Statistics and Its Interface Volume 18 (2025) 93–105

Variable selection for doubly robust causal inference

Eunah Cho* and Shu Yang

Confounding control is crucial and yet challenging for causal inference based on observational studies. Under the typical unconfoundness assumption, augmented inverse probability weighting (AIPW) has been popular for estimating the average causal effect (ACE) due to its double robustness in the sense it relies on either the propensity score model or the outcome mean model to be correctly specified. To ensure the key assumption holds, the effort is often made to collect a sufficiently rich set of pretreatment variables, rendering variable selection imperative. It is well known that variable selection for the propensity score targeted for accurate prediction may produce a variable ACE estimator by including the instrument variables. Thus, many recent works recommend selecting all outcome predictors for both confounding control and efficient estimation. This article shows that the AIPW estimator with variable selection targeted for efficient estimation may lose the desirable double robustness property. Instead, we propose controlling the propensity score model for any covariate that is a predictor of either the treatment or the outcome or both, which preserves the double robustness of the AIPW estimator. Using this principle, we propose a two-stage procedure with penalization for variable selection and the AIPW estimator for estimation. We show the proposed procedure benefits from the desirable double robustness property. We evaluate the finite-sample performance of the AIPW estimator with various variable selection criteria through simulation and an application.

1. INTRODUCTION

1.1 Ignorability and the need for variable selection

Unlike experimental studies, treatment assignments in observational studies are not random. As a result, distributions of the covariates differ between treatment arms, and direct comparisons between the treatment groups may be biased. Most causal inference methods rely on the ignorability assumption (also referred to as no unmeasured confounders), which indicates the treatment assignment can be ignored when conditioning the observed covariates. Under the ignorability assumption, researchers have proposed various methods to estimate the average causal effect (ACE), including regression imputation estimator, matching estimator, inverse propensity score weighted (IPW) estimator, augmented IPW (AIPW) estimator [e.g., 13]. Among them, the AIPW estimator has been popular because it is locally efficient and doubly robust in the sense that its consistency relies on either the correctly specified propensity score (PS) model or the outcome mean (OM) model, but not necessarily both [21, 23, 28, 17, 2].

In the past, it was desirable to include all possible pretreatment variables to avoid the risk of excluding related variables and satisfy the ignorability assumption [17, 29]. With the advances in technology, a rich set of pretreatment covariates can be collected. In these high-dimensional settings, including all variables can be computationally unstable, burdensome, or sometimes impossible. Thus, variable selection is indispensable for handling high-dimensional covariates.

1.2 Existing variable selection strategies

Typically, there are four main types of pretreatment variables: 1) instrumental variables, 2) confounders, 3) precision variables, and 4) spurious variables. We refer to variables that are only predictors of the treatment but not the outcome as the instrumental variables, variables that are predictors of both the treatment and outcome as the confounders, variables that are only predictors of the outcome but not the treatment as the precision variables, and others as the spurious variables.

Lunceford and Davidian [17] show that containing the precision variables in the PS model helps reduce standard errors while maintaining consistency. Following this result, many researchers suggest outcome predictor approaches, which include the precision variables and the confounders for the ACE estimation, since other variables, including the instrumental variables, may inflate the variance of the ACE estimators or introduce bias to the estimator [5, 20].

Shortreed and Ertefaie [29] suggest the outcome-adaptive lasso and provide simulation studies showing that including precision variables in the PS model increases efficiency. Ertefaie et al. [7] propose a penalized objective function that simultaneously considers the outcome and treatment assignment models for variable selection. Tang et al. [30] propose

arXiv: 2301.11094
*Corresponding author.

the causal ball screening that targets confounders and other outcome predictors as an adjustment set for the PS. Henckel et al. [11] provide a pruning procedure determining the optimal adjustment set. They show the adjusted least-squares treatment effect estimator based on the identified set has the smallest asymptotic variance among consistent adjusted least square estimators. Rotnitzky and Smucler [24] demonstrate that Henkel's results can also be extended to nonparametric estimators. However, it is challenging to distinguish confounders from instrumental variables, and hence the outcome predictor approach may exclude the true confounders from subsequent estimation. As a result, the omission of important confounders may lead to bias. To avoid such bias, VanderWeele and Shpitser [33] suggest adjusting for any covariates that are causes of either the treatment or outcome because those variables constitute a sufficient set adjusting for confounding. Belloni et al. [3] suggest a post-double-selection method where they consider the union of the covariates considered important in two equations from a partially linear model and estimate the ACE using linear square regression. Wilson and Reich [35] estimate the standard Bayesian regression model and then the posterior distribution using a confounder-specific loss function. They target the set of all confounders and the outcome-related covariates, but instrumental variables are included to avoid omitting confounder variables. Although the idea of using the union of the selected variables has already been proposed to avoid bias due to the exclusion of confounders, we provide another reason for using the union approach in terms of maintaining the double-robustness of the AIPW estimator.

1.3 Contribution and outline

This article shows that the AIPW estimator with variable selection targeted for efficient estimation (referred to as the outcome predictor approach) may lose the desirable double-robustness property. Generally, the outcome predictor approach is shown to be efficient, provided the postulated models are known or correctly specified. However, if the PS model is correctly specified, and the OM model is misspecified, the estimation of the PS model restricted to the outcome predictors may not be consistent; see Example 1. Thus, although the PS working model is correctly specified, the PS model based on the wrong set from the OM model is not consistent for the true PS model, and thus the AIPW estimator becomes not consistent. Given the above reasons, considering the selected instrumental variables in subsequent estimation aids in protecting the AIPW estimator's double-robustness property.

Using this principle, we propose a two-stage procedure with penalization for variable selection and the AIPW estimator for estimation. In the first stage, we select a set of variables considered important predictors of either the treatment or outcome using penalized estimating equations. In this paper, we used the smoothly clipped absolute deviation (SCAD) proposed by [8], but other penalized methods also can be applied. After variable selection, we employ the AIPW estimator to estimate the ACE with the nuisance models refitted based on the selected variables. We show the proposed procedure benefits from the desirable statistical properties, including selection consistency and double robustness.

The rest of the paper is organized as follows. Section 2 presents the basic setup. Section 3 illustrates the wishlist consisting of various variable selection criteria. Section 4 presents the asymptotic properties of our procedure. In Section 5, we compare our approach to common variable selection strategies for confounding control or efficient estimation. The simulation suggests that the AIPW estimator is still doubly robust with our variable selection procedure but is not with other selection strategies. In Section 6, we apply our procedure to an application, maternal smoking on birth weight data. We conclude the paper with a discussion in Section 7.

2. BASIC SETUP

2.1 Potential outcomes framework

Following [19] and [26], we adopt the potential outcomes framework. Denote X to be a vector of p-dimensional pretreatment covariates. Suppose that the treatment is a binary variable $A \in \{0,1\}$, with 0 and 1 being labels for control and active treatments, respectively. Under the common Stable Unit Treatment Value assumption [27], for each level of the treatment a, we assume that there exists a potential outcome $Y(a)$, representing the outcome had the unit, possibly contrary to the fact, been given the treatment a. We make the consistency assumption that links the observed outcome with the potential outcomes, i.e., the observed outcome Y is the potential outcome under the treatment regime actually following $Y(A)$. We focus on estimating the ACE, $\tau = E\{Y(1)-Y(0)\}$. The ACE is the target causal estimand in many scientific applications, generating important policy implications. Our methodology also applies to a broader class of causal estimands in [16].

The fundamental problem in estimating the ACE is that one may observe at most one of $Y(0)$ and $Y(1)$ for each unit. Throughout, we make the ignorability assumption widely used in the causal inference literature.

Assumption 1 (Ignorability)**.** $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.

Assumption 2 (Overlap)**.** *There exist constants c_1 and c_2 such that $0 < c_1 \leq e(X) \leq c_2 < 1$ almost surely, where $e(X) = P(A = 1|X)$ is the PS.*

Assumption 1 holds when all confounders are identified and measured. That is, this assumption requires X to include all factors related to both treatment and outcomes. Assumption 1 indicates that treatment assignment is independent of the potential outcomes given X. This assumption is automatically guaranteed in the experimental study because the treatment is assigned to each unit at random. In observational studies, researchers often collect a rich set of pretreatment covariates to make this assumption plausible, leading to possibly high-dimensional X.

Assumption 2 holds when there exists a sufficient overlap between the covariate distributions of the treatment and control group. This means that the distributions of the treatment and control groups need to be similar to each other. When Assumption 2 is not satisfied at a specific value of X, the unit at the value would be only treated or controlled. This leads to the extrapolation of one of two potential outcomes at that value and makes the inference about the ACE inappropriate.

2.2 Doubly robust estimator of the ACE

It is well known that under Assumption 1 and Assumption 2, the ACE can be identifiable and estimated through the outcome regression or the augmented/inverse probability weighting (AIPW/IPW) estimator. See [12] and [22] for surveys of these estimators.

Define $\mu_a(X) = E\{Y(a)|X\}$ for $a = 0, 1$. Then, under Assumption 1, $\mu_a(X) = E(Y|A = a, X)$. In practice, the outcome distribution and the PS are often unknown and therefore have to be modeled and estimated.

Assumption 3 (Outcome mean model)**.** *The parametric model $\mu_a(X) = E(Y|A = a, X)$ is a correct specification for $\mu_a(X)$, for $a = 0, 1$; i.e., $\mu_a(X) = \mu_a(X; \beta_{a0})$, where β_{a0} is the true OM model parameter for $a = 0, 1$.*

Assumption 4 (Propensity score model)**.** *The parametric model $e(X; \alpha_0)$ is a correct specification for $e(X)$; i.e., $e(X) = e(X; \alpha_0)$, where α_0 is the true model parameter.*

Under Assumption 3, let $\hat{\beta}_a$ be a consistent estimator of β_{a0}. Under Assumption 4, let $\hat{\alpha}$ be a consistent estimator of α_0. The *Augmented Inverse Propensity score Weighting (AIPW) estimator* is

$$\widehat{\tau}_{n,\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{A_iY_i}{e(X_i;\hat{\alpha})} + \bigg\{1 - \frac{A_i}{e(X_i;\hat{\alpha})}\bigg\}\mu_1(X_i;\hat{\beta}_1) - \frac{(1-A_i)Y_i}{1-e(X_i;\hat{\alpha})} - \bigg\{1 - \frac{(1-A_i)}{1-e(X_i;\hat{\alpha})}\bigg\}\mu_0(X_i;\hat{\beta}_0)\Bigg].$$

The AIPW estimator is doubly robust in the sense that it is consistent if either Assumption 3 or 4 holds and locally efficient if both assumptions hold [25].

Without loss of generality, we assume all covariates have a mean of zero and a common standard deviation so that we apply the penalty equally to all covariates. Define $\alpha^* = \mathrm{argmin}_{\alpha\in\mathbb{R}^p} E[\{A - e(X^T\alpha)\}^2]$ and $\beta_a^* = \mathrm{argmin}_{\beta\in\mathbb{R}^p} E[\{Y - \mu_a(X^T\beta_a)\}^2]$, $a = 0, 1$. As is common in the empirical literature, we assume a generalized linear model for the OM model in Assumption 3 and a logistic regression model for the PS model in Assumption 4. If working models $e(X^T\alpha)$ and $\mu_a(X^T\beta_a)$ are correctly specified, we have $e(X) = e(X^T\alpha^*)$ and $\mu_a(X) = \mu_a(X^T\beta_a^*)$, respectively. However, the working models may be misspecified. $\mathcal{M}_\alpha$ denotes the set of true important variables in the PS model, and $\mathcal{M}_\beta$ denotes the set of true important variables in the OM model. Define the union set of true important variables in the PS and the OM model as $\mathcal{U}$, i.e., $\mathcal{U} = \mathcal{M}_\alpha \cup \mathcal{M}_\beta$, and the intersection set of true important variables in the PS and the OM model as $\mathcal{I}$, i.e., $\mathcal{I} = \mathcal{M}_\alpha \cap \mathcal{M}_\beta$. We refer to true important sets as oracle sets. We use the hat for the set consisting of the selected variables by variable selection procedures. The set of variables selected from the PS model is denoted by $\widehat{\mathcal{M}}_\alpha$, and the set of variables selected from the OM model is denoted by $\widehat{\mathcal{M}}_\beta$. Likewise, we define $\widehat{\mathcal{U}} = \widehat{\mathcal{M}}_\alpha \cup \widehat{\mathcal{M}}_\beta$ and $\widehat{\mathcal{I}} = \widehat{\mathcal{M}}_\alpha \cap \widehat{\mathcal{M}}_\beta$.

3. VARIABLE SELECTION CRITERIA

3.1 Classification of pretreatment variables

In the presence of a considerable number of spurious covariates, including unnecessary covariates in the model can lead to statistical inefficiency of the estimation or sometimes be computationally infeasible. For this reason, variable selection is essential to exclude unnecessary covariates. We investigate the variable selection approaches for the AIPW estimator of the ACE, given its desirable double robustness property. As mentioned in the introduction, typically, there are four main types of pretreatment variables: instrumental variables (X_I), confounder variables (X_C), precision variables (X_P), and spurious variables (X_S). Figure 1 displays relationships of pretreatment variables.

Since the pretreatment variables play different roles in the estimation, which variables to be selected depends on the goal for variable selection. We categorize four goals for variable selection: (1) variable selection for prediction modeling; (2) variable selection for confounding control; (3) variable selection for efficient estimation; and (4) variable selection for double robustness. We discuss them in the following subsections.

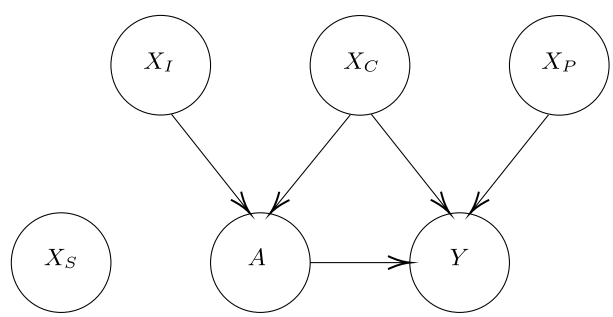

Figure 1. A diagram of pretreatment covariate structure. Abbreviations: X_I, instrumental variables; X_C, confounding variables; X_P, precision variables; X_S, spurious variables.

3.2 Variable selection for prediction modeling

Traditionally, variable selection aims to gain prediction accuracy. It is crucial to decide which variables to include in models because the choice of variables strongly influences the model's performance. The exclusion of essential variables from the models fails to identify a genuine relationship between outcomes, treatment, and covariates.

Backward elimination, forward selection, stepwise selection, p values, Akaike information criterion, Bayesian information criterion, and Mallows' C_p statistic are traditional variable selection procedures [6]. These methods can be computationally expensive and do not consider stochastic errors caused in the stages of variable selections [8]. Tibshirani [31] proposes the least absolute shrinkage and selection operator (LASSO), which is the penalized least squares estimate with the L_1 penalty in the squares and likelihood settings. Zou [41] demonstrate situations where the LASSO selection is inconsistent and present an alternative method, the Adaptive LASSO, where adaptive weights are employed to penalize different coefficients in the l_1 penalty. Fan and Li [8] show that the LASSO shrinkage gives a biased estimator for the large coefficients and propose the SCAD that takes advantage of the oracle properties. The procedure has the oracle properties if it identifies the right subset model and has the optimal estimation rate. The change in estimate criterion method attempts to obtain a low-bias estimator with a minimal covariate set. The method removes a covariate one at a time from a covariate set and then compares the estimates between the reduced set and the original set. If the change in estimate is significant enough, the covariate is removed from the adjustment set.

However, only focusing on the high prediction accuracy of the PS model or the OM model is not practically helpful for obtaining reasonable ACE estimates. If one cares about only the PS predictor, the AIPW estimator may lose efficiency, excluding X_P. We will discuss this in Subsection 3.4. If one cares about only the outcome predictor, the AIPW estimator may no longer retain a double robustness property. We will discuss this in Subsection 3.5.

3.3 Variable selection for confounding control

The omission of X_C that appears in both OM and PS models makes the ACE inconsistent. Thus, all X_C should be measured and included for the unbiased ACE. However, identifying all the confounders is not easy in practice. In the real world, the relationship between covariates and the treatment, as well as the relationship between covariates and the outcome, are not fully known. For this reason, a richly parameterized PS model is preferred to ensure the inclusion of X_C. In this process, too many covariates are involved in the model, which leads to a complicated model to estimate. An alternative approach is to adjust for covariates that are common causes of exposure and outcome. However, if the information on the common cause of treatment and outcome is not clear, there might be a missed set of covariates which is required to adjust for confounding [32]. Besides, this approach does not take into account X_P helpful for efficiency. Thus, this variable selection approach is not suitable for efficient and accurate estimation.

3.4 Variable selection for efficient estimation

One of the primary motivations for variable selection is to gain efficiency in estimating the ACE. Lunceford and Davidian [17] show that when including X_P as well as X_C into the PS model, all weighted estimators for the ACE are consistent, and the variance of the AIPW estimator based on X_P and X_C has a smaller variance relative to the one based only on X_C. Brookhart et al. [5] suggest using only outcome predictors for the PS model since X_P reduces the variance while X_I and X_S inflate the variance without changing bias. Henckel et al. [11] introduce a new graphical criterion. Using their Theorem 3.1, it is shown that the outcome predictor-based asymptotic variance is smaller than the treatment predictor-based asymptotic variance. Rotnitzky and Smucler [24] extend the results to non-parametric causal graphical models. In their settings, X_C affects outcomes through X_P. They point out that the stronger association between X_C and the PS model and the weaker association between X_C and X_P are, the less efficient it is to include X_C. Tang et al. [30] also show that adjusting for outcome predictors improves efficiency, and adjusting for X_I increases the variance. To show these results, they quantify the difference in the asymptotic variance of the estimators based on the X_P and X_I.

Many methods have been developed to obtain efficiency, including outcome predictors and excluding X_I. Shortreed and Ertefaie [29] derive the outcome-adaptive lasso. Ertefaie et al. [7] propose a penalized objective function, which selects outcome predictors while excluding X_I and X_S. Henckel et al. [11] establish a procedure to prune a valid adjustment to get a valid subset with a smaller asymptotic variance. Tang et al. [30] propose the causal ball screening for selecting all outcome predictors from modern ultra-high dimensional data sets and excluding X_I and X_S.

3.5 Variable selection for double robustness

As mentioned in the previous subsection, outcome predictor approaches make the ACE estimator more efficient under the correct models. However, there is a lack of information on the relationship between covariates or on both model specifications in the real world. The insufficient information makes it difficult to obtain the correct models and to identify relevant covariates that increase the efficiency of the ACE estimator. In this case, a robust estimation can be more reliable, although a bit of efficiency is sacrificed. In particular, this is obvious when one does not have complete knowledge of the OM model and uses the AIPW estimator for the ACE estimation. Suppose linear models are assumed for the non-linear OM models. Then, no variables may be selected, or some variables that are not associated with the true outcomes may be chosen for the OM model. In this case, including X_I can be helpful in terms of robustness. To illustrate, consider the following example.

Example 1. *Let $X = (X_1, X_2, X_3)$ be the vector of pretreatment covariates. Assume the true OM model is non-linear, e.g., $\mu_a(X) = 0.1X_1^2 + X_2^2 + 2X_2$; the true PS model follows a logistic model, e.g., $logit\{e(X)\} = 1 + X_1 + X_3$. Then, $\mathcal{M}_\beta = \{X_1, X_2\}$ and $\mathcal{M}_\alpha = \{X_1, X_3\}$. Thus, $X_1 \in X_C$, $X_2 \in X_P$, and $X_3 \in X_I$. Assume treatment assignment follows a correct logistic regression model, and the OM model may be misspecified as linear. In this case, since X_1 is non-linearly related to the OM models, X_1 is unlikely to be selected for the OM model in most cases. Consequently, we have $\widehat{\mathcal{M}}_\beta = \{X_2\}$ and $\widehat{\mathcal{M}}_\alpha = \{X_1, X_3\}$.*

In Example 1, a linear model is assumed for the non-linear OM model, whereas the PS model is correctly specified. That is, Assumption 3 is not satisfied and $\mu(X, \beta_a^*) \neq \mu(X) = E(Y|X)$. On the other hand, Assumption 4 is satisfied and $e(X, \alpha^*) = e(X) = E(A|X)$. Generally, the AIPW estimator should be consistent with the ACE by its doubly robust property because Assumption 4 is satisfied. However, in Example 1, we cannot obtain the consistent PS model with the outcome predictor approach because $X_C = \{X_1\}$ is likely to be dropped due to its weak linear relationship and $\widehat{\mathcal{M}}_\beta = \{X_2\}$ is not involved with the PS model. Thus, both the PS and OM models using a set based on the outcome predictors are not consistent. As a result, the AIPW estimator loses its doubly robust property, and it is no longer consistent for ACE. However, if $\widehat{\mathcal{U}} = \widehat{\mathcal{M}}_\alpha \cup \widehat{\mathcal{M}}_\beta = \{X_1, X_2, X_3\}$ is used to estimate the PS model, we can obtain the consistent PS model, and the AIPW estimator is consistent with ACE. Although $X_I = X_1$ may reduce efficiency, it is necessary to get a consistent AIPW estimator. Therefore, we suggest using $\widehat{\mathcal{U}}$ to retain the double robustness of the AIPW estimator. In the simulation study, we will show that the AIPW estimator based on $\widehat{\mathcal{I}}$ or $\widehat{\mathcal{M}}_\beta$ is not doubly robust when the PS model is correctly specified, but the OM model is misspecified.

3.6 Proposed procedure for variable selection and estimation

The proposed method has three steps: we separately select important variables for the OM model in Step 1 and the PS model in Step 2. In Step 3, we estimate the ACE using a doubly robust estimator based on the union set of selected covariates in Step 1 and Step 2. In Step 1 and Step 2, we employ a penalized estimating equation for determining important covariates. The SCAD is used here. We specify $q_\lambda(x)$ to be a folded concave SCAD penalty function [9, 37]. The SCAD penalty is defined by

$$(1) \qquad q_\lambda(|\theta|) = \lambda\left\{I(|\theta| < \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| \geq \lambda)\right\}$$

for $a > 0$, where $(\cdot)_+$ is the truncated linear function; i.e., if $x \geq 0$, $(x)_+ = x$, and if $x < 0$, $(x)_+ = 0$. We use $a = 3.7$ following the suggestion of [8].

In Step 1, we run the SCAD for each of the OM models separately. For β_1, we conduct the SCAD only using the observations (Y_i, X_i) with $A_i = 1$. Likewise, the observations (Y_i, X_i) with $A_i = 0$ are used for β_0. The penalized estimating functions for β_a, $a = 0, 1$, are defined as

$$(2) \quad U_1(\beta_1) = \frac{1}{n_1}\sum_{i=1}^{n} A_i\{Y_i - \mu_1(X_i^T\beta_1)\}X_i - q_{\lambda_{\beta_1}}(|\beta_1|)\cdot \mathrm{sign}(\beta_1),$$

$$(3) \quad U_2(\beta_0) = \frac{1}{n_0}\sum_{i=1}^{n} (1-A_i)\{Y_i - \mu_0(X_i^T\beta_0)\}X_i - q_{\lambda_{\beta_0}}(|\beta_0|)\cdot \mathrm{sign}(\beta_0),$$

where $n_1 = \sum_{i=1}^n I(A_i = 1)$ and $n_0 = \sum_{i=1}^n I(A_i = 0)$. $q_{\lambda_{\beta_a}}(|\beta_a|)$ is defined in Eq. (1).

In Step 2, we implement the SCAD for the PS model. The corresponding penalized estimating function for α is

$$(4) \quad U_3(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\{A_i - e(X_i;\alpha)\}X_i - q_{\lambda_\alpha}(|\alpha|)\cdot \mathrm{sign}(\alpha).$$

Let $(\tilde{\alpha}, \tilde{\beta}_0, \tilde{\beta}_1)$ denote the solution for the penalized joint estimating equation $U = 0$. In this procedure, $\widehat{\mathcal{M}}_\alpha$ is the set of variables that correspond to the nonzero coefficients in the PS model, and $\widehat{\mathcal{M}}_\beta$ is the set of variables that correspond to the nonzero coefficients in the OM model. Likewise, $\widehat{\mathcal{U}} = \widehat{\mathcal{M}}_\alpha \cup \widehat{\mathcal{M}}_\beta$ and $\widehat{\mathcal{I}} = \widehat{\mathcal{M}}_\alpha \cap \widehat{\mathcal{M}}_\beta$. Fan and Li [8] show that the SCAD estimators perform the oracle procedure in variable selection, which means they behave as if the correct submodels were known. Thus, the set $\widehat{\mathcal{U}}$ includes the true important variables in either the PS model or the OM model with probability approaching one.

In Step 3, we re-estimated the coefficients using the variables in the set $\widehat{\mathcal{U}}$ and then derive the AIPW estimator of the ACE. At this time, we use estimating functions without penalty. Under Assumption 3, for the OM model, let

$$(5) \qquad S_a(X, Y; \beta_a) = \frac{\partial \mu_a(X;\beta_a)}{\partial \beta_a}\{Y - \mu_a(X;\beta_a)\}$$

be the estimating function for β_a^* for $a = 0, 1$. Under Assumption 4, for the PS model, let

$$(6) \qquad S(A, X; \alpha) = \frac{A - e(X;\alpha)}{e(X;\alpha)\{1 - e(X;\alpha)\}}\frac{\partial e(X;\alpha)}{\partial \alpha}$$

be the estimating function for α^*. $\widehat{U}^C$ is defined as the complement of $\widehat{U}$. We use the estimating equations (Eq. 5 and 6) restricted to the parameter space $\{\beta_a : \beta_{a,\hat{\mathcal{U}}^C} = 0\}$ and $\{\alpha : \alpha_{\hat{\mathcal{U}}^C} = 0\}$, respectively. Let $(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$ denote the solution for Eq. (5) and (6). To summarize, our two-stage procedure for variable selection and estimation is as follows.

- Step 1: Under Assumption 3, use the penalization method to select important variables in the OM model using the SCAD, i.e., $\widehat{\mathcal{M}}_\beta = \{j : \tilde{\beta}_{a,j} \neq 0\}$, where $\tilde{\beta}_{a,j}$ is the solution for Eq. (2) and (3), for $a = 0, 1$.
- Step 2: Under Assumption 4, use the penalization method to select important variables in the PS model using the SCAD, i.e., $\widehat{\mathcal{M}}_\alpha = \{j : \tilde{\alpha}_j \neq 0\}$, where $\tilde{\alpha}_j$ is the solution for Eq. (4).
- Step 3: Let the set of variables for estimation be $\widehat{\mathcal{U}} = \widehat{\mathcal{M}}_\alpha \cup \widehat{\mathcal{M}}_\beta$. The proposed estimator is

$$\begin{aligned}(7)\quad &\hat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_iY_i}{e(X_i;\hat{\alpha})} + \left\{1 - \frac{A_i}{e(X_i;\hat{\alpha})}\right\}\mu_1(X_i;\hat{\beta}_1)\right.\\ &\left. - \frac{(1-A_i)Y_i}{1-e(X_i;\hat{\alpha})} - \left\{1 - \frac{(1-A_i)Y_i}{1-e(X_i;\hat{\alpha})}\right\}\mu_0(X_i;\hat{\beta}_0)\right],\end{aligned}$$

where $\hat{\alpha}$ and $\hat{\beta}_a$ are obtained by fitting the OM and PS models for α and β with $X_i \in \widehat{\mathcal{U}}, i = 1, \dots, n$.

In Steps 1 and 2, the choice of the regularization parameter λ is crucial because it controls the model's sparsity level. In many research, λ is chosen by cross-validation. However, according to [18], λ chosen from cross-validation selects too many noise variables in a high-dimensional setting. We modify the R function `cv.ncvreg` in the `ncvreg` package so that cross-validation selects the regularization parameter $(\lambda_a,\ \lambda_b)$ from a pre-range of λ. In this way, we can prevent over-selecting. `cv.ncvreg` solves the estimating function using a coordinate descent algorithm. The coordinate descent algorithms minimize the target function with respect to a single parameter at a time, with other components of the variable vector X being fixed at their current values. If specifying $(\lambda_a,\ \lambda_b)$ is difficult, another approach for resolving the overselecting problem is to utilize other variable selection methods, such as the Adaptive LASSO [41]. See Section S3 in the supplementary material https://link.intlpress.com/suppfile/sii/SII-18-1-a7-CHO-supplement.pdf.

4. ASYMPTOTIC RESULTS FOR VARIABLE SELECTION AND ESTIMATION

Under certain regularity conditions given in [8], $\tilde{\alpha}$ and $\tilde{\beta}_a$ satisfy the selection consistency and the oracle properties under penalized likelihood for both linear regression and logistic regression. Hence, we can obtain $\|\tilde{\alpha} - \alpha^*\|_2 = O_p\{(p/n)^{1/2}\}$ and $\|\tilde{\beta}_a - \beta_a^*\|_2 = O_p\{(p/n)^{1/2}\}$, for $a = 1, 2$. Now we focus on the asymptotic behavior of the AIPW estimator based on $(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$, which are obtained by fitting the OM and PS models. We consider an influence function to study the asymptotic properties of the proposed estimator. Under mild regularity conditions [e.g., 21],

$$\hat{\tau}(\alpha^*, \beta_0^*, \beta_1^*) - \tau = \frac{1}{n}\sum_{i=1}^{n}\psi(A_i, X_i, Y_i) + o_p(1),$$

where $\psi(A, X, Y)$ is the influence function of $\hat{\tau}$ with $E(\psi) = 0$ and $E(\psi^2) < \infty$ [4].

Let

$$\begin{aligned}\Sigma_\alpha &= E\left\{S^{\otimes 2}(A, X; \alpha)\right\}\\ &= E\left[\frac{1}{e(X;\alpha^*)\{1-e(X;\alpha^*)\}}\left\{\frac{\partial e(X;\alpha^*)}{\partial \alpha}\right\}^{\otimes 2}\right]\end{aligned}$$

be the Fisher information matrix for α in the PS model. For simplicity, denote

$$\begin{aligned}e_i^* &= e(X_i;\alpha^*),\\ \dot{e}_i^* &= \partial e(X_i;\alpha^*)/\partial \alpha^T,\\ S_i^* &= S(A_i, X_i; \alpha^*),\\ \mu_{ai}^* &= \mu_a(X_i;\beta_a^*),\\ \dot{\mu}_{ai}^* &= \partial \mu_a(X_i;\beta_a^*)/\partial \beta_a^T,\\ S_{ai}^* &= S_a(A_i, X_i, Y_i; \beta_a^*),\\ \dot{S}_{ai}^* &= \partial S_a(A_i, X_i, Y_i; \beta_a^*)/\partial \beta_a^T,\end{aligned}$$

for $a = 0, 1$. Under Assumption 3 or Assumption 4, the influence function for $\hat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$ can be written as following:

$$\begin{aligned}(8)\quad &\psi(A_i, X_i, Y_i)\\ &= \frac{A_iY_i}{e_i^*} + \left(1 - \frac{A_i}{e_i^*}\right)\mu_{1i}^* - \frac{(1-A_i)Y_i}{1-e_i^*} - \left(1 - \frac{1-A_i}{1-e_i^*}\right)\mu_{0i}^*\\ &+ E\left[\left\{\frac{A(Y-\mu_1^*)}{(e^*)^2} + \frac{(1-A)(Y-\mu_0^*)}{(1-e^*)^2}\right\}\dot{e}^*\right]\Sigma_\alpha^{-1}S_i^*\\ &- E\left\{\left(1 - \frac{A}{e^*}\right)\dot{\mu}_1^*\right\}\left\{E(\dot{S}_1^*)\right\}^{-1}S_{1i}^*\\ &+ E\left\{\left(1 - \frac{1-A}{1-e^*}\right)\dot{\mu}_0^*\right\}\left\{E(\dot{S}_0^*)\right\}^{-1}S_{0i}^* - \tau.\end{aligned}$$

The regularity condition for the SCAD and the details of Eq. (8) are presented in the supplementary material. Note that if e^* and μ_a^* is correctly specified, $e^* = e(X;\alpha^*) = e(X)$ and $\mu_a^* = \mu_a(X;\beta_a^*) = \mu_a(X)$, for $a = 1, 2$. Hence, if e^* or μ_a^* is the correct model, the right-hand side of Eq. (8) is zero, as shown in the supplementary material. It follows that $E(\psi) = 0$ and $E(\psi\psi^T) < \infty$. Thus, we have

$$\hat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) - \tau = \frac{1}{n}\sum_{i=1}^{n}\psi(A_i, X_i, Y_i) + o_p(1),$$

As a result, $\sqrt{n}\{\hat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) - \tau\}$ is also asymptotically normal.

Theorem 1. *Under Assumptions 1–4 and the regularity conditions specified in the supplementary material, if either $e(X^T\alpha)$ or $m(X^T\beta)$ is correctly specified,*

$$\sqrt{n}\{\hat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) - \tau\} \to \mathcal{N}\{0, E(\psi\psi^T)\}, \tag{9}$$

in distribution, as $n \to \infty$, where ψ is defined in Eq. (8).

Additionally, when the propensity score and the regression function are correctly specified, $\hat{\tau}$ achieves the semiparametric efficiency bound [10].

Theorem 2 (Double robustness of the proposed estimator)**.** *Under Assumptions 1–2, if either Assumption 3 or Assumption 4 holds, not necessarily both, $\hat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$ in Equation (7) is consistent with τ.*

The proof is given in the supplementary material.

There are two approaches to estimating the asymptotic variance: (a) conducting the bootstrap and (b) using the point estimation for the asymptotic variance, replacing ψ with $\widehat{\psi}$ in (9). For the former, since $\hat{\tau}$ is asymptotically linear and normal, we can estimate the valid variance estimator of $\hat{\tau}$ by bootstrapping original observations. For the latter, we use $\widehat{\psi}$ as an estimator for ψ in Eq. (8), substituting $(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$ for $(\alpha^*, \beta_0^*, \beta_1^*)$. Then, we can estimate the asymptotic variance of $\widehat{\tau}_n(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$, $V\{\hat{\tau}_n(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)\}$, by

$$\widehat{V}\{\hat{\tau}_n(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)\} = \frac{1}{n}\widehat{E}(\widehat{\psi}\widehat{\psi}^T) = \frac{1}{n^2}\sum_{i=1}^{n}\widehat{\psi}\widehat{\psi}^T.$$

Considering the long time it takes to conduct bootstrap, we use the latter in Sections 5 and 6.

5. SIMULATION STUDY

In this section, we conduct a simulation study to evaluate the finite sample performances of the doubly robust ACE estimator with different variable selection strategies. Additionally, the simulation study is done under model misspecification to manifest the robustness of the proposed estimator.

5.1 Simulation setup

We generate the dataset with size $n = 5000$. The covariate $X_i = (1, X_{1,i}, \ldots, X_{p-1,i})^T$ is p-dimensional, where p is set to be 50. The first component is one, and the others are independently generated from the standard normal with mean 0 and variance 1. Table 1 summarizes the structure of pretreatment variables for four scenarios. In all scenarios, the last $p-6$ coefficients were set to 0 in the PS and OM models, representing $p-6$ spurious covariates, and $X_C = \{X_{3,i}, X_{4,i}\}$. In Scenarios 1 and 4, $X_I = \{X_{1,i}, X_{2,i}\}$. In Scenarios 1 and 3, $X_P = \{X_{5,i}, X_{6,i}\}$.

We generate a binary treatment, A_i, from a Bernoulli distribution with the PS. For the PS model, we consider both a linear model (PSM I) and a non-linear model (PSM II):

- PSM I: $\text{logit}(e_i) = \alpha_1^T X_i$,
- PSM II: $\text{logit}(e_i) = 3.5 + \alpha_2^T \log(X_i^2) - \cos(X_{3,i} + X_{4,i})$,

where α_1 and α_2 are $(p-1)$-dimensional vectors of coefficients in the PS model. The true values of α_1 and α_2 are defined differently depending on the scenario structure:

- Scenario 1: $\alpha_1 = (0, 1, 1, 1, 1, 0, \ldots, 0)^T$ and $\alpha_2 = (0, 3, 3, 3, 3, 0, \ldots, 0)^T$,
- Scenario 2: $\alpha_1 = (0, 0, 0, 1, 1, 0, \ldots, 0)^T$ and $\alpha_2 = (0, 0, 0, 3, 3, 0, \ldots, 0)^T$,
- Scenario 3: $\alpha_1 = (0, 0, 0, 1, 1, 0, \ldots, 0)^T$ and $\alpha_2 = (0, 0, 0, 3, 3, 0, \ldots, 0)^T$,
- Scenario 4: $\alpha_1 = (0, 1, 1, 1, 1, 0, \ldots, 0)^T$ and $\alpha_2 = (0, 3, 3, 3, 3, 0, \ldots, 0)^T$.

For generating continuous outcome variable, Y_i, we consider both linear (OM I) and non-linear OM models (OM II):

- OM I: $Y_i = \beta_a^T X_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$, where $a = 0, 1$,
- OM II:
  $Y_0 = 1 + \exp(\sin(\beta_0^T X_i)) - 2\cos(\beta_{0,4}X_{3,i} + \beta_{0,5}X_{4,i}) + \beta_{0,6}X_{5,i} - \beta_{0,7}X_{6,i} + \epsilon_{0,i}$, $\epsilon_{0,i} \sim \mathcal{N}(0, 1)$,
  $Y_1 = 1 + \exp(2\sin(\beta_1^T X_i)) - \cos(\beta_{1,4}X_{3,i} + \beta_{1,5}X_{4,i}) + \beta_{1,6}X_{5,i} - \beta_{1,7}X_{6,i} + \epsilon_{1,i}$, $\epsilon_{1,i} \sim \mathcal{N}(0, 1)$,

where β_0 and β_1 are $(p-1)$-dimensional vectors of coefficients in the OM model, and $\beta_{a,j}$ is the j-th coefficient of β_a $(a = 0, 1)$. The true values of β_0 and β_1 are defined differently depending on the scenario structure:

- Scenario 1: $\beta_0 = (1, 0, 0, 1, 1, 1, 1, \ldots, 0)^T$ and $\beta_1 = (1, 0, 0, 2, 2, 2, 2, \ldots, 0)^T$,
- Scenario 2: $\beta_0 = (1, 0, 0, 1, 1, 0, 0, \ldots, 0)^T$ and $\beta_1 = (1, 0, 0, 2, 2, 0, 0, \ldots, 0)^T$,
- Scenario 3: $\beta_0 = (1, 0, 0, 1, 1, 1, 1, \ldots, 0)^T$ and $\beta_1 = (1, 0, 0, 2, 2, 2, 2, \ldots, 0)^T$,
- Scenario 4: $\beta_0 = (1, 0, 0, 1, 1, 0, 0, \ldots, 0)^T$ and $\beta_1 = (1, 0, 0, 2, 2, 0, 0, \ldots, 0)^T$.

Using the different α and β for each Scenario is to make the true important covariates different according to Scenarios. For example, in Scenario 1, $X_{1,i}, X_{2,i}, X_{3,i}$, and $X_{4,i}$ are important covariates in the PS models and $X_{3,i}, X_{4,i}, X_{5,i}$, and $X_{6,i}$ are in the OM models. In doing so, we can perform the

Table 1. The pretreatment variables in four scenarios

	X_I	X_C	X_P
Scenario 1	$X_{1,i}, X_{2,i}$	$X_{3,i}, X_{4,i}$	$X_{5,i}, X_{6,i}$
Scenario 2		$X_{3,i}, X_{4,i}$	
Scenario 3		$X_{3,i}, X_{4,i}$	$X_{5,i}, X_{6,i}$
Scenario 4	$X_{1,i}, X_{2,i}$	$X_{3,i}, X_{4,i}$	

simulations using different combinations of X_I, X_C, and X_P as in Table 1. Under the above data-generating mechanisms, the true ACE is zero for OM I. For OM II, the true ACE is 1.6031 in Scenarios 1 and 3 and 1.4280 in Scenarios 2 and 4.

For each scenario specified in Table 1, we consider four settings:

- Setting (a): a linear PS model (PSM I) + a linear OM model (OM I),
- Setting (b): a non-linear PS model (PSM II) + a linear OM model (OM I),
- Setting (c): a linear PS model (PSM I) + a non-linear OM model (OM II),
- Setting (d): a non-linear PS model (PSM II) + a non-linear OM model (OM II).

Although in case the data are generated from the non-linear PS or OM models, we estimate the coefficients based on a linear model. Thus, the PS model in Setting (b), the OM model in Setting (c), and both the PS and OM model in Setting (d) are misspecified. In the variable selection steps, the regularization parameter is chosen by cross-validation. To avoid over-selecting, we select λ from λ_{min} to λ_{max}. We set $\lambda_{min} = 0.1, 0.3$, and 0.02 for the linear OM model, the non-linear OM model, and the PS model, respectively. We set λ_{max} to be the default value provided by the package `ncvreg`, which is the largest eigenvalue of the design matrix. In this simulation, we use the same set for the PS model and the OM model in the AIPW estimation. Our proposed estimator is the AIPW estimator using the PS and OM models adjusted by $\widehat{\mathcal{U}}$. We compare the performance of our suggested estimator with the AIPW estimator using the PS and OM models adjusted by $\widehat{\mathcal{I}}$ and $\widehat{\mathcal{M}}_\beta$. Also, our suggested estimator is compared with oracle sets, $\mathcal{U}$, $\mathcal{I}$ and $\mathcal{M}_\beta$, which are the sets consisting of covariates used in the true models. We consider the following estimators:

- `TRUE`: $E(Y_1 - Y_0)$,
- `O-UNI`: the AIPW estimator based on $\mathcal{U}$ for comparison purpose,
- `O-INT`: the AIPW estimator based on $\mathcal{I}$ for comparison purpose,
- `O-OUT`: the AIPW estimator based on $\mathcal{M}_\beta$ for comparison purpose,
- `UNI`: the AIPW estimator based on $\widehat{\mathcal{U}}$ selected by the SCAD (the proposed approach),
- `INT`: the AIPW estimator based on $\widehat{\mathcal{I}}$ selected by the SCAD (the confounder-only approach),
- `OUT`: the AIPW estimator based on $\widehat{\mathcal{M}}_\beta$ selected by the SCAD (the outcome predictor approach).

Each simulation is based on 2000 Monte Carlo runs. We compute the proportion of over-selecting, under-selecting, average false negatives (the average number of not selected covariates that have true nonzero coefficients), and the average false positives (the average number of selected covariates that have true zero coefficients) for each simulation. We estimate the ACE and obtain the coverage rates of the 95% confidence interval.

5.2 Simulation results

Table 2 summarizes the selection performance of the proposed penalization procedure for all scenarios in terms of the proportion of over-selecting (Over), under-selecting (Under), average false negative (FN), and average false positives (FP). The under-selecting proportions for the proposed method are all zeros under the true model specification, implying that most of the true nonzero coefficients are selected by Steps I and II in the proposed procedure.

Figure 2 displays the distribution of the ACE estimates for all scenarios. `UNI` and `OUT` maintain the efficiency as much as their oracle estimators, while `INT` is more variable than its oracle estimator under the misspecification. All methods are more unstable when the OM model is misspecified than when it is correct. `O-UNI` and `UNI` have more considerable variability than `O-OUT` and `OUT` in Setting (a) and (c) of Scenarios 1 and 4, where X_I exists, and the PS model is linear. However, `UNI` has the advantage that when X_I are not useful for estimation, it does not use those variables for the AIPW estimator. In the setting (b), where the PS model is misspecified, most of X_I are dropped from $\widehat{\mathcal{U}}$, and `UNI` keeps efficiency as much as `OUT`. `INT` is more variable under Setting (c) in Scenarios 1 and 3 since the X_P are not considered. This phenomenon is consistent with the findings in the previous research that including X_I inflates standard errors while including X_P reduces standard errors [29, 30, 24].

In Setting (a)–(c), `O-UNI` and `UNI` are doubly robust in the sense that it is unbiased, provided that either the OM model or the PS model is correctly specified. Contrarily, `INT` and `OUT` show large biases compared to `O-INT` and `O-OUT` in Setting (c). Theoretically, `INT` and `OUT` should be unbiased as in `O-INT` and `O-OUT` since the PS model is correct. However, `INT` and `OUT` are no longer unbiased because they do not consider X_I necessary to maintain the double robustness of the AIPW estimator. These results show why we need to consider not only X_C and X_P but X_I as well for the estimation when the OM model may be misspecified. When both models are unknown, as in Setting (d), there is not much difference among `UNI, INT`, and `OUT` in terms of bias and efficiency because most of the variables are not selected in the models.

Table 3 displays the coverage rates for all scenarios. The results show that our coverage rates are close to the nominal coverage if either the OM or PS model is correctly specified. In contrast, the coverage rates for other approaches fail to reach the nominal coverage rate if the OM model is misspecified.

Table 2. Simulation results for the selection performance for the proposed penalization procedure

	β^*				α^*			
	Over ($\times 10^2$)	Under ($\times 10^2$)	FN	FP	Over ($\times 10^2$)	Under ($\times 10^2$)	FN	FP
Scenario 1								
(a) OM I and PSM I	0.05	0	0	0.0005	10.4	0	0	0.1145
(b) OM I and PSM II	4.25	0	0	0.0475	12.3	100	4	0.14
(c) OM II and PSM I	0	100	2	0	10.4	0	0	0.1145
(d) OM II and PSM II	0.05	100	2	0.0005	12.3	100	4	0.14
Scenario 2								
(a) OM I and PSM I	0	0	0	0	16.95	0	0	0.2355
(b) OM I and PSM II	0	0	0	0	29.7	99.8	2	0.5075
(c) OM II and PSM I	0	100	2	0	16.95	0	0	0.2355
(d) OM II and PSM II	0	100	2	0	29.7	99.8	2	0.5075
Scenario 3								
(a) OM I and PSM I	0	0	0	0	16.95	0	0	0.2355
(b) OM I and PSM II	0	0	0	0	29.7	99.8	2	0.5075
(c) OM II and PSM I	0	99.65	2	0	16.95	0	0	0.2355
(d) OM II and PSM II	0	100	2	0	29.7	99.8	2	0.5075
Scenario 4								
(a) OM I and PSM I	0.05	0	0	0.0005	10.4	0	0	0.1145
(b) OM I and PSM II	3.75	0	0	0.042	12.3	100	4	0.14
(c) OM II and PSM I	0	100	2	0	10.4	0	0	0.1145
(d) OM II and PSM II	0	100	2	0	12.3	100	4	0.14

Note: X_I exists in Scenarios 1 and 4, X_C exists in all Scenarios, and X_P exists in Scenarios 1 and 3. Under OM I (II), the OM model is correctly specified (misspecified), and under PSM I (II), the PS model is correctly specified (misspecified). The results include the proportion of over-selecting, under-selecting, average false negatives, and average false positives for each setting.

6. AN APPLICATION

Low birth weight infants undergo severe health and developmental difficulties, which incurs enormous societal costs. Thus, considerable attention has been focused on finding the causal determinant of an infant's birth weight. Maternal smoking is a significant risk factor for low birth weight infants [14, 34]. Many studies were carried out to determine the relationship between maternal smoking during pregnancy and low birth weight infants. Almond et al. [1] implement a program evaluation approach. Lee et al. [15] obtain a uniformly valid confidence band to show how smoking changes across different age groups of mothers.

The data is available on the STATA website.[1] The sample size for the data is 4262. The outcome of interest Y is infant birth weight measured in grams. The treatment variable A is a binary variable equal to 1 if the mother smokes and 0 otherwise. We are interested in getting the ACE of maternal smoking during pregnancy on infant birth weight using the proposed method. We consider 17 covariates for analysis. The included covariates are an indicator of being married (`mmarried`), an indicator of Hispanic (`mhisp, fhisp`), an indicator of foreign (`foreign`), an indicator of alcohol consumed during pregnancy (`alcohol`), an indicator of newborns died in previous births (`deadkids`), age (`mage, fage`), education attainment (`medu, fedu`), the number of prenatal care visits (`nprenatal`), months since last birth (`monthslb`), the order of birth of the infant (`order`), race (`mrace, frace`), trimester of first prenatal care visit (`prenatal`), and the month of birth (`birthmonth`). Additionally, we add quadratic terms of the five continuous variables and 26 interaction terms significant in either the PS model or the OM model. Therefore, the total number of covariates is 48.

Figure 3 displays the standardized mean difference for the covariates without an asterisk and the raw difference in means for the covariate with an asterisk. Note that the distribution of covariates is not balanced, which indicates the simple difference between the two treatment groups can introduce bias for the ACE. To estimate the ACE with our estimator, we assume the PS model to be a logistic regression model and the OM model to be a linear regression model. We estimate the standard errors using the asymptotic variance of Eq. (8).

Table 4 summarizes the selection results. There are 15 instrumental variables, 21 confounding variables, and four instrumental variables, which is similar to Scenario 1 in Section 5.

Table 5 displays the point estimates, the standard errors, and the 95% Wald confidence intervals. The result shows a similar pattern to Setting (c) in Scenario 1. `UNI` has a larger standard error than `INT` and `OUT`. Also, the estimate of `UNI`

[1] http://www.stata-press.com/data/r13/cattaneo2.dta

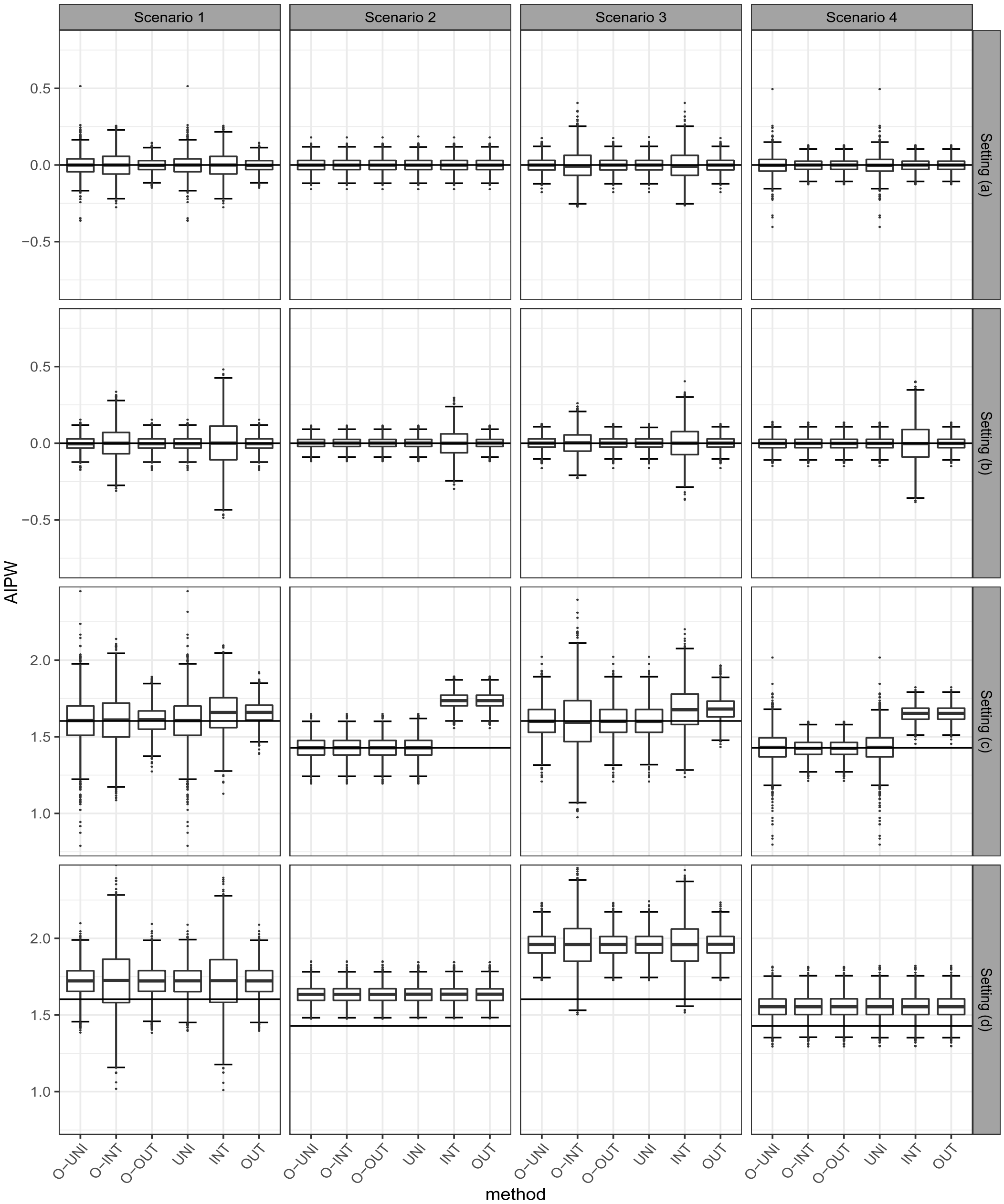

Figure 2. Estimation results under four Scenarios. X_I exists in Scenarios 1 and 4, X_C exists in all Scenarios, and X_P exists in Scenarios 1 and 3. The OM model is correctly specified (misspecified) in Settings (a) and (b) (Settings (c) and (d)), and the PS model is correctly specified (misspecified) in Settings (a) and (c) (Setting (b) and (d)).

is different from `INT` and `OUT`. As seen by the simulation in Section 5, `INT` and `OUT` may be biased due to the use of a wrong set, while the proposed method may correct the bias by its doubly robust property. With the proposed estimator, maternal smoking reduces birth weight by 218.67g on average, which is a smaller decrease than those with `INT` and `OUT`. All 95% confidence intervals do not include 0, which means it is significant at the 0.05 level that maternal smoking has a negative effect on birth weight.

7. CONCLUDING REMARKS

We establish the two-stage procedure to estimate the ACE with variable selection and the AIPW estimator. We compare the robustness of the AIPW estimator coupled with the union, intersection, and outcome predictor strategies using extensive simulation. Our method is most robust, remaining consistent if either the OM model or the PS model is correctly specified. Other methods fail to be doubly robust under the misspecification of the OM model. When the

Table 3. Simulation results for coverage rates

	O-UNI	O-INT	O-OUT	UNI	INT	OUT
			Scenario 1			
(a) OM I and PSM I	0.939	0.944	0.942	0.938	0.946	0.942
(b) OM I and PSM II	0.954	0.945	0.952	0.953	0.960	0.952
(c) OM II and PSM I	0.950	0.950	0.945	0.950	0.929	0.873
(d) OM II and PSM II	0.791	0.918	0.788	0.788	0.922	0.786
			Scenario 2			
(a) OM I and PSM I	0.955	0.955	0.955	0.955	0.955	0.955
(b) OM I and PSM II	0.947	0.947	0.947	0.947	0.967	0.947
(c) OM II and PSM I	0.949	0.949	0.949	0.948	0.000	0.000
(d) OM II and PSM II	0.042	0.042	0.042	0.041	0.039	0.039
			Scenario 3			
(a) OM I and PSM I	0.947	0.951	0.947	0.946	0.953	0.947
(b) OM I and PSM II	0.950	0.954	0.950	0.949	0.962	0.950
(c) OM II and PSM I	0.948	0.953	0.948	0.949	0.916	0.805
(d) OM II and PSM II	0.004	0.393	0.004	0.004	0.391	0.004
			Scenario 4			
(a) OM I and PSM I	0.941	0.945	0.945	0.941	0.945	0.945
(b) OM I and PSM II	0.945	0.945	0.945	0.945	0.962	0.945
(c) OM II and PSM I	0.940	0.952	0.952	0.941	0.008	0.008
(d) OM II and PSM II	0.622	0.621	0.621	0.627	0.623	0.623

Note: X_I exists in Scenarios 1 and 4, X_C exists in all Scenarios, and X_P exists in Scenarios 1 and 3. Under OM I (II), the OM model is correctly specified (misspecified), and under PSM I (II), the PS model is correctly specified (misspecified).

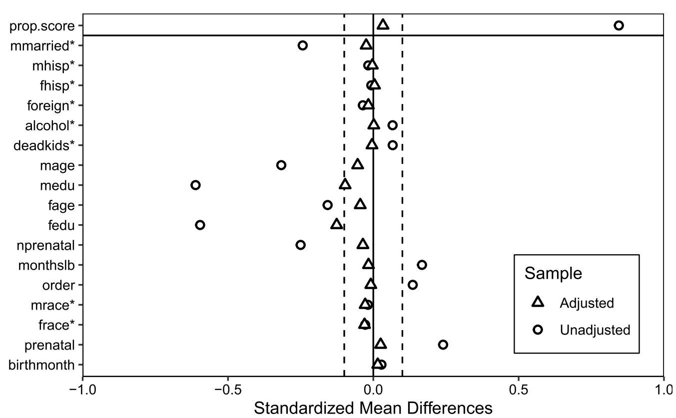

Figure 3. Balance check: standardized mean difference for covariates of birth weight data. The dashed vertical lines are drawn at 0.1 SMD.

instrumental variables are selected for estimation, our procedure may be more variable than other approaches. However, the inefficiency is offset by including precision variables. Thus, when there are precision variables, the AIPW estimator based on the proposed variable selection strategy is less variable than that based on the intersection strategy. Our simulation results also imply that all strategies are badly biased in the case when the OM model and the PS model are misspecified. Thus, we still need a correct specification of either the PS or OM model for consistent estimation of our procedure. Although we employ the SCAD for penalization, our method is flexible in the sense that other penalization methods, such as LASSO or Minimax concave penalty, can be used to select variables at the first stage. In particular, when there is a high correlation among variables,

Table 4. Selection result for birth weight data

	Selected variables
X_I	`mhisp,medu,fage,frace,birthmonth,mage,fedu, nprenatal:monthslb, nprenatal:order,alcohol:medu, mmarried:foreign,mmarried:mage,foreign:nprenatal, alcohol:fedu, medu:fedu`
X_C	`mmarried,mage,fedu,nprenatal,order,mrace,prenatal, age,nprenatal,mmarried:mrace,alcohol:mage, alcohol:nprenatal,deadkids:medu,monthslb:prenatal, fhisp:order,foreign:mage,foreign:mrace, mhisp:order,mmarried:fhisp,deadkids:prenatal, frace:birthmonth`
X_P	`fhisp,foreign,deadkids:order,deadkids:birthmonth`
	$p = 48$, $\lvert U\rvert = 40$, $\lvert I\rvert = 21$, $\lvert M_\beta\rvert = 25$

Table 5. Point estimate, standard error, and 95% Wald confidence interval for birth weight data

	Est	SE	CI
UNI	−218.67	50.54	(−317.73, −119.62)
INT	−229.38	28.19	(−284.62, −174.13)
OUT	−229.94	27.31	(−283.46, −176.43)

introducing the L_2-penalty may improve the performance of Steps 1 and 2. Elastic net proposed by [42] outperforms the LASSO, and the SCAD-L_2 performs better in terms of minimizing prediction error and maintaining variable selection precision than the SCAD [40].

There are several directions for future work: (i) we will extend the results to the causal analysis of longitudinal observational studies [36] and survival outcomes [38]; and (ii) we will develop variable selection procedures when confounders are subject to missingness [39], which is common-place in practice.

ACKNOWLEDGEMENTS

Yang is partially supported by the NIA grant 1R01AG066883 and NIEHS grant 1R01ES031651.

Received 28 December 2021

REFERENCES

[1] Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

[2] Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973. MR2216189

[3] Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650. MR3207983

[4] Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore. MR1245941

[5] 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(12):1149–1156.

[6] Chowdhury, M. Z. I. and Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8(1):1–7.

[7] Ertefaie, A., Asgharian, M., and Stephens, D. A. (2018). Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, 6(1):1–16. MR4351484

[8] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360. MR1946581

[9] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484. MR2849368

[10] Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.

[11] Henckel, L., Perković, E., and Maathuis, M. H. (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*. MR4412998

[12] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

[13] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. MR3309951

[14] Kramer, M. S. (1987). Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization*, 65(5):663–737.

[15] Lee, S., Okui, R., and Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32(7):1207–1225. MR3734484

[16] Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400. MR3803473

[17] 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(19):2937–2960.

[18] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473. MR2758523

[19] Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay de principle. English translation of excerpts (1990) by D. Dabrowska and T. Speed. *Statistical Science*, 5:463–472. MR1092985

[20] 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(6):551–559.

[21] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866. MR1294730

[22] Rosenbaum, P. R. (2002). Overt bias in observational studies. In *Observational Studies*, pages 71–104. Springer. MR1353914

[23] Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339. MR1666631

[24] 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(188):1–86. MR4209474

[25] Rotnitzky, A. and Vansteelandt, S. (2014). Double-robust methods. In *Handbook of Missing Data Methodology*, pages 185–212. CRC Press.

[26] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

[27] Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

[28] Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120. MR1731478

[29] Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122. MR3744525

[30] Tang, D., Kong, D., Pan, W., and Wang, L. (2020). Outcome model free causal inference with ultra-high dimensional covariates. *arXiv preprint arXiv:2007.14190*.

[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. MR1379242

[32] VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219.

[33] VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413. MR2872391

[34] Vogler, G. P. and Kozlowski, L. T. (2002). Differential influence of maternal smoking on infant birth weight: gene-environment interaction and targeted intervention. *JAMA*,

287(2):241–242.

[35] Wilson, A. and Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, 70(4):852–861. MR3295746

[36] Yang, S. (2021). Semiparametric efficient estimation of structural nested mean models with irregularly spaced observations. *Biometrics*, 10.1111/biom.13471:in press. MR4493499

[37] Yang, S., Kim, J. K., and Song, R. (2020a). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82:445–465. MR4084171

[38] Yang, S., Pieper, K., and Cools, F. (2020b). Semiparametric estimation of structural failure time models in continuous-time processes. *Biometrika*, 107:123–136. MR4064144

[39] Yang, S., Wang, L., and Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika*, 106:875–888. MR4031203

[40] Zeng, L. and Xie, J. (2014). Group variable selection via scad-l 2. *Statistics*, 48(1):49–66. MR3175757

[41] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429. MR2279469

[42] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320. MR2137327

Eunah Cho
AI/Big Data Analysis Team
LG Display
245, LG-ro
Wollong-myeon
Paju-si, Gyeonggi-do,
The Republic of Korea
E-mail address: eunah.cho@lgdisplay.com

Shu Yang
Department of Statistics
2311 Stinson Drive
5109 SAS Hall
Campus Box 8203
NC State University
Raleigh, North Carolina 27695
United States
E-mail address: syang24@ncsu.edu

Supporting Information

The online supporting information includes technical details and proofs.

Supporting Information for "Variable Selection for Doubly Robust Causal Inference" by Cho and Yang

The supporting information provides technical details. Section S1 lists the regularity conditions. Section S2 provides proof of the double robustness of the proposed estimator.

S1 Regularity conditions

Assumption S1 *Following Fan and Li (2001), we assume that the following regularity conditions hold the selection consistency and the oracle property.*

i) $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n} > 0$.

ii) *The observation V_i are independent and identically distributed with probability density $f(V,\beta)$ with respect to some measure μ. In this paper, The observation V_i is (Y_i, X_i) or (A_i, X_i). The parameter of interest β in $f(V,\beta)$ is α in the PS model or β in the OM model. λ_n is a regularization parameter indexed by the sample size of n. The probability density has common support, and the model is identifiable. Furthermore, the first and second logarithmic derivatives of f satisfying the equations*

$$E_\beta\left[\frac{\partial \log f(V,\beta)}{\partial \beta_j}\right] = 0 \ \ for\, j = 1, ..., p$$

and

$$\begin{aligned} I_{jk}(\beta) &= E_\beta\left[\frac{\partial}{\partial \beta_j}\log f(V,\beta)\frac{\partial}{\partial \beta_k}\log f(V,\beta)\right] \\ &= E_\beta\left[-\frac{\partial^2}{\partial \beta_j \partial \beta_k}\log f(V,\beta)\right] \end{aligned}$$

iii) *The Fisher information matrix $I(\beta) = E\left\{\left[\frac{\partial}{\partial \beta}\log f(V,\beta)\right]\left[\frac{\partial}{\partial \beta}\log f(V,\beta)\right]^T\right\}$ is finite and positive definite at $\beta = \beta_0$.*

iv) *There exists an open subset ω of Ω that contains the true parameter point β_0 such that for almost all V the density $f(V,\beta)$ admits all third derivatives $(\partial(V,\beta))/(\partial\beta_j\partial\beta_k\partial\beta_l)$ for all $\beta \in \omega$. Furthermore, there exist function M_{jkl} such that*

$$\left|\frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \beta_l}\log f(V,\beta)\right| \leq M_{jkl}(V) \ for\ all\ \beta \in \omega$$

where $m_{jkl} = E_{\beta_0}[M_{jkl}(V)] < \infty$ for j, k, l.

v) $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$

S2 Double robustness

By the Taylor expansion following Yang and Ding (2020), we obtain

$$\begin{aligned} \widehat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) - \tau = \widehat{\tau}(\alpha^*, \beta_0^*, \beta_1^*) - \tau & \\ + n^{-1}\sum E\left[\left\{\frac{A(Y-\mu_1^*)}{(e*)^2} + \frac{(1-A)(Y-\mu_0^*)}{(1-e^*)^2}\right\}\dot{e}^*\right]\Sigma_\alpha^{-1}S_i^* & \\ - n^{-1}\sum E\left\{\left(1-\frac{A}{e^*}\right)\dot{\mu}_1^*\right\}\left\{E(\dot{S}_1^*)\right\}^{-1}S_{1i}^* & \\ + n^{-1}\sum E\left\{\left(1-\frac{1-A}{1-e^*}\right)\dot{\mu}_0^*\right\}\left\{E(\dot{S}_0^*)\right\}^{-1}S_{0i}^*. & \end{aligned} \qquad \text{(S1)}$$

We can define the summand of the right-hand side of (S1) as the influence function $\psi(A_i, X_i, Y_i)$. Then, we have

$$\widehat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) - \tau = n^{-1} \sum \psi(A_i, X_i, Y_i) + o_p(1),$$

where

$$\begin{aligned}
\psi(A_i, X_i, Y_i) = & \frac{A_i Y_i}{e_i^*} + \left(1 - \frac{A_i}{e_i^*}\right) \mu_{1i}^* - \frac{(1-A_i)Y_i}{1-e_i^*} - \left(1 - \frac{1-A_i}{1-e_i^*}\right) \mu_{0i}^* - \tau \\
& + E\left[\left\{\frac{A(Y-\mu_1^*)}{(e^*)^2} + \frac{(1-A)(Y-\mu_0^*)}{(1-e^*)^2}\right\} \dot{e}^*\right] \Sigma_\alpha^{-1} S_i^* && \text{(S2)} \\
& - E\left\{\left(1 - \frac{A}{e^*}\right) \dot{\mu}_1^*\right\} \left\{E(\dot{S}_1^*)\right\}^{-1} S_{1i}^* && \text{(S3)} \\
& + E\left\{\left(1 - \frac{1-A}{1-e^*}\right) \dot{\mu}_0^*\right\} \left\{E(\dot{S}_0^*)\right\}^{-1} S_{0i}^*. && \text{(S4)}
\end{aligned}$$

In order for $\sqrt{n}(\widehat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) - \tau)$ to be asymptotically normal and for $\widehat{\tau}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$ to be consistent for τ, it is suffice to show $E(\psi)$=0. Note that by Assumptions 3 and 4, if the PS model is correctly specified, but the OM model is misspecified, then $e(X, \alpha^*) = e(X) = E(A|X)$ and $\mu_a(X, \beta_a^*) \neq \mu_a(X)$, for $a = 1, 2$. If the OM model is correctly specified, but the PS model is misspecified, then $e(X, \alpha^*) \neq e(X) = E(A|X)$ and $\mu_a(X, \beta_a^*) = \mu_a(X)$, for $a = 1, 2$. In addition, by Assumption 1, $\mu_a(X) = E(Y|A = a, X) = E(Y(a)|X)$. Following Tsiatis (2007) and Glynn and Quinn (2010),

$$E\left\{\frac{AY}{e^*} + \left(1 - \frac{A}{e^*}\right) \mu_1^* - \frac{(1-A)Y}{1-e^*} - \left(1 - \frac{1-A}{1-e^*}\right) \mu_{0i}^*\right\} \tag{S5}$$

can be written as

$$\begin{aligned}
& E\left[Y(1) - Y(0)\right] \\
& + E\left[\frac{(A-e^*)[E\{Y(1)|A,X\} - \mu^*]}{e^*}\right] + E\left[\frac{(A-e^*)[E\{Y(0)|A,X\} - \mu^*]}{1-e^*}\right].
\end{aligned} \tag{S6}$$

It is well known that Eq. (S5) is doubly robust because if either the PS model or the OM model is specified correctly, then Eq. (S6) are τ. The rest of the part is to show Eq. (S2) - (S4) are zero if either the PS model or the OM model is correct. Using the law of iterated conditional expectations, we obtain

$$\begin{aligned}
E(S^*) &= E\left\{\frac{A-e^*}{e^*(1-e^*)} \frac{\partial e^*}{\partial \alpha}\right\} \\
&= E\left[E\left\{\frac{A-e^*}{e^*(1-e^*)} \dot{e}_i^* \middle| X\right\}\right] \\
&= E\left[\left\{\frac{E(A|X) - e^*}{e^*(1-e^*)}\right\} \dot{e}_i^*\right],
\end{aligned} \tag{S7}$$

$$\begin{aligned}
E(S_a^*) &= E\left\{\frac{\partial \mu_a^*}{\partial \beta_a} (Y - \mu_a^*)\right\} \\
&= E\left[E\left\{\dot{\mu}_{ai}^* (Y - \mu_a^*) \middle| X\right\}\right] \\
&= E\left[\dot{\mu}_{ai}^* \left\{E(Y|X) - \mu_a^*\right\}\right], \\
& \text{for } a = 1, 2,
\end{aligned} \tag{S8}$$

$$
\begin{aligned}
&E\Bigg[\Bigg\{\frac{A(Y-\mu_1^*)}{(e*)^2}+\frac{(1-A)(Y-\mu_0^*)}{(1-e^*)^2}\Bigg\}\dot{e}^*\Bigg] \\
&=E\Bigg[\Bigg[\Bigg\{\frac{A(Y-\mu_1^*)}{(e*)^2}+\frac{(1-A)(Y-\mu_0^*)}{(1-e^*)^2}\Bigg\}\dot{e}^*\Bigg|A,X\Bigg]\Bigg] \\
&=E\Bigg[\frac{A}{(e^*)^2}\left\{E(Y|A=1,X)-\mu_1^*\right\}\dot{e}^*\Bigg] \\
&\quad+E\Bigg[\frac{(1-A)}{(e^*)^2}\left\{E(Y|A=0,X)-\mu_0^*\right\}\dot{e}^*\Bigg],
\end{aligned} \tag{S9}
$$

$$
\begin{aligned}
E\Bigg\{\Bigg(1-\frac{A}{e^*}\Bigg)\dot{\mu}_1^*\Bigg\} &= E\Bigg[E\Bigg\{\Bigg(1-\frac{A}{e^*}\Bigg)\dot{\mu}_1^*\Bigg|X\Bigg\}\Bigg] \\
&= E\Bigg\{\Bigg(1-\frac{E(A|X)}{e^*}\Bigg)\dot{\mu}_1^*\Bigg\},
\end{aligned} \tag{S10}
$$

and

$$
\begin{aligned}
E\Bigg\{\Bigg(1-\frac{1-A}{1-e^*}\Bigg)\dot{\mu}_0^*\Bigg\} &= E\Bigg[E\Bigg\{\Bigg(1-\frac{1-A}{1-e^*}\Bigg)\dot{\mu}_0^*\Bigg|X\Bigg]\Bigg\} \\
&= E\Bigg\{\Bigg(1-\frac{1-E(A|X)}{1-e^*}\Bigg)\dot{\mu}_0^*\Bigg\}.
\end{aligned} \tag{S11}
$$

If the OM model is correctly specified, but the PS model is misspecified, then Eq. (S8) and (S9) are zero, but Eq. (S7), (S10), and (S11) are not. In this case, E(ψ) of `UNI`, `INT`, and `OUT` are all zero, and $\widehat{\tau}(\hat{\alpha},\hat{\beta}_0,\hat{\beta}_1)$ is consistent for τ. If the PS model is correctly specified, but the OM model is misspecified, Eq. (S8) and (S9) do not disappear as expected, and Eq. (S7), (S10), and (S11) are also always not zero. Since `OUT` and `INT` do not consider the instrumental variables and cannot fit the PS model well, only $E(\psi)$ of `UNI` is zero. Therefore $\widehat{\tau}(\hat{\alpha},\hat{\beta}_0,\hat{\beta}_1)$ is doubly robust only for `UNI`.

S3 Over-selection problem

As mentioned in subsection 3.6, when using the default range provided by the `ncvreg` package, the cross-validation overselects variables. Table 1 presents the selection performance when the default range is used.

Table 1: Simulation results for the selection performance for the proposed penalization procedure with the default value

	β^*				α^*			
	Over ($\times 10^2$)	Under ($\times 10^2$)	FN	FP	Over ($\times 10^2$)	Under ($\times 10^2$)	FN	FP
Scenario 1								
(a) OM I and PSM I	55.0	0.0	0.00	2.65	31.7	0.0	0.00	1.33
(b) OM I and PSM II	53.6	0.0	0.00	2.63	36.1	99.8	3.73	1.41
(c) OM II and PSM I	67.3	0.0	0.00	3.45	31.7	0.0	0.00	1.33
(d) OM II and PSM II	89.8	11.7	0.17	6.05	36.1	99.8	3.73	1.41
Scenario 2								
(a) OM I and PSM I	53.1	0.0	0.00	2.79	33.7	0.0	0.00	1.46
(b) OM I and PSM II	54.5	0.0	0.00	2.68	38.7	98.7	1.88	1.49
(c) OM II and PSM I	96.9	4.1	0.05	9.37	33.7	0.0	0.00	1.46
(d) OM II and PSM II	91.3	0.0	0.00	6.07	38.7	98.7	1.88	1.49
Scenario 3								
(a) OM I and PSM I	53.3	0.0	0.00	2.78	33.7	0.0	0.00	1.46
(b) OM I and PSM II	54.2	0.0	0.00	2.62	38.7	98.7	1.88	1.49
(c) OM II and PSM I	62.2	0.0	0.00	3.33	33.7	0.0	0.00	1.46
(d) OM II and PSM II	93.0	5.4	0.07	6.34	38.7	98.7	1.88	1.49
Scenario 4								
(a) OM I and PSM I	56.1	0.0	0.00	2.78	31.7	0.0	0.00	1.33
(b) OM I and PSM II	54.5	0.0	0.00	2.70	36.1	99.8	3.73	1.41
(c) OM II and PSM I	70.5	84.6	1.37	4.02	31.7	0.0	0.00	1.33
(d) OM II and PSM II	68.8	0.0	0.00	3.83	36.1	99.8	3.73	1.41

Note: X_I exists in Scenarios 1 and 4, X_C exists in all Scenarios, and X_P exists in Scenarios 1 and 3. Under OM I (II), the OM model is correctly specified (misspecified), and under PSM I (II), the PS model is correctly specified (misspecified). The results include the proportion of over-selecting, under-selecting, average false negatives, and the average false positives for each setting.

If specifying (λ_a, λ_b) is difficult, then other strategies, such as the Adaptive LASSO can be used to avoid over-selection. Table 2 presents the selection performance of the Adaptive LASSO. The Adaptive LASSO resolves the over-selection problem of β and improves the over-selection problem of α in the linear PS model.

Table 2: Simulation results for the selection performance for the proposed penalization procedure with the default value

	β^* Over ($\times 10^2$)	β^* Under ($\times 10^2$)	β^* FN	β^* FP	α^* Over ($\times 10^2$)	α^* Under ($\times 10^2$)	α^* FN	α^* FP
Scenario 1								
(a) OM I and PSM I	0	0	0.00	0.00	0	0	0.00	0.00
(b) OM I and PSM II	0	0	0.00	0.00	87.5	99.6	3.36	3.63
(c) OM II and PSM I	0	0.6	0.01	0.00	0	0	0.00	0.00
(d) OM II and PSM II	0	100	2.00	0.00	87.5	99.6	3.36	3.63
Scenario 2								
(a) OM I and PSM I	0	0	0.00	0.00	0	0	0.00	0.00
(b) OM I and PSM II	0	0	0.00	0.00	94.4	97.8	1.75	4.17
(c) OM II and PSM I	0	98.2	1.94	0.00	0	0	0.00	0.00
(d) OM II and PSM II	0.7	0	0.00	0.01	94.4	97.8	1.75	4.17
Scenario 3								
(a) OM I and PSM I	0	0	0.00	0.00	0	0	0.00	0.00
(b) OM I and PSM II	0	0	0.00	0.00	94.4	97.8	1.75	4.17
(c) OM II and PSM I	0	0	0.00	0.00	0	0	0.00	0.00
(d) OM II and PSM II	0	99.2	1.97	0.00	94.4	97.8	1.75	4.17
Scenario 4								
(a) OM I and PSM I	0	0	0.00	0.00	0	0	0.00	0.00
(b) OM I and PSM II	0	0	0.00	0.00	87.5	99.6	3.36	3.63
(c) OM II and PSM I	0	100	2.00	0.00	0	0	0.00	0.00
(d) OM II and PSM II	0.5	0	0.00	0.02	87.5	99.6	3.36	3.63

Note: X_I exists in Scenarios 1 and 4, X_C exists in all Scenarios, and X_P exists in Scenarios 1 and 3. Under OM I (II), the OM model is correctly specified (misspecified), and under PSM I (II), the PS model is correctly specified (misspecified). The results include the proportion of over-selecting, under-selecting, average false negatives, and the average false positives for each setting.

Figure 1 displays the distribution of the ACE estimates when using the Adaptive LASSO with the default range of λ provided by `R` package `glmnet`. The results are similar to that in the main text when using the SCAD with a prespecified range of λ for cross-validation. In setting (c) in Scenarios 2 and 4, `INT` and `OUT` is not doubly-robust, while `UNI` is doubly-robust.

References

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554.

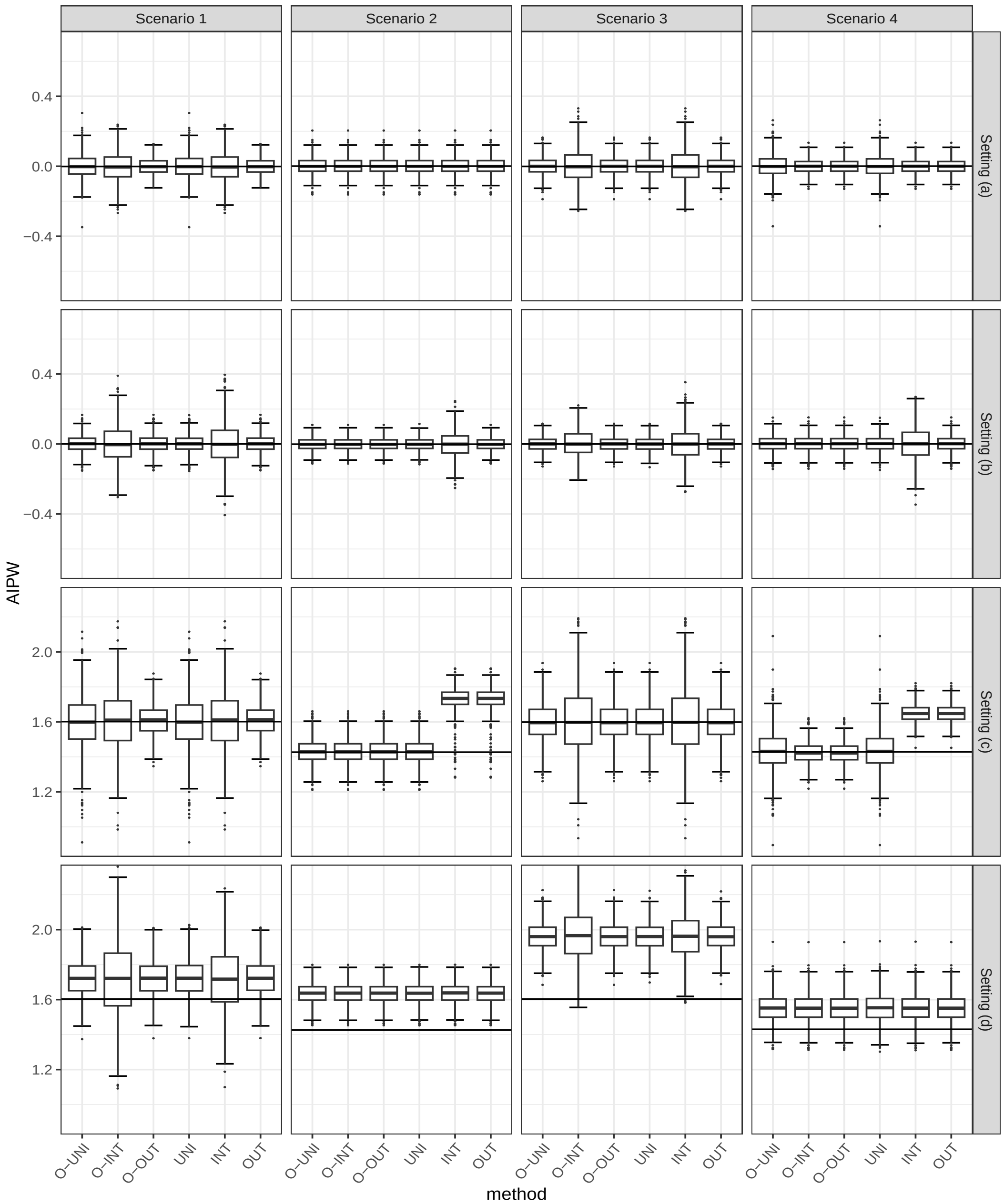

Figure 1: Estimation results under four Scenarios. X_I exists in Scenarios 1 and 4, X_C exists in all Scenarios, and X_P exists in Scenarios 1 and 3. The OM model is correctly specified (misspecified) in Settings (a) and (b) (Settings (c) and (d)), and the PS model is correctly specified (misspecified) in Settings (a) and (c) (Setting (b) and (d)).