{U}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\gamma}^{\mathcal{P}'}}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{(\partial \boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}}\} & E\{\mathcal{U}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\gamma}^{\mathcal{P}'}}\frac{\partial \tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})}{(\partial \boldsymbol{\gamma}^{\mathcal{P}'})^{\mathrm{T}}}\}, \end{bmatrix} \\
\mathcal{U} =& \frac{1}{\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})\{1-\tilde{e}(\boldsymbol{X};\boldsymbol{\beta}_0^{\mathcal{S}},\boldsymbol{\gamma}_0^{\mathcal{P}'})\}}
\end{aligned}$$

With (S18) and (S3), we have $D \perp\!\!\!\perp \boldsymbol{X}^{\mathcal{P}'} \mid \boldsymbol{X}^{\mathcal{S}}$, which implies

$$e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}_0^{\mathcal{S}}) = P(D=1 \mid \boldsymbol{X}^{\mathcal{S}}) = P(D=1 \mid \boldsymbol{X}^{\mathcal{S}_P}) = \tilde{e}(\boldsymbol{X}^{\mathcal{S}_P}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'}). \tag{S19}$$

Denote $H_{Ratio}^{\mathcal{S}_P}$, $E^{(\mathcal{S},\mathcal{P}'),(\mathcal{S},\mathcal{P}')}$ as

$$(H_{Ratio}^{\mathcal{S}}, H_{Ratio}^{\mathcal{P}'}), \quad \begin{bmatrix} E^{\mathcal{S},\mathcal{S}} & E^{\mathcal{S},\mathcal{P}'} \\ (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}} & E^{\mathcal{P}',\mathcal{P}'} \end{bmatrix},$$

respectively. Again by (S19) and simple algebra, we can show

$$\begin{aligned}
\sigma_0^2 - \sigma_P^2 =& \{H_{Ratio}^{\mathcal{P}'} - (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}H_{Ratio}^{\mathcal{S}}\}^{\mathrm{T}}\{E^{\mathcal{P}',\mathcal{P}'} - (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}E^{\mathcal{S},\mathcal{P}'}\}^{-1} \\
&\{H_{Ratio}^{\mathcal{P}'} - (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}H_{Ratio}^{\mathcal{S}}\} \geq 0.
\end{aligned}$$

Thus we finish the proof of this part. □

Part B For HT estimator (1), the proof is analogous.

When we use $\boldsymbol{X}^{\mathcal{S}}$ to estimate the average causal effect, the parameters of interest are $\boldsymbol{\theta}^{\mathcal{S}} = (\Delta^{\mathcal{S}}, (\boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}$, where $\boldsymbol{\beta}^{\mathcal{S}} \in \mathbb{R}^{d}$ is the coefficient of $\boldsymbol{X}^{\mathcal{S}}$ in the propensity score model, and $\Delta^{\mathcal{S}}$ is the average causal effect we aim to estimate. And the corresponding estimating equations are $\boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}}) = (\phi_0, \phi_1, \dots, \phi_d) : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$, where

$$\begin{cases} \phi_0 = -\Delta^{\mathcal{S}} + \frac{YD}{e} - \frac{Y(1-D)}{1-e} \\ \phi_i = (\frac{D}{e} - \frac{1-D}{1-e})\partial e/\beta_i, \end{cases}$$

$1 \leq i \leq d$, and $e = e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}^{\mathcal{S}})$ is the parametric propensity score. The M-estimator $\hat{\boldsymbol{\theta}}^{\mathcal{S}}$ satisfies

$$\sum_{i=1}^{n} \boldsymbol{\Phi}(Y_i, D_i, \boldsymbol{X}_i^{\mathcal{S}}; \hat{\boldsymbol{\theta}}^{\mathcal{S}}) = \mathbf{0}.$$

And the first element of $\hat{\boldsymbol{\theta}}^{\mathcal{S}}$ is our IPW estimator (1) $\hat{\Delta}^{\mathcal{S}}$. As the true value $\boldsymbol{\theta}_0^{\mathcal{S}} = (\Delta_0, (\boldsymbol{\beta}_0^{\mathcal{S}})^{\mathrm{T}})^{\mathrm{T}}$ satisfies $E\{\Phi(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}_0^{\mathcal{S}})\} = \mathbf{0}$, we have $V(\boldsymbol{\theta}_0^{\mathcal{S}}) = A(\boldsymbol{\theta}_0^{\mathcal{S}})^{-1} B(\boldsymbol{\theta}_0^{\mathcal{S}}) \, \{A(\boldsymbol{\theta}_0^{\mathcal{S}})^{-1}\}^{\mathrm{T}}$, $A(\boldsymbol{\theta}_0^{\mathcal{S}}) = E\{-\frac{\partial}{(\partial \boldsymbol{\theta}^{\mathcal{S}})^{\mathrm{T}}} \boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}}) \mid_{\boldsymbol{\theta}^{\mathcal{S}} = \boldsymbol{\theta}_0^{\mathcal{S}}}\}$, $B(\boldsymbol{\theta}_0^{\mathcal{S}}) = E\{\boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\theta}^{\mathcal{S}}) \boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}} ; \boldsymbol{\theta}^{\mathcal{S}})^{\mathrm{T}} \mid_{\boldsymbol{\theta}^{\mathcal{S}} = \boldsymbol{\theta}_0^{\mathcal{S}}} \}$. And σ_0^2 is the element of first row and first column of $V(\boldsymbol{\theta}_0^{\mathcal{S}})$, which satisfies

$$\sigma_0^2 = \sigma_{\mathcal{S}}^2 - (H_{HT}^{\mathcal{S}})^{\mathrm{T}} (E^{\mathcal{S},\mathcal{S}})^{-1} H_{HT}^{\mathcal{S}},$$

where

$$H_{HT}^{\mathcal{S}} = E\left\{ \left(\frac{Y(1)}{e(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}})} + \frac{Y(0)}{1 - e(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}})} \right) \frac{\partial e(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}})}{\partial \boldsymbol{\beta}^{\mathcal{S}}} \right\}, \quad \sigma_{\mathcal{S}}^2 = E\left[\frac{Y(1)^2}{e(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}})} + \frac{Y(0)^2}{1 - e(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}})} \right].$$

Similarly when we use $\boldsymbol{X}^{\mathcal{S}_P}$ to estimate the average causal effect, the parameters of interest are $\boldsymbol{\theta}^{\mathcal{S}_P} = (\Delta^{\mathcal{S}_P}, (\boldsymbol{\beta}^{\mathcal{S}})^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathcal{P}'})^{\mathrm{T}} \in \mathbb{R}^{p_0+1}$, where $\boldsymbol{\beta}^{\mathcal{S}} \in \mathbb{R}^{d}$ is the coefficient of $\boldsymbol{X}^{\mathcal{S}}$ and where $\boldsymbol{\gamma}^{\mathcal{P}'} \in \mathbb{R}^{p_0 - d}$ is the coefficient of $\boldsymbol{X}^{\mathcal{P}'}$ in the propensity score model, and $\Delta^{\mathcal{S}_P}$ is the average causal effect we aim to estimate. And the corresponding estimating equations are $\boldsymbol{\Phi}(Y, D, \boldsymbol{X}^{\mathcal{S}_P}; \boldsymbol{\theta}^{\mathcal{S}_P}) = (\phi_0, \phi_1, \ldots, \phi_{p_0})$: $\mathbb{R}^{p_0+1} \to \mathbb{R}^{p_0+1}$, where

$$\begin{cases} \phi_0 = -\Delta^{\mathcal{S}_P} + \frac{YD}{\tilde{e}} - \frac{Y(1-D)}{1-\tilde{e}} \\ \phi_i = (\frac{D}{\tilde{e}} - \frac{1-D}{1-\tilde{e}})\partial \tilde{e}/\beta_i \\ \phi_{d+j} = (\frac{D}{\tilde{e}} - \frac{1-D}{1-\tilde{\tilde{e}}})\partial \tilde{e}/\partial \gamma_j, \end{cases}$$

$1 \leq i \leq d$, $1 \leq j \leq p_0 - d$, $\tilde{e} = \tilde{e}(\boldsymbol{X}^{\mathcal{S}_P}; \boldsymbol{\beta}^{\mathcal{S}}, \boldsymbol{\gamma}^{\mathcal{P}'})$ is the parametric propensity score. The true value of the parameter of interest is $\boldsymbol{\theta}_0^{\mathcal{S}} = (\Delta_0, (\boldsymbol{\beta}_0^{\mathcal{S}})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, which satisfies $E\{\Phi(Y, D, \boldsymbol{X}^{\mathcal{S}_P}; \boldsymbol{\theta}_0^{\mathcal{S}_P}) = \mathbf{0}\}$. Follow a similar argument, we have

$$\sigma_P^2 = \sigma_{\mathcal{S}_P}^2 - (H_{HT}^{\mathcal{S}_P})^{\mathrm{T}} (E^{(\mathcal{S},\mathcal{P}'),(\mathcal{S},\mathcal{P}')})^{-1} H_{HT}^{\mathcal{S}_P},$$

where

$$\begin{aligned} H_{HT}^{\mathcal{S}_P} &= E\left[\left\{\frac{Y(1)}{\tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})} + \frac{Y(0)}{1 - \tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})}\right\} \frac{\partial \tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\beta}^{\mathcal{S}}}\right., \\ &\quad \left.\left\{\frac{Y(1)}{\tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})} + \frac{Y(0)}{1 - \tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})}\right\} \frac{\partial \tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})}{\partial \boldsymbol{\gamma}^{\mathcal{P}'}}\right], \\ \sigma_{\mathcal{S}_P}^2 &= E\left\{\frac{Y(1)^2}{\tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})} + \frac{Y(0)^2}{1 - \tilde{e}(\boldsymbol{X}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'})}\right\} = \sigma_{\mathcal{S}}^2, \end{aligned}$$

Similarly, by (S11), we have $D \perp\!\!\!\perp \boldsymbol{X}^{\mathcal{P}'} \mid \boldsymbol{X}^{\mathcal{S}}$, thus

$$e(\boldsymbol{X}^{\mathcal{S}}; \boldsymbol{\beta}_0^{\mathcal{S}}) = P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}}) = P(D = 1 \mid \boldsymbol{X}^{\mathcal{S}_P}) = \tilde{e}(\boldsymbol{X}^{\mathcal{S}_P}; \boldsymbol{\beta}_0^{\mathcal{S}}, \boldsymbol{\gamma}_0^{\mathcal{P}'}).$$

We can write $H_{HT}^{\mathcal{S}_P}$, $E^{(\mathcal{S},\mathcal{P}'),(\mathcal{S},\mathcal{P}')}$ as

$$(H_{HT}^{\mathcal{S}}, H_{HT}^{\mathcal{P}'}), \quad \begin{bmatrix} E^{\mathcal{S},\mathcal{S}} & E^{\mathcal{S},\mathcal{P}'} \\ (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}} & E^{\mathcal{P}',\mathcal{P}'} \end{bmatrix},$$

respectively. By simple algebra,

$$\sigma_0^2 - \sigma_P^2 = \{H_{HT}^{\mathcal{P}'} - (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}H_{HT}^{\mathcal{S}}\}^{\mathrm{T}}\{E^{\mathcal{P}',\mathcal{P}'} - (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}E^{\mathcal{S},\mathcal{P}'}\}^{-1}$$
$$\{H_{HT}^{\mathcal{P}'} - (E^{\mathcal{S},\mathcal{P}'})^{\mathrm{T}}(E^{\mathcal{S},\mathcal{S}})^{-1}H_{HT}^{\mathcal{S}}\} \geq 0.$$

Thus we finish the proof of this case. $\square$

2.2 Proof of Theorem 2

We firstly show that a subset of instrument variable $\mathcal{I}$ is independent of the potential outcome given $\boldsymbol{X}^{\mathcal{S}}$:

Lemma If $\boldsymbol{X}^{\mathcal{I}'} \subset \boldsymbol{X}^{\mathcal{I}}$, under assumption 6 we have

$$\boldsymbol{X}^{\mathcal{I}'} \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}}, \tag{S20}$$

where $d = 0, 1$. To see this, combining (S16) and the decomposition property (S4) yields (S20).

Part A For the Ratio estimator (2), based on simple calculation and transformation, we have

$$\sqrt{n}(\tilde{\Delta}_{Ratio}^{\mathcal{S}_I} - \Delta_0) = \left(n/\sum_{i=1}^{n}\frac{D_i}{\tilde{e}_i}\right)\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i(Y_i-\mu_1)}{\tilde{e}_i}\right\} - \left(n/\sum_{i=1}^{n}\frac{1-D_i}{1-\tilde{e}_i}\right)\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(1-D_i)(Y_i-\mu_0)}{1-\tilde{e}_i}\right\},$$

$$\sqrt{n}(\tilde{\Delta}_{Ratio}^{\mathcal{S}} - \Delta_0) = \left(n/\sum_{i=1}^{n}\frac{D_i}{e_i}\right)\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i(Y_i-\mu_1)}{e_i}\right\} - \left(n/\sum_{i=1}^{n}\frac{1-D_i}{1-e_i}\right)\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(1-D_i)(Y_i-\mu_0)}{1-e_i}\right\},$$

where

$$e_i = P(D = 1 \mid \boldsymbol{X}_i^{\mathcal{S}}), \quad \tilde{e}_i = P(D = 1 \mid \boldsymbol{X}_i^{\mathcal{S}_I}),$$

$$\mu_1 = E\{Y(1)\}, \quad \mu_0 = E\{Y(0)\}.$$

We use $q_1^{(n)}, q_0^{(n)}, \tilde{q}_1^{(n)}, \tilde{q}_0^{(n)}$ to denote $n/\sum_{i=1}^n (D_i/e_i), \quad n/\sum_{i=1}^n \{(1-D_i)/(1-e_i)\}$, $n/\sum_{i=1}^n (D_i/\tilde{e}_i), \quad n/\sum_{i=1}^n \{(1-D_i)/(1-\tilde{e}_i)\}$ respectively. As

$$\begin{aligned}
&E\left(\frac{D}{e}\right) = E\left\{E\left(\frac{D}{e} \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} = E\left\{\frac{1}{e}E\left(D \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} = E\left(\frac{e}{e}\right) = 1,\\
&E\left(\frac{1-D}{1-e}\right) = E\left\{E\left(\frac{1-D}{1-e} \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} = E\left\{\frac{1}{1-e}E(1-D \mid \boldsymbol{X}^{\mathcal{S}})\right\} = E\left\{\frac{1-e}{1-e}\right\} = 1,\\
&E\left(\frac{D}{\tilde{e}}\right) = E\left\{E\left(\frac{D}{\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}_I}\right)\right\} = E\left\{\frac{1}{\tilde{e}}E(D \mid \boldsymbol{X}^{\mathcal{S}_I})\right\} = E\left(\frac{\tilde{e}}{\tilde{e}}\right) = 1,\\
&E\left(\frac{1-D}{1-\tilde{e}}\right) = E\left\{E\left(\frac{1-D}{1-\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}_I}\right)\right\} = E\left\{\frac{1}{1-\tilde{e}}E(1-D \mid \boldsymbol{X}^{\mathcal{S}_I})\right\} = E\left\{\frac{1-\tilde{e}}{1-\tilde{e}}\right\} = 1,
\end{aligned}$$

we have $q_1^{(n)} \xrightarrow{p} 1$, $q_0^{(n)} \xrightarrow{p} 1$, $\tilde{q}_1^{(n)} \xrightarrow{p} 1$, $\tilde{q}_0^{(n)} \xrightarrow{p} 1$. By Central Limit Theorem, we have the following results:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{D_i(Y_i-\mu_1)}{e_i} \xrightarrow{d} N\left(0, E\left[\frac{D\{Y(1)-\mu_1\}^2}{e_i^2}\right]\right),\\
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{(1-D_i)(Y_i-\mu_0)}{1-e_i} \xrightarrow{d} N\left(0, E\left[\frac{(1-D)\{Y(0)-\mu_0\}^2}{(1-e_i)^2}\right]\right),\\
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\frac{D_i(Y_i-\mu_1)}{e_i} - \frac{(1-D_i)(Y_i-\mu_0)}{1-e_i}\right\} \xrightarrow{d}\\
&N\left(0, E\left[\frac{D\{Y(1)-\mu_1\}^2}{e_i^2} + \frac{(1-D)\{Y(0)-\mu_0\}^2}{(1-e_i)^2}\right]\right).
\end{aligned} \tag{S21}$$

Meanwhile, we can rewrite $\sqrt{n}(\Delta^{\mathcal{S}}_{Ratio} - \Delta_0)$ as

$$\begin{aligned}\sqrt{n}(\Delta^{\mathcal{S}}_{Ratio} - \Delta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{D_i(Y_i - \mu_1)}{e_i} - \frac{(1 - D_i)(Y_i - \mu_0)}{1 - e_i} \right\} + \\ \frac{q_1^{(n)} - 1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i(Y_i - \mu_1)}{e_i} - \frac{q_0^{(n)} - 1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(1 - D_i)(Y_i - \mu_0)}{1 - e_i}.\end{aligned} \tag{S22}$$

Combining (S21), (S22) and Slutsky's Theorem, we can obtain

$$\sqrt{n}(\tilde{\Delta}^{\mathcal{S}}_{Ratio} - \Delta_0) \xrightarrow{d} N(0, \sigma_0^2),$$

where

$$\begin{aligned}\sigma_0^2 &= E\left[\frac{D\{Y(1) - \mu_1\}^2}{e^2} + \frac{(1 - D)\{Y(0) - \mu_0\}^2}{(1 - e)^2}\right] \\ &= E\left[E\left[\frac{\{Y(1) - \mu_1\}^2}{e^2} \mid \boldsymbol{X}^{\mathcal{S}}\right] E(D \mid \boldsymbol{X}^{\mathcal{S}}) + E\left[\frac{(Y(0) - \mu_0)^2}{(1 - e)^2} \mid \boldsymbol{X}^{\mathcal{S}}\right] E\{(1 - D) \mid \boldsymbol{X}^{\mathcal{S}}\}\right] \\ &= E\left[E\left[\frac{\{Y(1) - \mu_1\}^2}{e^2} \mid \boldsymbol{X}^{\mathcal{S}}\right] e + E\left[\frac{\{Y(0) - \mu_0\}^2}{(1 - e)^2} \mid \boldsymbol{X}^{\mathcal{S}}\right](1 - e)\right] \\ &= E\{\frac{(Y(1) - \mu_1)^2}{e} + \frac{(Y(0) - \mu_0)^2}{1 - e}\}.\end{aligned}$$

The second equality holds since $\mathcal{S}$ is a sufficient set. Similarly, we have

$$\sqrt{n}(\tilde{\Delta}^{\mathcal{S}_I}_{Ratio} - \Delta_0) \xrightarrow{d} N(0, \sigma_I^2),$$

where

$$\sigma_I^2 = E\left[\frac{\{Y(1) - \mu_1\}^2}{\tilde{e}} + \frac{\{Y(0) - \mu_0\}^2}{1 - \tilde{e}}\right].$$

We will then show $\sigma_I^2 \geq \sigma_0^2$. As $\mathcal{I}' \subset \mathcal{I}$ and (S20), we have $E[\{Y(d) - \mu_d\}^2 \mid \boldsymbol{X}^{\mathcal{S}_I}] = E[\{Y(d) - \mu_d\}^2 \mid \boldsymbol{X}^{\mathcal{S}}]$ for $d = 0, 1$. Denote these conditional expectations as $\tilde{Y}(d)$ for $d = 0, 1$, we have

$$
\begin{aligned}
\sigma_I^2 &= E\left[E\left[\frac{\{Y(1)-\mu_1\}^2}{\tilde{e}} + \frac{\{Y(0)-\mu_0\}^2}{1-\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}_I}\right]\right] \\
&= E\left[\frac{1}{\tilde{e}}E[\{Y(1)-\mu_1\}^2 \mid \boldsymbol{X}^{\mathcal{S}_I}]\right] + E\left[\frac{1}{1-\tilde{e}}E[\{Y(0)-\mu_0\}^2 \mid \boldsymbol{X}^{\mathcal{S}_I}]\right] \\
&= E\left\{\frac{1}{\tilde{e}}\tilde{Y}(1)\right\} + E\left\{\frac{1}{1-\tilde{e}}\tilde{Y}(0)\right\} \\
&= E\left[E\left\{\frac{1}{\tilde{e}}\tilde{Y}(1) \mid \boldsymbol{X}^{\mathcal{S}}\right\}\right] + E\left[E\left\{\frac{1}{1-\tilde{e}}\tilde{Y}(0) \mid \boldsymbol{X}^{\mathcal{S}}\right\}\right] \\
&= E\left\{\tilde{Y}(1)E\left(\frac{1}{\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} + E\left\{\tilde{Y}(0)E\left(\frac{1}{1-\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} \\
&\geq E\left\{\tilde{Y}(1)\frac{1}{e} + \tilde{Y}(0)\frac{1}{1-e}\right\} \\
&= E\left[\frac{1}{e}E[\{Y(1)-\mu_1\}^2 \mid \boldsymbol{X}^{\mathcal{S}}] + \frac{1}{1-e}E[\{Y(0)-\mu_0\}^2 \mid \boldsymbol{X}^{\mathcal{S}}]\right] \\
&= E\left[E\left[\frac{\{Y(1)-\mu_1\}^2}{e} + \frac{\{Y(0)-\mu_0\}^2}{1-e} \mid \boldsymbol{X}^{\mathcal{S}}\right]\right] \\
&= \sigma_0^2.
\end{aligned}
$$

The inequality holds since

$$
1 = 1^2 = E^2(1 \mid \boldsymbol{X}^{\mathcal{S}}) = E^2\left(\frac{1}{\tilde{e}^{1/2}}\tilde{e}^{1/2} \mid \boldsymbol{X}^{\mathcal{S}}\right) \leq E\left(\frac{1}{\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}\right)E\left(\tilde{e} \mid \boldsymbol{X}^{\mathcal{S}}\right) = E\left(\frac{1}{\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}\right)e.
$$

Similarly, one has

$$
(1-e)E(\frac{1}{1-\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}) \geq 1.
$$

Thus we finish the proof for this part.

Part B For the Horvitz-Thompson estimator (1), based on simple calculation and transformation, we have

$$
\sqrt{n}(\tilde{\Delta}_{HT}^{\mathcal{S}_I} - \Delta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_iY_i}{\tilde{e}_i} - \frac{(1-D_i)Y_i}{1-\tilde{e}_i} - \Delta_0\right\},
$$

$$\sqrt{n}(\tilde{\Delta}_{HT}^{\mathcal{S}} - \Delta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{D_i Y_i}{e_i} - \frac{(1-D_i)Y_i}{1-e_i} - \Delta_0 \right\},$$

where $e_i = P(D = 1 \mid \boldsymbol{X}_i^{\mathcal{S}})$, $\tilde{e}_i = P(D = 1 \mid \boldsymbol{X}_i^{\mathcal{S}_I})$. We note that

$$\frac{D_i Y_i}{\tilde{e}_i} - \frac{(1-D_i)Y_i}{1-\tilde{e}_i} - \Delta_0, \quad \frac{D_i Y_i}{e_i} - \frac{(1-D_i)Y_i}{1-e_i} - \Delta_0$$

are i.i.d samples, respectively. By Central Limit Theorem,

$$\sqrt{n}(\tilde{\Delta}_{HT}^{\mathcal{S}} - \Delta_0) \xrightarrow{d} N(0, \sigma_0^2),$$

$$\sqrt{n}(\tilde{\Delta}_{HT}^{\mathcal{S}_I} - \Delta_0) \xrightarrow{d} N(0, \sigma_I^2),$$

where

$$\begin{aligned}
\sigma_0^2 &= E\left\{ \frac{DY(1)^2}{e^2} + \frac{(1-D)Y(0)^2}{(1-e)^2} \right\} - \Delta_0^2 \\
&= E\left[E\left\{ \frac{DY(1)^2}{e^2} \mid \boldsymbol{X}^{\mathcal{S}} \right\} \right] + E\left[E\left\{ \frac{(1-D)Y(0)^2}{(1-e)^2} \mid \boldsymbol{X}^{\mathcal{S}} \right\} \right] - \Delta_0^2 \\
&= E\left[\frac{1}{e^2} E\{D \mid \boldsymbol{X}^{\mathcal{S}}\} E\{Y(1)^2) \mid \boldsymbol{X}^{\mathcal{S}}\} \right] + E\left[\frac{1}{(1-e)^2} E\{(1-D) \mid \boldsymbol{X}^{\mathcal{S}}\} E\{Y(0) \mid \boldsymbol{X}^{\mathcal{S}}\} \right] - \Delta_0^2 \\
&= E\left[\frac{1}{e} E\{Y(1)^2 \mid \boldsymbol{X}^{\mathcal{S}}\} \right] + E\left[\frac{1}{1-e} E\{Y(0)^2 \mid \boldsymbol{X}^{\mathcal{S}}\} \right] - \Delta_0^2 \\
&= E\left[E\left\{ \frac{Y(1)^2}{e} + \frac{Y(0)^2}{1-e} \mid \boldsymbol{X}^{\mathcal{S}} \right\} \right] - \Delta_0^2 \\
&= E\left\{ \frac{Y(1)^2}{e} + \frac{Y(0)^2}{1-e} \right\} - \Delta_0^2.
\end{aligned}$$

The third equality holds because $\boldsymbol{X}^{\mathcal{S}}$ is a sufficient set, so $D \perp\!\!\!\perp Y(d) \mid \boldsymbol{X}^{\mathcal{S}}$. Similarly,

$$\sigma_I^2 = E\left\{ \frac{Y(1)^2}{\tilde{e}} + \frac{Y(0)^2}{1-\tilde{e}} \right\} - \Delta_0^2.$$

Next, we prove $\sigma_I^2 \geq \sigma_0^2$. As $\mathcal{I}' \subset \mathcal{I}$ and (S20), we have: $E[Y(d)^2 \mid \boldsymbol{X}^{\mathcal{S}}] = E[Y(d)^2 \mid \boldsymbol{X}^{\mathcal{S}_I}]$ for $d = 0, 1$. Denote these conditional expectations as $\tilde{Y}(d)$ for $d = 0, 1$ respectively. We have

$$
\begin{aligned}
\sigma_I^2 &= E\left[E\left\{\frac{Y(1)^2}{\tilde{e}} + \frac{Y(0)^2}{1-\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}_I}\right\}\right] - \Delta_0^2 \\
&= E\left[\frac{1}{\tilde{e}} E\{Y(1)^2 \mid \boldsymbol{X}^{\mathcal{S}_I}\}\right] + E\left[\frac{1}{1-\tilde{e}} E\{Y(0)^2 \mid \boldsymbol{X}^{\mathcal{S}_I}\}\right] - \Delta_0^2 \\
&= E\left\{\frac{1}{\tilde{e}} \tilde{Y}(1)\right\} + E\left\{\frac{1}{1-\tilde{e}} \tilde{Y}(0)\right\} - \Delta_0^2 \\
&= E\left[E\left\{\frac{1}{\tilde{e}} \tilde{Y}(1) \mid \boldsymbol{X}^{\mathcal{S}}\right\}\right] + E\left[E\left\{\frac{1}{1-\tilde{e}} \tilde{Y}(0) \mid \boldsymbol{X}^{\mathcal{S}}\right\}\right] - \Delta_0^2 \\
&= E\left\{\tilde{Y}(1) E\left(\frac{1}{\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} + E\left\{\tilde{Y}(0) E\left(\frac{1}{1-\tilde{e}} \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\} - \Delta_0^2 \\
&\geq E\left\{\tilde{Y}(1) \frac{1}{e} + \tilde{Y}(0) \frac{1}{1-e}\right\} - \Delta_0^2 \\
&= E\left[\frac{1}{e} E\{Y(1)^2 \mid \boldsymbol{X}^{\mathcal{S}}\} + \frac{1}{1-e} E\{Y(0)^2 \mid \boldsymbol{X}^{\mathcal{S}}\}\right] - \Delta_0^2 \\
&= E\left[E\left\{\frac{Y(1)^2}{e} + \frac{Y(0)^2}{1-e} \mid \boldsymbol{X}^{\mathcal{S}}\right\}\right] - \Delta_0^2 = \sigma_0^2.
\end{aligned}
$$

So we finish the proof of Part B. $\square$.

2.3 Proof of Theorem 3

Define $\alpha = BCov^2(X, Y \mid D)$, $\hat{\alpha} = BCov_n^2(X, Y \mid D)$, $\alpha_1 = BCov^2(X^{(1)}, Y^{(1)})$ and $\alpha_0 = BCov^2(X^{(0)}, Y^{(0)})$, where $(X^{(d)}, Y^{(d)})$ follow the same distribution as $(X, Y \mid D = d)$ for $d = 0, 1$ respectively. We use $\hat{\alpha}_1 = BCov_{n_1}^2, (X^{(1)}, Y^{(1)})$ $\hat{\alpha}_0 = BCov_{n_0}^2(X^{(0)}, Y^{(0)})$ to denote their sample estimators, respectively. We firstly show that there exists a constant $\tilde{c}$ such that

$$
P(|\alpha - \hat{\alpha}| > cn^{-k}) \leq O_P(\exp(-\tilde{c}n^{1-2k})), \tag{S23}
$$

where $c > 0$ is a constant and $0 < \kappa < 1/2$.

Recall $n_1 = \sum_{i=1}^n D_i$, $n_0 = n - n_1$, $\omega = P(D = 1)$ and $\hat{\omega} = n_1/n$.

We can write

$$\begin{aligned} \alpha - \hat{\alpha} &= \omega\alpha_1 + (1-\omega)\alpha_0 - \{\hat{\omega}\hat{\alpha}_1 + (1-\hat{\omega})\hat{\alpha}_0\} \\ &= \omega(\alpha_1 - \hat{\alpha}_1) + (1-\omega)(\alpha_0 - \hat{\alpha}_0) + (\hat{\alpha}_1 - \hat{\alpha}_0)(\omega - \hat{\omega}). \end{aligned}$$

Since $\alpha_1, \alpha_0, \hat{\alpha}_1, \hat{\alpha}_0 \in [0, 1]$, $|\alpha - \hat{\alpha}| \leq \omega|\alpha_1 - \hat{\alpha}_1| + (1-\omega)|\alpha_0 - \hat{\alpha}_0| + |\omega - \hat{\omega}|$. We have

$$\begin{aligned} P(|\alpha - \hat{\alpha}| \geq 2\epsilon) &\leq P(\omega|\alpha_1 - \hat{\alpha}_1| \geq \omega\epsilon) + P((1-\omega)|\alpha_0 - \hat{\alpha}_0| \geq (1-\omega)\epsilon) + P(|\omega - \hat{\omega}| \geq \epsilon) \\ &= P(|\alpha_1 - \hat{\alpha}_1| \geq \epsilon) + P(|\alpha_0 - \hat{\alpha}_0| \geq \epsilon) + P(|\omega - \hat{\omega}| \geq \epsilon). \end{aligned} \tag{S24}$$

We control the three terms respectively. To begin with, we handle the third term of (S24). We note that

$$\omega - \hat{\omega} = \frac{1}{n}\sum_{i=1}^n (\omega - D_i) = \sum_{i=1}^n Z_i,$$

where $Z_i = (\omega - D_i)/n$ are independent zero-mean random variables, and $|Z_i| \leq 1/n = M$, $E(Z_i^2) = \omega(1-\omega)/n^2$. Based on the Bernstein inequality, we have

$$P(\omega - \hat{\omega} \geq \epsilon) = P(\sum_{i=1}^n Z_i \geq \epsilon) \leq \exp\left(-\frac{\frac{1}{2}\epsilon^2}{\frac{\omega(1-\omega)}{n} + \frac{\epsilon}{3n}}\right).$$

So,

$$P(|\hat{\omega} - \omega| \geq \epsilon) = P(\sum_{i=1}^n Z_i \geq \epsilon) + P(-\sum_{i=1}^n Z_i \geq \epsilon) \leq 2\exp\left(-\frac{\frac{1}{2}\epsilon^2}{\frac{\omega(1-\omega)}{n} + \frac{\epsilon}{3n}}\right). \tag{S25}$$

Now we control the first and the second terms of (S24). Following equation (A.7) from the appendix of Pan et al. (2019), there exist two positive constants c_1 and c_2 such that

$$\begin{aligned} P(|\alpha_1 - \hat{\alpha}_1| \geq \epsilon) &\leq 2\exp(-c_1 n_1 \epsilon^2), \\ P(|\alpha_0 - \hat{\alpha}_0| \geq \epsilon) &\leq 2\exp(-c_0 n_0 \epsilon^2). \end{aligned}$$

We now show that $\exp(-c_1 n_1 \epsilon^2) = O_P(\exp(-c_1 n\omega\epsilon^2/2))$. As $\omega = P(D=1) > 0$, we have

$$\begin{aligned} &P\left(\left|\frac{\exp(-c_1 n_1\epsilon^2)}{\exp(-c_1\omega n\epsilon^2/2)}\right| > 1\right) = P(\frac{n\omega}{2} - n_1 > 0) = P(\omega - \hat{\omega} > \frac{\omega}{2}) \\ &= P(\sum_{i=1}^{n} Z_i > \frac{\omega}{2}) \leq \exp\left(-\frac{\frac{1}{8}\omega^2}{\frac{\omega(1-\omega)}{n} + \frac{\omega}{6n}}\right) \overset{n\to\infty}{\longrightarrow} 0. \end{aligned} \tag{S26}$$

Similarly, we have

$$\exp(-c_0 n_0 \epsilon^2) = O_P(\exp(-c_0 n(1-\omega)\epsilon^2/2)). \tag{S27}$$

When $\epsilon < 3\omega(1-\omega)$, we have

$$\exp\left(-\frac{\frac{1}{2}\epsilon^2}{\frac{\omega(1-\omega)}{n} + \frac{\epsilon}{3n}}\right) = \exp\left(-\frac{1}{2\omega(1-\omega) + 2\epsilon/3} n\epsilon^2\right) \leq \exp(-\tilde{c}_2 n\epsilon^2), \tag{S28}$$

where $\tilde{c}_2 = 1/\{4\omega(1-\omega)\}$. Let $\epsilon = cn^{-\kappa}/2$, where $0 < \kappa < 1/2$, $\tilde{c}_1 = c_1\omega/2$, $\tilde{c}_0 = c_0(1-\omega)/2$, combining (S25)–(S28), we have

$$\begin{aligned} P(|\alpha - \hat{\alpha}| \geq cn^{-\kappa}) \leq\ & O_P(\exp(-\tilde{c}_1 cn^{1-2\kappa})) + O_P(\exp(-\tilde{c}_0 cn^{1-2\kappa})) \\ & + O_P(\exp(-\tilde{c}_2 cn^{1-2\kappa})). \end{aligned}$$

Let $\tilde{c} = \min(c\tilde{c}_1, c\tilde{c}_0, c\tilde{c}_2)$, we have

$$P(|\alpha - \hat{\alpha}| \geq cn^{-\kappa}) \leq O_P(\exp(-\tilde{c}n^{1-2\kappa})).$$

Hence we finish the proof of equation (S23). Now let $\rho_j = BCov^2(X^{(j)}, Y \mid D)$ and $\hat{\rho}_j = BCov_n^2(X^{(j)}, Y \mid D)$ for $j = 1, 2, \ldots, p$. From equation (S23) we know that $P(|\hat{\rho}_j - \rho_j| > cn^{-\kappa}) = O_P(\exp(-c_1 n^{1-2\kappa}))$.

As $\tau_n = cn^{-\kappa}$ and $\{(\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}}) \not\subset \mathcal{K}\} \subset \{|\hat{\rho}_j - \rho_j| > cn^{-\kappa},$ for some $j \in (\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}})\}$, we have

$$P(\{(\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}}) \subset \mathcal{K}\}) \geq 1 - \eta P(|\hat{\rho}_j - \rho_j| > cn^{-\kappa}) \geq 1 - \eta O(\exp(-\tilde{c}n^{1-2\kappa})),$$

where η is the cardinality of $(\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}})$. Hence

$$P(\{(\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}}) \subset \mathcal{K}\}) \to 1.$$

2.4 Proof of Theorem 4

We assume p is fixed while n tends to infinity. In the following, we use $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ to denote the solution of

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left\{ D_i \log\left(\frac{1 - e(\boldsymbol{X}_i; \boldsymbol{\beta})}{e(\boldsymbol{X}_i; \boldsymbol{\beta})} \right) - \log\left(1 - e(\boldsymbol{X}_i; \boldsymbol{\beta})\right) \right\} + \lambda_n \sum_{j=1}^{p} \frac{1}{\hat{\omega}_j} |\beta_j| \right].$$

For parts (a) and (b), the proof is very similar to the ones in the proof of Theorem 1 in Shortreed and Ertefaie (2017), so we omit the proof here.

We will now prove the oracle efficiency results in part (c).

We use $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$ to denote the coefficient estimates when we use $\mathcal{A}$ as a prior, and we set $\tilde{\boldsymbol{\beta}}^{\mathcal{A}^c} = \mathbf{0}$. Let $\hat{e}_i = e(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})$, $\tilde{e}_i = e(\boldsymbol{X}_i; \tilde{\boldsymbol{\beta}})$. We use Δ_{HT}, Δ_{Ratio}, Δ_{DR} to denote IPW estimators (1), (2), (3) respectively. Without loss of generality, we assume $\mathcal{A} = \{1, 2, 3, \ldots, p_0\}$. Denote $\hat{\Delta}_{HT}$, $\hat{\Delta}_{Ratio}$ and $\hat{\Delta}_{DR}$ the IPW estimator by plugging $\hat{e}_i$ in (1), (2), (3), respectively; and $\tilde{\Delta}_{HT}$, $\tilde{\Delta}_{Ratio}$ and $\tilde{\Delta}_{DR}$ the IPW estimator by plugging $\tilde{e}_i$ in (1), (2), (3), respectively. Since we focus on Logit propensity score model, we use $e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$ instead of $e(\boldsymbol{X}; \boldsymbol{\beta})$ to denote PS model, that is Logit$\{e(\boldsymbol{X}; \boldsymbol{\beta}) =$ Logit$\{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$.

We will prove the following results:

$$\sqrt{n}(\hat{\Delta}_{HT} - \tilde{\Delta}_{HT}) \xrightarrow{p} 0,$$
$$\sqrt{n}(\hat{\Delta}_{Ratio} - \tilde{\Delta}_{Ratio}) \xrightarrow{p} 0,$$
$$\sqrt{n}(\hat{\Delta}_{DR} - \tilde{\Delta}_{DR}) \xrightarrow{p} 0.$$

We first show

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \xrightarrow{p} 0. \tag{S29}$$

From part (a) we know that for any $j \notin \mathcal{A}$, $\lim_{n\to\infty} P(\hat{\beta}_j \neq 0) = 0$. And when we estimate $\boldsymbol{\beta}$, we only use variables in $\mathcal{A}$, thus $\tilde{\beta}_j = 0$ for any $j \notin \mathcal{A}$. One has

$$\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^{\mathcal{A}^c} - \tilde{\boldsymbol{\beta}}^{\mathcal{A}^c} \neq 0) = 0,$$

which implies $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathcal{A}^c} - \tilde{\boldsymbol{\beta}}^{\mathcal{A}^c}) \xrightarrow{p} 0$. We then show $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}) \xrightarrow{d} 0$. By the KKT conditions, we have

$$|\sum_{i=i}^{n} X_i^{(j)} \left\{ D_i - e(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}) \right\}| \leq \frac{\lambda_n}{\hat{\omega}_j^{(n)}},$$
$$\sum_{i=i}^{n} X_i^{(j)} \left\{ D_i - e(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) \right\} = 0,$$

where $j \in \mathcal{A}$, $e = e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$ is the logistic regression model we specified before. One has

$$\frac{1}{\sqrt{n}} |\sum_{i=1}^{n} X_i^{(j)} \{ e(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}) - e(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) \}| \leq \frac{\lambda_n}{\sqrt{n}\hat{\omega}_j^{(n)}}. \tag{S30}$$

Denote $\boldsymbol{\beta}_0$ the true coefficient of the oracle propensity score model, which satisfies $e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0) = P(D = 1 \mid \boldsymbol{X}_{\mathcal{A}})$. Let

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \frac{\hat{\boldsymbol{u}}}{\sqrt{n}},$$
$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \frac{\tilde{\boldsymbol{u}}}{\sqrt{n}},$$
$$\boldsymbol{u} = \hat{\boldsymbol{u}} - \tilde{\boldsymbol{u}}.$$

By Taylor expansion at the point $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$:

$$e(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}) = e(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0) + e'(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \frac{\hat{\boldsymbol{u}}}{\sqrt{n}} + e''(U_i) \frac{(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{u}})^2}{n},$$
$$e(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) = e(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0) + e'(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \frac{\tilde{\boldsymbol{u}}}{\sqrt{n}} + e''(V_i) \frac{(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{u}})^2}{n},$$

where U_i is between $\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$, and V_i is between $\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$. The lefthand side of (S30) can be written as $A_{1j}^{(n)} + A_{2j}^{(n)}$, where

$$
\begin{aligned}
A_{1j}^{(n)} &= \sum_{i=1}^{n} \frac{X_i^{(j)}}{n} e'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{u},\\
A_{2j}^{(n)} &= \sum_{i=1}^{n} \frac{X_i^{(j)}}{n^{3/2}} \left\{e''(U_i)(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{u}})^2 - e''(V_i)(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{u}})^2\right\},
\end{aligned}
$$

We can rewrite (S30) as vector forms

$$
\begin{aligned}
&\frac{\lambda_n}{\sqrt{n}}\boldsymbol{w} \geq |\boldsymbol{A}_{\mathbf{1}}^{(n)} + \boldsymbol{A}_{\mathbf{2}}^{(n)}|,\\
&\boldsymbol{A}_1^{(n)} = \frac{1}{n}\sum_{i=1}^{n} e'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{X}_i^{\mathcal{A}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{u},\\
&\boldsymbol{A}_2^{(n)} = \frac{1}{n^{3/2}}\sum_{i=1}^{n} \boldsymbol{X}_i^{\mathcal{A}}\left\{e''(U_i)(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{u}})^2 - e''(V_i)(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{u}})^2\right\},
\end{aligned}
$$

where $\boldsymbol{w} = (1/\hat{\omega}_1^{(n)}, \ldots, 1/\hat{\omega}_{p_0}^{(n)})^{\mathrm{T}}$, $\boldsymbol{X}_i^{\mathcal{A}} = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p0)})^{\mathrm{T}}$. When $n \to \infty$, for any $j \in \mathcal{A}$, $\hat{\omega}_j^{(n)} \xrightarrow{p} c_j > 0$ and $\lambda_n/\sqrt{n} \xrightarrow{p} 0$, by the Continuous mapping theorem, we have $\lambda_n\boldsymbol{w}/\sqrt{n} \xrightarrow{p} 0$. We also have

$$
\frac{1}{n}\sum_{i=1}^{n} e'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{X}_i^{\mathcal{A}}\boldsymbol{X}_i^{\mathrm{T}} \xrightarrow{p} E(e'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{X}^{\mathcal{A}}\boldsymbol{X}^{\mathrm{T}}).
$$

As $\boldsymbol{u}^{\mathcal{A}^c} = \sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathcal{A}^c} - \tilde{\boldsymbol{\beta}}^{\mathcal{A}^c}) \xrightarrow{p} \mathbf{0}$, if we can show $\boldsymbol{A}_2^{(n)} \xrightarrow{p} \mathbf{0}$, by Slutsky's Theorem, we have $\boldsymbol{u}^{\mathcal{A}} \xrightarrow{p} \mathbf{0}$. We will then show $\boldsymbol{A}_2^{(n)} \xrightarrow{p} \mathbf{0}$. More precisely, we will show that for all $j \in \mathcal{A}$, we have $\sum_{i=1}^{n} X_i^{(j)} e''(V_i)(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{u}})^2/n^{3/2} \xrightarrow{p} 0$ and $\sum_{i=1}^{n} X_i^{(j)} e''(V_i)(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{u}})^2/n^{3/2} \xrightarrow{p} 0$.

We note that for logistic model, we have $0 < |e| < 1$, $|e'| = |e(1-e)| < 1$, $|e''| = |e(1-e)(1-2e)| < 1$. We note the maximum likelihood estimator is asymptotically normal, so we have $\tilde{\boldsymbol{u}} = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \hat{\Sigma})$. And from part (b) we have $\hat{\boldsymbol{u}} = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \Sigma^*)$. We now show that $\sum_{i=1}^n X_i^{(j)} e''(U_i)(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{u}})^2 / n^{3/2} \xrightarrow{p} 0$. We observe that

$$
\begin{aligned}
&\frac{1}{n^{3/2}} \sum_{i=1}^n |X_i^{(j)} e''(V_i)(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{u}})^2| \\
&\leq \sum_{i=1}^n \frac{|X_i^{(j)}||\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i|}{n} \frac{|\hat{\boldsymbol{u}}^{\mathrm{T}} \hat{\boldsymbol{u}}|}{n^{1/2}} \\
&\leq \sum_{i=1}^n \frac{|X_i^{(j)}||\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i|}{n} \frac{|\hat{\boldsymbol{u}}^{\mathrm{T}}|}{n^{1/4}} \frac{|\hat{\boldsymbol{u}}|}{n^{1/4}}.
\end{aligned}
$$

By the Weak Law of Large Numbers, we have $\sum_{i=1}^n |X_i^{(j)}||\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i|/n \xrightarrow{p} E(|X_i||\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}|) < \infty$. Besides, we have $|\hat{\boldsymbol{u}}|/n^{1/4} \xrightarrow{p} \mathbf{0}$. Thus, by the Continuous mapping theorem, we see that $\sum_{i=1}^n X_i^{(j)} e''(U_i)(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{u}})^2 / n^{3/2} \xrightarrow{p} 0$. Similarly, we have $\sum_{i=1}^n X_i^{(j)} e''(V_i)(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{u}})^2 / n^{3/2} \xrightarrow{p} 0$. So far, we have finished the proof for (S29).

Now we will show that the CBS propensity score estimator, $\hat{e}_i$, and the oracle propensity score estimator, $\tilde{e}_i$ are asymptotically equivalent, i.e.

$$
\sqrt{n}\{e(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}) - e(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}})\} \xrightarrow{p} 0.
$$

By Taylor expansion at the point $\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}$, we have $e(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}) = e(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) +$ $e'(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u} / \sqrt{n} + e''(T_i)(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u})^2 / n$, where T_i is between $\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}$. We have

$$
\begin{aligned}
\sqrt{n}|e(\boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}) - e(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}})| &= \left| e'(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u} + \frac{e''(T_i)(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u})^2}{\sqrt{n}} \right| \\
&\leq |e'(\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u}| + \left| \frac{e''(T_i)(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u})^2}{\sqrt{n}} \right| \\
&\leq |\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u}| + \left| \frac{(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u})^2}{\sqrt{n}} \right| \\
&\leq |\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{u}| + \frac{(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i)(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{u})}{\sqrt{n}}.
\end{aligned}
$$

As $\boldsymbol{u} \xrightarrow{p} 0$, we have $\sqrt{n}|e(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - e(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})| \xrightarrow{p} 0$.

We will now present the proof for HT, Ratio, and DR estimators respectively.

(i) Proof For Horvitz-Thompson Estimator:

We have

$$
\begin{aligned}
&\sqrt{n}|\hat{\Delta}_{HT} - \tilde{\Delta}_{HT}| \\
&= \left|\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left\{Y_iD_i\left(\frac{1}{\hat{e}_i} - \frac{1}{\tilde{e}_i}\right) - Y_i(1-D_i)\left(\frac{1}{1-\hat{e}_i} - \frac{1}{1-\tilde{e}_i}\right)\right\}\right| \\
&\leq \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left|\frac{Y_iD_i}{\hat{e}_i\tilde{e}_i} - \frac{Y_i(1-D_i)}{(1-\hat{e}_i)(1-\tilde{e}_i)}\right| \cdot |\tilde{e}_i - \hat{e}_i| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\{\left|\frac{Y_iD_i}{\hat{e}_i^2}\right| + \left|\frac{Y_iD_i}{\tilde{e}_i^2}\right| + \left|\frac{Y_i(1-D_i)}{(1-\hat{e}_i)^2}\right| + \left|\frac{Y_i(1-D_i)}{(1-\tilde{e}_i)^2}\right|\right\} \cdot \sqrt{n}|\tilde{e}_i - \hat{e}_i| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\{\left|\frac{Y_iD_i}{\hat{e}_i^2}\right| + \left|\frac{Y_iD_i}{\tilde{e}_i^2}\right| + \left|\frac{Y_i(1-D_i)}{(1-\hat{e}_i)^2}\right| + \left|\frac{Y_i(1-D_i)}{(1-\tilde{e}_i)^2}\right|\right\} \cdot \left(|\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{u}| + \frac{(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})}{\sqrt{n}}\right),
\end{aligned}
\tag{S31}
$$

where $\hat{e}_i = e(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, $\tilde{e}_i = e(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})$.

We now show that

$$
\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_iD_i\boldsymbol{X}_{\boldsymbol{i}}^{\mathrm{T}}}{\tilde{e}_i^2}\right| \xrightarrow{p} E\left\{\frac{|YD\boldsymbol{X}^{\mathrm{T}}|}{e(\boldsymbol{X}\boldsymbol{\beta}_0)^2}\right\}, \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_iD_i\boldsymbol{X}_{\boldsymbol{i}}^{\mathrm{T}}}{\hat{e}_i^2}\right| \xrightarrow{p} E\left\{\frac{|YD\boldsymbol{X}^{\mathrm{T}}|}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2}\right\}.
$$

We define $L(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n}|Y_iD_i\boldsymbol{X}_i^{\mathrm{T}}|/e(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2$, $D(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - L(\boldsymbol{\beta}_0)$. Assume $\boldsymbol{\beta}$ is a consistent estimator of $\boldsymbol{\beta}_0$, we have

$$
D(\boldsymbol{\beta}) = \left\{\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} + o_p(1)\right\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}
$$

So if $\boldsymbol{\beta} \xrightarrow{p} \boldsymbol{\beta}_0$, we have $D(\boldsymbol{\beta}) \xrightarrow{p} 0$. Thus we have $D(\hat{\boldsymbol{\beta}}) \xrightarrow{p} 0$ and $D(\tilde{\boldsymbol{\beta}}) \xrightarrow{p} 0$, which imply $(1/n)\sum_{i=1}^{n}|Y_iD_i\boldsymbol{X}_i^{\mathrm{T}}|/e(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2 = L(\hat{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}_0) + D(\hat{\boldsymbol{\beta}}) \xrightarrow{p} E\{|YD\boldsymbol{X}^{\mathrm{T}}|/e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2\}$ and

$(1/n)\sum_{i=1}^{n}|Y_iD_i\boldsymbol{X}_i^{\mathrm{T}}|/e(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})^2 = L(\tilde{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}_0) + D(\tilde{\boldsymbol{\beta}}) \xrightarrow{p} E\{|YD\boldsymbol{X}^{\mathrm{T}}|/e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2\}$. The following relationships could be shown analogously:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i(1-D_i)\boldsymbol{X}_i^{\mathrm{T}}|}{(1-\hat{e}_i)^2} \xrightarrow{p} E\left[\frac{|Y(1-D)\boldsymbol{X}^{\mathrm{T}}|}{\{1-e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2}\right],$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i(1-D_i)\boldsymbol{X}_i^{\mathrm{T}}|}{(1-\tilde{e}_i)^2} \xrightarrow{p} E\left[\frac{|Y(1-D)\boldsymbol{X}^{\mathrm{T}}|}{\{1-e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2}\right],$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_iD_i\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i|}{\tilde{e}_i^2} \xrightarrow{p} E\left\{\frac{|YD\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}|}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2}\right\},$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_iD_i\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i|}{\tilde{e}_i^2} \xrightarrow{p} E\left\{\frac{|YD\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}|}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2}\right\},$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i(1-D_i)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}|}{(1-\hat{e}_i)^2} \xrightarrow{p} E\left[\frac{|Y(1-D)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}|}{\{1-e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2}\right],$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i(1-D_i)\boldsymbol{X}\boldsymbol{X}_i^{\mathrm{T}}|}{(1-\tilde{e}_i)^2} \xrightarrow{p} E\left[\frac{|Y(1-D)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}|}{\{1-e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2}\right]$$

Given condition B3, these expectations are finite. As $\boldsymbol{u} \xrightarrow{p} 0$, by the Continuous mapping theorem, we can conclude that the righthand of (S31) converges to zero in probability, which implies $\sqrt{n}(\hat{\Delta}_{HT} - \tilde{\Delta}_{HT}) \xrightarrow{p} 0$. The result indicates $\sqrt{n}(\hat{\Delta}_{HT} - \Delta_0) \xrightarrow{p} N(0, \sigma_{HT}^2)$.

(ii) Proof For Ratio Estimator:

We have

$$\begin{aligned}
&\sqrt{n}(\hat{\Delta}_{Ratio} - \tilde{\Delta}_{Ratio}) \\
&= \sqrt{n}\left(\sum_{i=1}^{n}\frac{D_i}{\hat{e}_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{D_iY_i}{\hat{e}_i}\right) - \sqrt{n}\left(\sum_{i=1}^{n}\frac{D_i}{\tilde{e}_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{D_iY_i}{\tilde{e}_i}\right) \\
&- \sqrt{n}\left(\sum_{i=1}^{n}\frac{1-D_i}{1-\hat{e}_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{e}_i}\right) + \sqrt{n}\left(\sum_{i=1}^{n}\frac{1-D_i}{1-\tilde{e}_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\tilde{e}_i}\right) \\
&= B_1^{(n)} + B_2^{(n)} + B_3^{(n)} + B_4^{(n)},
\end{aligned}$$

where

$$B_1^{(n)} = \sqrt{n}\left(\sum_{i=1}^{n}\frac{D_i}{\hat{e}_i}\right)^{-1}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{e}_i}-\frac{D_iY_i}{\tilde{e}_i}\right),$$

$$B_2^{(n)} = \sqrt{n}\left(\sum_{i=1}^{n}\frac{D_iY_i}{\tilde{e}_i}\right)\left\{\left(\sum_{i=1}^{n}\frac{D_i}{\hat{e}_i}\right)^{-1}-\left(\sum_{i=1}^{n}\frac{D_i}{\tilde{e}_i}\right)^{-1}\right\},$$

$$B_3^{(n)} = \sqrt{n}\left(\sum_{i=1}^{n}\frac{1-D_i}{1-\hat{e}_i}\right)^{-1}\sum_{i=1}^{n}\left\{\frac{(1-D_i)Y_i}{1-\tilde{e}_i}-\frac{(1-D_i)Y_i}{1-\hat{e}_i}\right\},$$

$$B_4^{(n)} = \sqrt{n}\left\{\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\tilde{e}_i}\right\}\left\{\left(\sum_{i=1}^{n}\frac{1-D_i}{1-\tilde{e}_i}\right)^{-1}-\left(\sum_{i=1}^{n}\frac{1-D_i}{1-\hat{e}_i}\right)^{-1}\right\}.$$

Here we only show that $B_1^{(n)}, B_2^{(n)} \xrightarrow{p} 0$, the proof for $B_3^{(n)}, B_4^{(n)} \xrightarrow{p} 0$ is similar so we simply omit it. For $B_1^{(n)}$ and $B_2^{(n)}$,

$$B_1^{(n)} = \left(\frac{1}{n}\sum_{i}^{n}\frac{D_i}{\hat{e}_i}\right)^{-1}\cdot\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{e}_i}-\frac{D_iY_i}{\tilde{e}_i}\right),$$

$$B_2^{(n)} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\tilde{e}_i}\right)\cdot\left(\frac{1}{n}\sum_{i}^{n}\frac{D_i}{\hat{e}_i}\right)^{-1}\left(\frac{1}{n}\sum_{i}^{n}\frac{D_i}{\tilde{e}_i}\right)^{-1}\cdot\frac{\sqrt{n}}{n}\left(\sum_{i}^{n}\frac{D_i}{\tilde{e}_i}-\frac{D_i}{\hat{e}_i}\right).$$

Firstly, from proof of (i), we know that

$$\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{e}_i}-\frac{D_iY_i}{\tilde{e}_i}\right)\xrightarrow{p}0,\quad \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\hat{e}_i}-\frac{D_i}{\tilde{e}_i}\right)\xrightarrow{p}0. \tag{S32}$$

We will use the same technique in the proof of (i) to show that

$$\frac{1}{n}\sum_{i}^{n}\frac{D_i}{\hat{e}_i}\xrightarrow{p}E\left\{\frac{D}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)}\right\}=1,\quad \frac{1}{n}\sum_{i}^{n}\frac{D_i}{\tilde{e}_i}\xrightarrow{p}E\left\{\frac{D}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)}\right\}=1, \tag{S33}$$

$$\frac{1}{n}\sum_{i}^{n}\frac{D_iY_i}{\tilde{e}_i}\xrightarrow{p}E\left\{\frac{DY}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)}\right\}=E\{Y(1)\}. \tag{S34}$$

With (S32), (S33) and (S34), by the Continuous mapping theorem, we can conclude that $B_1^{(n)}, B_2^{(n)} \xrightarrow{p} 0$. And analogously we have $B_3^{(n)}, B_4^{(n)} \xrightarrow{p} 0$. So $\sqrt{n}(\hat{\Delta}_{Ratio} - \tilde{\Delta}_{Ratio}) \xrightarrow{p} 0$, which implies $\sqrt{n}(\hat{\Delta}_{Ratio} - \Delta_0) \xrightarrow{p} N(0, \sigma^2_{Ratio})$.

We define $L(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n D_i/e(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, $D(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - L(\boldsymbol{\beta}_0)$. Assume $\boldsymbol{\beta}$ is a consistent estimator of $\boldsymbol{\beta}_0$, we have

$$D(\boldsymbol{\beta}) = \left\{\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} + o_p(1)\right\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}$$

So if $\boldsymbol{\beta} \xrightarrow{p} \boldsymbol{\beta}_0$, we have $D(\boldsymbol{\beta}) \xrightarrow{p} 0$. Thus we have $D(\hat{\boldsymbol{\beta}}) \xrightarrow{p} 0$ and $D(\tilde{\boldsymbol{\beta}}) \xrightarrow{p} 0$, which imply $(1/n)\sum_{i=1}^n D_i/e(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = L(\hat{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}_0) + D(\hat{\boldsymbol{\beta}}) \xrightarrow{p} E\{D/e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} = 1$ and $(1/n)\sum_{i=1}^n D_i/e(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}) = L(\tilde{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}_0) + D(\tilde{\boldsymbol{\beta}}) \xrightarrow{p} E\{D/e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} = 1$.

So far, we have proved (S33). And we can obtain (S34) using similar arguments.

(iii) Proof For Doubly Robust Estimator:

From the definition of doubly robust estimator (3), we know that

$$\Delta_{DR} = \Delta_{HT} - \frac{1}{n}\sum_{i=1}^n (D_i - e_i)\left(\frac{b_1(\boldsymbol{X}_i)}{e_i} + \frac{b_0(\boldsymbol{X}_i)}{1 - e_i}\right).$$

Thus we have

$$\begin{aligned}\sqrt{n}(\hat{\Delta}_{DR} - \tilde{\Delta}_{DR}) &= \sqrt{n}(\hat{\Delta}_{HT} - \tilde{\Delta}_{HT}) - \frac{\sqrt{n}}{n}\sum_{i=1}^n (D_i - \hat{e}_i)\left(\frac{\hat{b}_1(\boldsymbol{X}_i)}{\hat{e}_i} + \frac{\hat{b}_0(\boldsymbol{X}_i)}{1 - \hat{e}_i}\right) \\ &\quad + \frac{\sqrt{n}}{n}\sum_{i=1}^n (D_i - \tilde{e}_i)\left(\frac{\tilde{b}_1(\boldsymbol{X}_i)}{\tilde{e}_i} + \frac{\tilde{b}_0(\boldsymbol{X}_i)}{1 - \tilde{e}_i}\right) \\ &= C_n^{(1)} + C_n^{(2)} + C_n^{(3)},\end{aligned}$$

where

$$
\begin{aligned}
C_n^{(1)} &= \sqrt{n}(\hat{\Delta}_{HT} - \tilde{\Delta}_{HT}), \\
C_n^{(2)} &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}(D_i - \hat{e}_i)\left(\frac{-\hat{b}_1(\boldsymbol{X}_i) + \tilde{b}_1(\boldsymbol{X}_i)}{\hat{e}_i} + \frac{-\hat{b}_0(\boldsymbol{X}_i) + \tilde{b}_0(\boldsymbol{X}_i)}{1-\hat{e}_i}\right), \\
C_n^{(3)} &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}(D_i - \tilde{e}_i)\left(\frac{\tilde{b}_1(\boldsymbol{X}_i)}{\tilde{e}_i} + \frac{\tilde{b}_0(\boldsymbol{X}_i)}{1-\tilde{e}_i}\right) - \frac{\sqrt{n}}{n}\sum_{i=1}^{n}(D_i - \hat{e}_i)\left(\frac{\tilde{b}_1(\boldsymbol{X}_i)}{\hat{e}_i} + \frac{\tilde{b}_0(\boldsymbol{X}_i)}{1-\hat{e}_i}\right) \\
&= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\frac{1}{\tilde{e}_i} - \frac{1}{\hat{e}_i}\right)D_i\tilde{b}_1(\boldsymbol{X}_i) - \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\frac{1}{1-\tilde{e}_i} - \frac{1}{1-\hat{e}_i}\right)(1-D_i)\tilde{b}_0(\boldsymbol{X}_i).
\end{aligned}
$$

From (i), we know that $C_n^{(1)} \xrightarrow{p} 0$. We note that for any $\epsilon > 0$,

$$
P(|C_n^{(2)}| > \epsilon) \le P(C_n^{(2)} \neq 0) \le P(\mathcal{A}_n \neq \mathcal{A}) \to 0.
$$

$P(\mathcal{A}_n \neq \mathcal{A}) \to 0$ as our variable selection procedure is consistent. We only need that $C_n^{(3)} \xrightarrow{p} 0$, then we finish the proof of theorem 4. To show this, notice that

$$
\begin{aligned}
|C_n^{(3)}| \le & \frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_i}{\tilde{e}_i\hat{e}_i}\right| \cdot \sqrt{n}|\hat{e}_i - \tilde{e}_i| \cdot |\tilde{b}_1(X_i)| \\
& + \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1-D_i}{(1-\tilde{e}_i)(1-\hat{e}_i)}\right| \cdot \sqrt{n}|\hat{e}_i - \tilde{e}_i| \cdot |\tilde{b}_0(X_i)| \\
& \le \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{e}_i^2} + \frac{1}{\tilde{e}_i^2}\right) \cdot f_1(\boldsymbol{X}_i) \cdot g_1(\tilde{\boldsymbol{\alpha}}_1) \cdot \left(|\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{u}| + \frac{(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})}{\sqrt{n}}\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{(1-\hat{e}_i)^2} + \frac{1}{(1-\hat{e}_i)^2}\right) \cdot f_0(\boldsymbol{X}_i) \cdot g_0(\tilde{\boldsymbol{\alpha}}_0) \cdot \left(|\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{u}| + \frac{(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})}{\sqrt{n}}\right) \\
& = D_n^{(1)} + D_n^{(2)}.
\end{aligned}
$$

We will use the same technique in the proof of (i) to show that

$$\frac{1}{n}\sum_{i}^{n}\frac{\boldsymbol{X}_i^{\mathrm{T}}f_1(\boldsymbol{X}_i)}{\hat{e}_i^2}\xrightarrow{p}E\left(\frac{\boldsymbol{X}^{\mathrm{T}}f_1(\boldsymbol{X})}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2}\right),\ \ \frac{1}{n}\sum_{i}^{n}\frac{\boldsymbol{X}_i^{\mathrm{T}}f_1(\boldsymbol{X}_i)}{\tilde{e}_i^2}\xrightarrow{p}E\left(\frac{\boldsymbol{X}^{\mathrm{T}}f_1(\boldsymbol{X})}{e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2}\right). \tag{S35}$$

$$\frac{1}{n}\sum_{i}^{n}\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i f_0(\boldsymbol{X}_i)}{(1-\hat{e}_i)^2}\xrightarrow{p}E\left(\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}f_0(\boldsymbol{X})}{\{1-e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2}\right), \tag{S36}$$

$$\frac{1}{n}\sum_{i}^{n}\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i f_0(\boldsymbol{X}_i)}{(1-\tilde{e}_i)^2}\xrightarrow{p}E\left(\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}f_0(\boldsymbol{X})}{\{1-e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2}\right). \tag{S37}$$

By condition (B4) and the continuous mapping theorem, we can conclude that $g_d(\tilde{\boldsymbol{\alpha}}_d)\xrightarrow{p}g_d(\boldsymbol{\alpha}_d)$. By condition (B5), we know that these expectations are finite. Combining (S35), (S36), (S37) and $\boldsymbol{u}\xrightarrow{p}\mathbf{0}$, by the Slutsky's Theorem have $D_n^{(1)}\xrightarrow{p}0$. And analogously we have $D_n^{(2)}\xrightarrow{p}0$. So $\sqrt{n}(\hat{\Delta}_{DR}-\tilde{\Delta}_{DR})\xrightarrow{p}0$, which implies $\sqrt{n}(\hat{\Delta}_{DR}-\Delta_0)\xrightarrow{p}N(0,\sigma_{DR}^2)$.

We define $L(\boldsymbol{\beta})=(1/n)\sum_{i=1}^{n}\{\boldsymbol{X}_i^{\mathrm{T}}f_1(\boldsymbol{X}_i)\}/e(\boldsymbol{X}_i;\boldsymbol{\beta})^2$, $D(\boldsymbol{\beta})=L(\boldsymbol{\beta})-L(\boldsymbol{\beta}_0)$. For any $\boldsymbol{\beta}\xrightarrow{p}\boldsymbol{\beta}_0$, we have

$$D(\boldsymbol{\beta})=\left\{\frac{\partial L(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}_0}+o_p(1)\right\}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}},$$

which indicates $D(\boldsymbol{\beta})\xrightarrow{p}0$. Thus we have $D(\hat{\boldsymbol{\beta}})\xrightarrow{p}0$ and $D(\tilde{\boldsymbol{\beta}})\xrightarrow{p}0$, which imply $(1/n)\sum_{i=1}^{n}\{\boldsymbol{X}_i^{\mathrm{T}}f_1(\boldsymbol{X}_i)\}/e(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2=L(\hat{\boldsymbol{\beta}})=L(\boldsymbol{\beta}_0)+D(\hat{\boldsymbol{\beta}})\xrightarrow{p}E\{\boldsymbol{X}^{\mathrm{T}}f_1(\boldsymbol{X})/e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2\}$ and $(1/n)\sum_{i=1}^{n}\{\boldsymbol{X}_i^{\mathrm{T}}f_0(\boldsymbol{X}_i)\}/e(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})^2=L(\tilde{\boldsymbol{\beta}})=L(\boldsymbol{\beta}_0)+D(\tilde{\boldsymbol{\beta}})\xrightarrow{p}E\{\boldsymbol{X}^{\mathrm{T}}f_0(\boldsymbol{X})/e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2\}$. So far, we have proved (S35), and we can prove (S36) analogously. □

3 Details in the real data application

3.1 Data usage acknowledgement

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. For up-to-date information, see www.adni-info.org.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Pri-

vate sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

3.2 Details in preprocessing the genetics data

For these genetic data, we applied the following preprocessing technique. The first line quality control steps include (i) call rate check per subject and per Single Nucleotide Polymorphism (SNP) marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy-Weinberg equilibrium test, (v) marker removal by the minor allele frequency, and (vi) population stratification. The second line preprocessing steps include removal of SNPs with (i) more than 5% missing values, (ii) minor allele frequency (MAF) smaller than 10%, and (iii) Hardy-Weinberg equilibrium p-value $< 10^{-6}$. $503,892$ SNPs obtained from 22 chromosomes were included in for further processing. MACH-Admix software (http://www.unc.edu/$\sim$yunmli/MaCH-Admix/) (Liu et al., 2013) is applied to perform genotype imputation, using 1000G Phase I Integrated Release Version 3 haplotypes (http://www.1000genomes.org) (Consortium et al., 2012) as a reference panel. Quality control was also conducted after imputation, excluding markers with (i) low imputation accuracy (based on imputation output R^2), (ii) Hardy-Weinberg equilibrium p-value 10^{-6}, and (iii) minor allele frequency (MAF) $< 5\%$.