BIOMETRICS 000, 000–000 DOI: 000
000 0000

Ultra-high Dimensional Variable Selection for Doubly Robust Causal Inference

Dingke Tang[1,*], Dehan Kong[1,], Wenliang Pan[2,***], Linbo Wang[1,****]**

[1]Department of Statistical Sciences, University of Toronto, ON M5S 3G3, Canada.

[2]Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, China.

**email:* dingke.tang@mail.utoronto.ca

***email:* dehan.kong@utoronto.ca

****email:* panwliang@mail.sysu.edu.cn

*****email:* linbo.wang@utoronto.ca

SUMMARY: Causal inference has been increasingly reliant on observational studies with rich covariate information. To build tractable causal procedures, such as the doubly robust estimators, it is imperative to first extract important features from high or even ultra-high dimensional data. In this paper, we propose causal ball screening for confounder selection from modern ultra-high dimensional data sets. Unlike the familiar task of variable selection for prediction modeling, our confounder selection procedure aims to control for confounding while improving efficiency in the resulting causal effect estimate. Previous empirical and theoretical studies suggest excluding causes of the treatment that are not confounders. Motivated by these results, our goal is to keep all the predictors of the outcome in both the propensity score and outcome regression models. A distinctive feature of our proposal is that we use an outcome model-free procedure for propensity score model selection, thereby maintaining double robustness in the resulting causal effect estimator. Our theoretical analyses show that the proposed procedure enjoys a number of properties, including model selection consistency and point-wise normality. Synthetic and real data analysis show that our proposal performs favorably with existing methods in a range of realistic settings. Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database.

KEY WORDS: Alzheimer's disease; Average causal effect; Ball covariance; Confounder selection; Variable screening.

This paper has been submitted for consideration for publication in *Biometrics*

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.13625

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 conditional on baseline covariates, can be used to remove bias due to observed confounders. Robins et al. (1994) proposed a doubly robust method that combines outcome regression and propensity score modeling. Their estimator has been shown to enjoy favorable theoretical properties under correct specification of the outcome regression and/or propensity score models.

Traditionally, specifications of the propensity score and outcome regression models are typically driven by expert knowledge. However, this is becoming increasingly difficult in modern applications, where researchers are often presented with high or even ultra-high dimensional covariates. For example, a popular database for studying potential risk factors of Alzheimer's disease is available through the Alzheimers Disease Neuroimaging Initiative (ADNI). The ADNI study collects rich covariate information such as clinical and behavioral covariates, and genetic information including millions of SNPs. The dimension of covariates p is much larger than the sample size n. This is known as ultra-high dimensional data in the literature, as classical variable selection methods for high-dimensional data such as the Lasso are not feasible due to computational complexity.

In response to these challenges, there has been a growing interest in developing data-driven procedures for covariate selection in causal inference. A central aim of these methods is to reduce bias and improve efficiency in the final causal effect estimator (Witte and Didelez, 2019). This is in sharp contrast to covariate selection in prediction modeling (e.g. Tibshirani, 1996; Fan and Lv, 2008), where the goal is to find a sparse representation of the association structure with good prediction accuracy. In particular, a good prediction model for the propensity score includes all

strong predictors of the treatment. However, theoretical results and empirical evidence imply that inclusion of variables associated with treatment but not the outcome may inflate the variance of the resulting causal effect estimates (e.g. Brookhart et al., 2006; de Luna et al., 2011; Schnitzer et al., 2016; Rotnitzky and Smucler, 2020). Such variables are commonly referred to as instrumental variables (IVs). In particular, Hahn (1998, 2004) showed that the semiparametric efficiency bound for estimating the average causal effect may be reduced if some covariates are known to be instrumental variables. Although selection of instrumental variables may lead to non-uniform inference (e.g. Leeb and Pötscher, 2005; Moosavi et al., 2021), it has been reported that in many practical settings, selection of instrumental variables improves the estimate in a mean squared error sense (Brookhart et al., 2006); see the end of Section 4 for a detailed discussion.

Motivated by these results, various procedures have been developed for selecting proper variables into the propensity score and outcome regression models. 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 confounder selection methods

relies on correct specification of the outcome regression model. If the same outcome model was used for both PS model selection and causal effect estimation, then the resulting "doubly robust" estimator (Robins et al., 1994) is no longer doubly robust. In particular, if the outcome regression model is incorrect, then the selected PS model may miss important confounders, so that estimates from the PS model, outcome regression model, and hence the "doubly robust" estimator may all be biased. An additional pitfall of existing methods is that none of them is well-suited for covariate selection from an ultra-high dimensional feature set, such as the one collected by the ADNI study. For ultra-high dimensional data, penalization or Bayesian selection methods face challenges in computational cost and estimation accuracy (Fan and Lv, 2008).

In this paper, we propose Causal Ball Screening (CBS), a novel doubly robust causal effect estimating procedure that combines an outcome model-free screening step motivated by the ball covariance (Pan et al., 2018, 2020) with a refined selection and doubly robust estimation step. In contrast to aforementioned approaches that aim to exclude instruments, our proposal for propensity score model selection 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. As a result, the resulting causal effect estimator is doubly robust. Furthermore, to the best of our knowledge, our method is the first in the causal inference literature that applies to ultra-high dimensional settings.

The rest of the article is organized as follows. Section 2 introduces background on the target adjustment set in doubly robust causal effect estimation, and the ball covariance. In Section 3, we introduce our CBS procedure for doubly robust causal effect estimation with ultra-high dimensional covariates. Section 4 provides theoretical justifications for the CBS. Simulation studies in Section 5 compare our proposal with several state-of-art methods in their finite-sample performance. In Section 6, we apply our method to the ADNI study and estimate the causal effect of tau

protein level in cerebrospinal fluid on Alzheimer's behavioral score while accounting for ultra-high dimensional covariates. We end with a brief discussion in Section 7.

2. Background

2.1 *The propensity score*

Following the potential outcome framework, we use D to denote a binary treatment assignment, $X = (X^{(1)}, \ldots, X^{(p)})$ 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 assume that the covariates X are ultra-high dimensional in the following sense.

DEFINITION 1 (Ultra-high dimensionality): We say covariates X are ultra-high dimensional if the number of covariates $p = O\left\{\exp(n^{\iota})\right\}$ for some constant $\iota > 0$, where n is the sample size.

We make the stable unit treatment value assumption (Rubin, 1980).

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 $Y = DY(1)+(1-D)Y(0)$ (VanderWeele and Hernan, 2013). Suppose we observe n independent samples from the joint distribution of (X, D, Y), denoted by $(X_i, D_i, Y_i), i = 1, \ldots, n$. We are interested in estimating the average causal effect (ACE) $\Delta = E\{Y(1) - Y(0)\}$. The ACE can be non-parametrically identified under the following assumptions.

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

ASSUMPTION 3 (Positivity): $0 < c \leqslant P(D = 1 \mid X^{\mathcal{S}}) \leqslant 1 - c < 1$, where $c \in (0, 1)$.

Rosenbaum and Rubin (1983) introduced the notion of propensity score $e(X^{\mathcal{S}}) = P(D = 1 \mid X^{\mathcal{S}})$ and showed that under Assumptions 2 and 3, adjusting for the propensity score is sufficient to remove confounding: $D \perp\!\!\!\perp Y(d) \mid e(X^{\mathcal{S}}), d = 0, 1$.

2.2 *Doubly robust estimation*

Let $\widehat{e}_i = \widehat{e}(X_i^{\mathcal{S}})$ be the estimated propensity score, and $\widehat{b}_d(X_i^{\mathcal{S}})$ be the estimate of outcome regression $E(Y \mid D = d, X_i^{\mathcal{S}})$. The classical doubly robust estimator (Robins et al., 1994) is defined as

$$\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i - (D_i - \widehat{e}_i)\widehat{b}_1(X_i^{\mathcal{S}})}{\widehat{e}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - D_i)Y_i + (D_i - \widehat{e}_i)\widehat{b}_0(X_i^{\mathcal{S}})}{1 - \widehat{e}_i}. \tag{1}$$

It has been shown that $\widehat{\Delta}$ is locally semiparametric efficient in the sense that if both the propensity score and outcome regression model estimates converge to their corresponding true values at sufficiently fast rates, then asymptotically the variance of $\widehat{\Delta}$ achieves the semiparametric efficiency bound (e.g. Chernozhukov et al., 2018). Furthermore, it is consistent if either the outcome regression or the propensity model is correctly specified.

2.3 *Target adjustment set*

We now discuss the target adjustment set of variables to include in the propensity score and outcome regression models, in order to eliminate bias and reduce variance in the resulting doubly robust causal effect estimates. We first divide the covariates into four disjoint subsets under the framework of a causal directed acyclic graph (DAG) (Pearl, 2009). Relevant background on the DAG framework is provided in Section S1 of the Supplementary Material.

Let $X^{\mathcal{C}} = \{X^{(j)} \in pa(Y) : D \text{ and } X^{(j)} \text{ are d-connected given } pa(Y) \setminus \{X^{(j)}\}\}$, where $pa(Y)$ denotes the parents of Y. In the following we shall refer to $X^{\mathcal{C}}$ as confounders, $X^{\mathcal{P}} = pa(Y) \setminus X^{\mathcal{C}}$ as precision variables, $X^{\mathcal{I}} = pa(D) \setminus X^{\mathcal{C}}$ as instrumental variables, and $X^{\mathcal{N}} = X \setminus (X^{\mathcal{C}} \cup X^{\mathcal{P}} \cup X^{\mathcal{I}})$ as null variables. Figure 1 provides the simplest causal diagram associated with these definitions.

[Figure 1 about here.]

REMARK 1: Under Assumptions S1 and S2, our definition of confounders is a special case of the reduced covariate set labelled as Z in de Luna et al. (2011). The instrumental variable is commonly used to estimate causal effects when Assumption 2 may be violated. Under the causal sufficiency assumption in the Supplementary Material, our definition of instrumental variable here coincides with that in the literature (e.g. Wang and Tchetgen Tchetgen, 2018).

If one already adjusts for all the confounders, then conditioning on additional precision variables and instrumental variables still renders the treatment D and potential outcome $Y(d)$ conditional independent; see Proposition S2 in the Supplementary Material and de Luna et al. (2011, Proposition 3) for detailed arguments. Adjusting for the null variables may, however, introduce bias. Consider the causal DAG with $D \rightarrow Y$ and $D \leftarrow X^{\mathcal{I}} \rightarrow X^{collider} \leftarrow X^{\mathcal{P}} \rightarrow Y$. Under our definitions, $X^{collider}$ is a null variable and the empty set is a valid adjustment set. But adjusting for $X^{collider}$ may lead to a biased causal effect estimate due to collider bias (Pearl, 2009).

Previous theoretical and empirical findings suggest that inclusion of instrumental variables in addition to confounding variables in the adjustment set may result in efficiency loss (Hahn, 1998, 2004; Brookhart et al., 2006; de Luna et al., 2011).

PROPOSITION 1: (Hahn, 2004) Let Assumptions 1–3 and Assumptions S1–S2 in the Supplementary Material hold, and suppose we have two restrictions (R.1) $\mathcal{P} \neq \emptyset$ and (R.2) $\mathcal{I} \neq \emptyset$. Then the semiparametric efficiency bound for estimating Δ when (R.1) holds is equal to the bound without the restriction, and the semiparametric efficiency bound for estimating Δ when (R.2) holds is lower than the bound without the restriction.

Following these results, our target adjustment set for propensity score and outcome modeling is $\mathcal{A} = \mathcal{C} \cup \mathcal{P}$. We include the confounders in $\mathcal{A}$ to avoid confounding bias, and exclude null variables to avoid potential collider bias. Motivated by Proposition 1, we shall exclude instruments in $\mathcal{A}$ to reduce variance in the resulting doubly robust causal effect estimator. As inclusion of any subset

of precision variables $\mathcal{P}$ in the adjustment set does not introduce bias or affect efficiency, we do not attempt to exclude precision variables in the adjustment set.

REMARK 2: In the case that the precision variables are not sparse in the covariate set (but the confounders are), one would need to further exclude precision variables in the adjustment set.

2.4 *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 entirely model-free for data in Euclidean spaces, and its empirical version is easy to compute as a test statistic of independence. Furthermore, compared with other measures of dependence between two random variables such as the mutual information (Cover and Thomas, 2012) and distance correlation (Székely et al., 2007), the ball correlation does not require the random variables to have finite moments. It hence provides robustness for data with a heavy-tailed distribution (Pan et al., 2018).

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 induced 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 2: 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 $\delta^Y_{ij,k}, \delta^Y_{ij,kl}, \xi^Y_{ij,klst}$.

PROPOSITION 2: (Separability property, Pan et al., 2020) Let (X_i, Y_i), i = 1, 2, ..., 6 be i.i.d samples from the joint distribution of (X, Y). Then $BCov^2(X, Y) = E(\xi^X_{12,3456}\xi^Y_{12,3456})$.

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 the independent test.

We next introduce the empirical version of ball covariance.

DEFINITION 3: (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. Causal ball screening

In this section, we develop causal ball screening (CBS), a two-step procedure for covariate selection and causal effect estimation. 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., 2018, 2020). The second step is a refined selection and estimation step that excludes the null and instrument variables and estimates the average causal effect.

3.1 *Conditional ball covariance screening*

When the candidate feature set is ultra-high dimensional, a common strategy is to use the sure independence screening procedure 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) (in)dependences:

$$X^{\mathcal{C}} \not\perp\!\!\!\perp Y, \qquad X^{\mathcal{P}} \not\perp\!\!\!\perp Y, \qquad X^{\mathcal{I}} \not\perp\!\!\!\perp Y, \qquad X^{\mathcal{N}} \perp\!\!\!\perp Y; \tag{2}$$

$$X^{\mathcal{C}} \not\perp\!\!\!\perp Y \mid D, \;\; X^{\mathcal{P}} \not\perp\!\!\!\perp Y \mid D, \;\; X^{\mathcal{I}} \not\perp\!\!\!\perp Y \mid D, \;\; X^{\mathcal{N}} \perp\!\!\!\perp Y \mid D. \tag{3}$$

On the surface, it seems that independence screening based on (2) or (3) works equally well. In

practice, however, note that the dependence between $X^{\mathcal{I}}$ and Y after conditioning on D is induced by collider bias. Previous qualitative analyses (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 perform our screening based on (3) under the assumption that the instrumental variables $X^{\mathcal{I}}$ have weaker dependence with the outcome Y after conditioning on the treatment variable D, and hence are more likely to be screened out by the conditional independence screening.

To perform conditional independence screening based on (3), 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 $(X^{(d)}, Y^{(d)}) \stackrel{d}{=} (X, Y \mid D = d), d = 0, 1$.

The conditional 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 subjects who receive treatment and $n_0 = n - n_1$. Let $\widehat{\omega} = n_1/n$ be the empirical estimator of ω.

DEFINITION 4: The empirical conditional ball covariance $BCov_n(X, Y \mid D)$ is defined as the square root of: $BCov_n^2(X, Y \mid D) = \widehat{\omega} \sum_{(i,j,k,l,s,t):D_i,D_j,D_k,D_l,D_s,D_t=1} \xi^X_{ij,klst} \xi^Y_{ij,klst} / n_1^6 + (1 - \widehat{\omega}) \sum_{(i,j,k,l,s,t):D_i,D_j,D_k,D_l,D_s,D_t=0} \xi^X_{ij,klst} \xi^Y_{ij,klst} / n_0^6$.

The following proposition is an extension of Lemma 2.1 in Pan et al. (2018).

PROPOSITION 3: $BCov(X, Y \mid D) = 0 \Leftrightarrow X \perp\!\!\!\perp Y \mid D$.

To perform conditional ball covariance screening, as summarized in the first two steps of Algorithm 1, 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 q baseline covariates with the largest ball covariance into the next step. Let $\mathcal{K}$ be the selected set after the screening step. Without loss of generality, we assume $\mathcal{K} = \{1, 2, \ldots, q\}$.

3.2 *Refined selection and doubly robust estimation*

There may be instrumental variables and null variables remaining in the set $\mathcal{K}$ after the screening step. We now propose a second refined selection step to further exclude these variables. To estimate parameters in the outcome regression models on $b_d(X),\ d = 0, 1$, we use the Lasso estimator (Tibshirani, 1996) that

$$\widehat{\alpha}_{\mathcal{K}}^{(d)} = \underset{\alpha_{\mathcal{K}}}{\operatorname{argmin}} \left\{ \sum_{i:D_i=d} (Y_i - X_{i,\mathcal{K}}^{\mathrm{T}} \alpha_{\mathcal{K}})^2 + \lambda_Y^{(d)} ||\alpha_{\mathcal{K}}||_1 \right\}, d = 0, 1. \tag{4}$$

In practice, we use 10-fold cross-validation to select the tuning parameters $\lambda_Y^{(d)}, d = 0, 1$.

Refined selection for the propensity score model is more involved. As we explained in the introduction, for double robustness of the resulting causal effect estimator, selection of the PS model should not depend on the outcome regression model. So we shall apply the idea of adaptive Lasso (Zou, 2006; Shortreed and Ertefaie, 2017) to our setting. Specifically, let

$$\widehat{\beta}_{\mathcal{K}} = \underset{\beta_{\mathcal{K}}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left[D_i \log \left\{ \frac{1 - e(X_{i,\mathcal{K}}; \beta_{\mathcal{K}})}{e(X_{i,\mathcal{K}}; \beta_{\mathcal{K}})} \right\} - \log \left\{ 1 - e(X_{i,\mathcal{K}}; \beta_{\mathcal{K}}) \right\} \right] + \lambda_D \sum_{j=1}^{q} \frac{1}{\widehat{\omega}_j} |\beta_j| \right), \tag{5}$$

where $\widehat{\omega}_j$ is a nonparametric estimator of the importance of covariate $X^{(j)}$ in the outcome model, λ_D is a tuning parameter and the propensity score follows a logistic regression model that $e(X; \beta) = \text{expit}(X^T\beta) = \exp(X^T\beta)/\{1 + \exp(X^T\beta)\}$. In practice, $\widehat{\omega}_j$ 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 3.1. In the simulations and data analysis, we let $\widehat{w}_j = |\widehat{z}\, BCov_n^2(X^{(j)}, Y \mid D)|^\gamma$, where γ is a tunning parameter and $\widehat{z} = 1/\max_j |BCov_n^2(X^{(j)}, Y \mid D)|$ is a scale constant. We then select the pair of parameters (γ, λ_D) that minimize the weighted absolute mean difference (Shortreed and Ertefaie, 2017)

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

where $\widehat{\tau}_i^{\lambda_D,\gamma} = D_i/\widehat{e}_i^{\lambda_D,\gamma} + (1 - D_i)/(1 - \widehat{e}_i^{\lambda_D,\gamma})$ and $\widehat{e}_i^{\lambda_D,\gamma}$ is the estimated propensity score with pair of parameters (λ_D, γ).

Finally, we use plug-in estimators $\widehat{e}_i = \widehat{e}(X_i; \widehat{\beta})$ and $\widehat{b}_{d,i} = X_i^{\mathrm{T}} \widehat{\alpha}^{(d)}$ to construct a doubly robust

Algorithm 1 Causal Ball Screening: A doubly robust estimator of average causal effect with ultra-high dimensional covariates

Input: $(X_i, Y_i, D_i)_{i=1}^n$ **Output:** $\widehat{\Delta}$

1: For $j = 1, \ldots, p$, calculate $\widehat{\rho}_j = BCov_n^2(X^{(j)}, Y \mid D)$.

2: Select the q variables with the largest $\widehat{\rho}_j$, and denote them as $\mathcal{K}$; without loss of generality, let $\mathcal{K} = \{1, \ldots, q\}$.

3: For $d = 0, 1$, set $\{\widehat{\alpha}^{(d)}\}^{\mathrm{T}} = \left(\left(\widehat{\alpha}_{\mathcal{K}}^{(d)}\right)^{\mathrm{T}}, 0^{\mathrm{T}}\right)$, where $\widehat{\alpha}_{\mathcal{K}}^{(d)}$ is the Lasso estimator obtained via (4).

4: Set $\widehat{\beta}^{\mathrm{T}} = (\widehat{\beta}_{\mathcal{K}}^{\mathrm{T}}, 0^{\mathrm{T}})$, where $\widehat{\beta}_{\mathcal{K}}$ is the adaptive Lasso estimator obtained via (5).

5: For $i = 1, \ldots, n$, calculate $\widehat{e}_i = \widehat{e}(X_i; \widehat{\beta})$ and $\widehat{b}_{d,i} = b_d(X_i; \widehat{\alpha}^{(d)})$.

6: Plug $\widehat{b}_{d,i}, \widehat{e}_i, d = 0, 1, i = 1, \ldots, n$ into equation (1) to obtain a doubly robust estimator $\widehat{\Delta}$.

estimator of Δ based on equation (1). These procedures are summarized in Steps 3–6 of Algorithm 1.

4. Theoretical Properties

In this section, we study theoretical properties of the proposed CBS procedure as outlined in Algorithm 1. We first show the sure independence screening property, which guarantees that the set $\mathcal{K}$ selected by the conditional ball covariance screening procedure in Section 3.1 includes all the confounders and precision variables with high probability. The following two assumptions are common in the sure independence screening literature.

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

(A2): (Ultra-high dimensional covariates) $\log(p) = o(n^{1-2\kappa})$, where κ is defined in (A1).

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 exponentially with the sample size.

Theorem 1 shows that we may select the top q variables with the largest $\widehat{\rho}_j$.

THEOREM 1: *(Sure independence screening property) Assume that* $W = \{j : \rho_j = 0\}$ *and* $|W^c| \leqslant q$, *where* $|\cdot|$ *denotes the cardinality. Then under conditions (A1) and (A2), we have* $P(\max_{j\in W} \widehat{\rho}_j < \min_{j\in\mathcal{A}} \widehat{\rho}_j) \to 1$ *and hence* $P((\boldsymbol{X}^{\mathcal{C}} \cup \boldsymbol{X}^{\mathcal{P}}) \subset \mathcal{K}) \to 1$ *as* $n \to \infty$.

We then present theoretical guarantees for the variable selection and estimation step. We assume that the data-driven weights $\widehat{w}_j$ and tuning parameters satisfy the following conditions:

(B1): (Convergence inside the target set) For each $j \in \mathcal{A}$, $\widehat{w}_j \xrightarrow{p} c_j$, 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 $j \in \mathcal{A}^c$, $\widehat{w}_j = O_p(n^{-s})$;

(B3): $\lambda_D/\sqrt{n} \to 0$, $\lambda_D n^{s-1} \to \infty$, and for $d = 0, 1$, $\lambda_Y^{(d)}/n^{1/2} \to \infty$, $\lambda_Y^{(d)}/n^{2/3} \to 0$;

(B4): $X_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}}/n \longrightarrow C$, where C is a positive definite matrix;

(B5): $\mathcal{A} \subset \mathcal{K}$.

Conditions (B1) and (B2) are rate conditions on the data-driven weight $\widehat{w}_j$ that are standard in the adaptive Lasso literature (e.g. Zou, 2006). Condition (B3) requires that the tuning parameters $\lambda_D, \lambda_Y^{(d)}, d = 0, 1$ satisfy some rate conditions. Condition (B4) holds as long as $X_i's$ are i.i.d. with finite second moments. Condition (B5) assumes that the set $\mathcal{K}$ we select in the screening step includes all the variables in the target adjustment set. A necessary condition for (B5) is that the cardinality of the target adjustment set $\mathcal{A}$ is no larger than q. Under this condition, due to Corollary 1, Condition (B5) holds with high probability.

THEOREM 2: *Under Conditions (B1) – (B5) and Assumptions 1–3, we have:*

(a). (Variable selection consistency for the outcome model) Let $\widehat{\mathcal{A}}_{OR} = \{j : \widehat{\alpha}_j^{(d)} \neq 0 \ for \ d = 0 \ or \ 1\}$. *Suppose that the outcome regression model satisfies a linear relationship:*

$$Y = DX_{\mathcal{A}}^{\mathrm{T}}\alpha_{\mathcal{A}}^{(1)*} + (1 - D)X_{\mathcal{A}}^{\mathrm{T}}\alpha_{\mathcal{A}}^{(0)*} + \epsilon, \tag{6}$$

where ϵ *is a random noise with mean* 0 *and variance* σ^2. *Then* $\lim_{n\to\infty} P(\widehat{\mathcal{A}}_{OR} = \mathcal{A}) = 1$.

(b). (Variable selection consistency for the propensity score model) Let $\widehat{\mathcal{A}}_{PS} = \{j : \widehat{\beta}_j \neq 0\}$. *If the underlying propensity score model* $e(X_{\mathcal{A}})$ *is such that*

$$e(X_{\mathcal{A}}) = P(D = 1 \mid X_{\mathcal{A}}) = \textit{expit}(X_{\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{A}}^{*}). \tag{7}$$

Then $\lim_{n\to\infty} P(\widehat{\mathcal{A}}_{PS} = \mathcal{A}) = 1$.

(c). (Double robustness) Assume the estimated propensity score model $e(X_i; \widehat{\beta})$ *and outcome model* $b_d(X_i; \widehat{\alpha}^{(d)})$ *converge to some* $e^0(X)$ *and* $b_d^0(X)$ *in the sense that*

$$\frac{1}{n}\sum_{i=1}^{n}\{e(X_i;\widehat{\beta}) - e^0(X_i)\}^2 = o_p(1),\ \frac{1}{n}\sum_{i=1}^{n}\{b_d(X_i;\widehat{\alpha}^{(d)}) - b_d^0(X_i)\}^2 = o_p(1).$$

Then $(\widehat{\Delta} - \Delta) \xrightarrow{p} 0$ *if either* (6) *holds and* $\widehat{\mathcal{A}}_{OR} = \mathcal{A}$, *or* (7) *holds and* $\widehat{\mathcal{A}}_{OR} = \mathcal{A}$.

(d). (Oracle asymptotic distribution) Assume that $\widehat{\mathcal{A}}_{PS} = \widehat{\mathcal{A}}_{OR} = \mathcal{A}$. *Then under models* (6) *and* (7),

1. $\sqrt{n}(\widehat{\Delta} - \Delta) = \sum_{i=1}^{n}\{\phi(Y_i, b_1(X_{i,\mathcal{A}}), b_0(X_{i,\mathcal{A}}), e(X_{i,\mathcal{A}}), D_i, \Delta)\}/\sqrt{n} + o_p(1)$;
2. $\sqrt{n}(\widehat{\Delta} - \Delta)V_{\Delta}^{-1/2} \xrightarrow{d} N(0,1)$;
3. $V_{\Delta} - \widehat{V}_{\Delta} = o_p(1)$;
4. $|P\{\Delta \in [\widehat{\Delta} - c_m\widehat{V}_{\Delta}^{1/2}/\sqrt{n}, \widehat{\Delta} + c_m\widehat{V}_{\Delta}^{1/2}/\sqrt{n}]\} - (1-m)| \to 0$,

where $\phi(Y, b_1(X), b_0(X), e(X), D, \Delta) = \dfrac{D\{Y - b_1(X)\}}{e(X)} - \dfrac{(1-D)\{Y - b_0(X)\}}{1 - e(X)} + b_1(X) - b_0(X) - \Delta$, $\mathbb{E}_n(O) = \dfrac{1}{n}\sum_{i=1}^{n} O_i$, $m \in (0,1)$ is the significance level, $c_m = \Phi^{-1}(1 - m/2)$, $V_{\Delta} = \mathbb{E}\{\phi^2(Y, b_1(X_{\mathcal{A}}), b_0(X_{\mathcal{A}}), e(X_{\mathcal{A}}), D, \Delta)\}$ and $\widehat{V}_{\Delta} = \mathbb{E}_n\{\phi^2(Y_i, b_1(X_i;\widehat{\alpha}^{(1)}), b_0(X_i;\widehat{\alpha}^{0}), e(X_i;\widehat{\beta}), D_i, \widehat{\Delta})\}$.

REMARK 3: Theorem 2(a) is parallel to Theorem 1 in Zhao and Yu (2006). Theorem 2(b) is parallel to Theorem 4 in Zou (2006) and Theorem 1 in Shortreed and Ertefaie (2017).

Our results in Theorem 2 (c) and (d) depend on correct selection of the target adjustment set $\mathcal{A}$. As such, the resulting uncertainty estimates do not take into account of the uncertainty in the selection of target adjustment set. In other words, as we aim to exclude instruments in our adjustment set, our procedure does not permit uniform valid inference. See Leeb and Pötscher (2005) for discussions of similar phenomena in other contexts involving variable selection. Alternative procedures that include instruments in the adjustment set are available (e.g. Van der Laan, 2014;

Farrell, 2015; Chernozhukov et al., 2018). In theory, these procedures are uniformly valid and should perform better in the worst-case scenario in which confounders are only weakly related to the outcome. However, in many practical situations, they pay a high price in terms of variance due to inclusion of instrumental variables. For example, Brookhart et al. (2006) show that with small samples, the inclusion of variables that are strongly related to the exposure but only weakly related to the outcome can be detrimental to an estimate in a mean-squared error sense. We refer readers to Moosavi et al. (2021) for a recent discussion on the dilemma between uniform validity and efficiency in performing variable selection in causal inference problems. We also illustrate this tradeoff via simulation studies in Section 5.

5. 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: $(n, p) = (300, 100)$, $(300, 1000)$, $(600, 200)$, $(600, 2000)$. The covariates $X^{(j)}, j = 1, \ldots, p$ are independently generated from the uniform distribution on $(-1, 1)$. The binary treatment D is then generated from a Bernoulli distribution with $P(D = 1 \mid X) = \text{expit}(X^{\mathrm{T}}\beta)$, where $\beta \in \mathbb{R}^p$ such that $\beta_1 = \beta_2 = 0.2$, $\beta_5 = \beta_6 = 0.3$ and $\beta_j = 0, j \notin \{1, 2, 5, 6\}$. Given D and X, the outcome Y is generated from the following model: $Y = \alpha_0 + X^{\mathrm{T}}\alpha + D\Delta + \epsilon$, where $\epsilon \sim N(0, 1)$, $\Delta = 2$, $\alpha_0 = 0$, $\alpha_j = 2,\ 1 \leqslant j \leqslant 4$ and $\alpha_j = 0,\ j \geqslant 5$. Note that in our simulations, both the outcome regression and propensity score models are sparse.

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

(i) CBS: The proposed Algorithm 1, where we select the top 30 covariates in Step 2;

(ii) Outcome adaptive Lasso (OAL): We use the `R` code provided in Shortreed and Ertefaie (2017) to estimate the propensity score. Note that the method by Shortreed and Ertefaie (2017) cannot directly handle the cases with $p > n$. For those scenarios, we first apply a conditional sure-

independence screening procedure on D (Barut et al., 2016) and select the top 30 covariates. We then apply the method of Shortreed and Ertefaie (2017) on the selected set. We use Lasso to estimate $\widehat{b}_d(X_i)$, and the tuning parameter is selected via 10-fold cross-validation.

(iii) Robust Inference (RI, Farrell, 2015): We use Lasso to fit the propensity score model, and the group-Lasso method implemented in `grpreg` to fit the outcome model. The estimates $\widehat{e}(X_i)$ and $\widehat{b}_d(X_i)$ are then plugged into (1) to obtain the causal effect estimate. The tuning parameters are selected via 10-fold cross-validation.

(iv) Double/Debiased Machine Learning (DBML, Chernozhukov et al., 2018): We implement the cross-fitting estimator with finite-sample adjustment using the median method; here we repeat 10-fold random partitions five times and take their median.

[Table 1 about here.]

Table 1 reports the biases and mean squared errors of various ACE estimators, as well as the empirical coverage of 95% Wald-type confidence intervals based on 1,000 Monte Carlo runs. For Table 1, we fit the correct propensity score and outcome regression models for all four estimating methods. The CBS and OAL perform much better than the RI and DBML methods, suggesting that at least for the simulation settings we consider, excluding instrumental variables significantly improve the efficiency and reduce the MSE of the resulting doubly robust estimator. We also notice that the performance of the RI method deteriorates quickly with the covariate dimension p, while DBML performs the worst under the settings we consider.

We further compare these four estimators in terms of their double robustness. In the following, we consider a low-dimensional setting with $(n, p) = (2000, 100)$. The treatment D and outcome Y are generated via the following equations:

$$\text{(Outcome model):} \qquad Y = 2(X_1 + X_2) + 2(X_3^2 + X_4^2) + D\Delta + \epsilon;$$

$$\text{(Propensity score model):} \qquad \text{logit}\{P(D = 1 \mid X)\} = 0.2(X_1 + X_2) + 0.3(X_5 + X_6) + \epsilon.$$

To examine double robustness, we consider settings where the outcome regression or the propensity

score model may be misspecified. In these settings, the analyst assumes that the propensity score model is a logistic regression with predictors $\left\{\left(X^{(j)}\right)^2; j = 1, \ldots, p\right\}$, and/or that the outcome model is linear with predictors D and $\{X^{(j)^2}; j = 1, \ldots, p\}$. Figure 2 shows boxplots of estimates from the four estimators under different combinations of correct/incorrect specifications of the outcome regression and propensity score models. One can see that $\widehat{\Delta}^{CBS}$, $\widehat{\Delta}^{RI}$, and $\widehat{\Delta}^{DBML}$ are consistent as long as at least one of the outcome regression or the propensity score model is correctly specified, thus exhibiting double robustness. In contrast, $\widehat{\Delta}^{OAL}$ is not consistent when the propensity score model is correctly specified but the outcome regression model is not, as it relies on the outcome regression model for confounder selection in the propensity score model.

[Figure 2 about here.]

6. 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 Data Availability Statement Section. We consider the clinical, genetic, and behavioral measures in the ADNI data set. 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 stabilization (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 a 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 Alzheimer's 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 Alzheimer's 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 and behavioral covariates, including baseline age, gender, and education length, as they are widely considered as the main risk factors for AD (e.g. Guerreiro and Bras, 2015; Vina and Lloret, 2010). 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, 82 have CSF-tau protein level above the cut-off point. The mean (SD) age in the high/low tau-protein group is $75.8(7.28)$ and $75.3(6.55)$ years old, respectively, and the mean (SD) education length in the high/low tau-protein group is $15.3(3.05)$ and $15.9(2.99)$ years, respectively. The two groups are unbalanced in terms of gender: 54.9% of study participants in the high tau-protein group are female, while 64.0% of study participants in the low tau-protein group are female. We nevertheless include all three covariates as they were determined *a priori*. In the Supplementary Material, we report sensitivity analysis in which we adjust for more baseline clinical and behavioral covariates. Analysis results show that adjusting for these additional covariates has minimal effects on the results we obtained.

Denote as Z the covariates age, gender and education length. We first fit a linear regression $Y \sim Z$ to adjust for these clinical covariates. We then apply Steps 1-2 of our CBS procedure using the

fitted residuals from the linear regression as the outcome and select the top 30 genetic covariates. We list the top 10 SNPs in Table 2. All of them are located on Chromosome 19 and have previously been found to be strongly associated with Alzheimer's. See Table S1 in the Supplementary Material for a list of references for the SNPs reported in Table 2.

[Table 2 about here.]

Since some SNPs selected through our first step screening are perfectly correlated, we only keep one among a group of SNPs whose genotypes are identical to each other for the subsequent analysis. We further apply our refined selection procedure in Section 3.2 on these selected covariates and the covariates age, gender and education length, where the coefficients corresponding to the three clinical and behavioral covariates are not penalized. Finally, we use the doubly robust estimator (1) to estimate the average causal effect of CSF-tau protein level on the ADAS-11 score. Analysis results suggest that, on average, being in the high-level CSF-tau group will raise the ADAS-11 score by 5.96 (95% CI = $[4.15, 7.76]$) points.

In addition to varying the adjusted confounders, in Section S4.3 of the Supplementary Material, we also describe sensitivity analysis varying the number of covariates selected in the first screening step. Tables S2 and S3 suggest that our selection and estimation procedures are relatively robust to the choice of adjusted confounders and the number of covariates selected in the screening step.

7. Discussion

In this paper, we propose a novel selection and estimation procedure for doubly robust causal inference with ultra-high dimensional covariates, called causal ball screening. In comparison to previous approaches that use the same outcome model for propensity score model selection and causal effect estimation, the estimator we propose is doubly robust. Moreover, as we illustrate in the real data analysis, it can be applied to select variables important for causal inference from millions of baseline covariates.

We have so far considered causal effect estimation using the classical doubly robust estimator by Robins et al. (1994). Our developments can also be combined with other techniques for causal effect estimation, such as the covariate balancing propensity score (Imai and Ratkovic, 2014) and the subclassification weights (Wang et al., 2016). This is left as future work.

Acknowledgement

Pan was partially supported by the National Natural Science Foundation of China (71991474, 12071494), and the Science and Technology Program of Guangzhou, China (202002030129, 202102080481). Kong and Wang were partially supported by the Natural Science and Engineering Research Council of Canada and the CANSSI Collaborative Research Team Grant.

Data Availability Statement

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`http://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`.

References

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

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., and Yuan, M. (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.

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

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

de Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika **98,** 861–875.

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

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

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

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica **66,** 315–331.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. Review of Economics

and Statistics **86,** 73–76.

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.

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

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. Econometric Theory **21,** 21–59.

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.

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.

Moosavi, N., Häggström, J., and de Luna, X. (2021). The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters. https://arxiv.org/abs/2105.02071.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018). 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.

Pearl, J. (2009). Causality. Cambridge University Press.

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.

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. Journal of Machine Learning Research **21,** 1–86.

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

Schnitzer, M. E., Lok, J. J., and Gruber, S. (2016). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. The International Journal of Biostatistics **12,** 97–115.

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

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

Tapiola, T., Alafuzoff, I., Herukka, S.-K., Parkkinen, L., Hartikainen, P., Soininen, H., et al. (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.

Van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. The International Journal of Biostatistics **10,** 29–57.

VanderWeele, T. J. and Hernan, M. A. (2013). Causal inference under multiple versions of treatment. Journal of Causal Inference **1,** 1–20.

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.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. The Journal of Machine Learning Research **7,** 2541–2563.

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.

Supporting Information

Web Appendices, Tables, Figures and Proofs referenced in Sections 2, 3, 4 and 6 are available with this paper at the Biometrics website on Wiley Online Library. The proposed CBS method is available both on Wiley Online Library and on Github: `https://github.com/dingketang/ultra-high-DRCI`.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

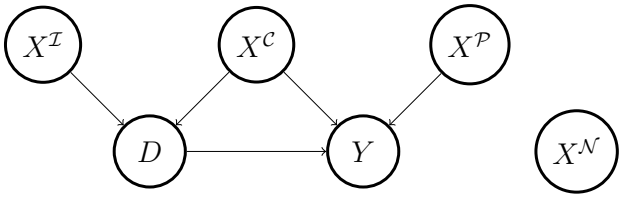

Figure 1. A causal directed acyclic graph illustrating the four types of baseline covariates: confounders $X^{\mathcal{C}}$, precision variables $X^{\mathcal{P}}$, instrumental variables $X^{\mathcal{I}}$ and null variables $X^{\mathcal{N}}$

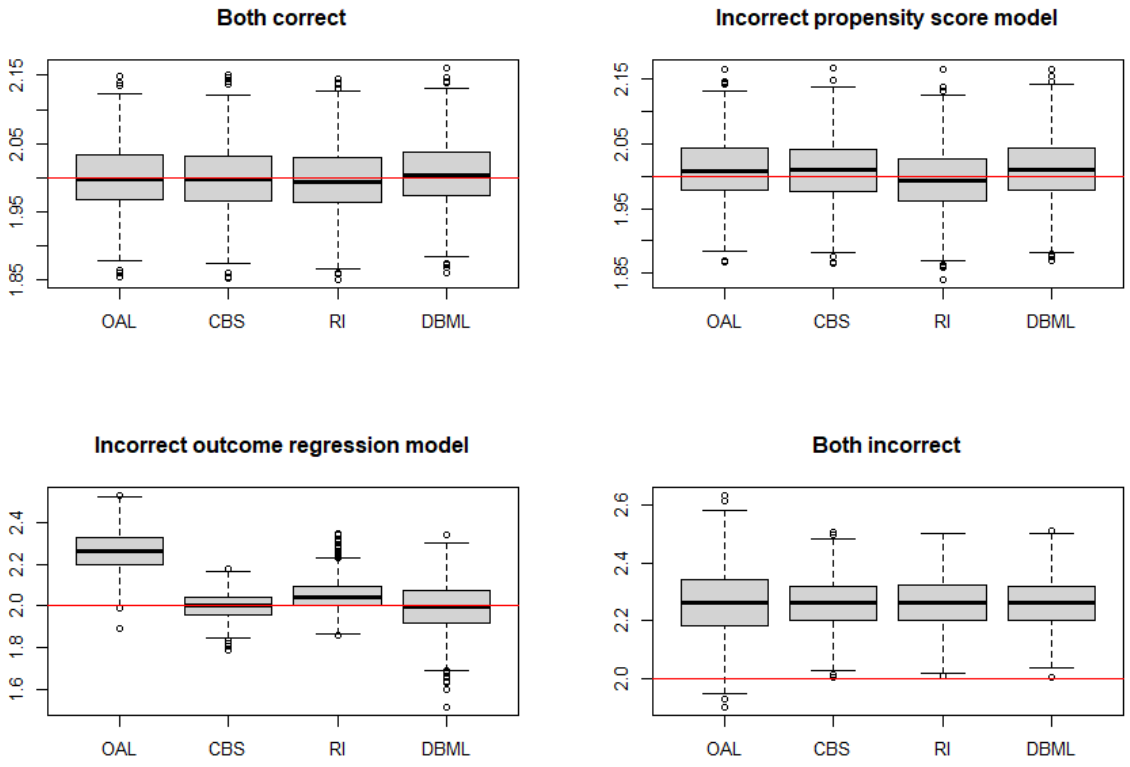

Figure 2. Boxplots of causal effect estimates obtained by $\widehat{\Delta}^{OAL}$, $\widehat{\Delta}^{CBS}$, $\widehat{\Delta}^{RI}$ and $\widehat{\Delta}^{DBML}$. The horizontal red lines correspond to the true causal effect $\Delta = 2$. Results are based on 1,000 Monte Carlo runs. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1

Simulation results based on $1,000$ *Monte Carlo runs. Both the propensity score and outcome regression models are correctly specified. We report bias* $\times 100$*, MSE* $\times 100$*, and the empirical coverage probability for each estimator. The nominal coverage probability is 95%. Standard errors of bias and MSE are reported in parentheses. Bold numbers represent the best result in each scenario*

(n,p)	Method	$Bias \times 100 (SE \times 100)$	$MSE \times 100 (SE \times 100)$	$Empirical\ coverage \times 100\%$
$(300, 100)$	OAL	$1.3(0.37)$	**1.4(0.064)**	93.5
	CBS	**0.97(0.38)**	$1.5(0.066)$	**94.3**
	RI	$1.5(0.76)$	$5.8(0.96)$	92.5
	DBML*	$8.5(39)$	$1.5 \times 10^4 (2.9 \times 10^3)$	98.3
	DBML	$2.7 \times 10^5 (2.8 \times 10^5)$	$7.7 \times 10^{11} (7.7 \times 10^{11})$	98.4
$(300, 1000)$	OAL	$1.8(0.39)$	**1.6(0.075)**	**92.6**
	CBS	**1.6(0.40)**	$1.6(0.076)$	92.2
	RI	$19(0.91)$	$12(2.2)$	51.7
	DBML*	$10\ (37)$	$1.3 \times 10^4 (2.2 \times 10^3)$	98.5
	DBML	$7.3 \times 10^2 (6.6 \times 10^2)$	$4.4 \times 10^6 (3.5 \times 10^6)$	98.5
$(600, 200)$	OAL	$0.28(0.26)$	$0.68(0.030)$	**94.9**
	CBS	**0.04(0.26)**	**0.68(0.030)**	95.6
	RI	$2.3(0.59)$	$3.6(1.1)$	91.6
	DBML*	$16\ (23)$	$5.3 \times 10^3 (1.4 \times 10^3)$	98.3
	DBML	$8.1 \times 10^2 (7.1 \times 10^2)$	$5.0 \times 10^6 (5.0 \times 10^6)$	98.3
(600,2000)	OAL	$0.58(0.27)$	$0.71(0.030)$	94.0
	CBS	**0.22(0.27)**	**0.71(0.030)**	**94.2**
	RI	$13(0.51)$	$4.3(0.55)$	55.9
	DBML*	$16(1.9 \times 10^2)$	$2.1 \times 10^4 (1.4 \times 10^2)$	98.2
	DBML	$1.2 \times 10^2 (3.1 \times 10^2)$	$9.5 \times 10^5 (3.6 \times 10^5)$	98.4

For DBML*, we exclude Monte Carlo runs for which the bias of DBML estimate is greater than 100. For $(n, p) = (300, 100), (300, 1000), (600, 200)$ and $(600, 2000)$, we exclude $(35, 23, 11, 26)$ runs out of 1000.

Table 2
The top ten SNPs selected by Steps 1-2 of our CBS procedure

Rank	SNP name	Gene	Chromosome number
1	rs429358	ApoE	19
2	rs56131196	ApoC1	19
3	rs4420638	ApoC1	19
4	rs12721051	ApoC1	19
5	rs769449	ApoE	19
6	rs10414043	ApoC1	19
7	rs7256200	ApoC1	19
8	rs73052335	ApoC1	19
9	rs111789331	ApoC1	19
10	rs6857	NECTIN2	19