

# Variable Selection in Causal Inference Using Penalization

Ashkan Ertefaie †

*University of Michigan, Ann Arbor, USA.*

Masoud Asgharian

David A. Stephens

*McGill University, Montreal, Canada.*

**Summary.** In the causal adjustment setting, variable selection techniques based on either the outcome or treatment allocation model can result in the omission of confounders or the inclusion of spurious variables in the propensity score. We propose a variable selection method based on a penalized likelihood which considers the response and treatment assignment models simultaneously. The proposed method facilitates confounder selection in high-dimensional settings. We show that under some conditions our method attains the oracle property. The selected variables are used to form a double robust regression estimator of the treatment effect. Simulation results are presented and economic growth data are analyzed.

**Keywords:** Causal inference, Average treatment effect, Propensity score, Variable selection, Penalized likelihood, Oracle estimator.

## 1. Introduction

In the analysis of observational data, when attempting to establish the magnitude of the causal effect of treatment (or exposure) in the presence of confounding, the practitioner is faced with certain modeling decisions that facilitate estimation. Should one take the parametric approach, at least one of two statistical models must be proposed; (i) the *conditional mean model* that models the expected response as a function of predictors, and (ii) the *treatment allocation model* that describes the mechanism via which treatment is allocated to (or, at least, received by) individuals in the study, again as a function of the predictors (Rosenbaum & Rubin, 1983; Robins & Brumback, 2000).

Predictors that appear in both mechanisms (i) and (ii) are termed *confounders*, and their omission from model (ii) is typically regarded as a serious error, as it leads to inconsistent estimators of the treatment effect. Thus practitioners usually adopt a conservative approach, and attempt to ensure that they do not omit confounders by fitting a richly parameterized treatment allocation model. The conservative approach, however, can lead to predictors of treatment allocation, but not response, being included in the treatment allocation model. The inclusion of such “spurious” variables in model (ii) is usually regarded as harmless. However, the typical reported forfeit for this conservatism is inflation of variance of the effect estimator (Greenland, 2008; Schisterman et al., 2009). This problem also applies to the conditional mean model, but is in practice less problematic,

†Address for correspondence: Department of Statistics, University of Michigan, Ann Arbor, 48105, Michigan, USA. E-mail: ertefaie@umich.edu

as practitioners seem to be more concerned with bias removal, and therefore more liable to introduce the spurious variables in model (ii). Little formal guidance as to how the practitioner should act in this setting has been provided.

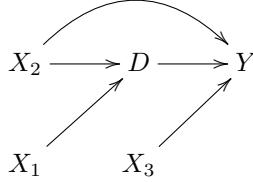
As has been conjectured and studied in simulation by Brookhart et al. (2006a), it is plausible that judicious variable selection may lead to appreciable efficiency gains. However, confounder selection methods based on either just the treatment assignment model, or just the response model, may fail to account for non-ignorable confounders which barely predict the treatment or the response, respectively (Crainiceanu et al., 2008). In this manuscript, we use the term *weak confounder* for these variables. Vansteelandt et al. (2010) shows that confounder selection procedures based on AIC and BIC can be sub-optimal and introduce a method based on the focused information criterion (FIC) which targets the treatment effect by minimizing a prediction mean square error (see also the cross-validation method of Brookhart & van der Laan (2006b)). Van der Laan et al. (2007) introduces a *Super Learner* estimator which is computed by selecting a candidate from a set of estimators obtained from different models using a cross-validation risk (Van der Laan et al., 2004; Sinisi et al., 2007).

Van der Laan & Gruber (2010) selects the sufficient and minimal variables necessary for the propensity score model to estimate an unbiased causal effect by inspecting the efficient influence function (Porter et al., 2011). De Luna et al. (2011) discusses the variance inflation caused by adding the spurious variables in the model and show that it may cause bias as well. Under some assumptions, they also characterize the minimal set of covariates needed for consistent estimation of the treatment effect. Bayesian adjustment for confounding (BAC) is another method introduced by Wang et al. (2012). BAC specifies a prior distribution for a set of possible models which includes a dependence parameter,  $w \in [1, \infty]$ , representing the odds of including a variable in the outcome model given that the same variable is in the treatment mechanism model. Assuming that we *know* a priori that all the predictors of the treatment are in fact confounders, then  $w$  can be set to  $\infty$  (Crainiceanu et al., 2008; Zigler et al., 2013). However, in practice, none of these methods can be used in high-dimensional settings where the number of covariates are larger than sample size.

It is known that *asymptotically* penalizing the conditional outcome model, given treatment and covariates, results in a valid variable selection strategy in causal inference. However, for small to moderate sample sizes this may result in missing weak non-ignorable confounders, which barely predict the outcome but strongly predict the treatment mechanism. The objective of this manuscript is to improve the small sample performance of the outcome penalization strategy while maintaining its asymptotic performance (Table 4). We present a covariate selection procedure which facilitates the estimation of the treatment effect in the *high-dimensional* cases. We parametrize the conditional joint likelihood of the outcome and treatment given covariates such that penalizing this joint likelihood has the ability to select even weak confounders, i.e., confounders which are non-ignorable even if they are barely associated with the outcome or treatment mechanism. This likelihood is just used to identify the set of important covariates, i.e., non-ignorable confounders and predictors of outcome, and, in general, the estimated parameters do not have any causal interpretation. We derive the asymptotic properties of the maximum penalized likelihood estimator using a method that does not require the second derivative of the joint density function. We utilize the selected covariates to estimate the causal effect of interest using our proposed doubly robust estimator.

We restrict our attention to the *unmediated* causal effect (where the effect of exposure on outcome is not mediated by an intermediate variable); in the presence of mediation, direct and indirect effects may not in general be identifiable (Robins & Greenland, 1992; Petersen et al., 2006; Robins et al., 2010; Hafeman & VanderWeele, 2010).

Our  
new  
method  
is  
good

**Fig. 1.** Covariate types: Type-I:  $X_1$ , Type-II:  $X_2$  and Type-III:  $X_3$ .

## 2. Preliminaries & Notation

Let  $Y(d)$  denote the (potential) response to treatment  $d$ , and let  $D$  denote the treatment received. The observed response,  $Y$ , is defined as  $DY(1) + (1 - D)Y(0)$ . We will assume three types of predictors:

- (I) *treatment predictors* ( $X_1$ ), which are related to treatment and not to outcome.
- (II) *confounders* ( $X_2$ ), which are related to both outcome and treatment.
- (III) *outcome predictors* ( $X_3$ ), which are related to outcome and not to treatment;

see the directed acyclic graph (DAG) in Figure 1. We restrict our attention here to the situation where each predictor can be classified into one of these three types, and to single time-point studies. In addition, as is usual, we will make the assumption of *no unmeasured confounders*, that is, that treatment received  $D$  and potential response to treatment  $d$ ,  $Y(d)$ , are independent, given the measured predictors  $X$ . In any practical situation, to facilitate causal inference, the analyst must make an assessment as to the structural nature of the relationships between the variables encoded by the DAG in Figure 1.

### 2.1. The Propensity Score for binary treatments

The propensity score,  $\pi$ , for binary treatment  $D$  is defined as  $\pi(x) = \Pr(D = 1|x)$ , where  $x$  is a  $p$ -dimensional vector of (all) covariates. Rosenbaum & Rubin (1983) show that  $\pi$  is the coarsest function of covariates that exhibits the balancing property, that is,  $D \perp X|\pi$ . As a consequence, the causal effect  $\mu = \mathbb{E}[Y(1) - Y(0)]$  can be computed by iterated expectation

$$\mu = \mathbb{E}_X[\mathbb{E}\{Y(1)|X\} - \mathbb{E}\{Y(0)|X\}] = \mathbb{E}_\pi[\mathbb{E}\{Y(1)|\pi\} - \mathbb{E}\{Y(0)|\pi\}], \quad (1)$$

where  $\mathbb{E}_\pi$  denotes the expectation with respect to the distribution of  $\pi$ . For more details see Rubin (2008) and Rosenbaum (2010).

**Remark 1:** In the standard formulation of the propensity score, no distinction is made between our three types of covariates. Note that, however, for consistent estimation of  $\mu$ , it is not necessary to balance on covariates that are not confounders. Covariates  $X_1$  that predict  $D$  but not  $Y$  may be unbalanced in treated and untreated groups, but will not affect the estimation of the effect of  $D$  on  $Y$ ,

as  $D$  will be conditioned upon, thereby blocking any effect of  $X_1$  (De Luna et al., 2011). Covariates  $X_3$  are unrelated to  $D$ , so will by assumption be in balance in treated and untreated groups in the population. Therefore, the propensity score need only be constructed from confounding variables  $X_2$ ; in this case, it is easy to see that the propensity score,  $\pi_2 = \pi_2(x_2)$ , say, is a balancing score in the sense that  $D \perp X_2 | \pi_2$ : we have  $\Pr(D = 1 | \pi_2(x_2) = t, X_2 = x_2) = t = \Pr(D = 1 | \pi_2(x_2) = t)$ , independent of  $x_2$ , in the usual way. Then, in the presence of outcome predictors  $X_3$  of  $Y$ , the sequel to equation (1) takes the form

$$\begin{aligned}\mu &= \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_{X_2, X_3}[\mathbb{E}\{Y(1)|X_2, X_3\} - \mathbb{E}\{Y(0)|X_2, X_3\}] \\ &= \mathbb{E}_{\pi_2, X_3}[\mathbb{E}\{Y(1)|\pi_2, X_3\} - \mathbb{E}\{Y(0)|\pi_2, X_3\}].\end{aligned}\quad (2)$$

**Remark 2:** Inclusion of covariates that are just related to the outcome in the propensity score model increases the covariance between the fitted  $\pi$  and  $Y$ , decreases the variance of the estimated causal effect, in line with the simulation of Brookhart et al. (2006a).

## 2.2. Penalized Estimation

In a given parametric model, if  $\eta$  is a  $r$ -dimensional regression coefficient,  $p_\lambda(\cdot)$  is a penalty function and  $l_m(\eta)$  is the negative log-likelihood, the maximum penalized likelihood (MPL) estimator  $\hat{\eta}_{ml}$  is defined as

$$\hat{\eta}_{ml} = \arg \min_{\eta} \left[ l_m(\eta) + n \sum_{j=1}^r p_\lambda(|\eta_j|) \right].$$

MPL estimators are shrinkage estimators, and as such, they have more bias, though less variation than unpenalized analogues. Commonly used penalty functions include LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), EN (Zou & Hastie, 2005) and HARD (Antoniadis, 1997).

The remainder of this paper is organized as follows. Section 3 presents our two step variable selection and estimation procedure; we establish its theoretical properties. The performance of the proposed method is studied via simulation in Section 4. We analyze a real data set in Section 5, and Section 6 contains concluding remarks. All the proofs are relegated to the Appendix.

## 3. Penalization and Treatment Effect Estimation

In this section, we develop the methodology which facilitates the estimation of the treatment effect in high-dimensional cases. We separate the covariate selection and the treatment effect estimation procedure. First, we present a reparametrized penalized likelihood which is used to identify the important covariates, and establish the theoretical properties of the resulting MPL estimators. Note that since the likelihood is reparametrized the MPL estimators do not have any causal interpretation. Second, the treatment effect estimation is performed using our doubly robust estimator with the selected covariates in the previous step.

Doubly robust estimation for various outcome covariates and no need for

### 3.1. Likelihood construction

Consider the parametric likelihood  $\mathcal{L}(\eta; y, d, x)$  proportional to

$$\prod_{i=1}^n f(y_i|d_i, g(x_i; \alpha), \beta) P(D = 1|h(x_i, \alpha))^{d_i} P(D = 0|h(x_i, \alpha))^{1-d_i}, \quad (3)$$

where  $\beta$  is an  $r_1$ -dimensional vector parametrizing the association between the outcome and the treatment and  $\alpha$  is an  $r_2$ -dimensional vector containing parameters that appear in the model for  $Y(d)|X$  and  $D|X$ . The functions  $g()$  and  $h()$  used in our joint likelihood have the same form as one would use when modeling the outcome model and treatment mechanism separately. For example, assuming linear working models,  $g(x; \alpha) = \sum_{j=1}^{r_2} \alpha_j x_j$  and  $h(x; \alpha) = \sum_{j=1}^{r_2} \alpha_j x_j$ . Note that for each  $j$ , the parameter  $\alpha_j$  corresponding to  $x_j$  is the same in both models. This is why we call (3) a reparametrized likelihood. We explain the rational behind this reparametrization in section 3.2.

Since our goal is to select the minimal set of covariates necessary for a consistent estimation of the causal effect, we impose a penalty on the parameters  $\alpha$  only; there is no penalization of the  $\beta$  parameters. The penalized pseudo-density for  $z_i = (y_i, d_i, x_i)$  is  $f_p(z_i; \eta) = f(z_i; \eta)f(\alpha)$ , where  $f(z_i; \eta)$  is the joint density used in (3) and  $f(\alpha) = \exp\{-p_{\lambda_n}(\alpha)\}$ . Accordingly, the MPL estimator,  $\hat{\eta}$ , can be defined by

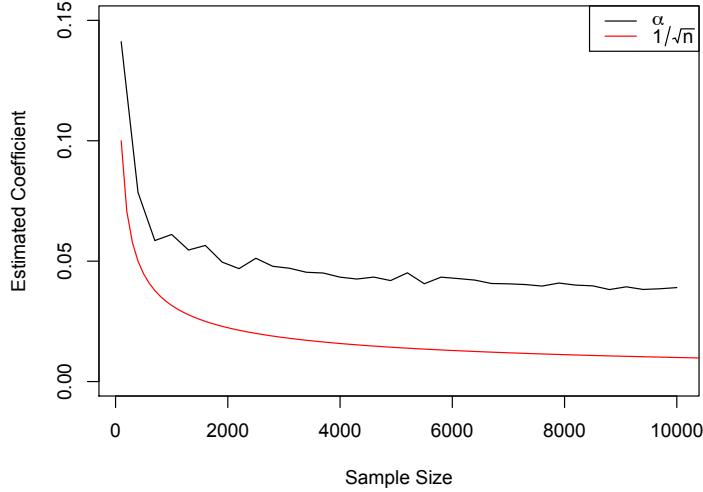
$$\hat{\eta} = \arg \sup_{\eta} \prod_{i=1}^n f_p(z_i; \eta) = \arg \sup_{\eta} \sum_{i=1}^n \log f_p(z_i; \eta).$$

Note the joint density (3) is a misspecification of the true data density. As such, the corresponding penalized likelihood just checks whether  $\alpha_k = 0$  for  $k = 1, \dots, r_2$ , and is not for other estimation purposes. This is discussed in detail in the following subsection.

### 3.2. Avoiding omission of confounders during selection

Standard variable selection techniques based on the conditional outcome/treatment model have the tendency to omit important confounders by ignoring covariates that are weakly associated with the outcome/treatment but strongly associated with treatment/outcome (Vansteelandt et al., 2010). However, our likelihood parametrization in (3), which has the parameter  $\alpha$  in both response and propensity score models, allows us to select such weak confounders. More specifically, our parametrization gives each covariate two chances to appear in the model; once in the response model and once in the treatment allocation model and thus considers both the covariate-exposure and the covariate-outcome association. Our reparametrization has a drawback of setting  $\alpha = 0$  if the tradeoff between the value of the coefficient in the two parts of the likelihood somehow cancel out. In other words, when the association parameter of a variable with the outcome and treatment have opposite signs, then for particular association values, the reparametrized likelihood sets the parameter corresponding to the variable to zero. However, in Appendix B, we show that this particular data generating low has zero measure.

Our proposed parametrization, however, has another drawback that needs to be taken care of. Figure 2 shows that this strategy sets  $\alpha \neq 0$  if a covariate is related to either the outcome or treatment. This figure presents a case where there is just one covariate and the coefficient of this covariate in outcome and treatment models are  $1/\sqrt{n}$  and 0.3, respectively, where  $n$  is the sample size. As it is expected, the estimated parameter  $\alpha$  corresponding to this covariate does not converge to zero when estimated using the likelihood (3) as sample size increases. Hence, our parametrization



**Fig. 2.** Performance of the misspecified likelihood for different sample sizes  $n$ . Red and black lines are  $1/\sqrt{n}$  and the estimated coefficient  $\alpha$  using the reparametrized likelihood.

These types  
are likely:  
- related to treatment  
- related to outcome  
- related to both

gives an equal chance to Type-I and Type-III covariates for selection as key covariates. This may result in over-representing the Type-I variables which is against our goal of keeping variables which are either predictors of the response or non-ignorable confounders. To deal with this problem, we introduce the *boosting* parameter  $\nu$  which boosts covariates Type-III relative to Type-I. The boosting parameter can be defined as  $\nu = \frac{1}{|\tilde{\alpha}_Y|(1+|\tilde{\alpha}_D|)}$ , where  $\tilde{\alpha}_Y$  and  $\tilde{\alpha}_D$  are the least squares (or ridge) estimate of the parameters in the response and treatment models, respectively. Our penalty function is proportional to the boosting parameter,

$$p_{\lambda_n}(\cdot) = \nu p_{\lambda_n}^*(\cdot),$$

where  $p_{\lambda_n}^*(\cdot)$  is a conventional penalty function. Therefore, the magnitude of the penalty on each parameter is proportional to its contribution to the response model. Note that as  $\tilde{\alpha}_Y \rightarrow 0$ , our penalty function puts more penalty on the parameters while considering the covariate-treatment association. For example, when a covariate barely predicts the outcome and treatment, our proposed penalty function imposes a stronger penalty on the parameter compared to a case where a covariate barely predicts the outcome and is strongly related to treatment. For example, when  $p_{\lambda_n}^*(\cdot)$  is lasso, our penalty is  $p_{\lambda_n}(|\alpha_j|) = \lambda_n \nu_j |\alpha_j|$ . A similar argument can be found in the adaptive LASSO (Zou, 2006).

### 3.3. Main Theorems

The following conditions guarantee a consistent penalized estimating procedure for the parameter  $\eta$  with respect to the likelihood (3) which sets the small coefficients to zero for covariate selection.

- P1. For all  $n$ ,  $p_{\lambda_n}(0) = 0$  and  $p_{\lambda_n}(\alpha)$  is non-negative, symmetric about 0 and it is non-decreasing on both  $\mathcal{R}^+$  and  $\mathcal{R}^-$ , i.e. on positive and negative half line. Moreover, it is twice differentiable with derivatives  $p'_{\lambda_n}(\alpha)$  and  $p''_{\lambda_n}(\alpha)$  exist everywhere except at  $\alpha = 0$ .
- P2. As  $n \rightarrow \infty$ ,  $\max_{\alpha \neq 0} [p''_{\lambda_n}(\alpha)] \rightarrow 0$  and  $\max_{\alpha \neq 0} [\sqrt{n} p'_{\lambda_n}(\alpha)] \rightarrow 0$ .
- P3. For  $N_n \equiv (0, B_n)$ ,  $\liminf_{\alpha \in N_n} p'_{\lambda_n}(\alpha) = \infty$ , where  $B_n \rightarrow 0$  as  $n \rightarrow \infty$ .

Assumption P1 is used to prove Theorem 5 given in Appendix E, while P2 prevents the  $j$ th element of the penalized likelihood from being dominated by the penalty function since it vanishes when  $n \rightarrow \infty$ . If  $\alpha_j = 0$ , condition P3 allows the penalty function to dominate the penalized likelihood which leads to the sparsity property.

Suppose the  $r$ -dimensional vector of parameters  $\eta_0 = (\eta_{01}, \eta_{02} = 0)$  is the true values of the parameter  $\eta$ , such that  $\eta_{02} = (\eta_j) = 0$  for  $j = s+1, \dots, r$ ;  $s$  denotes the true number of predictors present in the model (*exact sparsity assumption*). Note that since there is no penalty on the  $\beta$ s,  $\eta_{02}$  consists of those  $\alpha$  that should be shrunk to zero ( $\alpha_j = 0$  for  $j = s', \dots, r_2$ ). Let  $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$  be the vector of MPL estimators corresponding to (3).

Theorem 5 in Appendix E establishes the existence of the consistent penalized maximum likelihood estimator with respect to the joint likelihood (3) under standard regularity conditions (Ibragimov & Has' Minskii (1981)) given as C1-C4 in Appendix A.

The next theorem proves the sparsity and asymptotic normality of the MPL estimators. Let  $I(\eta)$  be the Fisher information matrix derived from the constructed likelihood.

**THEOREM 1. (*Oracle properties*)** Suppose assumptions C1-C4 and P1-P3 are fulfilled and further  $\det[I(\eta)] \neq 0$  for  $\eta \in \Xi$ . Then

- (a)  $\Pr(\hat{\eta}_2 = 0) \rightarrow 1$  as  $n \rightarrow \infty$   
Under additional assumption C5,
- (b)  $\sqrt{n}(\hat{\eta}_{01} - \eta_{01}) \xrightarrow{d} N(0, I^{-1}(\eta_{01}))$ ,

where  $\eta_{01} = (\beta, \alpha_{01})$  and  $\alpha_{01}$  is the true vector of non-zero coefficients.

**Remark 3:** As long as the postulated response and treatment model identify the true non-zero coefficients in each model as  $n \rightarrow \infty$ , the proposed variable selection method consistently identifies the set of non-ignorable confounders. Assuming linear working models, a sufficient but *not* necessary condition for selecting non-ignorable confounders is the linearity of the true models in their parameters. In Appendix F, we conducted simulation studies under different misspecification scenarios where the true models are non-linear in parameters and working models are linear.

### 3.4. Choosing the Tuning Parameter

We select the tuning parameter using the *Generalized Cross Validation* (GCV) method suggested by Tibshirani (1996) and Fan & Li (2001). Let  $\mathbf{W} = (D, \mathbf{X})$ , then

$$\text{GCV}(\lambda) = \frac{\text{RSS}(\lambda)/n}{[1 - d(\lambda)/n]^2},$$

where  $\text{RSS}(\lambda) = \|Y - \mathbf{W}\hat{\eta}\|^2$ ,  $d(\lambda) = \text{trace}[\mathbf{X}(\mathbf{X}'\mathbf{X} + n\Sigma_\lambda(\hat{\eta}))^{-1}\mathbf{X}']$  is the effective number of parameters and  $\Sigma_\lambda(\eta) = \text{diag}[p'_\lambda(|\eta_1|)/|\eta_1|, \dots, p'_\lambda(|\eta_{r_2}|)/|\eta_{r_2}|]$ . The selected tuning parameter  $\hat{\lambda}$  is defined by  $\hat{\lambda} = \arg \min_\lambda \text{GCV}(\lambda)$ .

### 3.5. Estimation

In the treatment effect estimation, we fit the following model using the set of selected covariates in the previous step. Note that a user may want to use other causal adjustment models such as IPTW or propensity score matching.

Our model is a slight modification of the conventional propensity score regression approach of Robins et al. (1992), and specifies

$$\mathbb{E}[Y_i|S_i = s_i, \mathbf{X}_i = \mathbf{x}_i] = \theta s_i + g(\mathbf{x}; \gamma), \quad (4)$$

where  $S_i = D_i - \mathbb{E}[D_i|x_i] = D_i - \pi(\mathbf{x}_i)$ ,  $g(\mathbf{x}; \gamma)$  is a function of covariates and  $\pi$  is the propensity score. The quantity  $S_i$  is used in place of  $D_i$ ; if  $D_i$  is used the fitted model may result in a biased estimator for  $\theta$  since  $g(x; \gamma)$  may be incorrectly specified. By defining  $S_i$  in this way, we restore  $\text{cor}[S_i, X_{ij}] = 0$  for  $j = 1, 2, \dots, p$  where  $p$  is the number of the selected variables (if  $\pi(x_i) = \mathbb{E}[D_i|x_i]$  is correctly specified), as  $\pi(x_i)$  is the (fitted) expected value of  $D_i$ , and hence  $\mathbf{x}'_j(D - \pi(x)) = 0$ , where  $\mathbf{x}'_j = (x_{1j}, \dots, x_{nj})$ . Therefore, misspecification of  $g(\cdot)$  will not result in an inconsistent estimator of  $\theta$ .

In general, this model results in a *doubly robust* estimator (see Davidian et al. (2005), Schafer & Kang (2005) and Bang & Robins (2005)); it yields a consistent estimator of  $\theta$  if *either* the propensity score model or conditional mean model (4) is correctly specified, and is the most efficient estimator (Tsiatis (2006)) when both are correctly specified. For additional details on the related asymptotic and finite sample behavior, see Kang & Schafer (2007), Neugebauer & van der Laan (2005), van der Laan & Robins (2003) and Robins (1999).

The model chosen for estimation of the treatment effect is data dependent. Owing to the inherited uncertainty in the selected model, making statistical inference about the treatment effect becomes “post-selection inference”. Hence, inference about the treatment effect obtained in the estimation step needs to be done cautiously. The weak consistency of the estimator results from the following theorem.

**THEOREM 2.** *Let  $\zeta(\hat{\theta}_{M_n}, M_n)$  be a smooth function of  $\hat{\theta}_{M_n}$  where  $M_n$  is a set of selected variables using our method. Then  $\zeta(\hat{\theta}_{M_n}, M_n) \xrightarrow{P} \zeta(\theta_{M_0}, M_0)$  as  $n \rightarrow \infty$  where  $M_0$  is the set of non-zero coefficients.*

Although, in this paper, we do not derive the asymptotic variance of the treatment effect estimator, in the simulation section, we provide some empirical results about the performance of a bootstrap estimator which is based on a method introduced by Chatterjee & Lahiri (2011).

### 3.6. The Procedure Summary

The penalized treatment effect estimation process explained in sections 3.1 to 3.5 can be summarized as follows:

- (a) Estimate the vector of parameter  $\hat{\eta}$  as  $\arg \sup_{\eta} \sum_{i=1}^n \log f_p(z_i; \eta)$  where  $f_p(\cdot)$  is defined in section 3.1.
- (b) Using the covariates with  $\eta \neq 0$ , fit a propensity score  $\pi(\mathbf{X})$ .
- (c) Define a random variable  $S_i = D_i - \pi(\mathbf{X}_i)$  and fit the response model  $\mathbb{E}[Y_i|d, \mathbf{x}] = \theta s_i + g(\mathbf{x}_i; \gamma)$ . The vector of parameters  $(\theta, \gamma)$  is estimated using standard least square method. For simplicity, we assume the linear working model for  $g(\mathbf{x}_i; \gamma) = \gamma' \mathbf{x}_i$ . The design matrix  $\mathbf{X}$  includes a subset of variables with  $\eta \neq 0$ .

**Table 1.** Performance of the proposed method when  $r_2 > n$  and in the presence of a weak confounder. S.D<sup>emp</sup>: empirical standard error; S.D<sup>tb</sup>: sandwich standard error.

Method	Bias	S.D <sup>emp</sup>	S.D <sup>tb</sup>	MSE	Bias	S.D <sup>emp</sup>	S.D <sup>tb</sup>	MSE
Scenario 1.								
$n = 300$								
SCAD	0.010	0.515	0.502	0.266	0.012	0.386	0.381	0.149
LASSO	0.067	0.522	0.509	0.277	0.057	0.425	0.421	0.184
PS-fit	0.164	5.575	—	31.104	0.101	4.295	—	18.453
Oracle	0.017	0.510	—	0.260	0.007	0.373	—	0.139
$n = 500$								
SCAD	0.062	0.606	0.592	0.372	0.019	0.483	0.456	0.234
LASSO	0.037	0.612	0.593	0.375	0.012	0.481	0.460	0.232
Y-fit	0.710	0.598	—	0.862	0.818	0.453	—	0.875
PS-fit	0.381	6.722	—	45.326	0.094	5.117	—	26.189
Oracle	0.045	0.638	—	0.409	0.018	0.459	—	0.211

#### 4. Simulation Studies

In this section, we study the performance of our proposed variable selection method using simulated data when the number of covariates ( $r_2$ ) is larger than the sample size. This also includes a scenario in which there is a weak non-ignorable confounder that is strongly related to the treatment but weakly to the outcome. We consider linear working models for both  $g()$  and  $h()$  functions throughout this section.

We generate 500 data sets of sizes 300 and 500 from the following two models:

1.  $D \sim \text{Bernoulli} \left( \frac{\exp\{0.5x_1 + 0.5x_6 - 0.5x_7 - 0.5x_8\}}{1 + \exp\{0.5x_1 + 0.5x_6 - 0.5x_7 - 0.5x_8\}} \right)$   
 $Y \sim \text{Normal}(d + 2x_1 + 0.5x_2 + 5x_3 + 5x_4, 2)$
2.  $D \sim \text{Bernoulli} \left( \frac{\exp\{0.5x_1 + x_2 + 0.5x_6 - 0.5x_7 - 0.5x_8\}}{1 + \exp\{0.5x_1 + x_2 + 0.5x_6 - 0.5x_7 - 0.5x_8\}} \right),$   
 $Y \sim \text{Normal}(d + 2x_1 + 0.2x_2 + 5x_3 + 5x_4, 2)$

where  $\mathbf{X}_k$  has a  $N(1, 2)$  for  $k = 1, \dots, 550$ . Note that in the second scenario,  $x_2$  is considered as a weak confounder. Results are summarized in Table 1; the *Y-fit* row refers to the estimator obtained by penalizing the outcome model using *SCAD* penalty.

We estimate the standard error of the treatment effect using an idea similar to Chatterjee & Lahiri (2011). We bootstrap the sample and in each bootstrap force the components of the penalized estimator  $\hat{\eta}$  to zero whenever they are close to zero and estimate the treatment effect using the selected covariates. More specifically, we define  $\tilde{\eta} = \hat{\eta}I(|\hat{\eta}| > 1/\sqrt{n})$ . We utilize this thresholded bootstrap method to estimate the standard error of the treatment effect (S.D<sup>tb</sup>). Although more investigation is required to validate the asymptotic behaviour of this method, our simulation results in Table 1 show that the estimated standard error S.D<sup>tb</sup> is close to the empirical estimator S.D<sup>emp</sup> (slightly underestimated).

In the first scenario there is no weak confounder and the *Y-fit* is omitted since the result is similar to the SCAD row. The variance of the estimator in the *PS-fit* is too large due to the inclusion of spurious variables that are not related to the response. The SCAD and LASSO estimators, however, are unbiased and perform as well as the oracle model. In the second scenario, the *Y-fit* estimator is bias because of under selecting the confounder  $X_2$  while the proposed estimators using both SCAD and LASSO remain unbiased. Table 4 presents the average number of coefficients set to zero

**Table 2.** Penalized ATE estimators based on the SCAD and LASSO penalty functions.

Method	Correct		Incorrect	
	$n = 300$	$n = 500$	$n = 300$	$n = 500$
SCAD	546	0.05	546	0.05
LASSO	545	0.00	546	0.00
Y-fit	546	0.90	546	0.92

correctly or incorrectly under the second scenario. This, in fact, highlights the importance of our proposed method.

In Appendix F, we examine the performance of our covariate selection estimation procedure when either of the working models of  $g()$  or  $h()$  is misspecified. Our results show that the proposed method outperforms both *Y-fit* and *PS-fit*.

## 5. Application to Real Data

In this section we examine the performance of our proposed method on the cross-country economic growth data used by Doppelhofer et al. (2003). For illustration purposes, we focus on a subset of the data which includes 88 countries and 35 variables. Additional details are provided in Doppelhofer & Weeks (2009). We are interested in selecting non-ignorable variables which confound the effect of *life expectancy* (exposure variable) on the *average growth rate of gross domestic product per capita in 1960-1996* (outcome).

The causal effect of life expectancy on economic growth is controversial. Acemoglu & Johnson (2006) find no evidence of increasing the life expectancy on economic growth while Husain (2012) shows that it might have positive effect. We dichotomize the life expectancy based on the observed median, which is 50 years. Hence, the exposure variable  $D=1$  if life expectancy is below 50 years in that country and 0 otherwise.

We select the significant covariates for the conditional mean and the treatment allocation models using the penalized likelihood (3). After covariate selection, we fit the model  $\mathbb{E}[Y] = \theta s + g(x; \gamma)$ , where  $\theta$  is the treatment effect parameter (the function  $g()$  assumed to be linear). Interaction or the higher order of the propensity score can be added to the response model if needed.

In our analysis, *PS-fit* and *Y-fit* refer to the cases where just the propensity score model and the conditional outcome models are penalized using SCAD to select the significant covariates (LASSO has a similar performance). Table 5 presents the list of variables and their estimated coefficients which are selected at least by one of the methods.

The proposed method selects 11 variables while *Y-fit* and *PS-fit* select 7 and 10 variables, respectively. This is mainly because of non-ignorable confounders which either barely predict the outcome or treatment. More specifically, *Population Density 1960*, *Initial Income*, *Public Education Spending Share*, and *Investment Price* are such non-ignorable confounders. Table 5 shows that although the effect of life expectancy is positive, it is not significant. Hence our results are consistent with Acemoglu & Johnson (2006). As we expected *PS-fit* results in inflating the standard error because instrumental variables such as *Higher Education Enrollment*, *Land Area Near Navigable Water* and *Colony Dummy* are included. Also, including these variables in the propensity score causes bias. This is a confirmatory example of the result given by De Luna et al. (2011) and Abadie & Imbens (2006).

**Table 3.** The economic growth data: List of significant variables. Penalized ATE estimators based on the SCAD and LASSO penalty functions. The two estimators PS-fit and Y-fit are obtained by penalizing the propensity score and outcome model via SCAD penalty, respectively.

Variable	Y-fit	PS-fit	SCAD	LASSO
Ethnolinguistic Fractionalization	-0.39	-0.43	<b>-0.42</b>	<b>-0.33</b>
Population Density 1960	-0.01	0.00	<b>-0.16</b>	0.00
East Asian Dummy	0.48	0.13	<b>0.53</b>	<b>0.45</b>
Initial Income (Log GDP in 1960)	0.00	0.96	<b>0.19</b>	<b>0.15</b>
Public Education Spending Share	0.05	0.00	<b>0.13</b>	0.00
Nominal Government Share	0.00	0.00	<b>-0.18</b>	0.00
Higher Education Enrolment	0.00	0.23	0.00	0.00
Investment Price	-0.25	0.00	<b>-0.24</b>	<b>-0.16</b>
Land Area Near Navigable Water	0.00	0.52	0.00	0.00
Fraction GDP in Mining	0.00	0.00	<b>0.11</b>	0.00
Fraction Muslim	0.00	-0.05	0.00	0.00
Timing of Independence	0.00	-0.11	0.00	0.00
Political Rights	0.00	-0.52	0.00	0.00
Real Exchange Rate Distortions	-0.04	-0.04	<b>-0.20</b>	0.00
Colony Dummy	0.00	-0.09	0.00	0.00
European Dummy	0.00	0.00	<b>0.59</b>	<b>0.25</b>
Latin American Dummy	-0.18	0.00	0.00	0.00
Landlocked Country Dummy	0.00	0.00	<b>-0.21</b>	0.00

**Table 4.** The economic growth data: Penalized ATE estimators based on the SCAD and LASSO penalty functions. The two estimators PS-fit and Y-fit are obtained by penalizing the propensity score and outcome model via SCAD penalty, respectively.

Method	ATE	S.D.	C.I.(%95)
SCAD	0.438	0.405	(-0.372,1.248)
LASSO	0.451	0.400	(-0.348,1.252)
Y-fit	0.394	0.337	(-0.280,1.068)
PS-fit	0.774	0.890	(-1.006,2.554)

## 6. Discussion

We establish a two-step procedure for estimating the treatment effect in high-dimensional settings. First, we deal with the sparsity by penalizing a reparametrized conditional joint likelihood of the outcome and treatment given covariates. Then, the selected variables are used to form a double robust regression estimator of the treatment effect by incorporating the propensity score in the conditional expectation of the response. The selected covariates may be used in other causal techniques as well as the proposed regression method.

Although, in high-dimensional cases, asymptotically penalizing the conditional outcome model given treatment and covariates is a valid variable selection approach in causal inference, it may perform poorly in finite sample by underselecting non-ignorable confounders which are weakly associated with outcome. Our proposed method improves the finite sample performance of the outcome penalization approach while maintaining the same asymptotic performance. The selected variables are used in a double robust regression estimator for estimating the treatment effect by incorporating the propensity score in the conditional expectation of the response.

Any covariate selection procedure which involves the outcome variable affects the subsequent inference of the selected coefficients. This is because the selected model itself is stochastic and it needs to be accounted for. This is often referred to as “post-selection inference” in the statistical literature. Berk et al. (2012) proposes a method to produce a valid confidence interval for the coefficients of the selected model. In our setting, although we do not penalize the treatment effect, the randomness of the selected model affects the inference about the causal effect parameter through confounding. Moreover, note that the oracle property of the penalized regression estimators is a pointwise asymptotic feature and does not necessarily hold for all the points in the parameter space (Leeb & Pötscher, 2005, 2008). In this manuscript, we assume that the parameter dimension ( $r_2$ ) is fixed while the number of observation tends to infinity. One important extension to our work is to generalize the framework to cases where the tuple  $(n, r_2)$  tends to infinity (Negahban et al., 2009). Analyzing the convergence of the estimated vector of parameters in the more general setting requires an adaptation of restricted eigenvalue condition (Bickel et al., 2009) or restricted isometry property (Candes & Tao, 2007).

## Acknowledgment

This research was supported in part by NIDA grant P50 DA010075. The second and third authors acknowledge the support of Discovery Grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors are grateful to Professor Dylan Small for enlightening discussion.

## Appendix A. Required conditions

In this Appendix, we prove the results stated in the text. Here is the list of the regularity assumptions:

- C1. The parameter space  $\Xi$  is a bounded open set in  $\mathcal{R}^p$ .
- C2. The joint penalized density  $f_p(z; \eta)$ , where  $z_i = (y_i, d_i, x_i)$  is a continuous function of  $\eta$  on  $\Xi^c$  for almost all  $z \in \mathcal{Z}$ , where  $\mathcal{Z}$  and  $\Xi^c$  represent the sample space  $(y_i, d_i, x_i)$  and the closure of  $\Xi$  respectively.

- C3. For all  $\eta \in \Xi$  and all  $\gamma > 0$ ,  $\kappa_\eta(\gamma) = \inf_{\|\eta - \eta^*\| > \gamma} r^2(\eta, \eta^*) > 0$ , where  $r^2(\eta, \eta^*) = \int_{\mathcal{Z}} [f^{1/2}(z; \eta) - f^{1/2}(z; \eta^*)]^2 d\tau$ .
- C4. For  $\eta \in \Xi^c$ ,  $w_\eta(\delta) = \left[ \int_{\mathcal{Z}} \sup_{\|h\| \leq \delta} \{f^{1/2}(z; \eta) - f^{1/2}(z; \eta + h)\}^2 d\tau \right] \rightarrow 0$  as  $\delta \rightarrow 0$ .
- C5.  $f(z; \eta)$  has finite Fisher information at each  $\eta \in \Xi$ .

Assumption *C3* is the identifiability condition, essentially requiring that the distance between the averaged densities over the response and the covariates for two different values of the parameters  $\eta$  and  $\eta^*$  be positive. Assumption *C4* is referred to as the smoothness condition; it states that the distance of the joint densities over  $\eta$  and  $\eta^*$  when  $\eta \rightarrow \eta^*$  should approach zero as the sample size goes to infinity.

## Appendix B. Cases where $\alpha = 0$

Assume that  $X$  is the only confounder/covariate. We conceptualize the following (true) Gaussian structural equation model:

$$\begin{aligned} X &= \epsilon_1 \\ Z &= a_{12}X + \epsilon_2 \\ Y &= a_{13}X + a_{23}Z + \epsilon_3 \end{aligned}$$

where  $(\epsilon_1, \epsilon_2, \epsilon_3)$  are generated from a standard normal distribution. Since we are considering cases where  $\alpha = 0$ , the penalty function can be ignored by assumption *P1*. Assume that the parameter  $\beta$  in the reparametrized likelihood (3) is known and let  $g(x, \alpha) = h(x, \alpha) = \alpha x$ . Then by taking a derivative with respect to  $\alpha$  of the likelihood (3),  $\alpha$  is defined as

$$\alpha = \frac{\text{cov}(x, y) + \text{cov}(x, z)[1 - \beta]}{\text{cov}(x, x)}$$

Hence  $\alpha = 0$  iff 1)  $\text{cov}(x, y) = \text{cov}(x, z) = 0$ , 2)  $\text{cov}(x, y) = 0$  &  $\beta = 1$ , or 3)  $\text{cov}(x, y) + \text{cov}(x, z)[1 - \beta] = 0$ . The latter is a drawback of our method, however, this particular data generating low has zero measure. Note that  $\text{cov}(x, y) + \text{cov}(x, z)[1 - \beta] = 0$  implies that  $a_{13} + a_{12}a_{23} + a_{12}[1 - \beta] = 0$ . This is a hypersurface in the space of  $(a_{13}, a_{12}, a_{23}, \beta)$  and the set of distributions that satisfy this restriction has measure zero in  $\mathbb{R}^4$ .

The same argument can be extended to the cases with more than one confounder. Then we have a union of a finite set of hypersurfaces. Also, the same idea can be generalized to settings where variables are not normally distributed.

## Appendix C. Lemmas

LEMMA 3. *Let  $Z_1, \dots, Z_n$  be independent and identically distributed with a density  $f(Z, \eta)$  that satisfies the conditions of C1-C4. If the penalty function satisfies P3, then as  $n \rightarrow \infty$*

$$R_n(\eta_2) = \prod_{i=1}^n \left[ \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} \right] < 1, \quad \text{for } \eta_2 \neq 0. \quad (\text{Appendix C.1})$$

PROOF.  $R_n(\eta_2)$  can be written as

$$\prod_{i=1}^n \left[ \frac{f(z_i; \eta_1, \eta_2) e^{-\sum_{j=s}^p p_{\lambda_n}(|\eta_j|)}}{f(z_i; \eta_1, 0)} \right].$$

By theorem 1.1 in Chapter II of Ibragimov & Has' Minskii (1981), it can be written as

$$R_n(\eta_2) = \exp \left[ \left\{ \sum_{i=1}^n \frac{\partial \ln f_p(z_i; \eta_1, 0)}{\partial \eta_2} \right\} ||\eta_2|| - n \sum_{j=s}^p p'_{\lambda_n}(|\eta_j|) - \frac{1}{2} \eta_2 I(\eta_1, 0) \eta_2 + \psi_n(\eta_2) \right],$$

where  $p(|\psi_n(\eta_2)| > \epsilon) \rightarrow 0$ . Since  $\sum_{i=1}^n \partial \ln f(z_i; \eta_1, 0) / \partial \eta_2 = O_p(n)$ , equivalent to the condition P3, the desired inequality holds if

$$\sum_{j=s}^p p'_{\lambda_n}(|\eta_j|) > ||\eta_2|| = O_p(1) \quad \blacksquare$$

Note that in our setting,  $p_{\lambda_n}(|\eta_j|) = \frac{1}{|\eta_j|} p_{\lambda_n}^*(|\eta_j|) = \frac{\sqrt{n}}{O_p(1)} p_{\lambda_n}^*(|\eta_j|)$  where  $p_{\lambda_n}^*(.)$  is one of the standard penalty functions such as LASSO or SCAD.

The following Lemma is an adaptation of the results given by Ibragimov & Has' Minskii (1981), page 36.

LEMMA 4. Suppose assumption C1-C4 are satisfied. Then for any fixed  $\eta \in \Xi$

$$\mathbb{E}_\eta \left[ \sup_{\Gamma} \prod_{i=1}^n \frac{f_p^{1/2}(z_i; \eta + b)}{f_p^{1/2}(z_i; \eta)} \right] \leq \exp \left[ -\frac{n}{2} \left\{ \kappa_{\eta, n} \left( \frac{\gamma}{2} \right) - 2w_{\eta+b_0, n}(\delta) + p_{\lambda_n}(|\eta + b^m|) - p_{\lambda_n}(|\eta|) \right\} \right], \quad (\text{Appendix C.2})$$

where  $\Gamma$  is the sphere of radius  $\delta$ , situated in its entirety in the region  $||b|| > \gamma/2$ ,  $b_0$  is the center of  $\Gamma$  and  $\inf_{\Gamma} p_{\lambda_n}(|\eta + b|) = p_{\lambda_n}(|\eta + b^m|)$ .

PROOF. The proof follows from the proof of Theorem 1.4.3 in Ibragimov & Has' Minskii (1981). Let

$$R_n(b) = \prod_{i=1}^n \frac{f_p(z_i; \eta + b)}{f_p(z_i; \eta)} = \prod_{i=1}^n \frac{f(z_i; \eta + b) e^{-p_{\lambda_n}(|\eta + b|)}}{f(z_i; \eta) e^{-p_{\lambda_n}(|\eta|)}}.$$

We want to find an upper bound for the expectation  $\mathbb{E}_\eta \left[ \sup_{\Gamma} R_n^{1/2}(b) \right]$ , where  $\Gamma$  is the sphere of a radius  $\delta$  situated in its entirety in the region  $||b|| > \frac{1}{2}\gamma$ . If  $b_0$  is the center of  $\Gamma$ , then

$$\begin{aligned} \sup_{\Gamma} R_n^{1/2}(b) &= \sup_{\Gamma} \prod_{i=1}^n \left[ \frac{f(z_i; \eta + b) e^{-p_{\lambda_n}(|\eta + b|)}}{f(z_i; \eta) e^{-p_{\lambda_n}(|\eta|)}} \right]^{1/2} \leq \prod_{i=1}^n \sup_{\Gamma} e^{\frac{1}{2} p_{\lambda_n}(|\eta|) - \frac{1}{2} p_{\lambda_n}(|\eta + b|)} \\ &\quad \prod_{i=1}^n f^{-1/2}(z_i; \eta) \left[ f^{1/2}(z_i; \eta + b_0) + \sup_{h \leq \delta} |f^{1/2}(z_i; \eta + b_0 + h) - f^{1/2}(z_i; \eta + b_0)| \right]. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}_\beta \left[ \sup_{\Gamma} R_n^{1/2}(b) \right] &\leq \prod_{i=1}^n \sup_{\Gamma} e^{\frac{1}{2} p_{\lambda_n}(|\eta|) - \frac{1}{2} p_{\lambda_n}(|\eta+b|)} \left[ \int_{\mathcal{Z}} f^{1/2}(z_i; \eta) f^{1/2}(z_i; \eta + b_0) d\tau \right. \\ &\quad \left. + \int_{\mathcal{Z}} \sup_{|h| \leq \delta} f^{1/2}(z_i; \eta) |f^{1/2}(z_i; \eta + b_0 + h) - f^{1/2}(z_i; \eta + b_0)|^n d\tau \right]. \end{aligned}$$

We further note that

$$\begin{aligned} \int_{\mathcal{Z}} f^{1/2}(z; \eta) f^{1/2}(z; \eta + b_0) d\tau &= \frac{1}{2} \left[ \int_{\mathcal{Z}} f(z; \eta) d\tau + \int_{\mathcal{Z}} f(z; \eta + b_0) d\tau \right. \\ &\quad \left. - \int_{\mathcal{Z}} [f^{1/2}(z; \eta) - f^{1/2}(z; \eta + b_0)]^2 d\tau \right] \quad (\text{Appendix C.3}) \end{aligned}$$

and

$$\begin{aligned} \int_{\mathcal{Z}} \sup_{|h| \leq \delta} f^{1/2}(z_i; \eta) |f^{1/2}(z_i; \eta + b_0 + h) - f^{1/2}(z_i; \eta + b_0)| d\tau &\leq 1 - \frac{1}{2} r^2(\eta + b_0) \leq 1 - \frac{\kappa_\eta(\frac{\gamma}{2})}{2} \\ &\leq w_{b_0}(\delta). \quad (\text{Appendix C.4}) \end{aligned}$$

The last inequality follows from the Cauchy-Schwarz inequality. Finally,

$$\mathbb{E}_\beta \left[ \sup_{\Gamma} R_n^{1/2}(b) \right] \leq \exp \left[ -\frac{n}{2} \left\{ \kappa_\eta(\frac{\gamma}{2}) - 2w_{b_0}(\delta) + p_{\lambda_n}(|\eta + b^m|) - p_{\lambda_n}(|\eta|) \right\} \right]$$

where  $\sup_{\Gamma} e^{-p_{\lambda_n}(|\eta+b|)} = e^{-p_{\lambda_n}(|\eta+b^m|)}$ , using the inequality  $1 + a \leq e^a$ .  $\blacksquare$

## Appendix D. Proofs of Theorem 1 & 2

*Proof of Theorem 1: Part (a)* Consider  $\eta_0 = (\eta_{01}, 0)$  and partition  $\eta = (\eta_1, \eta_2)$ . We need to show that in the neighbourhood  $\|\eta - \eta_0\| < O(h_n)$  where  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ ,

$$\prod_{i=1}^n \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \widehat{\eta}_1, 0)} < 1.$$

It can be written as

$$\prod_{i=1}^n \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \widehat{\eta}_1, 0)} = \prod_{i=1}^n \left[ \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} \right] \left[ \frac{f_p(z_i; \eta_1, 0)}{f_p(z_i; \widehat{\eta}_1, 0)} \right] < \prod_{i=1}^n \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} < 1.$$

By the result of Lemma 3, the last inequality holds with probability one as  $n \rightarrow \infty$ .

*Part (b)* : Under the conditions listed in the Theorem, we have

$$\begin{aligned} \sum_{i=1}^n \log f_p(z_i; \eta_{01} + \frac{c}{\sqrt{n}}) - \log f_p(z_i; \eta_{01}) &= \frac{1}{\sqrt{n}} c' \sum_{i=1}^n \frac{\partial \log f(z_i; \eta_{01})}{\partial \eta_{01}} - c' \sqrt{n} p'_{\lambda_n}(\alpha_{01}) \\ &\quad - c' p''_{\lambda_n}(\alpha_{01}) c - \frac{1}{2} c' I(\eta_{01}) c + R_n(\eta_{01}, c), \quad \text{for } |c| < M, \end{aligned}$$

where  $f_p(\cdot)$  is the penalized density defined in §4.2 and  $c$  is a constant vector. Note that  $\eta_{01} = (\beta, \alpha_{01})$  and  $\alpha_{01}$  is the true vector of non-zero coefficients. Using the proof of Theorem 2.1.1 in Ibragimov & Has' Minskii (1981), one can show that  $R_n(\eta_{01}, c) \rightarrow 0$  in probability.

Using the proof of Theorem 2.5.2 in Bickel et al. (1993), we can show that for any  $\epsilon > 0$

$$P\left(\left|\sqrt{n}(\hat{\eta}_{01} - \eta_{01}) - \frac{1}{\sqrt{n}}I^{-1}(\eta_{01})\sum_{i=1}^n \frac{\partial \log f(z_i; \eta_{01})}{\partial \eta_{01}} + \sqrt{n}p'_{\lambda_n}(\alpha_{01}) + p''_{\lambda_n}(\alpha_{01})\right| > \epsilon\right) \rightarrow 0,$$

as  $n \rightarrow \infty$ . Under assumption P2, it completes the proof of part (b).  $\blacksquare$

*Proof of Theorem 2:* Using the triangle inequality,

$$|\zeta(\hat{\theta}_{M_n}, M_n) - \zeta(\theta_0, M_0)| \leq |\zeta(\hat{\theta}_{M_n}, M_n) - \zeta(\hat{\theta}_{M_0}, M_0)| + |\zeta(\hat{\theta}_{M_0}, M_0) - \zeta(\theta_0, M_0)|.$$

By differentiability of the  $\zeta(\cdot, \cdot)$  function in  $\theta$ , we have  $\zeta(\hat{\theta}_{M_0}, M_0) \xrightarrow{P} \zeta(\theta_0, M_0)$ . Also,  $\forall t > 0$ , we have

$$\begin{aligned} p(|\zeta(\hat{\theta}_{M_n}, M_n) - \zeta(\hat{\theta}_{M_0}, M_0)| > t) &\leq p(\{M_n = M_0\} \cap \{|\zeta(\hat{\theta}_{M_n}, M_n) - \zeta(\hat{\theta}_{M_0}, M_0)| > t\}) \\ &\quad + p(\{M_n \neq M_0\} \cap \{|\zeta(\hat{\theta}_{M_n}, M_n) - \zeta(\hat{\theta}_{M_0}, M_0)| > t\}) \\ &\leq p(M_n \neq M_0) = 0 \end{aligned}$$

The last inequality follows by the oracle property of our procedure (Theorem 2). See also Theorem 4.2 in Wasserman & Roeder (2009). This completes the proof of weak consistency.

## Appendix E. Existence of the consistent penalized maximum likelihood estimator

**THEOREM 5.** *Under assumptions C1-C4 and P1-P3, the penalized maximum pseudo-likelihood estimator  $\hat{\eta}_n$  converges to  $\eta_0$  as  $n \rightarrow \infty$  almost surely where  $\eta_0$  is the true parameter value with respect to (3).*

**PROOF.** For fixed  $\gamma > 0$ , the exterior of the sphere  $\|\eta - \eta_0\| = \|b\| \leq \gamma$  can be covered by  $N$  spheres  $\Gamma_k$ ,  $k = 1, \dots, N$  of radius  $\delta$  with centers  $b_k$ . The small value  $\delta$  is chosen such that (i) all the  $N$  spheres are located in the  $\|b\| > \gamma/2$ , (ii)  $2w_{b_k}(\delta) \leq \kappa_\eta(\gamma/2)/4$  where  $w(\cdot)$  and  $\kappa(\cdot)$  are defined in C3 and C4, respectively, and (iii)  $\forall b \in \Gamma_k$ ,  $|p_{\lambda_n}(|\eta + b_k|) - p_{\lambda_n}(|\eta|)| \leq \kappa_\eta(\gamma/2)/4$ . Let  $u_k \in \Gamma_k$  so that  $R(\hat{u}_k) = \sup_{u_k \in \Gamma_k} R(u_k)$ . Then, in view of the result of Lemma 4, we have

$$\begin{aligned} P(|\hat{\eta}_n - \eta_0| > \gamma) &\leq \sum_{k=1}^N P(|\hat{\eta}_n - \eta_0| \in \Gamma_k) \leq \sum_{k=1}^N P(\sup_{u_k \in \Gamma_k} R_n(u_k) \geq R(0)) \\ &\leq \sum_{k=1}^N \exp\left[-\frac{n}{2}\left\{\kappa_\eta\left(\frac{\gamma}{2}\right) - 2w_{b_k}(\delta) + p_{\lambda_n}(|\eta + b_k^m|) - p_{\lambda_n}(|\eta|)\right\}\right] \\ &\leq N \exp\left[-\frac{n}{2}\left\{\kappa_\eta\left(\frac{\gamma}{2}\right) - \frac{1}{4}\kappa_\eta\left(\frac{\gamma}{2}\right) - \frac{1}{4}\kappa_\eta\left(\frac{\gamma}{2}\right)\right\}\right], \\ &\leq N \exp\left[-\frac{n}{2}\left\{\frac{1}{2}\kappa_\eta\left(\frac{\gamma}{2}\right)\right\}\right], \end{aligned}$$

where  $\sup_{b \in \Gamma_k} e^{-p_{\lambda_n}(|\eta+b|)} = e^{-p_{\lambda_n}(|\eta+b_k^m|)}$ . Note that  $R(0) = 1$ . The second inequality follows from the fact that when the MPL estimator  $\hat{\eta}_n$  falls in at least one of the spheres  $\Gamma_k$  where  $\Gamma_k$  covers outside of the neighborhood  $\gamma/2$  of  $\eta_0$ , it means  $\sup_{u_k \in \Gamma_k} \prod_{i=1}^n f_p(z_i; \eta_0 + u_k) \geq \prod_{i=1}^n f_p(z_i; \eta_0)$ . By the definition of  $R_n(u)$ , this inequality can be written as  $\sup_{u_k \in \Gamma_k} R_n(u_k) \geq 1$ . Also the third inequality follows from Lemma S1 and Markov's inequality.

Thus,

$$P(|\hat{\eta}_n - \eta_0| > \gamma) \leq N \exp \left[ -\frac{n}{4} \kappa_\eta \left( \frac{\gamma}{2} \right) \right],$$

and hence we have strong consistency, as

$$P \left( \bigcup_{m=n}^{\infty} |\hat{\eta}_{2m}| \right) \leq \frac{N \exp \left[ -\frac{n}{4} \kappa_\eta \left( \frac{\gamma}{2} \right) \right]}{1 - \exp \left[ -\frac{1}{4} \kappa_\eta \left( \frac{\gamma}{2} \right) \right]} \rightarrow 0 \text{ as } n \rightarrow \infty. \blacksquare$$

## Appendix F. Performance under model misspecifications

In this simulation study, we want to examine the performance of our proposed method when 1) either of the working models  $g()$  or  $h()$  are misspecified and 2) the number of potential confounders ( $r_2$ ) is larger than the sample size.

1.  $D \sim \text{Bernoulli} \left( \frac{\exp\{0.1x_1+x_2+0.7\frac{x_{10}+x_9}{1+|x_8|}\}}{1+\exp\{0.1x_1+x_2+0.7\frac{x_{10}+x_9}{1+|x_8|}\}} \right)$   
 $Y \sim \text{Normal}(d + 0.5x_1 + 0.1x_2 + 2x_3 + 2x_4, 2)$
2.  $D \sim \text{Bernoulli} \left( \frac{\exp\{x_1-x_2-0.1x_8-x_9+x_{10}\}}{1+\exp\{x_1-x_2-0.1x_8-x_9+x_{10}\}} \right)$   
 $Y \sim \text{Normal}(d + 2x_8 + 2 \frac{\exp\{0.2x_3+0.2x_4\}}{\exp\{0.2|x_1|+0.2|x_2|\}}, 2)$

where  $\mathbf{X}_k$  has a  $N(0, 2)$  for  $k = 1, \dots, 550$ .

In both scenarios, we consider linear working models for  $g()$  and  $h()$ . Thus, at least one of them is misspecified. Table 5 summarized the results.  $PS\text{-fit}$  refers to the propensity score model including only the variables affecting treatment allocation (commonly done by practitioners) and  $Y\text{-fit}$  refers to the estimator obtained by penalizing the outcome model using SCAD penalty. We applied our variable selection procedure using the SCAD and LASSO penalties.

In scenario 1,  $x_2$  is a non-ignorable confounder which is weakly associated with the outcome. Ignoring this variable by  $Y\text{-fit}$  method results in bias which does not go zero by increasing the sample size.  $PS\text{-fit}$  method, in scenarios 1 & 2, ignores the non-ignorable confounders  $x_1$  and  $x_8$ , respectively, which leads to a bias treatment effect estimate. Our proposed method using SCAD and LASSO outperforms all the other methods by increasing the chance of including all the confounders (weak or strong) in the model.

## References

- ABADIE, A. & IMBENS, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74 235–267.
- ACEMOGLU, D. & JOHNSON, S. (2006). Disease and development: the effect of life expectancy on economic growth. Tech. rep., National Bureau of Economic Research.

**Table 5.** Performance of the proposed method when either of the response or treatment models are misspecified and  $r_2 > n$ .

Method	Bias	S.D.	MSE	Bias	S.D.	MSE
Scenario 1.	$n = 300$				$n = 500$	
SCAD	0.036	0.453	0.206	0.047	0.309	0.097
LASSO	0.209	0.456	0.252	0.057	0.425	0.184
Y-fit	0.109	0.360	0.142	0.121	0.290	0.099
PS-fit	0.227	1.033	1.118	0.157	0.805	0.665
Scenario 2.	$n = 300$				$n = 500$	
SCAD	0.116	0.358	0.142	0.091	0.319	0.110
LASSO	0.110	0.377	0.154	0.016	0.288	0.083
Y-fit	0.185	0.398	0.193	0.205	0.291	0.126
PS-fit	0.737	0.918	1.387	0.673	0.768	1.044

ANTONIADIS, A. (1997). Wavelets in statistics: a review. *Statistical Methods and Applications* 6 97–130.

BANG, H. & ROBINS, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61 962–972.

BERK, R., BROWN, L., BUJA, A., ZHANG, K. & ZHAO, L. (2012). Valid post-selection inference. *Submitted Ann. Statist.* <http://>.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. & WELLNER, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Johns Hopkins University Press.

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 1705–1732.

BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. & STURMER, T. (2006a). Variable selection for propensity score models. *American Journal of Epidemiology* 163 1149–1156.

BROOKHART, M. A. & VAN DER LAAN, M. J. (2006b). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics & Data Analysis* 50 475–498.

CANDES, E. & TAO, T. (2007). The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics* 2313–2351.

CHATTERJEE, A. & LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association* 106 608–625.

CRAINICEANU, C., DOMINICI, F. & PARMIGIANI, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* 95 635.

DAVIDIAN, M., TSIATIS, A. & LEON, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science* 20 261.

- DE LUNA, X., WAERNBAUM, I. & RICHARDSON, T. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* 98 861–875.
- DOPPELHOFER, G., MILLER, R. & SALA-I MARTIN, X. (2003). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *American Economic Review* .
- DOPPELHOFER, G. & WEEKS, M. (2009). Jointness of growth determinants. *Journal of Applied Econometrics* 24 209–244.
- FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 1348–1261.
- GREENLAND, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology* 167 523.
- HAFEMAN, D. M. & VANDERWEELE, T. J. (2010). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* 21 1531–5487.
- HUSAIN, M. J. (2012). Alternative estimates of the effect of the increase of life expectancy on economic growth. *Economics Bulletin* 32 3025–3035.
- IBRAGIMOV, I. A. & HAS' MINSKII, R. Z. (1981). *Statistical Estimation–Asymptotic Theory*. Springer.
- KANG, J. & SCHAFER, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22 523–539.
- LEEB, H. & PÖTSCHER, B. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* 21 21–59.
- LEEB, H. & PÖTSCHER, B. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* 142 201–211.
- NEGAHBAN, S., RAVIKUMAR, P. D., WAINWRIGHT, M. J., YU, B. ET AL. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*. 1348–1356.
- NEUGEBAUER, R. & VAN DER LAAN, M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* 129 405–426.
- PETERSEN, M. L., SINISI, S. E. & VAN DER LAAN, M. J. (2006). Estimation of direct causal effects. *Epidemiology* 17 276–284.
- PORTER, K., GRUBER, S., VAN DER LAAN, M. & SEKHON, J. (2011). The relative performance of targeted maximum likelihood estimators. *UC Berkeley Division of Biostatistics Working Paper Series* 279.
- ROBINS, J. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, vol. 6.

- ROBINS, J. M. & BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 11 550–560.
- ROBINS, J. M. & GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3 143–155.
- ROBINS, J. M., MARK, S. D. & NEWHEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48 479–495.
- ROBINS, J. M., RICHARDSON, T. S. & SPIRITES, P. (2010). On identification and inference for direct effects. *Epidemiology* In Press.
- ROSENBAUM, P. (2010). Causal inference in randomized experiments. *Design of Observational Studies* 21–63.
- ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 41–55.
- RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2 808–840.
- SCHAFFER, J. L. & KANG, J. D. Y. (2005). Discussion of “semi-parametric estimation of treatment effect in a pretest–posttest study with missing data” by m. davidian et al. *Statistical Science* 20 292—295.
- SCHISTERMAN, E. F., COLE, S. & PLATT, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 20 488.
- SINISI, S., POLLEY, E., PETERSEN, M., RHEE, S. & VAN DER LAAN, M. (2007). Super learning: an application to the prediction of HIV-1 drug resistance. *Statistical applications in genetics and molecular biology* 6 7.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58 267–288.
- TSIATIS, A. A. (2006). *Semiparametric theory and missing data*. Springer Verlag.
- VAN DER LAAN, M., DUDOIT, S. & VAN DER VAART, A. (2004). The cross-validated adaptive epsilon-net estimator. *UC Berkeley Division of Biostatistics Working Paper Series* 142.
- VAN DER LAAN, M. & GRUBER, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics* 6 17.
- VAN DER LAAN, M., POLLEY, E. & HUBBARD, A. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* 6 25.
- VAN DER LAAN, M. & ROBINS, J. (2003). *Unified methods for censored longitudinal data and causality*. Springer Verlag.
- VANSTEELANDT, S., BEKAERT, M. & CLAESKENS, G. (2010). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* 1477–0334.

- WANG, C., PARMIGIANI, G. & DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* 68 661–671.
- WASSERMAN, L. & ROEDER, K. (2009). High dimensional variable selection. *Annals of statistics* 37 2178.
- ZIGLER, C. M., WATTS, K., YEH, R. W., WANG, Y., COULL, B. A. & DOMINICI, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics* .
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101 1418–1429.
- ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67 301–320.