Ashkan Ertefaie*, Masoud Asgharian and David A. Stephens

Variable Selection in Causal Inference using a Simultaneous Penalization Method

https://doi.org/10.1515/jci-2017-0010
Received March 08, 2017; accepted September 26, 2017

Abstract: In the causal adjustment setting, variable selection techniques based only on the outcome or only on the treatment allocation model can result in the omission of confounders and hence may lead to bias, or the inclusion of spurious variables and hence cause variance inflation, in estimation of the treatment effect. We propose a variable selection method using a penalized objective function that is based on both the outcome and treatment assignment models. The proposed method facilitates confounder selection in high-dimensional settings. We show that under some mild conditions our method attains the oracle property. The selected variables are used to form a doubly robust regression estimator of the treatment effect. Using the proposed method we analyze a set of data on economic growth and study the effect of life expectancy as a measure of population health on the average growth rate of gross domestic product per capita.

Keywords: causal inference, variable selection, propensity score

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 outcome 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 [1, 2].

Predictors that appear in both mechanisms (i) and (ii) are termed *confounders*, and their omission from models (i) and (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 only – and not outcome – being included in the treatment allocation model. The inclusion of such “spurious" variables in model (ii) is usually regarded as harmless. However, the typical forfeit for this conservatism is inflation of variance of the effect estimator [3, 4]. Moreover, [5] and [6] showed that inclusion of variables that are only predictors of treatment may cause bias [7, 8]. This problem also applies to the conditional mean model, but is in practice less problematic, as practitioners seem to be more concerned with bias removal, and therefore more likely to introduce the spurious variables in model (ii). Little formal guidance as to how the practitioner should act in this setting has been provided. In this paper, we refer to variables that only predict the treatment and are not associated with the outcome as treatment predictors.

As has been shown by [9], it is plausible that judicious variable selection may lead to appreciable efficiency gains, and several approaches with this aim have been proposed [10, 11]. However, confounder

***Corresponding author: Ashkan Ertefaie,** University of Rochester Medical Center, Biostatistics and Computational Biology, 265 Crittenden Boulevard, Rochester, New York 14642, USA, E-mail: ashkan_ertefaie@urmc.rochester.edu, ertefaie@gmail.com
Masoud Asgharian, David A. Stephens, Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada

selection methods based on either just the treatment assignment model or just the outcome model may fail to account for non-ignorable confounders which barely predict the treatment or the outcome, respectively [12, 13]: in this manuscript, we use the term *weak confounder* for these variables. [14] 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 [15]). [16] 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 [17, 18].

Bayesian adjustment for confounding (BAC) is a parametric variable selection approach introduced by [19]; 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 propensity score model [20]. If we *know* a priori that a predictor of treatment is in fact a confounder, then w can be set to ∞ [13, 21]. [22] proposes a decision-theoretic approach to confounder selection that can handle high-dimensional cases; a Bayesian regression model is fit and using the posterior credible region of the regression parameters, a set of candidate models is formed. A sparse model is then found by penalizing models that do not include confounders. This method is conservative in the sense that it may include treatment predictors that may inflate the variance of the treatment effect [23]. Also, tuning the penalty function can be challenging. Doubly robust procedures have also been proposed where variables are selected using both the outcome and the treatment assignment models [12, 24].

Asymptotically, it is known that penalizing the conditional outcome model, given treatment and covariates, results in a valid variable selection strategy for causal effect estimation; however, for small to moderate sample sizes, it may result in the omission of weak confounders [25]. The objective of this manuscript is to improve the small sample performance of the outcome penalization strategy while maintaining its asymptotic performance (Table 2). We present a covariate selection procedure which facilitates the estimation of the treatment effect in the high-dimensional cases. Specifically, we propose a penalized objective function which considers both covariate-treatment and covariate-outcome associations and has the ability to select even weak confounders. This objective function is used to identify the set of non-ignorable confounders and predictors of outcome; the resulting parameter estimates do not have any causal interpretation. We derive the asymptotic properties of procedure and show that under some mild conditions the estimators have oracle properties (specifically, are consistent and asymptotically normally distributed). We utilize the selected covariates to estimate the causal effect of interest using a doubly robust estimator.

2 Preliminaries & notation

In standard notation, let $Y(d)$ denote the (potential) outcome arising from treatment d, and let D denote the treatment received. We consider for illustration the case of binary treatment. The observed outcome, Y, is defined as $DY(1) + (1 - D)Y(0)$. We restrict attention here to the situation where each predictor can be classified into one of three types, and to single time-point studies. We consider

(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. In addition, as is usual, we will make the assumption of *no unmeasured confounders*, that is, that treatment received D and potential outcome to treatment d, $Y(d)$, are independent, given the measured confounders, i.e., X_2. 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.

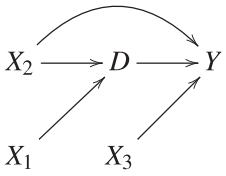

Figure 1: Covariate types: Type-I: X_1, Type-II: X_2 and Type-III: X_3.

2.1 The propensity score for binary treatments

The *propensity score*, $\pi(.)$, for binary treatment D is defined as $\pi(x) = \Pr(D = 1|\boldsymbol{x})$, where $\boldsymbol{x}$ is a p-dimensional vector of (all) covariates. In its random variable form, [2] show that $\pi(X)$ is the coarsest function of covariates that exhibits the balancing property, that is, $D \perp \boldsymbol{X}|\pi(\boldsymbol{X})$. As a consequence, the causal effect $\mu = \mathbb{E}[Y(1)-Y(0)]$ can be computed by iterated expectation

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

where $\mathbb{E}_\pi$ denotes the expectation with respect to the distribution of $\pi(\boldsymbol{X})$. For more details see [26] and [27].

Remark 1: Inclusion of covariates that are just related to the outcome in the propensity score model increases the covariance between the fitted π and Y, and decreases the variance of the estimated causal effect, in line with the results of [9] and [12].

2.2 Penalized estimation

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

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

MPL estimators are shrinkage estimators, and as such, they typically have more finite sample bias, though less variation than unpenalized analogues. Commonly used penalty functions include LASSO [28], SCAD [29], Elastic Net [30] and HARD [31].

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 proofs are given in Appendix A of the Supplementary Materials.

3 Penalization and treatment effect estimation

In this section, we present our proposed method for estimating the treatment effect in high-dimensional cases. We separate the covariate selection and treatment effect estimation procedure. First, we form a penalized objective function which is used to identify the important covariates, and establish the theoretical properties of the resulting estimators (i.e., minimizers of the penalized objective function). Note that because of the special characteristics of this function the estimators do not have any causal interpretation and are used just to *prioritize* variables. Second, treatment effect estimation is performed using a doubly robust estimator with the selected covariates. We use a simple model structure to illustrate the methodology, and assume that a random sample of size n of observations of outcome, exposure and covariates is available.

In order to present the method, we initially assume that columns of $\boldsymbol{X}$ are orthogonal; this simplifying assumption is relaxed in Appendix C of the Supplementary Materials.

3.1 Penalized objective function

Consider the following linear outcome model under a binary exposure d

$$Y = \theta d + \boldsymbol{x}\alpha_y + \epsilon,$$

where $\boldsymbol{x}$ is a $1 \times r$ standardized covariate vector, and ϵ is standard normal residual error. Assuming a logit model for the propensity score, we have

$$\pi(\boldsymbol{x}, \alpha_d) = p(D = 1|\boldsymbol{x}, \alpha_d) = \frac{\exp\{\boldsymbol{x}\alpha_d\}}{1 + \exp\{\boldsymbol{x}\alpha_d\}}.$$

Let $\boldsymbol{X}$ denote the corresponding $n \times r$ design matrix. In the formulation, α_y and α_d denote $r \times 1$ vectors of parameters and θ is the treatment effect.

First, we form an objective function, $M(\alpha)$, in which the coefficients α_y and α_d are replaced by a single common vector of coefficients α that is proportional to a weighted sum of $|\alpha_y|$ and $|\alpha_d|$, where $|.|$ denotes the componentwise absolute value. This will guarantee that the objective function satisfies the following condition:

Argmin Condition: An element, $\widehat{\alpha}_j$, of the minimizer $\widehat{\alpha}$ of $M(\alpha)$ converges to zero as $n \to \infty$ if the corresponding covariate is not associated with both treatment and outcome.

In standard linear regression where we regress Y on a vector of covariates $\boldsymbol{X}$ (with no treatment variable) that are presumed orthogonal, an example of an objective function that estimates $|\alpha_y|$ is

$$M(\alpha) = \frac{1}{2n}\left(\left|\sum_{i=1}^{n} \boldsymbol{x}_i^\top y_i\right| - n\alpha\right)^\top \left(\left|\sum_{i=1}^{n} \boldsymbol{x}_i^\top y_i\right| - n\alpha\right)$$

yielding the estimating equation and estimate

$$\frac{\partial M(\alpha)}{\partial \alpha} = \left|\sum_{i=1}^{n} \boldsymbol{x}_i^\top y_i\right| - n\alpha = 0 \qquad \Longrightarrow \qquad \widehat{\alpha} = \frac{1}{n}\left|\sum_{i=1}^{n} \boldsymbol{x}_i^\top y_i\right|.$$

We have used the fact $n^{-1}\boldsymbol{x}^\top\boldsymbol{x}$ is an identity matrix due to standardization. It is clear that as $n \longrightarrow \infty$, $\widehat{\alpha} \xrightarrow{p} |\mathbb{E}[x_i^\top Y_i]|$ Similarly, in a binary exposure setting, $|\alpha_d|$ can be estimated using the objective function $\left[-\left|\sum_{i=1}^{n} \boldsymbol{x}_i d_i\right| \alpha + \sum_{i=1}^{n} \log(1 + \exp\{\boldsymbol{x}_i\alpha\})\right]$ which is based on minus the log likelihood for a logistic regression. Now, we combine these two objective functions and show that the resulting objective function have some interesting features. First, we obtain the least squares (or ridge) $\widetilde{\theta}$ of θ by regressing Y on D and the vector of covariates $\boldsymbol{X}$. Because of the inclusion of spurious and treatment predictors in the model, $\widetilde{\theta}$ is not efficient but is but is consistent for θ under the no unmeasured confounders assumption and a correctly specified outcome model. Define $\widetilde{Y} = Y - \widetilde{\theta}D$. Let

$$M(\alpha) = \frac{1}{2}\Big[\Big|\sum_{i=1}^{n} \boldsymbol{x}_i\widetilde{y}_i\Big| - \alpha^\top \sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{x}_i)\Big]\left(\sum_{i=1}^{n} \boldsymbol{x}_i^\top\boldsymbol{x}_i\right)^{-1}\Big[\Big|\sum_{i=1}^{n} \boldsymbol{x}_i\widetilde{y}_i\Big| - \alpha^\top \sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{x}_i)\Big]^\top + \frac{1}{\tau}\Big[-\Big|\sum_{i=1}^{n} \boldsymbol{x}_i d_i\Big|\alpha + \sum_{i=1}^{n} \log(1 + \exp\{\boldsymbol{x}_i\alpha\})\Big], \tag{2}$$

where τ is a positive constant. Under orthogonality of $\boldsymbol{X}$ and standardization, we have

$$M(\alpha) = \frac{1}{2n}\Big[\Big|\sum_{i=1}^{n}\boldsymbol{x}_i\widetilde{y}_i\Big| - n\alpha^{\top}\Big]\Big[\Big|\sum_{i=1}^{n}\boldsymbol{x}_i\widetilde{y}_i\Big| - n\alpha^{\top}\Big]^{\top} + \frac{1}{\tau}\Big[-\Big|\sum_{i=1}^{n}\boldsymbol{x}_i d_i\Big|\alpha + \sum_{i=1}^{n}\log(1+\exp\{\boldsymbol{x}_i\alpha\})\Big].$$

Note that for each j, the parameter α_j corresponding to x_j is the same in both models. Now, we show that in this function the absolute values play a critical rule in satisfying the *Argmin* condition. Let $\widehat{\alpha} = \arg\min_{\alpha} M(\alpha)$ and $\widetilde{\alpha}_y$ be the least squares estimate of the parameters in the outcome model. By convexity of $M(\alpha)$, $\widehat{\alpha}$ must be the unique solution to $\partial M(\alpha)/\partial\alpha = 0$, which implies that

$$\left[n\widehat{\alpha} + \frac{1}{\tau}\sum_{i=1}^{n}\boldsymbol{x}_i^{\top}\frac{\exp\{\boldsymbol{x}_i\widehat{\alpha}\}}{1+\exp\{\boldsymbol{x}_i\widehat{\alpha}\}}\right] = \left|\sum_{i=1}^{n}\boldsymbol{x}_i^{\top}\widetilde{y}_i\right| + \frac{1}{\tau}\left|\sum_{i=1}^{n}\boldsymbol{x}_i^{\top}d_i\right| = n\left|\widetilde{\alpha}_y\right| + \frac{1}{\tau}\left|\sum_{i=1}^{n}\boldsymbol{x}_i^{\top}d_i\right|. \tag{3}$$

This equation shows that $\widehat{\alpha} = 0$ (i.e. the *Argmin* condition is satisfied) if $cov(\boldsymbol{X}, Y) = cov(\boldsymbol{X}, D) = \boldsymbol{0}$. Note that $cov(\boldsymbol{X}, D) = \boldsymbol{0}$ implies $cov(\mathbf{X}, \widetilde{Y}) = cov(\mathbf{X}, Y)$. Let $\widetilde{\alpha}_d$ be the maximum likelihood estimate of the parameters in the treatment model. Then, using the first two terms of a Taylor expansion, we have $\left|\sum_{i=1}^{n}\mathbf{x}_i^{\top}d_i\right| \approx n\left|\widetilde{\alpha}_d\right|/4$, $\exp\{\boldsymbol{x}_i\widehat{\alpha}\}/(1+\exp\{\boldsymbol{x}_i\widehat{\alpha}\}) \approx 1/2 + \widehat{\alpha}\boldsymbol{x}_i/4$, and thus

$$\widehat{\alpha} \approx \frac{4\tau}{4\tau+1}\left|\widetilde{\alpha}_y\right| + \frac{1}{4\tau+1}\left|\widetilde{\alpha}_d\right|.$$

Hence, $\widehat{\alpha}$ is a weighted sum of $|\widetilde{\alpha}_y|$ and $|\widetilde{\alpha}_d|$. The constant τ controls the contribution of components $\left|\widetilde{\alpha}_y\right|$ and $|\widetilde{\alpha}_d|$ to the estimator. For example, for $\tau = 2$, $\widehat{\alpha} \approx \frac{8}{11}\left|\widetilde{\alpha}_y\right| + \frac{1}{11}|\widetilde{\alpha}_d|$; while for $\tau = 0.1$, $\widehat{\alpha} \approx \frac{0.4}{1.4}\left|\widetilde{\alpha}_y\right| + \frac{1}{1.4}|\widetilde{\alpha}_d|$. Thus, τ gives a flexibility to our objective function such that as it decreases to zero the proposed estimate $\widehat{\alpha}$ converges to $|\widetilde{\alpha}_d|$. See Section 3.2 for more detailed discussion.

Figure 2 visually presents how $\widehat{\alpha}_j$ behaves, for a fixed $\tau = 0.5$, when $\widehat{\alpha}_{jy}$ converges to zero as sample size increases, i.e., the jth variable becomes insignificant. Specifically, this figure presents a case where there is just one covariate (i.e., $j = 1$) and the coefficient of this covariate in outcome and treatment models are $\alpha_{jy} = 1/\sqrt{n}$ and $\alpha_{jd} = 0.3$, respectively, where n is the sample size. As expected, $\widehat{\alpha}$ corresponding to this covariate does not converge to zero as sample size increases. The same behavior would be observed if $\alpha_{jd} \to 0$ as $n \to \infty$ and α_{jy} is a non-zero constant.

Therefore, penalizing objective function (2) results in selecting covariates that are either related to outcome (Type-III) or treatment (Type-I). However, this is against our goal of keeping variables that are either predictors of the outcome or non-ignorable confounders and excluding treatment predictors (Type-I). To deal with this problem, we present a weighted lasso penalty function that is tailored specifically for causal inference variable selection:

$$\lambda\sum_{j=1}^{r}\frac{|\alpha_j|}{(\widetilde{\alpha}_{jy})^2(1+|\widetilde{\alpha}_{jd}|)^2},$$

where $\widetilde{\alpha}_y$ and $\widetilde{\alpha}_d$ are the least squares (or ridge) and maximum likelihood estimates of the parameters in the outcome and treatment models, respectively, and λ is a tuning parameter. Thus, the proposed modified penalized objective function is given by

$$M_p(\alpha) = M(\alpha) + \lambda\sum_{j=1}^{r}\frac{|\alpha_j|}{(\widetilde{\alpha}_{jy})^2(1+|\widetilde{\alpha}_{jd}|)^2}. \tag{4}$$

We refer to $\widehat{\alpha} = \arg\min_{\alpha} M_p(\alpha)$ as penalized modified objective function estimators (PMOE). The magnitude of the penalty on each parameter is proportional to its contribution to the outcome and treatment model.

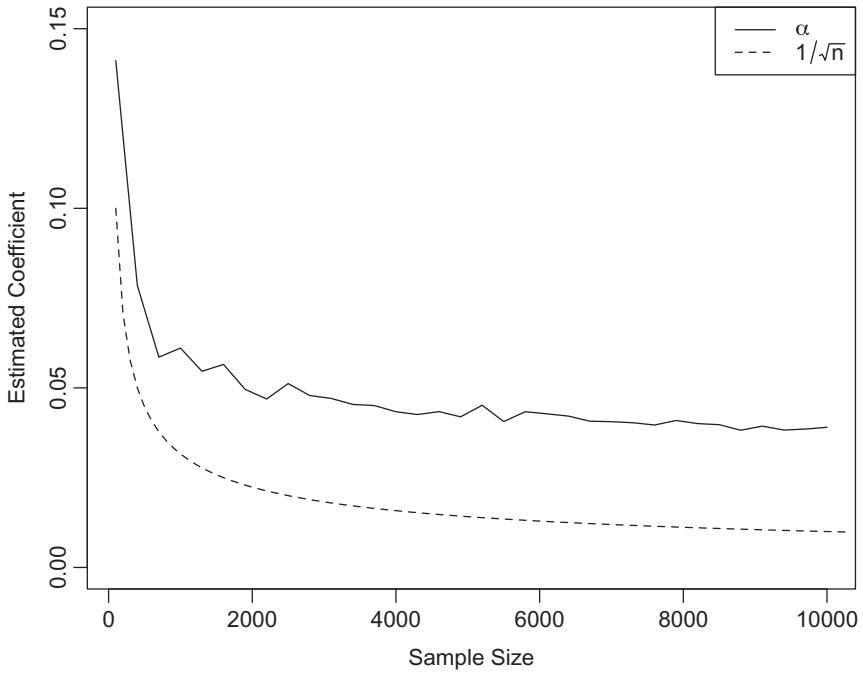

Figure 2: Performance of the modified objective function based estimator for different sample sizes n.
The dashed and solid lines are $1/\sqrt{n}$ and the estimated coefficient $\widehat{\alpha}$ using the modified objective function.

Note that as $\widehat{\alpha}_{jy} \to 0$, our penalty function puts more penalty on the jth parameter 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. This is an important feature of the proposed penalty function which allows selecting such weak confounders for small sample sizes. The proposed penalized estimator asymptotically selects the same covariates as the adaptive lasso on the outcome model [32, 25]. Thus, the proposed method improves the small sample performance of the outcome model penalization strategy while maintaining the asymptotic properties of this strategy.

3.2 The role of τ in PMOE

The constant τ reflects investigators belief about the importance of including variables that are weakly(strongly) related to outcome and strongly(weakly) related to treatment. For smaller values of τ, variables that are weakly related to the outcome but strongly to the treatment have more contribution to $\widehat{\alpha}$ and have more chance to be selected. Also, when $\tau = \infty$, the treatment mechanism does not have any contribution in the objective function (2). Thus, our procedure performs similarly to the outcome model based variable selection that may exclude non-ignorable confounders that are weakly related to the outcome. Often investigators do not have an idea a priori about how to weigh the importance of different types of confounders. In Section 3.4, a method that helps the investigators to properly choose this tuning parameter is presented. Our simulation studies shed more light on the role of τ in our procedure.

3.3 Main theorem

Suppose $\alpha_0 = (\alpha_{01}, \alpha_{02})$ is the true parameter value of the r-dimensional vector of parameters where $\alpha_{02} = \{\alpha_j, j = s+1, ..., r\} \equiv \mathbf{0}$ contains those elements of α that are in fact zero, so that the corresponding

predictors are not confounders; s denotes the true number of predictors present in the model (*exact sparsity* assumption). Let $\widehat{\alpha} = (\widehat{\alpha}_1, \widehat{\alpha}_2)$ be the vector of estimators corresponding to (4). The next theorem proves the sparsity and asymptotic normality of the proposed penalized estimators under the following two conditions:

(a) Let

$$\sum_{i=1}^{n} \epsilon_{i\widetilde{y}} = |\sum_{i=1}^{n} \boldsymbol{x}_i \widetilde{y}_i| - n\alpha_0^{\top} \qquad \text{and} \qquad \sum_{i=1}^{n} \epsilon_{id} = |\sum_{i=1}^{n} d_i \boldsymbol{x}_i| - \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\exp\{\boldsymbol{x}_i\alpha_0\}}{1+\exp\{\boldsymbol{x}_i\alpha_0\}}.$$

We assume that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\epsilon_{i\widetilde{y}} + \frac{1}{\tau}\epsilon_{id}]$ converges in distribution a multivariate normal $N_r(0, \Sigma(\alpha_0))$.

(b) Let

$$\frac{1}{n}\left[n\mathbb{1} + \frac{1}{\tau}\sum_{i=1}^{n} \frac{\exp\{\boldsymbol{x}_i\alpha_0\}}{[1+\exp\{\boldsymbol{x}_i\alpha_0\}]^2} \boldsymbol{x}_i^{\top}\boldsymbol{x}_i\right] \xrightarrow{p} \Omega(\alpha_0),$$

where $\Omega(\alpha_0)$ is a $r \times r$ positive definite matrix and $\mathbb{1}$ is the identity matrix.

Theorem 1. ***(Oracle properties)*** *Suppose conditions (a) & (b) are fulfilled, further* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n\sqrt{n} \to \infty$. *Then*

(a) $Pr(\widehat{\alpha}_2 = \mathbf{0}) \to 1$ *as* $n \to \infty$

(b) $\sqrt{n}(\widehat{\alpha}_{01} - \alpha_{01}) \xrightarrow{d} N(0, \Omega_{11}^{-1}\Sigma_{11}\Omega_{11}^{-1})$,

where $\alpha_{01} = \alpha_{01}$ *is the true vector of non-zero coefficients. Also,* Ω_{11} *and* Σ_{11} *are corresponding elements of* Ω *and* Σ, *respectively*.

3.4 Choosing the tuning parameters

We select the tuning parameter $\xi = (\tau, \lambda)$ using the *Generalized Cross Validation* (GCV) method suggested by [28] and [29]. Let $\boldsymbol{X}_\xi$ be the selected covariates corresponding to a specific value of ξ, we first regress $\widetilde{Y}$ on $\boldsymbol{X}_\xi$ and calculate the residual sum of square (RSS) of this this model. Then

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

where $d(\xi) = \text{trace}[\boldsymbol{X}_\xi(\boldsymbol{X}_\xi^{\top}\boldsymbol{X}_\xi + n\Sigma_\xi(\widehat{\alpha}))^{-1})\boldsymbol{X}_\xi^{\top}]$ is the effective number of parameters and $\Sigma_\xi(\alpha) = \text{diag}[p'_\xi(|\alpha_1|)/|\alpha_1|, ..., p'_\xi(|\alpha_r|)/|\alpha_r|]$. The selected tuning parameter $\widehat{\xi}$ is defined by $\widehat{\xi} = \arg\min_\xi \text{GCV}(\xi)$.

3.5 Estimation of the causal effect

For treatment effect estimation, we fit the following model using the set of covariates selected in the previous step; note that a user may want to use other causal adjustment models such as inverse probability weighting or propensity score matching, and the selection approach can be used for these procedures also.

Our model is a slight modification of the conventional propensity score regression approach of [33], and specifies

$$\mathbb{E}[Y_i|S_i = s_i, \boldsymbol{X}_i = \boldsymbol{x}_i] = \theta s_i + g(\boldsymbol{x}; \gamma), \tag{5}$$

where $S_i = D_i - \mathbb{E}[D_i|\boldsymbol{X}_i] = D_i - \pi(\boldsymbol{X}_i)$, $g(\boldsymbol{x}; \gamma)$ is a function of covariates and π 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 θ since $g(\boldsymbol{x}; \gamma)$ may be incorrectly specified. By defining S_i in this way, we restore $\text{corr}(S_i, \boldsymbol{X}_{ij}) = 0$ for $j = 1, 2, .., p$

where p is the number of selected variables (if $\pi(\mathbf{x}_i) = \mathbb{E}[D_i|\boldsymbol{X}_i = \boldsymbol{x}_i]$ is correctly specified), as $\pi(\boldsymbol{x}_i)$ is the (fitted) expected value of D_i, and hence $\boldsymbol{x}_j^\top(D - \pi(\boldsymbol{x})) = 0$, where $\boldsymbol{x}_j^\top = (x_{1j}, \ldots, x_{nj})$. Therefore, misspecification of $g(.)$ will not result in an inconsistent estimator of θ provided that $\pi(\boldsymbol{x})$ is correctly specified. Moreover, [33] showed that the proposed estimator attains the semiparametric bound when both $g(\boldsymbol{x};\gamma)$ and $\pi(\boldsymbol{x})$ are correctly specified [34].

In general, this model results in a *doubly robust* estimator (see [35], [36] and [37]); it yields a consistent estimator of θ if *either* the propensity score model or conditional mean model (5) is correctly specified, and is the most efficient estimator [38] when both are correctly specified. For additional details on the related asymptotic and finite sample behavior, see [39–41] and [42].

Remark 2: The importance of the doubly robust estimator is that, provided the postulated outcome and treatment model identify the true non-zero coefficients in each model, the proposed estimation procedure will consistently estimate the treatment effect given that at least one of the propensity score or outcome models are correctly specified. 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.

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 "post-selection inference". Hence, inference about the treatment effect obtained in the estimation step needs to be done with caution. Under certain regularity conditions, [43] developed the weak consistency of the post-selection estimator [44, 45].

The post-selection inference issue implies that the standard error of $\widehat{\theta}$ obtained by fitting a linear model (5) that includes the selected variables are, in general, under estimated, and thus the standard confidence intervals will be under covered. To mitigate this problem, we propose to approximate the standard errors of our estimator using an idea similar to [46]. Specifically, we bootstrap the sample and in each bootstrap, force the components of the penalized estimator $\widehat{\alpha}$ to zero whenever they are close to zero and estimate the treatment effect using the selected covariates, i.e., we define $\widehat{\alpha}^\dagger = \widehat{\alpha}1(|\widehat{\alpha}| > 1/\sqrt{n})$. Although our simulation results suggest that the thresholded bootstrap method produces valid confidence intervals, developing the theoretical properties of the method merits further investigation that is beyond the scope of this paper.

3.6 The procedure summary

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

1. Estimate the vector of parameters $\widehat{\alpha} = \arg\min_\alpha M_p(\alpha)$ where $M_p(\alpha)$ is defined in (4).
2. Using the covariates with $\alpha \neq 0$, estimate the propensity score $\widehat{\pi}(\boldsymbol{X})$.
3. Define a random variable $S_i = D_i - \widehat{\pi}(\boldsymbol{X}_i)$ and fit the outcome model $\mathbb{E}[Y_i|d, \boldsymbol{x}] = \theta s_i + g(\boldsymbol{x}_i;\gamma)$. The vector of parameters (θ, γ) is estimated using ordinary least squares. For simplicity, we assume the linear working model for $g(\boldsymbol{x}_i;\gamma) = \gamma^\top \boldsymbol{x}_i$. The design matrix $\boldsymbol{X}$ includes a subset of variables with $\alpha \neq 0$.

4 Simulation studies

In this section, we study the performance of our proposed variable selection method using simulated data. We compare our results with BAC method introduced by [19], the Bayesian credible region (Cred. Reg.) introduced by [22], the collaborative targeted maximum likelihood estimators (CTMLE) introduced by [24], a doubly robust approach (Belloni) proposed by [12] and outcome penalized estimator (Y-fit) where we regress the outcome against the treatment and covariates, but only penalize the covariates. Our simulation includes a scenario in which there is a weak confounder that is strongly related to the treatment but weakly to the outcome. We consider linear working models for $g()$ throughout this section.

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

1. $Y \sim \text{Normal}(D + 2X_1 + 0.5X_2 + 5X_3 + 5X_4, 4)$
2. $Y \sim \text{Normal}(D + 2X_1 + 0.2X_2 + 5X_3 + 5X_4, 4)$

where in both models

$$D \sim \text{Bernoulli}\left(\pi(\boldsymbol{X}) = \frac{\exp\{0.5X_1 - X_2 + 0.3X_5 - 0.3X_6 + 0.3X_7 - 0.3X_8\}}{1 + \exp\{0.5X_1 - X_2 + 0.3X_5 - 0.3X_6 + 0.3X_7 - 0.3X_8\}}\right),$$

and $\boldsymbol{X}_k$ has a $N(1, 4)$ for $k = 1, \ldots, 100$. 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 *LASSO* penalty, and the *Oracle* row refers to estimator obtained by including (X_1, X_2, X_3, X_4) in the propensity score and the outcome model (5). We studied the performance of PMOE for different values of constant $\tau = 0.1, 0.5, 1, 2, 4, 100$ and the optimal one that is chosen by the GCV criterion, i.e., $\text{PMOE}^{\tau=opt}$. In the first scenario there is no weak confounder and the Y-fit is omitted since the results are similar to the $\text{PMOE}^{\tau=20}$ row. The Bayesian adjustment for confounding (BAC) has been implemented using the R package BEAU with $\omega = \infty$ and the Bayesian credible region (Cred. Reg.) has been implemented using the R package BayesPen with flat prior. The doubly robust variable selection method, i.e., Belloni, is also implemented using the R package hdm. Finally, we implemented CTMLE.

Figure 3 shows the proportion of times that our algorithm selects a certain value of tuning parameter τ. In scenario 1, where there is no weak confounder, the algorithm tends to select higher values for τ which means the procedure adaptively selects the important covariates mostly based on the outcome model; while in scenario 2, where X_2 is a non-ignorable weak confounder, the algorithm puts more weights on the treatment model by selecting smaller values for τ.

Table 1: Simulation results for scenarios 1 & 2. Bias, S.D. and MSE are for the treatment effect.

Method	Bias	S.D	MSE	Covg	Bias	S.D	MSE	Covg
Scenario 1.		$n = 300$				$n = 500$		
$\text{PMOE}^{\tau=0.1}$	0.028	0.652	0.426	0.951	0.023	0.502	0.252	0.942
$\text{PMOE}^{\tau=0.5}$	0.008	0.628	0.395	0.955	0.031	0.517	0.268	0.953
$\text{PMOE}^{\tau=1}$	0.010	0.640	0.410	0.961	0.015	0.503	0.253	0.951
$\text{PMOE}^{\tau=20}$	0.005	0.643	0.414	0.938	0.040	0.505	0.257	0.962
$\text{PMOE}^{\tau=opt}$	0.012	0.633	0.400	0.943	0.010	0.503	0.253	0.949
Cred. Reg.	0.004	0.780	0.608	0.949	0.024	0.585	0.342	0.950
BAC (ω=∞)	0.064	1.332	1.781	0.963	0.040	1.062	1.130	0.952
Belloni	0.201	2.065	4.305	0.897	0.015	1.414	1.999	0.902
CTMLE	0.050	0.874	0.766	0.797	0.033	0.625	0.392	0.841
Oracle	0.009	0.630	0.396	0.945	0.028	0.504	0.258	0.956
Scenario 2.		$n = 300$				$n = 500$		
$\text{PMOE}^{\tau=0.1}$	0.039	0.648	0.422	0.933	0.028	0.490	0.241	0.948
$\text{PMOE}^{\tau=0.5}$	0.070	0.665	0.447	0.942	0.005	0.512	0.262	0.957
$\text{PMOE}^{\tau=1}$	0.132	0.685	0.486	0.956	0.025	0.539	0.292	0.934
$\text{PMOE}^{\tau=20}$	0.878	0.654	1.198	0.948	0.898	0.504	1.065	0.144
$\text{PMOE}^{\tau=opt}$	0.021	0.656	0.431	0.958	0.031	0.503	0.254	0.948
Cred. Reg.	0.010	0.765	0.582	0.955	0.027	0.595	0.355	0.946
Y-fit	0.710	0.598	0.862	0.102	0.818	0.453	0.875	0.058
BAC (ω=∞)	0.070	1.242	1.542	0.963	0.008	1.011	1.022	0.966
Belloni	0.043	1.598	2.556	0.893	0.086	1.363	1.865	0.901
CTMLE	0.049	0.839	0.706	0.831	0.053	0.741	0.552	0.821
Oracle	0.026	0.657	0.432	0.953	0.034	0.500	0.251	0.947

S.D: empirical standard error; Covg: coverage of confidence intervals; CTMLE: collaborative-TMLE; Belloni: doubly robust approach; Cred. Reg.: bayesian credible region; BAC: bayesian adjustment for confounding; Y-fit: penalized outcome model via the LASSO; PMOE: proposed method; Oracle: true model.

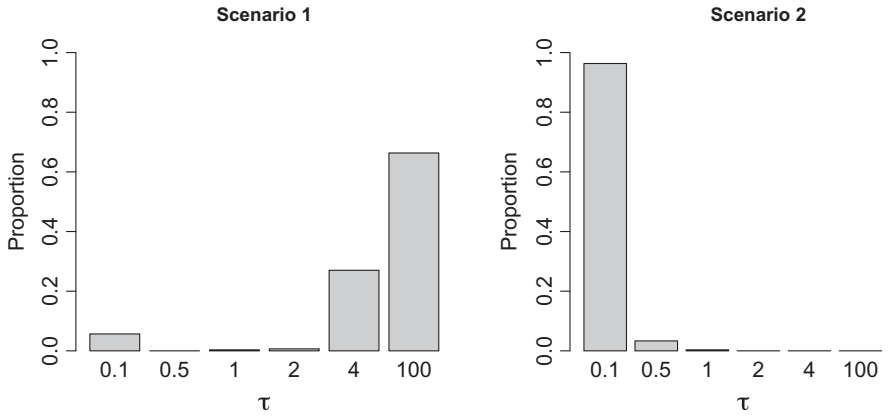

Figure 3: Proportion of times that the algorithm selects certain values of τ.
The left panel presents the first simulation scenario where there is no weak confounders while the right panel presents the second simulation scenario where there is a weak non-ignorable confounder.

The variance of the estimator in the BAC is too large due to the inclusion of spurious variables that are not related to the outcome. The PMOE and Cred. Reg. estimators, however, are unbiased and have smaller variance. In fact, PMOE has the lowest variance compared to BAC and Cred. Reg. methods regardless of the value of τ. In the second scenario, the Y-fit estimator is biased because of under selecting the confounder X_2. Also, the bias of the proposed PMOE estimator increases by increasing τ that is expected because as τ increases the proposed method should perform similarly to the outcome based variable selection methods such as Y-fit. In general, doubly robust procedures, e.g., CTMLE and Belloni, are sensitive to the positivity assumption. Specifically, when the estimated propensity score using the selected variables is close to boundaries, such methods lead to unstable estimates and possibly under covered confidence intervals. Both CTMLE and Belloni produce confidence intervals that are under covered, and CTMLE seems to be affected more seriously by the issue discussed above. The main drawback of Belloni's method is that it fails to eliminate variables that only predict the treatment from the set of selected variables. As our results in Table 1 shows this deficiency leads to inflating the standard error of the corresponding estimates. The estimator PMOE^{opt} outperforms all the other estimators and has similar performance as the oracle estimator which confirms that our method successfully tunes τ.

Table 2 presents the average number of coefficients set to zero correctly and incorrectly under the second scenario. There are four non-zero coefficients in our generative model so the number in the correct column should be 96 and in the incorrect column should be 0. This table shows that Cred. Reg., and BAC are somewhat conservative and include some of the variables that should not be included which can be the source of

Table 2: Simulation results for scenario 2.

Method	Correct	Incorrect	Correct	Incorrect
	n = 300		n = 500	
PMOE$^{\tau=0.1}$	95.91	0.03	95.97	0.01
PMOE$^{\tau=0.5}$	96.00	0.12	96.00	0.03
PMOE$^{\tau=1}$	96.00	0.22	96.00	0.08
PMOE$^{\tau=20}$	96.00	0.89	96.00	0.87
PMOE$^{\tau=opt}$	95.33	0.01	95.99	0.00
Cred. Reg.	92.62	0.00	92.83	0.00
BAC (ω=∞)	91	0.00	91	0.00
Y-fit	96	0.90	96	0.92

Number of coefficients that are correctly or incorrectly set to zero. Cred. Reg.: bayesian credible region; BAC: bayesian adjustment for confounding; Y-fit: penalized outcome model via the LASSO; PMOE: proposed method.

the observed variance inflation in Table 1. This is mostly due to inclusion of variables that are predictors of treatment model but have no association with the outcome. Also, increasing τ in PMOE, increases the chance of setting the coefficient of X_2 to zero. The Y-fit row shows that this method is setting a nonzero coefficient to zero (i.e., coefficient of X_2) which explains the bias in Table 1. The tuned estimator PMOE^{opt} has satisfactory performance in ignoring covariates that are only predictors of the treatment mechanism and selecting all the non-ignorable confounders. This, in fact, highlights the importance of our proposed method. We couldn't obtain the information needed in Table 2 using the available CTMLE's and Belloni's R codes, and thus, they are omitted.

Simulation studies presented in Appendix D of the Supplementary Materials study the performance of our covariate selection and estimation procedure when the covariates are non-orthogonal. Web Tables 2 and 3 presented in Appendix D show that the proposed method is still outperforming the other methods. Moreover, in Appendix B of the Supplementary Materials, we study cases that the number of covariates is larger than the sample size ($r > n$). We also investigate cases where either of the working models of the propensity score or the outcome model is misspecified. Our results show that the proposed method performs well, and outperforms Y-fit (Web Table 1).

5 Application: the effect of life expectancy on economic growth

In this section, we use the proposed method to analyse the cross-country economic growth data reported by [47] and compare our findings with similar studies on the same data set. For illustration purposes, we focus on a subset of the data which includes 88 countries and 35 variables. Additional details are provided in [48]. We are interested in selecting non-ignorable variables which confound the effect of *life expectancy* (exposure variable) as a measure of population health on the *average growth rate of gross domestic product per capita in 1960-1996* (outcome).

The causal (or, at least, *unconfounded*) effect of life expectancy on economic growth is controversial. [49] find no evidence of increasing life expectancy on economic growth while [50] shows that it might have a 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 applied our method described in Section 3 to selection important confounders. In our analysis, Y-fit refers to the case where just the outcome model is penalized using LASSO to select the significant covariates. We also implement the BAC with $\omega = \infty$ and Credible region with flat prior methods. Belloni's and CTMLE methods are also implemented. Our procedure selects $\tau = 0.1$ as the optimal value for this tuning parameter. The estimates obtained by PMOE$^{\tau=0.1}$, i.e., 0.617 with CI: (−1.27,1.361), and CTMLE, i.e., 0.719 with CI: (0.512,0.927) are fairly close. However the standard error of the latter one is 70% smaller that the former, which raises concern about the validity of the corresponding confidence intervals (see Table 1). Y-fit and Belloni lead to the smallest and largest effects, respectively. Based on our simulation studies, we conjecture that this is because Y-fit may ignore some of non-ignorable confounders and Belloni's may include some of the treatment predictors in the model. The BAC estimate, i.e., 0.524 with CI:(−0.228,1.286), is also close to the PMOE$^{\tau=0.1}$. The Credible region method leads to the second highest effect estimate of 0.820 with CI: (−0.203,1.341).

Table 3 presents the list of variables that are selected at least by one of the methods. For brevity, we focus on Y-fit, BAC, Cred. Reg., and PMOE. The proposed method selects 12 and 9 variables depending on the value of τ while Y-fit, BAC and Cred. Reg. select 7, 6 and 11 variables, respectively. Y-fit and PMOE$^{\tau=5}$ are susceptible to under selecting some of non-ignorable confounders that barely predict the outcome. Specifically, our results suggest that *Population Density 1960*, and *Initial Income* are such non-ignorable confounders which are known to be important confounders in the economics literature. Table 4 shows that ignoring these variables leads to a substantially different treatment effect estimates. Moreover, because there are evidence that variables such as *Oil producing Country, Land Area Near Navigable Water, and Nominal Government Share* may confound the effect of the life expectancy on the economic growth [48, 49, 53–56, 51, 52], PMOE$^{\tau=0.1}$

Table 3: The economic growth data: List of significant variables.

Variable	Y-fit	BAC $\omega=\infty$	Cred. Reg.	PMOE $\tau = opt = 0.1$	PMOE $\tau = 0.5$	PMOE $\tau = 5$
Air Distance to Big Cities	–	–	–	√	√	√
Ethnolinguistic Fractionalization	√	√	√	√	√	√
Fraction of Catholics	–	–	–	√	√	√
Population Density 1960	–	–	√	√	√	–
East Asian Dummy	√	√	√	√	√	√
Initial Income (Log GDP in 1960)	–	√	–	√	√	√
Public Education Spending Share	√	–	√	–	–	–
Nominal Government Share	–	–	–	√	–	–
Investment Price	√	√	√	–	–	–
Land Area Near Navigable Water	–	–	√	–	–	–
Fraction GDP in Mining	–	–	√	√	√	√
Fraction Muslim	–	–	√	–	–	–
Political Rights	–	√	–	√	√	√
Real Exchange Rate Distortions	√	–	–	–	–	–
Colony Dummy	–	–	√	–	–	√
European Dummy	√	–	–	–	–	–
Latin American Dummy	√	√	–	√	√	√
Landlocked Country Dummy	–	–	√	–	–	–
Oil producing Country Dummy	–	–	–	√	–	–
Land Area Near Navigable Water	–	–	√	√	–	–

Cred. Reg.: bayesian credible region; BAC: bayesian adjustment for confounding; Y-fit: penalized outcome model via the LASSO; PMOE: proposed method.

seems to select more reasonable covariates than PMOE$^{\tau=0.5}$. Despite the small discrepancies in the selected variables between the BAC, Creg. Reg. and PMOE$^{\tau=opt}$, these methods lead to the same conclusion and suggest that while the life expectancy has positive effect on the average growth rate of gross domestic product per capita in 1960-1996, the effect is not statistically significant.

To gain insight into the effect of parameter τ on the selection results, we plot the estimated coefficients α for different values of τ for given tuning parameters. In Figure 4, the blue solid (dark dashed) lines correspond to coefficients that their estimated value is (not) greater than 0.05 for the entire range of τ. Also, displays (a)–(d) correspond to tuning parameter values λ = 0.0, 0.0001, 0.001, and 0.01, respectively. Figure 4 shows that, for different values of τ, the selected set of covariates may vary slightly. For example, in Figure 4(c) where a moderate penalty function is imposed, the proposed method suggests to include *Colony* (the dashed line) to the selected set for larger values of τ. Also, as the penalty becomes stronger, the estimated coefficients become more stable across values of τ (Figure 4(d)). This is because, for larger values of λ, the penalty function has a more dominant rule on the variable selection than τ.

Table 4 also reports the standard errors of the estimated effect of life expectancy using different penalization methods. The standard error of the PMOE estimator is approximated using an idea similar to [46]. Specifically, we bootstrap the sample and in each bootstrap force the components of the penalized estimator $\widehat{\alpha}$ to zero whenever they are close to zero and estimate the treatment effect using the selected covariates, i.e., we define $\widehat{\alpha}^{\dagger} = \widehat{\alpha}1(|\widehat{\alpha}| > 1/\sqrt{n})$. We utilize this thresholded bootstrap method to approximate the standard error of the treatment effect. Although more investigation is required to validate the asymptotic properties of this method, we only use this standard errors to shed light on the behavior of the penalized estimators. For example, the estimators correspond to PMOE$^{\tau=5}$ and Y-fit have the lowest standard errors that may support the possibility of under-selecting important covariates. Our results suggests that although the effect of life expectancy is positive, it is unlikely to be significant that is consistent with [49].

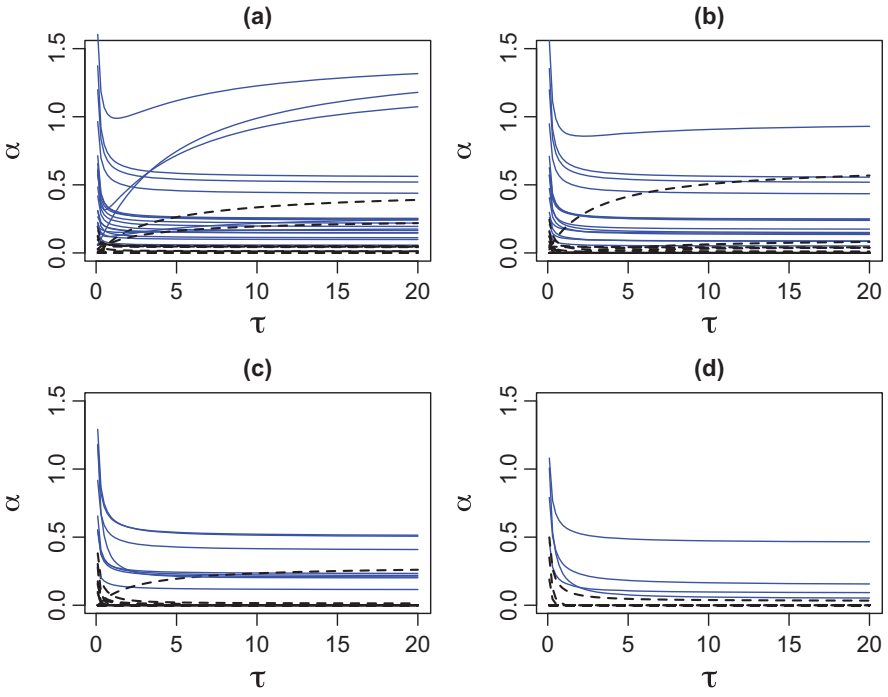

Figure 4: Economic growth data: The plot of the estimated parameters using the proposed method given different values of parameter τ.
The tuning parameter λ is fixed at 0.0, 0.0001, 0.001, and 0.01 in plots (a), (b), (c) and (d), respectively. The blue solid (dark dashed) lines correspond to coefficients that their estimated value is (not) greater than 0.05 for the entire range of τ.

Table 4: The economic growth data.

Method	ATE	S.D.	CI (%95)
PMOE$^{\tau=opt=0.1}$	0.617	0.372	(−0.127, 1.361)
PMOE$^{\tau=0.5}$	0.475	0.352	(−0.229, 1.179)
PMOE$^{\tau=5}$	0.454	0.340	(−0.226, 1.134)
Cred. Reg.	0.820	0.386	(−0.203, 1.341)
BAC ($\omega=\infty$)	0.524	0.381	(−0.228, 1.286)
Belloni	1.262	0.123	(0.999, 1.523)
CTMLE	0.719	0.106	(0.512, 0.927)
Y-fit	0.352	0.334	(−0.111, 1.345)

ATE: average treatment effect; CTMLE: collaborative-TMLE; Belloni: doubly robust approach; Cred. Reg.: bayesian credible region; BAC: bayesian adjustment for confounding; Y-fit: penalized outcome model via the LASSO; PMOE: proposed method; CI: confidence interval.

6 Discussion

We have established a two-step procedure for estimating an unconfounded treatment effect in high-dimensional settings. First, we deal with the sparsity by penalizing a modified objective function which considers both covariate-outcome and covariate-treatment associations. Then, the selected variables are used to form a doubly robust regression estimator of the treatment effect by incorporating the propensity score in the conditional expectation of the outcome. The selected covariates may be used in other causal techniques as well as the proposed regression method. The proposed method may also be used to identify *candidate* instrumental variables [57–60]. Specifically, one can penalize the treatment model first and record the selected variables and then apply the proposed method and identify the candidate instrumental variables as those that are selected by the treatment model but not the proposed method [23].

As described in section 3.5, any covariate selection procedure which involves the outcome variable affects the subsequent inference of the selected coefficients [61, 63, 65, 64, 62]. This is because the selected model itself is stochastic and it needs to be accounted for. [66] proposes a method to produce a valid confidence interval for the coefficients of the selected model in the post-selection context. 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 [67, 68]. In this manuscript, we assume that the parameter dimension (r) 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) tends to infinity [69]. Analyzing the convergence of the estimated vector of parameters in the more general setting requires an adaptation of restricted eigenvalue condition [70] or restricted isometry property [71].

Funding: This research was partly supported by the Natural Sciences and Engineering Council of Canada through Discovery Grants to Masoud Asgharian (NSERC RGPIN 217398-13).

References

1. Robins JM, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–560.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
3. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol. 2008;167:523.
4. Schisterman EF, Cole S, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 2009;20:488.
5. De Luna X, Waernbaum I, Richardson T. Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika. 2011;98:861–875.
6. Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol 2011;174:1223–1227.
7. Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, Stürmer T. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiology and drug safety 2011;20:551–559.
8. Pearl J. On a class of bias-amplifying variables that endanger effect estimates (2012). arXiv preprint arXiv:1203.3503.
9. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Meth. 2008;13:279.
10. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol. 2006a;163:1149–1156.
11. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology (Cambridge, Mass.) 2009;20:512.
12. Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. Rev Econ Stud. 2014;81:608–650.
13. Crainiceanu C, Dominici F, Parmigiani G. Adjustment uncertainty in effect estimation. Biometrika. 2008;95:635.
14. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. Stat Meth Med Res 2010;1477–0334.
15. Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Comput Stat Data Anal. 2006b;50:475–498.
16. Van der Laan M, Polley E, Hubbard A. Super learner. Stat Appl Genet Molec Biol. 2007;6:25.
17. Sinisi S, Polley E, Petersen M, Rhee S, Van Der Laan M. Super learning: an application to the prediction of HIV-1 drug resistance. Stat Appl Genetics Molecular Biol. 2007;6:7.
18. Van der Laan M, Dudoit S, Van der Vaart A. The cross-validated adaptive epsilon-net estimator. UC Berkeley Division of Biostatistics Working Paper Series, 2004:142.
19. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. Biometrics. 2012;68:661–671.
20. Wang C, Dominici F, Parmigiani G, Zigler CM. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. Biometrics. 2015.

21. Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F. Model feedback in Bayesian propensity score estimation. Biometrics, 2013.
22. Wilson A, Reich BJ. Confounder selection via penalized credible regions. Biometrics. 2014.
23. Lin W, Feng R, Li H. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. J Am Stat Assoc 2015;110:270–288.
24. Van der Laan M, Gruber S. Collaborative double robust targeted maximum likelihood estimation. Int J Biostat. 2010;6:17.
25. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. Biometrics 2017.
26. Rubin DB. For objective causal inference, design trumps analysis. Ann Appl Stat 2008;2:808–840.
27. Rosenbaum P. Causal inference in randomized experiments. Design of Observational Studies 2010;21–63.
28. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc, Series B. 1996;58:267–288.
29. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1261.
30. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67:301–320.
31. Antoniadis A. Wavelets in statistics: a review. Stat Meth Appl. 1997;6:97–130.
32. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–1429.
33. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics. 1992;48:479–495.
34. Chamberlain G. Asymptotic efficiency in estimation with conditional moment restrictions. J Econom 1987;34:305–334.
35. Davidian M, Tsiatis A, Leon S. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. Stat Sci. 2005;20:261.
36. Schafer JL, Kang JDY. Discussion of "semi-parametric estimation of treatment effect in a pretest–postest study with missing data" by M. Davidian et al. Stat Sci 2005;20:292—295.
37. Bang H, Robins J. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61:962–972.
38. Tsiatis AA. Semiparametric theory and missing data. Springer Verlag, 2006.
39. Kang J, Schafer J. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22:523–539.
40. Neugebauer R, van der Laan M. Why prefer double robust estimators in causal inference? J Stat Plann Inference 2005;129:405–426.
41. van der Laan M, Robins J. Unified methods for censored longitudinal data and causality. Springer Verlag, 2003.
42. Robins JM Robust estimation in sequentially ignorable missing data and causal inference models. In: Proceedings of the American Statistical Association Section on Bayesian Stat Sci, 1999, 2000:6–10.
43. Zhao P, Yu B. On model selection consistency of lasso. J Mach Learn Res. 2006;7:2541–2563.
44. Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models (2009).
45. Zhang J, Jeng XJ, Liu H. Some two-step procedures for variable selection in high-dimensional linear regression (2008). arXiv preprint arXiv:0810.1644.
46. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. J Am Stat Assoc. 2011;106:608–625.
47. Doppelhofer G, Miller R, Sala-i Martin X. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. Am Econ Rev. 2003.
48. Doppelhofer G, Weeks M. Jointness of growth determinants. J Appl Econometrics. 2009;24:209–244.
49. Acemoglu D, Johnson S. Disease and development: the effect of life expectancy on economic growth, Technical report, National Bureau of Economic Research. (2006).
50. Husain MJ. Alternative estimates of the effect of the increase of life expectancy on economic growth. Economics Bulletin 2012;32:3025–3035.
51. Doppelhofer G, Weeks M. Robust growth determinants. Technical report, CESifo working paper: Fiscal Policy, Macroeconomics and Growth (2011).
52. Eicher TS, Papageorgiou C, Raftery AE. Default priors and predictive performance in bayesian model averaging, with application to growth determinants. J Appl Econometrics. 2011;26:30–55.
53. Ley E, Steel MF. Jointness in bayesian variable selection with applications to growth regression. J Macroeconomics. 2007;29:476–493.
54. Ley E, Steel MF. Comments on jointness of growth determinants. J Appl Econometrics. 2009a;24:248–251.
55. Ley E, Steel MF. On the effect of prior assumptions in bayesian model averaging with applications to growth regression. J Appl Econometrics. 2009b;24:651–674.
56. Magnus JR, Powell O, Prüfer P. A comparison of two model averaging techniques with an application to growth empirics. J Econometrics. 2010;154:139–153.
57. Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J Am Stat Assoc. 1995;90:431–442.
58. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med 2014;33:2297–2340.
59. Kang H, Cai TT, Small DS. Robust confidence intervals for causal effects with possibly invalid instruments (2015). arXiv preprint arXiv:1504.03718.

60. Kang H, Zhang A, Cai TT, Small DS. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. J Am Stat Assoc. 2016;111:132–144.
61. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference with the lasso. Ann Stat. 2016;44:907–927.
62. Lee JD, Sun Y, Taylor JE, et al. On model selection consistency of regularized m-estimators. Electron J Stat. 2015;9:608–642.
63. Taylor J, Lockhart R, Tibshirani RJ, Tibshirani R. Exact post-selection inference for forward stepwise and least angle regression (2014). arXiv preprint arXiv:1401.3889.
64. Taylor J, Tibshirani RJ. Statistical learning and selective inference. Proc Nat Acad Sci. 2015;112:7629–7634.
65. Tibshirani R, Taylor J, Lockhart R, Tibshirani R. Exact post-selection inference for sequential regression procedures (2014). arXiv preprint arXiv:1401.3889.
66. Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. Submitted Ann. Statist. http (2012).
67. Leeb H, Pötscher B. Model selection and inference: Facts and fiction. Econ Theo 2005;21:21–59.
68. Leeb H, Pötscher B. Sparse estimators and the oracle property, or the return of Hodges' estimator. J Econometrics 142:201–211.
69. Negahban S, Ravikumar PD, Wainwright MJ, Yu B, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In: NIPS, 2009:1348–1356.
70. Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of lasso and Dantzig selector. Ann Stat. 2009;1705–1732.
71. Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Stat. 2007:2313–2351.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jci-2017-0010)