Partially Adaptive Regularized Multiple Regression Analysis for Estimating Linear Causal Effects

Abstract

This paper assumes that cause-effect relationships among variables can be described with a linear structural equation model. Then, a situation is considered where a set of observed covariates satisfies the back-door criterion but the ordinary least squares method cannot be applied to estimate linear causal effects because of multicollinearity/high-dimensional data problems. In this situation, we propose a novel regression approach, the "partially adaptive L_p-regularized multiple regression analysis" (PAL_pMA) method for estimating the total effects. Different from standard regularized regression analysis, PAL_pMA provides a consistent or less-biased estimator of the linear causal effect. PAL_pMA is also applicable to evaluating direct effects through the single-door criterion. Given space constraints, the proofs, some numerical experiments, and an industrial case study on setting up painting conditions of car bodies are provided in the Supplementary Material.

1 INTRODUCTION

1.1 BACKGROUND

The multicollinearity problem [Frisch, 1934], which occurs when two or more explanatory variables are highly correlated, is an important issue in regression analysis. If multicollinearity exists, because the performance of least squares/maximum likelihood estimators of regression coefficients is inadequate, valid results may not be obtained. The high-dimensional data problem occurs in the framework of regression analysis when the sample size is smaller than the number of explanatory variables. High-dimensional data analysis also suffers from multicollinearity, which causes overfitting and interferes with obtaining admissible solutions for regression coefficients. Recently, due to the development of technological advances that help collect data with a large number of variables to better understand a given phenomenon of interest, multicollinearity/high-dimensional data problems have become serious in many domains. To overcome this difficulty, numerous kinds of variable selection techniques based on regularized regression analysis, for example, the least absolute shrinkage and selection operator (LASSO), elastic net, smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) methods, have been proposed by many statistical and AI researchers and practitioners [Bühlmann and van de Geer, 2011; Efron et al, 2004; Fan and Li, 2001; Hoerl and Kennard, 1970; Tibshirani, 1996; van de Geer et al, 2014; Zhang, 2010; Zou, 2006; Zou and Hastie, 2005].

Currently, the role of regression analysis is not limited to the prediction of a response variable by explanatory variables; it also plays an important role in evaluating the linear causal effects of the treatment variable on the response variable. In particular, the total effect, which is one of the representative linear causal effects and the main interest in this paper, means the changes in the expected response variable by one unit through an external intervention [Pearl, 2009, 2013, 2017]. As has often been noted in the framework of statistical causal inference, to derive the consistent estimator of the total effect, in addition to the treatment variable, confounders must be included as explanatory variables in the regression model. However, there are many confounders that have an effect on both the treatment variable and the response variable and that are highly correlated in reality. This situation leads to the multicollinearity problem, which deteriorates the estimation accuracy of the total effects and formulates an unreliable plan that prevents us from conducting appro-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

priate policy decision making. On the other hand, the present countermeasures against the multicollinearity problem are formulated independently of the confounding problem. Thus, although stable results of regression analysis may be derived by these countermeasures from the viewpoint of the prediction, they may yield a highly biased estimate of the linear causal effect.

1.2 CONTRIBUTIONS

In this paper, when the cause-effect relationships among variables can be described with a linear structural equation model, we consider a situation where a set of observed covariates satisfies the back-door criterion but the ordinary least squares (OLS) method cannot be applied to estimate the total effects because of the multicollinearity/high-dimensional data problem. In this situation, to evaluate the total effect, we propose a novel regression approach, the "Partially Adaptive L_p-regularized Multiple regression Analysis" (PAL$_p$MA) method for $p = 1, 2$. In particular, PAL$_1$MA has the following desirable properties:

(1) In statistical causal inference, it is important not to remove a treatment variable or confounders from the regression model when estimating the total effects. However, even if some covariates are guaranteed to be important confounders from qualitative causal knowledge, standard regularized regression analysis may remove them and the treatment variable from the model, depending on the value of the regularization parameter. In contrast, PAL$_1$MA enables us to include both the treatment variable and such covariates in the regression model, regardless of the value taken by the regularization parameter. In particular, when we know that a set of covariates satisfies the back-door criterion, the solution path with such information can be utilized as the criteria of parameter tuning to estimate the total effects.

(2) Regarding PAL$_p$MA for $p = 1, 2$, we can derive a collapsibility condition, i.e., a sufficient condition that the L_p-regularized estimator of the regression coefficient of interest is consistent with the OLS estimator regardless of the value taken by the regularization parameter, and thus leads to the consistent estimator of the total effects under the condition. The collapsibility problem in regression analysis have been discussed by many researchers [Clogg et al, 1992; Geng and Asano, 1993; Guo and Geng, 1995; Wermuth, 1989ab]. However, to the best of our knowledge, there has been much less discussion of collapsibility problem in the context of regularized regression analysis.

(3) Compared to standard regularized regression analysis, PAL$_1$MA can reduce the bias or provide higher coincidence rates for the signs of the OLS estimator, even when the collapsibility conditions are violated. In contrast, in standard regularized regression analysis, the regression coefficients can flip from positive to negative values and from negative to positive values as they shrink toward zero, depending on the value of the regularization parameter. This phenomenon implies that standard regularized regression analysis may provide misleading qualitative results regarding the total effects compared to PAL$_1$MA.

From these properties, PAL$_1$MA contributes to solving the multicollinearity/high-dimensional data problems of evaluating linear causal effects in the context of statistical causal inference. Given space constraints, the proofs, some numerical experiments and an industrial case study on setting up painting conditions of car bodies [Kuroki, 2012] are provided in the Supplementary Material.

2 LINEAR STRUCTURAL CAUSAL MODEL

In the context of statistical causal inference, a directed acyclic graph that represents cause-effect relationships is called a causal diagram. A directed graph is a pair $G = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a finite set of vertices and the set $\boldsymbol{E}$ of directed arrows is a subset of the set $\boldsymbol{V} \times \boldsymbol{V}$ of ordered pairs of distinct vertices. In this paper, we refer to vertices in the directed acyclic graph and random variables of the linear structural equation model interchangeably. In addition, for the graph theoretic terminology used in this paper, we refer readers to Pearl [2009].

Definition 1 *(Linear Structural Causal Model) Suppose a directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$ with set $\boldsymbol{V} = \{V_1, V_2, \cdots, V_m\}$ of variables is given. The graph G is called a causal diagram when each child-parent family in the graph G represents a linear structural equation model*

$$V_i = \mu_{v_i} + \sum_{V_j \in pa(V_i)} \alpha_{v_i v_j} V_j + \epsilon_{v_i}, \quad i = 1, 2, \ldots, m \tag{1}$$

as the data generating process, where $pa(V_i)$ denotes a set of parents of V_i in G and random disturbances $\epsilon_{v_1}, \epsilon_{v_2}, \ldots, \epsilon_{v_m}$ are assumed to be independent and identically distributed with mean 0. In addition, μ_{v_i} is an intercept, and $\alpha_{v_i v_j} (\neq 0)$ is called a path coefficient or a direct effect of V_j on V_i $(i, j = 1, 2, \ldots, m\,;\, i \neq j)$. Then, equation (1) is called a linear structural causal model (SCM) in this paper.

To proceed with our discussion, we define some nota-

tion. For univariates X and Y and a set of variables $\boldsymbol{Z}$, let $\sigma_{xy\cdot z}$ be the conditional covariance between X and Y given $\boldsymbol{Z} = \boldsymbol{z}$, and let $\sigma_{xx\cdot z}$ be the conditional variance of X given $\boldsymbol{Z} = \boldsymbol{z}$. The regression coefficient of X in the regression model of Y on X and $\boldsymbol{Z}$ is denoted by $\beta_{yx\cdot z} = \sigma_{xy\cdot z}/\sigma_{xx\cdot z}$. For sets of variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ ($\boldsymbol{Y}$ can be univariate), let $\Sigma_{xy\cdot z}$ be the conditional cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z} = \boldsymbol{z}$, and let $\Sigma_{xx\cdot z}$ be the conditional variance-covariance matrix of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$. In addition, let $B_{yx\cdot z} = \Sigma_{xx\cdot z}^{-1}\Sigma_{xy\cdot z}$ denote the regression coefficient vector of $\boldsymbol{X}$ in the regression model of Y on $\boldsymbol{X}$ and $\boldsymbol{Z}$. The set of variables $\boldsymbol{Z}$ is omitted from these arguments if it is an empty set. Similar notation is used for the remaining statistical parameters. Furthermore, letting $\boldsymbol{X} = \{X_1, X_2, ..., X_q\}$, the i-th element of $B_{yx\cdot z}$ is denoted by $\beta_{yx_i\cdot x_{(i)}z}$, where $\boldsymbol{X}_{(i)} = \boldsymbol{X}\backslash\{X_i\}$ $(i = 1, 2, ..., q)$. $\boldsymbol{0}_q$ is a q-dimensional zero vector. Similar notation is used for other sets of variables.

The main purpose of this paper is to estimate total effects from observed data. The total effect τ_{yx} of X on Y is defined as the total sum of the products of the path coefficients on the sequence of directed arrows along all the directed paths from X to Y. To achieve our aim, we introduce the back-door criterion [Pearl, 2009] as one of the representative identifiability criteria for the total effects. Here, when a linear causal effect can be determined uniquely from the variance/covariance parameters of the observed variables, it is said to be identifiable, that is, it can be estimated consistently.

Definition 2 *(Back-Door Criterion) Let $\{X, Y\}$ and $\boldsymbol{Z}$ be disjoint subsets of $\boldsymbol{V}$ in a directed acyclic graph G. If a set $\boldsymbol{Z}$ of vertices satisfies the following conditions relative to an ordered pair (X, Y), then $\boldsymbol{Z}$ is said to satisfy the back-door criterion relative to (X, Y).*

1. *No vertex in $\boldsymbol{Z}$ is a descendant of X, and*
2. *$\boldsymbol{Z}$ d-separates X from Y in the graph obtained by deleting all the directed arrows emerging from X from graph G.*

If a set $\boldsymbol{Z}$ of observed variables satisfies the back-door criterion relative to (X, Y) in a causal diagram G, then, the total effect τ_{yx} is identifiable and is given by the formula $\beta_{yx\cdot z}$ [Pearl, 2009]. For other identification conditions of linear causal effects, refer to, for example, Brito [2004], Chen [2017], Chen et al [2017], Kuroki and Pearl [2014], Pearl [2009], Stanghellini [2004], Stanghellini and Pakpahan [2015] and Tian [2004, 2007ab].

Here, a covariate is defined as an element of non-descendants of X and Y. In addition, covariates in a minimal set of variables that satisfy the back-door criterion are called confounders. Note that such a minimal set is not unique and whether or not a certain covariate is considered a confounder depends on the selected minimal set. Furthermore, a set of covariates satisfying the back-door criterion is also called a sufficient set of confounders; otherwise, it is called an insufficient set of confounders. For details on the SCM, refer to the paper by Pearl [2009]. Finally, the direct effect is also known as one of the representative linear causal effects. However, we are concerned with the evaluation of the total effects because the direct effect can also be discussed in the framework of regression analysis through the "single-door criterion" [Pearl, 2009]. Thus, the total effects are identified with linear causal effects in this paper.

3 PAL$_p$MA

3.1 SETUP

Let X, Y, $\boldsymbol{Z}$ and $\boldsymbol{W}$ be a treatment variable (and an explanatory variable), a response variable, an r-dimensional vector of explanatory variables ($\boldsymbol{Z}$ can be empty) and a q-dimensional vector of explanatory variables ($\boldsymbol{W}$ can be empty), respectively. For a sample size of n, consider the linear regression model of Y on X, $\boldsymbol{Z}$ and $\boldsymbol{W}$

$$\boldsymbol{y} = \boldsymbol{x}\beta_{yx\cdot zw} + \boldsymbol{z}B_{yz\cdot xw} + \boldsymbol{w}B_{yw\cdot xz} + \boldsymbol{\epsilon}_{y\cdot xzw}, \quad (2)$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ represent n-dimensional observation vectors of X and Y, respectively. In addition, $\boldsymbol{z}$ and $\boldsymbol{w}$ are an $n \times r$ observation matrix of $\boldsymbol{Z}$ and an $n \times q$ observation matrix of $\boldsymbol{W}$, respectively. Furthermore, $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$ are the regression coefficient of X, the regression vector of $\boldsymbol{Z}$ and the regression vector of $\boldsymbol{W}$ in equation (2), respectively. $\boldsymbol{\epsilon}_{y\cdot xzw}$ is an n-dimensional vector of error variables. Here, we assume that elements of $\boldsymbol{\epsilon}_{y\cdot xzw}$ are independent and identically distributed with mean zero and variance $\sigma_{yy\cdot xzw} < \infty$. In this paper, we also assume that a treatment variable, a response variable and explanatory variables are standardized to a sample mean of zero and a variance of one in advance. Here, we consider a situation where (i) $\boldsymbol{Z} \cup \boldsymbol{W}$ is a set of covariates satisfying the back-door criterion relative to (X, Y), (ii) $\boldsymbol{Z}$ is a subset of confounders selected from prior causal knowledge (possibly an empty set, a sufficient set of confounders, or an insufficient set of confounders), and (iii) $\boldsymbol{W}$ is a set of covariates for which it is uncertain which covariate should be added to $\boldsymbol{Z}$ as a confounder, or we know that a given set of covariates satisfies the back-door criterion but the OLS method is not applicable to estimating total effects using such a set because of the multicollinearity/high-dimensional data problem.

Then, for a smaller subset of $\boldsymbol{Z} \cup \boldsymbol{W}$, if the signs of the regression coefficients of X are equivalent between

the regression models using $\boldsymbol{Z} \cup \boldsymbol{W}$ and a selected smaller set, the regression model using such a subset will not provide misleading qualitative results regarding the total effects. Under the above setting, the aim of this paper is to derive a consistent or less-biased estimator of the total effect.

This paper mainly focuses on a situation where the sum of squares matrix of $\{X\} \cup \boldsymbol{Z}$ is invertible but that of $\{X\} \cup \boldsymbol{Z} \cup \boldsymbol{W}$ is not, because if it is invertible then the total effect is estimable by the OLS method [Pearl, 2009].

3.2 PAL$_p$MA ESTIMATOR

We let s_{xy}, S_{zw} and S_{xz} be the sum of cross-products between X and Y, the sum of the cross-product matrix between $\boldsymbol{Z}$ and $\boldsymbol{W}$ and the sum of the cross-product vectors between X and $\boldsymbol{Z}$, respectively. In addition, we let s_{xx}, S_{zz} and $I_{q,q}$ be the sum of squares of X, the sum of squares matrix of $\boldsymbol{Z}$ and a $q \times q$ identity matrix, respectively. Furthermore, $s_{xx\cdot zw}$, $S_{xw\cdot z}$ and $S_{ww\cdot z}$ are the conditional sum of squares of X given $\boldsymbol{Z}$ and $\boldsymbol{W}$, the conditional sum of the cross-product vector between X and $\boldsymbol{W}$ given $\boldsymbol{Z}$ and the conditional sum of squares matrix of $\boldsymbol{W}$ given $\boldsymbol{Z}$, respectively. Similar notation is used for the remaining sum of squares/cross-products. Then, the proposed method, PAL$_p$MA, is formulated as follows:

Let

$$\begin{pmatrix} \hat{\beta}_{yx\cdot zw} \\ \hat{B}_{yz\cdot xw} \\ \hat{B}_{yw\cdot xz} \end{pmatrix} = \begin{pmatrix} s_{xx} & S_{xz} & S_{xw} \\ S_{xz}^T & S_{zz} & S_{zw} \\ S_{xw}^T & S_{zw}^T & S_{ww} \end{pmatrix}^{-1} \begin{pmatrix} s_{xy} \\ S_{zy} \\ S_{wy} \end{pmatrix} \quad (3)$$

when the sum of squares matrix of the explanatory variables is invertible, and

$$\begin{pmatrix} \tilde{\beta}_{yx\cdot zw} \\ \tilde{B}_{yz\cdot xw} \\ \tilde{B}_{yw\cdot xz} \end{pmatrix} = \begin{pmatrix} s_{xx} & S_{xz} & S_{xw} \\ S_{xz}^T & S_{zz} & S_{zw} \\ S_{xw}^T & S_{zw}^T & \lambda I_{q,q} + S_{ww} \end{pmatrix}^{-1} \begin{pmatrix} s_{xy} \\ S_{zy} \\ S_{wy} \end{pmatrix}, \quad (4)$$

for $\lambda > 0$ when the sum of squares matrix of the explanatory variables is not invertible. Then, for $p = 1, 2$, consider the loss function

$$\begin{aligned} &L_p(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_{yw\cdot xz}) \\ &= \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}B_{yw\cdot xz}||_2^2 \\ &\qquad + \lambda_p||\boldsymbol{\gamma} \odot B_{yw\cdot xz}||_p^p, \end{aligned} \quad (5)$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_q)^T$ is a weight vector such that

$$\boldsymbol{\gamma} = \left(|\tilde{\beta}_{yw_1\cdot xzw_{(1)}}|^{-\xi}, \ldots, |\tilde{\beta}_{yw_q\cdot xzw_{(q)}}|^{-\xi}\right)^T \quad (6)$$

for the non-invertible sum of squares matrix of the explanatory variables with tuning parameter $\xi \geq 0$, and

$$\boldsymbol{\gamma} = \left(|\hat{\beta}_{yw_1\cdot xzw_{(1)}}|^{-\xi}, \ldots, |\hat{\beta}_{yw_q\cdot xzw_{(q)}}|^{-\xi}\right)^T \quad (7)$$

for the invertible sum of squares matrix of the explanatory variables with tuning parameter $\xi \geq 0$, where the superscript "T" stands for a transposed vector/matrix. In addition, $\odot$ refers to the Hadamard product. $||\cdot||_p$ denotes the L_p norm, and λ_p is called a regularization parameter corresponding to the L_p norm ($\lambda_p \geq 0$). $|\cdot|$ stands for the absolute value. The loss function (equation (5)) is different from standard L_p loss functions in the sense that the regularization parameter λ_p is not assigned to $\beta_{yx\cdot zw}$ or $B_{yz\cdot xw}$. In this sense, equation (5) is called a partially adaptive L_p-regularized loss function in this paper. Here, under the assumption the sum of squares matrix of explanatory variables $\{X\} \cup \boldsymbol{Z} \cup \boldsymbol{W}$ is invertible, letting $\lambda_p = 0$, $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$ that minimize equation (5) yield equation (3), i.e., the OLS estimators $\hat{\beta}_{yx\cdot zw}$, $\hat{B}_{yz\cdot xw}$ and $\hat{B}_{yw\cdot xz}$ of equation (2), respectively. Letting $\lambda_2 = \lambda$ and $\xi = 0$, $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$ that minimize equation (5) yield equation (4), i.e., the ridge-type estimators $\tilde{\beta}_{yx\cdot zw}$, $\tilde{B}_{yz\cdot xw}$ and $\tilde{B}_{yw\cdot xz}$ of equation (2), respectively.

For $p = 1$ and $\lambda_1 > 0$, $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$ that minimize equation (5) are called PAL$_1$MA estimators, denoted by $\check{\beta}_{yx\cdot zw}^{\dagger}$, $\check{B}_{yz\cdot xw}^{\dagger}$ and $\check{B}_{yw\cdot xz}^{\dagger}$, respectively. If $\boldsymbol{W}$ is an active set for given $\lambda_1 > 0$, that is, a subset of explanatory variables with nonzero regression coefficients, but does not include any elements of $\{X\} \cup \boldsymbol{Z}$ (i.e., any i-th element of $\check{B}_{yw\cdot xz}^{\dagger}$ does not take the value zero for given $\lambda_1 > 0$), and letting q be the number of explanatory variables in the active set $\boldsymbol{W}$, then under the assumption that the sum of squares matrix of explanatory variables $\{X\} \cup \boldsymbol{Z} \cup \boldsymbol{W}$ is invertible, $\check{\beta}_{yx\cdot zw}^{\dagger}$ is given by

$$\check{\beta}_{yx\cdot zw}^{\dagger} = \hat{\beta}_{yx\cdot zw} + \frac{\lambda_1}{s_{xx\cdot zw}} \hat{B}_{xw\cdot z}^T \boldsymbol{\gamma} \odot \mathrm{sign}(\check{B}_{yw\cdot xz}^{\dagger}). \quad (8)$$

Here, $\hat{B}_{xw\cdot z}$ is given by $\hat{B}_{xw\cdot z} = S_{ww\cdot z}^{-1} S_{wx\cdot z}$. In addition, for a q-dimensional vector $\boldsymbol{a} = (a_1, a_2, ..., a_q)^T$, $\mathrm{sign}(\boldsymbol{a}) = (\mathrm{sign}(a_1), \mathrm{sign}(a_2), ..., \mathrm{sign}(a_q))^T$, where

$$\mathrm{sign}(a_i) = \begin{cases} 1 & : a_i > 0 \\ 0 & : a_i = 0 \\ -1 & : a_i < 0 \end{cases} \quad (9)$$

for $i = 1, 2, ..., q$.

For $p = 2$ and $\lambda_2 > 0$, $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$ that minimize equation (5) are called PAL$_2$MA estimators,

denoted by $\tilde{\beta}^{\dagger}_{yx\cdot zw}$, $\tilde{B}^{\dagger}_{yz\cdot xw}$ and $\tilde{B}^{\dagger}_{yw\cdot xz}$, respectively. Then, $\tilde{\beta}^{\dagger}_{yx\cdot zw}$ is given by

$$\tilde{\beta}^{\dagger}_{yx\cdot zw} = \hat{\beta}_{yx\cdot zw} + \frac{\lambda_2 \hat{B}^T_{yw\cdot xz}(S_{ww\cdot z} + \lambda_2 \mathrm{diag}(\boldsymbol{\gamma}))^{-1} S_{wx\cdot z}}{s_{xx\cdot z} - S_{xw\cdot z}(S_{ww\cdot z} + \lambda_2 \mathrm{diag}(\boldsymbol{\gamma}))^{-1} S_{wx\cdot z}}, \quad (10)$$

where $\mathrm{diag}(\boldsymbol{\gamma})$ is a diagonal matrix whose (i,i) element corresponds to the i-th element of $\boldsymbol{\gamma}$ $(i = 1, 2, ..., q)$.

3.3 L_p COLLAPSIBILITY

In this section, we extend the concept of collapsibility from the framework of traditional regression analysis to regularized regression analysis as follows:

Definition 3 *(L_p Collapsibility) For a given p, $\boldsymbol{W}$ is said to be L_p collapsible with the regression coefficient of X on Y in regression model (2) when the coefficient does not depend on $\boldsymbol{W}$ or the regularization parameter λ_p. In particular, when $\boldsymbol{W}$ is L_p collapsible with the regression coefficient of X on Y in regression model (2) for $p = 1, 2$, $\boldsymbol{W}$ is said to be collapsible with the regression coefficient of X on Y in regression model (2).*

From equations (8) and (10), the following theorem is derived immediately:

Theorem 1 *For $p = 1, 2$, when the sum of squares matrix of $\boldsymbol{Z} \cup \boldsymbol{W}$ is invertible, if $S_{xw\cdot z} = \mathbf{0}_q$ holds, $\boldsymbol{W}$ is collapsible with the regression coefficient of X on Y in regression model (2), i.e., we have*

$$\check{\beta}^{\dagger}_{yx\cdot zw} = \tilde{\beta}^{\dagger}_{yx\cdot zw} = \hat{\beta}_{yx\cdot zw} = \hat{\beta}_{yx\cdot z}. \quad (11)$$

Particularly, if X is conditionally independent of $\boldsymbol{W}$ given $\boldsymbol{Z}$, we have

$$E(\check{\beta}^{\dagger}_{yx\cdot zw}) = E(\tilde{\beta}^{\dagger}_{yx\cdot zw}) = E(\hat{\beta}_{yx\cdot zw}) = E(\hat{\beta}_{yx\cdot z}). \quad (12)$$

Note that $\boldsymbol{W}$ is assumed to be an active set for $p = 1$ in Theorem 1.

Theorem 2 *For $p = 2$, when the sum of squares matrix of $\{X\} \cup \boldsymbol{Z} \cup \boldsymbol{W}$ is invertible, if $S_{yw\cdot xz} = \mathbf{0}_q$ holds, $\boldsymbol{W}$ is L_2 collapsible with the regression coefficient of X on Y in regression model (2), i.e., we have*

$$\tilde{\beta}^{\dagger}_{yx\cdot zw} = \hat{\beta}_{yx\cdot zw} = \hat{\beta}_{yx\cdot z}. \quad (13)$$

Particularly, if Y is conditionally independent of $\boldsymbol{W}$ given X and $\boldsymbol{Z}$, we have

$$E(\tilde{\beta}^{\dagger}_{yx\cdot zw}) = E(\hat{\beta}_{yx\cdot zw}) = E(\hat{\beta}_{yx\cdot z}). \quad (14)$$

Generally, standard regularized regression analysis does not provide consistent estimators of the regression coefficients. In contrast, from Theorem 1, for $p = 1, 2$, $\mathrm{PAL}_p\mathrm{MA}$ provides the consistent estimator of the regression coefficient of X on Y if X and $\boldsymbol{W}$ are conditionally independent given $\boldsymbol{Z}$, regardless of the regularization parameter. In other words, when $\boldsymbol{W}$ is L_p collapsible with the regression coefficient of X on Y and $\boldsymbol{Z}$ satisfies the back-door criterion relative to (X, Y) in regression model (2), $\mathrm{PAL}_p\mathrm{MA}$ can provide a consistent estimator of the total effect. On the other hand, when X is not conditionally independent of $\boldsymbol{W}$ given $\boldsymbol{Z}$, $\mathrm{PAL}_p\mathrm{MA}$ may provide a biased estimator of the regression coefficient of X on Y.

To reduce the bias, consider a partially adaptive L_2-regularized loss function with a weight vector $\boldsymbol{\gamma}^*$ and a tuning parameter ξ^* such that $\boldsymbol{x}$ and $\boldsymbol{y}$ are replaced by an empty set and $\boldsymbol{x}$ in equation (5), respectively. Letting $\tilde{B}^{\dagger}_{xw\cdot z}$ and $\tilde{B}^{\dagger}_{xz\cdot w}$ be $\mathrm{PAL}_2\mathrm{MA}$ estimators of $B_{xw\cdot z}$ and $B_{xz\cdot w}$ derived from such a loss function, respectively, from equation (8), we formulate the modified $\mathrm{PAL}_1\mathrm{MA}$ estimator of $\beta_{yx\cdot zw}$ as

$$\check{\beta}^{*}_{yx\cdot zw} = \check{\beta}^{\dagger}_{yx\cdot zw} - \frac{\lambda_1}{\tilde{s}^{\dagger}_{xx\cdot zw}} \tilde{B}^{\dagger T}_{xw\cdot z} \boldsymbol{\gamma} \odot \mathrm{sign}(\check{B}^{\dagger}_{yw\cdot xz}), \quad (15)$$

$$\tilde{s}^{\dagger}_{xx\cdot zw} = ||\boldsymbol{x} - \boldsymbol{z}\tilde{B}^{\dagger}_{xz\cdot w} - \boldsymbol{w}\tilde{B}^{\dagger}_{xw\cdot z}||_2^2 \quad (16)$$

for an active set $\boldsymbol{W}$. When the sum of squares matrix of $\{X\} \cup \boldsymbol{Z} \cup \boldsymbol{W}$ is invertible, we have

$$\begin{aligned}\check{\beta}^{*}_{yx\cdot zw} &= \check{\beta}^{\dagger}_{yx\cdot zw} - \frac{\lambda_1}{\tilde{s}^{\dagger}_{xx\cdot zw}} \tilde{B}^{\dagger T}_{xw\cdot z} \boldsymbol{\gamma} \odot \mathrm{sign}(\check{B}^{\dagger}_{yw\cdot xz}) \\ &= \hat{\beta}_{yx\cdot zw} + \lambda_1 \left(\frac{1}{s_{xx\cdot zw}} \hat{B}_{xw\cdot z} - \frac{1}{\tilde{s}^{\dagger}_{xx\cdot zw}} \tilde{B}^{\dagger}_{xw\cdot z} \right)^T \\ &\quad \times \boldsymbol{\gamma} \odot \mathrm{sign}(\check{B}^{\dagger}_{yw\cdot xz}). \quad (17)\end{aligned}$$

Thus, when $\boldsymbol{Z} \cup \boldsymbol{W}$ satisfies the back-door criterion, if $\hat{B}_{xw\cdot z} = \tilde{B}^{\dagger}_{xw\cdot z}$ and $s_{xx\cdot zw} = \tilde{s}^{\dagger}_{xx\cdot zw}$ hold (i.e., there is a null regularization parameter is zero), then the total effect is estimated by $\check{\beta}^{*}_{yx\cdot zw}$. In addition, since we have

$$\begin{aligned}\check{\beta}^{*}_{yx\cdot zw} &= \hat{\beta}_{yx\cdot zw} + \lambda_1 \hat{B}^T_{xw\cdot z} \\ &\times \left(\frac{I_{q,q}}{s_{xx\cdot zw}} - \frac{S_{ww\cdot z}(S_{ww\cdot z} + \lambda_2 \mathrm{diag}(\boldsymbol{\gamma}^*))^{-1}}{\tilde{s}^{\dagger}_{xx\cdot zw}} \right) \\ &\quad \times \boldsymbol{\gamma} \odot \mathrm{sign}(\check{B}_{yw\cdot xz}), \quad (18)\end{aligned}$$

if $S_{xw\cdot z} = \mathbf{0}_q$, the total effect is also estimated by $\check{\beta}^{*}_{yx\cdot zw}$.

On the contrary, even when the sum of squares matrix of $\{X\} \cup \boldsymbol{Z} \cup \boldsymbol{W}$ is not invertible, by taking a small value of $\lambda_2 > 0$ such that $S_{ww\cdot z} + \lambda_2 \mathrm{diag}(\boldsymbol{\gamma}^*)$ is invertible in equation (18), the modified $\mathrm{PAL}_1\mathrm{MA}$ can provide the less-biased estimator of the total effects. Hereafter, the modified $\mathrm{PAL}_1\mathrm{MA}$ estimator is merely called the $\mathrm{PAL}_1\mathrm{MA}$ estimator.

3.4 I-PROGLES

Similar to standard regularized regression analysis such as LASSO, adaptive LASSO and elastic net, it is difficult to provide the explicit formula of the PAL$_1$MA estimator of the regression coefficient of X on Y, since equation (5) includes the non-differentiable term $||\boldsymbol{\gamma} \odot B_{yw\cdot xz}||_1^1$; the optimization algorithm is needed to derive the PAL$_1$MA estimator. Here, note that standard LASSO algorithms such as least angle regression [Efron et al, 2004] and generalized path seeking [Friedman, 2012] are not applicable to achieve our aim since neither $\beta_{yx\cdot zw}$ nor $B_{yz\cdot xw}$ are regularized in equation (5).

To derive the PAL$_1$MA estimator $\check{\beta}^*_{yx\cdot zw}$, we propose a novel optimization algorithm that adopts the idea of the block coordinate relaxation method [Sardy et al, 2000]: "integrated algorithm of PROximal Gradient method and LEast Squares method" (i-PROGLES). i-PROGLES, which is shown in Algorithm 1, can be considered the integrated iterative algorithm of the proximal gradient method [Daubechies et al, 2004] and the OLS method. i-PROGLES enables us to include both the treatment variable and some of important confounders in the regression model, regardless of the value taken by the regularization parameters. In addition, if we know that a set of covariates satisfies the back-door criterion, the solution path with such information can be utilized as the criteria of parameter tuning of i-PROGLES to include the set of covariates.

To construct i-PROGLES, let $\boldsymbol{w}_i$ be an n-dimensional observation vector of the i-th explanatory variable W_i of $\boldsymbol{W}$ ($W_i \in \boldsymbol{W} : i = 1, 2, ..., q$). In addition, based on the weight vector $\boldsymbol{\gamma}$ from equations (6) and (7), we define the $n \times q$ matrix $\boldsymbol{w}^\sharp$ and $B^\sharp_{yw\cdot xz}$ as $\boldsymbol{w}^\sharp = \left(\gamma_1^{-1}\boldsymbol{w}_1, \gamma_2^{-1}\boldsymbol{w}_2, \ldots, \gamma_q^{-1}\boldsymbol{w}_q\right)$ and $\boldsymbol{\gamma} \odot B_{yw\cdot xz}$, respectively. Then, for $p = 1$, equation (5) is reformulated as

$$\begin{aligned} &L_1^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B^\sharp_{yw\cdot xz}) \\ &= \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}^\sharp B^\sharp_{yw\cdot xz}||_2^2 \\ &\qquad + \lambda_1 ||B^\sharp_{yw\cdot xz}||_1^1. \end{aligned} \tag{19}$$

Here, $B^\sharp_{yw\cdot xz}[0]$ is defined as the solution of equation (19) given $\beta_{yx\cdot zw} = \hat{\beta}_{yx\cdot z}(= \beta_{yx\cdot zw}[0])$ and $B_{yz\cdot xw} = \hat{B}_{yz\cdot x}(= B_{yz\cdot xw}[0])$. Based on equation (19), in the first substep of the $k+1$-th step ($k \geq 0$), we evaluate $B^\sharp_{yw\cdot xz}$ as the solution of the naive LASSO given $\beta_{yx\cdot zw} = \beta_{yx\cdot zw}[k]$ and $B_{yz\cdot xw} = B_{yz\cdot xw}[k]$:

$$\begin{aligned} &B^\sharp_{yw\cdot xz}[k+1] \\ &= \underset{B}{\operatorname{argmin}} \left(L_1^\sharp(\beta_{yx\cdot zw}[k], B_{yz\cdot xw}[k], B) \right). \end{aligned} \tag{20}$$

Here, letting $S^\sharp_{ww}$, $S^\sharp_{yw}$, $S^\sharp_{wx}$ and $S^\sharp_{wz}$ be the sum of squares matrix of $\boldsymbol{W}^\sharp$, the sum of cross-products vector between Y and $\boldsymbol{W}^\sharp$, the sum of cross-products vector between $\boldsymbol{W}^\sharp$ and X and the sum of cross-products matrix between $\boldsymbol{W}^\sharp$ and $\boldsymbol{Z}$, respectively, and

$$\begin{aligned} &f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_{yw\cdot xz}) \\ &= \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}^\sharp B_{yw\cdot xz}||_2^2, \end{aligned} \tag{21}$$

$B^\sharp_{yw\cdot xz}[k+1]$ is formulated by

$$\begin{aligned} &B^\sharp_{yw\cdot xz}[k+1] = \operatorname{prox}_{\eta\lambda_1}\Big(B^\sharp_{yw\cdot xz}[k] - \eta \\ &\times \frac{\partial}{\partial B} f^\sharp\left(\beta_{yx\cdot zw}[k], B_{yz\cdot xw}[k], B\right)_{B = B^\sharp_{yw\cdot xz}[k]} \Big), \end{aligned} \tag{22}$$

which is straightforward from equation (20) through the proximal gradient method [Daubechies et al, 2004] given $\beta_{yx\cdot zw}[k]$ and $B_{yz\cdot xw}[k]$. In this paper, $\operatorname{prox}_a(b)$ is defined as

$$\operatorname{prox}_a(b) = \begin{cases} b - a & : \ b \geq a \\ 0 & : \ -a < b < a \\ b + a & : \ b \leq -a \end{cases}. \tag{23}$$

In addition, noting

$$\begin{aligned} &\frac{\partial}{\partial B} f^\sharp\left(\beta_{yx\cdot zw}[k], B_{yz\cdot xw}[k], B\right)_{B = B^\sharp_{yw\cdot xz}[k]} \\ &= S^\sharp_{wx}\beta_{yx\cdot zw}[k] + S^\sharp_{wz}B_{yz\cdot xw}[k] + S^\sharp_{ww}B^\sharp_{yw\cdot xz}[k] - S^\sharp_{wy}, \end{aligned} \tag{24}$$

we have

$$\begin{aligned} &B^\sharp_{yw\cdot xz}[k+1] = \operatorname{prox}_{\eta\lambda_1}(B^\sharp_{yw\cdot xz}[k] - \eta\ (S^\sharp_{wx}\beta_{yx\cdot zw}[k] \\ &\quad + S^\sharp_{wz}B_{yz\cdot xw}[k] + S^\sharp_{ww}B^\sharp_{yw\cdot xz}[k] - S^\sharp_{wy})), \end{aligned} \tag{25}$$

where η satisfies $\eta \leq (\lambda_{\max}(S^\sharp_{ww}))^{-1}$. Here, $\lambda_{\max}(S^\sharp_{ww})$, which is the maximum eigenvalue of $S^\sharp_{ww}$, corresponds to the Lipschitz constant with respect to $(\partial/\partial B_{yw\cdot xz})f^\sharp$.

In the second substep of the $k+1$-th step, we evaluate $\beta_{yx\cdot zw}[k+1]$ and $B_{yz\cdot xw}[k+1]$ by the OLS method given $B_{yw\cdot xz} = B_{yw\cdot xz}[k+1]$:

$$\begin{aligned} &\left(\beta_{yx\cdot zw}[k+1], B^T_{yz\cdot xw}[k+1]\right)^T \\ &\qquad = \underset{b,B}{\operatorname{argmin}} \left(f^\sharp(b, B, B_{yw\cdot xz}[k+1])\right) \\ &= \begin{pmatrix} s_{xx} & S_{xz} \\ S^T_{xz} & S_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{x}^T \\ \boldsymbol{z}^T \end{pmatrix} \left(\boldsymbol{y} - \boldsymbol{w}B_{yw\cdot xz}[k+1]\right). \end{aligned} \tag{26}$$

Regarding the convergence of i-PROGLES, the following theorem can be derived:

Algorithm 1 : i-PROGLES (both λ_2 and ξ_2 are used to derive $\tilde{B}^{\dagger}_{xw\cdot z}$ and $\tilde{s}^{\dagger}_{xx\cdot zw}$)

Input: $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ and $\boldsymbol{w}$, $k^* > 0$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\xi_1 > 0$, $\xi_2 > 0$

$$\beta_{yx\cdot zw}[0] = \hat{\beta}_{yx\cdot z}, \quad B_{yz\cdot xw}[0] = \hat{B}_{yz\cdot x}$$

$$B^{\sharp}_{yw\cdot xz}[0] = \underset{B}{\operatorname{argmin}} \left(\frac{1}{2} ||\boldsymbol{y} - \boldsymbol{x}\hat{\beta}_{yx\cdot z} - \boldsymbol{z}\hat{B}_{yz\cdot x} - \boldsymbol{w}^{\sharp} B||_2^2 + \lambda_1 ||B||_1 \right)$$

Calculate the weight vector: If the sum of squares matrix of the explanatory variables is not invertible, set

$$\boldsymbol{\gamma} = \left(|\tilde{\beta}_{yw_1\cdot xzw_{(1)}}|^{-\xi_1}, |\tilde{\beta}_{yw_2\cdot xzw_{(2)}}|^{-\xi_1}, \ldots, |\tilde{\beta}_{yw_q\cdot xzw_{(q)}}|^{-\xi_1} \right)^T$$

If the sum of squares matrix of the explanatory variables is invertible, set

$$\boldsymbol{\gamma} = \left(|\hat{\beta}_{yw_1\cdot xzw_{(1)}}|^{-\xi_1}, |\hat{\beta}_{yw_2\cdot xzw_{(2)}}|^{-\xi_1}, \ldots, |\hat{\beta}_{yw_q\cdot xzw_{(q)}}|^{-\xi_1} \right)^T$$

1: **for** $k = 0$ to k^* **do**

2: Set

$$\eta \leq (\lambda_{\max}(S^{\sharp}_{ww}))^{-1}$$

$$B^{\sharp}_{yw\cdot xz}[k+1] = \operatorname{prox}_{\eta\lambda_1}(B^{\sharp}_{yw\cdot xz}[k] - \eta(S^{\sharp}_{wx}\beta_{yx\cdot zw}[k] + S^{\sharp}_{wz}B_{yz\cdot xw}[k] + S^{\sharp}_{ww}B^{\sharp}_{yw\cdot xz}[k] - S^{\sharp}_{wy}))$$

3: Set

$$B_{yw\cdot xz}[k+1] = \left(\gamma_1^{-1}\beta^{\sharp}_{yw_1\cdot xzw_{(1)}}[k+1], \gamma_2^{-1}\beta^{\sharp}_{yw_2\cdot xzw_{(2)}}[k+1] \ldots, \gamma_q^{-1}\beta^{\sharp}_{yw_q\cdot xzw_{(q)}}[k+1] \right)^T$$

4: Set

$$\beta_{yx\cdot zw}[k+1] = \hat{\beta}_{yx\cdot z} - \hat{B}_{wx\cdot z}B_{yw\cdot xz}[k+1], \quad B_{yz\cdot xw}[k+1] = \hat{B}_{yz\cdot x} - \hat{B}_{wz\cdot x}B_{yw\cdot xz}[k+1]$$

5: **end for**

6: Set

$$\check{\beta}^*_{yx\cdot zw} = \beta_{yx\cdot zw}[k^*+1] - \frac{\lambda_1}{\tilde{s}^{\dagger}_{xx\cdot zw}} \tilde{B}^{\dagger T}_{xw\cdot z}\boldsymbol{\gamma} \odot \operatorname{sign}(B_{yw\cdot xz}[k^*+1])$$

7: **return** $\check{\beta}^*_{yx\cdot zw}$

Theorem 3 *Let $\{\beta_{yx\cdot zw}[k]\}_{k\geq 0}$, $\{B_{yz\cdot xw}[k]\}_{k\geq 0}$ and $\{B_{yw\cdot xz}[k]\}_{k\geq 0}$ be the sequences of $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$, respectively, generated by i-PROGLES, and let $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{z})$. When $\beta^*_{yx\cdot zw}$, $B^*_{yz\cdot xw}$ and $B^*_{yw\cdot xz}$ minimize equation (19) regarding $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$, respectively, there exists the natural number K for any $\epsilon > 0$ such that*

$$\begin{aligned} & L_1\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^*_{yw\cdot xz}\right) \\ & \quad - L_1\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B_{yw\cdot xz}[k+1]\right) \\ & \leq \frac{\lambda_{max}(S^{\sharp}_{ww})}{2k} ||B^{\sharp}_{yw\cdot xz}[0] - B^{\sharp *}_{yw\cdot xz}||_2^2 \\ & + \frac{\lambda_{max}(S_{uu})}{2} \lambda_{max}(S^{\sharp}_{wu}S^{-2}_{uu}S^{\sharp}_{uw})\epsilon. \end{aligned} \tag{27}$$

holds for any $k \geq K$, where $B^{\sharp}_{yw\cdot xz}[k] = \boldsymbol{\gamma} \odot B_{yw\cdot xz}[k]$ and $B^{\sharp *}_{yw\cdot xz} = \boldsymbol{\gamma} \odot B^*_{yw\cdot xz}$.

The proof is given in the Supplementary Material.

4 NUMERICAL EXPERIMENT

In this section, we present a numerical experiment to compare the performance of LASSO, adaptive LASSO, elastic net, SCAD, MCP, OLS and PAL$_1$MA. For simplicity, letting X and Y be the treatment variable and the response variable, respectively, consider the linear SCM with 42 explanatory variables for Y in the form of

$$\left. \begin{array}{l} Y = \alpha_{yx}X + \alpha_{yz_1}Z_1 + \alpha_{yz_2}Z_2 + A_{yw}\boldsymbol{W} + \epsilon_y \\ X = \alpha_{xz_1}Z_1 + \alpha_{xz_2}Z_2 + \epsilon_x \end{array} \right\} \tag{28}$$

where $\boldsymbol{W}$ includes 39 variables. In this setting, we assume that $\{Z_1, Z_2\}$ satisfies the back-door criterion relative to (X, Y) and the path coefficients of $\{Z_2\} \cup \boldsymbol{W}$ on Y are regularized but Z_1 is not. Then, Theorem 1 does not hold, and the estimated total effect may be biased.

To set up a simulation, we first construct the population variance-covariance matrix. To eliminate the arbitrariness, the true values of the path coefficients

Table 1. Results based on cross-validation.

	(a) $\tau_{yx} = 0.474$					parameter settings				
	mean	bias	mse	sd	sign	λ	ξ	ϕ	λ_1	ξ_1
LASSO	0.1812	0.2929	0.1012	0.1238	0.8824	0.0830	-	-	-	-
adaptive LASSO	0.2736	0.2006	0.0776	0.1934	0.8932	3.4300	1.7000	-	-	-
Elastic net	0.2101	0.2641	0.0807	0.1047	0.9664	0.0780	-	0.5500	-	-
MCP	0.2290	0.2451	0.0862	0.1617	0.8462	0.0600	19.5000	-	-	-
SCAD	0.1909	0.2832	0.1032	0.1517	0.8216	0.0860	15.5000	-	-	-
PAL_1MA	0.4486	0.0256	0.0640	0.2516	0.9746	-	-	-	0.0100	0.1000
OLS	0.4717	0.0025	0.2961	0.5441	0.8154	-	-	-	-	-

mean: sample mean; bias: bias between the true value and the sample mean; mse: mean squared error; sd: standard deviation; sign: coincidence rate between the signs of the true value and the estimates; λ, λ_1: regularization parameters; ξ, ξ_1: tuning parameters; ϕ: mixing parameter. The regularization parameter λ_2 and tuning parameter ξ_2 are selected as $\lambda_2 = 0.0014$, $\xi_2 = 0.0013$. Refer to the Supplementary Material for the selection of these parameters.

α_{yx}, α_{yz_1}, α_{yz_2}, $A_{yw} = (\alpha_{yw_1}, ..., \alpha_{yw_{39}})$, α_{xz_1} and α_{xz_2} are randomly and independently determined according to the uniform distribution with the interval $[-3, 3]$. In addition, we assume that (i) the random disturbances ϵ_x and ϵ_y independently follow the normal distribution with mean zero and variance one, and (ii) the random disturbances are also independent of their nondescendants. Furthermore, the population variance-covariance matrices of $\{Z_1, Z_2\} \cup \boldsymbol{W}$ are randomly determined according to Pourahmadi and Wang [2015].

We generated 30 random samples of 42 variables from a multivariate normal distribution with a zero mean vector and the above variance-covariance matrix for 5000 replications. Table 1 shows the basic statistics of the total effects estimated by LASSO, adaptive LASSO, elastic net, SCAD, MCP, OLS and PAL_1MA based on the given sample size of 30 for each parameter setting. Regarding the parameter tuning for regularized regression analysis, see the Supplementary Material. Here, for the OLS method, we select a set of covariates based on prior causal knowledge; i.e., $\{Z_1, Z_2\}$ are selected.

From Table 1, both the PAL_1MA estimators and the OLS estimators are almost consistent with the true values of the total effects, but the other regularized regression methods yield highly biased estimators. In addition, the coincidence rates between the signs of the estimated total effects and the true total effects for PAL_1MA are better than those for the other regression methods. From Fig. 1, the interquartile ranges of both PAL_1MA and OLS include the true value of the total effects, but the other regularized regression analyses do not include this value of the total effects. For further discussion on the simulation experiments, see the Supplementary Material.

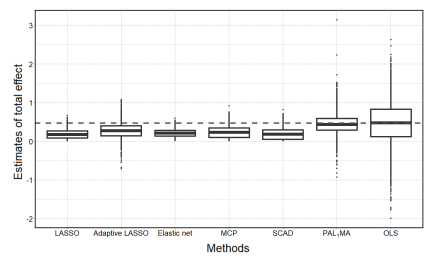


Fig. 1. Boxplots of the estimated total effects. The dashed lines show the true total effects.

5 CONCLUSION

In current situations where advanced artificial intelligence (AI) technology enables us to collect large datasets, it would not be so difficult to observe a large number of covariates. In such situations, it would be reasonable to consider that such a set of covariates satisfies the back-door criterion to estimate the total effects. However, when multicollinearity/high-dimensional data problems occur in even this situation, it is difficult to evaluate the linear causal effects reliably. To solve this problem, we established PAL_pMA to provide a consistent or less-biased estimator of the total effects. In addition, through numerical experiments and a case study in Supplementary Material, we confirmed that PAL_1MA is superior to other estimation methods. The results of this paper are applicable to evaluating the direct effect in the framework of regression models through the "single-door criterion" [Pearl, 2009].

Finally, although PAL_pMA is formulated based on linear regression models, it would be interesting to extend our approach to a wide variety of statistical models, including generalized linear models, generalized estimating equations and proportional hazards models. Such an extension would be straightforward — the objective function would be replaced with a more general form. This extension will be left for future work.

References

Brito, C. Graphical methods for identification in structural equation models. Computer Science Department, UCLA, PhD Thesis, 2004.

Bühlmann, P. and van de Geer, S. *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer Science and Business Media, 2011.

Chen, B. R. Graphical methods for linear structural equation modeling. Computer Science Department, UCLA, PhD Thesis, 2017.

Chen, B., Kumor, D. and Bareinboim, E. Identification and model testing in linear structural equation models using auxiliary variables, *Proceedings of the 34th International Conference on Machine Learning*, 757–766, 2017.

Clogg, C.C., Petkova, E. and Shihadeh, E. S. Statistical methods for analyzing collapsibility in regression models. *Journal of Educational Statistics*, **17**:51–74, 1992.

Daubechies, I., Defrise, M. and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, **57**:1413–1457, 2004.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. Least angle regression. *Annals of Statistics*, **32**:407–499, 2004.

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

Frisch, R. *Statistical Confluence Analysis by Means of Complete Regression Systems*, University Institute of Economics, 1934.

Friedman, J. H. Fast sparse regression and classification. *International Journal of Forecasting*, **28**:722-738, 2012.

Geng, Z. and Asano, C. Strong collapsibility of association measures in linear models. *Journal of the Royal Statistical Society: Series B*, **55**:741–747, 1993.

Guo, J. H. and Geng, Z. Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society: Series B*, **57**:263–267, 1995.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**:55–67, 1970.

Kuroki, M. Optimizing an external intervention using a structural equation model with an application to statistical process analysis. *Journal of Applied Statistics*, **39**:673-694, 2012.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, **101**:423–437, 2014.

Pearl, J. *Causality: Models, Reasoning, and Inference, 2nd edition*, Cambridge University Press, 2009.

Pearl, J. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, **1**:155–170, 2013.

Pearl, J. A linear 'microscope' for interventions and counterfactuals. *Journal of Causal Inference*, **5**:1–15, 2017.

Pourahmadi, M. and Wang, X. Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor. *Statistics and Probability Letters*, **106**:5-12, 2015.

Sardy, S., Bruce, A. G. and Tseng, P. Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics*, **9**:361-379, 2000.

Stanghellini, E. Instrumental variables in Gaussian directed acyclic graph models with an unobserved confounder. *Environmetrics*, **15**:463-469, 2004.

Stanghellini, E. and Pakpahan, E. Identification of causal effects in linear models: beyond instrumental variables. *Test*, **24**:489–509, 2015.

Tian, J. Identifying linear causal effects. *Proceeding of the 19th National Conference on Artificial Intelligence*, 104-111, 2004.

Tian, J. A criterion for parameter identification in structural equation models. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 392-399, 2007a.

Tian, J. On the identification of a class of linear models. *Proceedings of the 22nd National Conference on Artificial Intelligence*, 1284-1289, 2007b.

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

van de Geer, S., Bühlmann, P., Ritov, Y. A. and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, **42**:1166–1202, 2014.

Wermuth, N. Moderating effects of subgroups in linear models. *Biometrika*, **76**:81–92, 1989a.

Wermuth, N. Moderating effects in multivariate normal distributions. *Methodika*, **3**:74–93, 1989b.

Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**:894-942, 2010.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**:1418-1429, 2006.

Zou, H. and Hastie, T. Regularization and variable selection via the Elastic net. *Journal of the Royal Statistical Society: Series B*, **67**:301-320, 2005.

Partially Adaptive Regularized Regression for Estimating Linear Causal Effects: Supplementary Material

A The Proof of Theorem 3

A.1 Basic Theory

In this section, we provide a brief review of the basic theory of optimization, which is used to prove Theorem 3. Readers who are familiar with optimization theory can skip this section. For details, also refer to, for example, Beck [2017].

Throughout the Supplementary Material, let $f(\boldsymbol{x})$ be a proper, closed, and convex function. Here, $f(\boldsymbol{x})$ is called proper when the domain of $f(\boldsymbol{x})$, $\mathrm{dom}(f) = \{\boldsymbol{x}|f(\boldsymbol{x}) < \infty\}$, is not empty, and $f(\boldsymbol{x})$ takes values on the extended real number line—i.e., $(-\infty, \infty]$. In addition, $f(\boldsymbol{x})$ is called closed when $\liminf_{x \to x_0} f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$ holds, where "lim inf" is the limit inferior (of f at point $\boldsymbol{x}_0$). Furthermore, $f(\boldsymbol{x})$ is called σ-strongly convex for a given $\sigma > 0$ if $\mathrm{dom}(f)$ is convex, and the following inequality holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ and $\lambda \in [0, 1]$:

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\sigma}{2}\lambda(1-\lambda)||\boldsymbol{x} - \boldsymbol{y}||_2^2. \tag{A.1}$$

In particular, $f(\boldsymbol{x})$ is called convex if equation (A.1) holds for $\sigma = 0$. In addition, a set C is called convex if it holds that $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in C$ for any $\boldsymbol{x}$, $\boldsymbol{y} \in C$ and $\lambda \in [0, 1]$. Throughout the Supplementary Material, the function $g(\boldsymbol{x})$ is also a proper, closed, and convex function.

The function $f(\boldsymbol{x})$ is also assumed to satisfy the following conditions: (i) $\mathrm{dom}(f)$ is convex, (ii) $\mathrm{dom}(g) \subset \mathrm{int}(\mathrm{dom}(f))$ and (iii) $f(\boldsymbol{x})$ is l_f-smooth over $\mathrm{int}(\mathrm{dom}(f))$. Here, $f(\boldsymbol{x})$ is called an l_f-smooth function when $(\partial/\partial \boldsymbol{x})f$ is a Lipschitz continuous function with Lipschitz constant l_f. For a set A, $\mathrm{int}(A)$ is a set of all interior points of A. The function $h(\boldsymbol{x})$ is called Lipschitz continuous if there exists a positive real constant K such that $|h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2)| \leq K||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathrm{dom}(h)$ and such a K is called a Lipschitz constant. Finally, the minimizer of $f(\boldsymbol{x})$ is a point $\boldsymbol{a}$ for which $f(\boldsymbol{x}) > f(\boldsymbol{a})$ at $\boldsymbol{x}$ around $\boldsymbol{a}$.

For a p-dimensional vector $\boldsymbol{x} \in \mathbb{R}^p$, consider a problem that finds the minimizer of

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{A.2}$$

where we assume that the optimal set of $\underset{x}{\mathrm{argmin}}\,(f(\boldsymbol{x}) + g(\boldsymbol{x}))$ is nonempty in this Supplementary Material.

Under the preparation above, we introduce the following propositions.

Proposition 1 (Convergence rate of the proximal gradient method [Beck, 2017] *For the sequence $\{\boldsymbol{x}[k]\}_{k \geq 0}$ defined by*

$$\boldsymbol{x}[k+1] = \underset{x \in \mathbb{R}^p}{\mathrm{argmin}} \left(f(\boldsymbol{x}[k]) + \left\langle \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})_{x=x[k]}, \boldsymbol{x} - \boldsymbol{x}[k] \right\rangle + g(\boldsymbol{x}) + \frac{1}{2t}||\boldsymbol{x} - \boldsymbol{x}[k]||_2^2 \right) \tag{A.3}$$

$$= \underset{x \in \mathbb{R}^p}{\mathrm{argmin}} \left(tg(\boldsymbol{x}) + \frac{1}{2}||\boldsymbol{x} - (\boldsymbol{x}[k] - t\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})_{x=x[k]})||_2^2 \right) \tag{A.4}$$

for $t > 0$ and the initial vector $\boldsymbol{x}[0]$ from dom(F), $\boldsymbol{x}^*$ is the minimizer of $F(\boldsymbol{x})$:

$$F(\boldsymbol{x}[k]) - F(\boldsymbol{x}^*) \leq \frac{l_f}{2k}||\boldsymbol{x}[0] - \boldsymbol{x}^*||_2 \tag{A.5}$$

for $k \geq 0$ and $t \leq 1/l_f$.

Proposition 2 [Beck, 2017] *For D to be a Euclidean space, let $f : D \to (-\infty, \infty]$ be a proper, closed, and σ-strongly convex function ($\sigma > 0$). Then,*

(a) $f(\boldsymbol{x})$ has a unique minimizer $\boldsymbol{x}^*$ in dom(f),

(b) for all $\boldsymbol{x} \in$ dom(f),

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\sigma}{2}||\boldsymbol{x} - \boldsymbol{x}^*||_2^2.$$

Proposition 3 [Beck, 2017] *Let D be a Euclidean space and $f : D \to (-\infty, \infty]$ be a σ-strongly convex function if and only if the function $f(\boldsymbol{x}) - \frac{\sigma}{2}||\boldsymbol{x}||_2^2$ is convex.*

A.2 Preparation

For Section 3, let $\boldsymbol{w}_i$ be an n-dimensional observation vector of the i-th explanatory variable W_i of $\boldsymbol{W}$ ($W_i \in \boldsymbol{W} : i = 1, 2, ..., q$). In addition, based on the weight vector $\boldsymbol{\gamma}$ of equations (6) and (7), we define the $n \times q$ matrix $\boldsymbol{w}^\sharp$ and $B^\sharp_{yw\cdot xz}$ as

$$\boldsymbol{w}^\sharp = \left(\frac{\boldsymbol{w}_1}{\gamma_1}; \frac{\boldsymbol{w}_2}{\gamma_2}; \ldots; \frac{\boldsymbol{w}_q}{\gamma_q}\right) \tag{A.6}$$

and $\boldsymbol{\gamma} \odot B_{yw\cdot xz}$, respectively. Then, for $p = 1$, equation (5) is reformulated as

$$L_1^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B^\sharp_{yw\cdot xz}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}^\sharp B^\sharp_{yw\cdot xz}||_2^2 + \lambda_1||B^\sharp_{yw\cdot xz}||_1^1. \tag{A.7}$$

Then, to solve our problem, we adopt the idea of the block-coordinate-relaxation method [Sardy et al, 2000]. Intuitively, in the block-coordinate-relaxation method, a whole set of variables is divided into several blocks, and the original optimization problem is iteratively solved as a sequential optimization problem regarding some blocks under the assumption that the remaining blocks are constant. Based on this idea, first, we divide equation (A.7) into the following two kinds of functions:

$$f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B^\sharp_{yw\cdot xz}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}^\sharp B^\sharp_{yw\cdot xz}||_2^2 \tag{A.8}$$

$$g(B^\sharp_{yw\cdot xz}) = \lambda_1||B^\sharp_{yw\cdot xz}||_1^1. \tag{A.9}$$

Then, when we divide a whole set of variables into $\{X\} \cup \boldsymbol{Z}$ and $\boldsymbol{W}$ according to the block-coordinate-relaxation method, the minimum optimization for equation (A.7) includes the following two substep minimization procedures in the $k+1$-th step ($k \geq 0$):

$$\left.\begin{array}{l} B^\sharp_{yw\cdot xz}[k+1] = \underset{B}{\operatorname{argmin}} \left(L_1^\sharp(\beta_{yx\cdot zw}[k], B_{yz\cdot xw}[k], B)\right) \\ \left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1]^T\right)^T = \underset{b,B}{\operatorname{argmin}} \left(L_1^\sharp(\boldsymbol{b}, B, B^\sharp_{yw\cdot xz}[k+1])\right) \end{array}\right\}, \tag{A.10}$$

where

$$\beta_{yx\cdot zw}[0] = \hat{\beta}_{yx\cdot z}, \quad B_{yz\cdot xw}[0] = \hat{B}_{yz\cdot x}$$

$$B^\sharp_{yw\cdot xz}[0] = \underset{B}{\operatorname{argmin}} \left(\frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\hat{\beta}_{yx\cdot z} - \boldsymbol{z}\hat{B}_{yz\cdot x} - \boldsymbol{w}^\sharp B||_2^2 + \lambda_1||B||_1^1\right).$$

First, from equation (A.3), $B^\sharp_{yw\cdot xz}[k+1]$ can be expressed as follows:

$$\begin{aligned} B^\sharp_{yw\cdot xz}[k+1] = \underset{B}{\operatorname{argmin}} \Big(& f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B^\sharp_{yw\cdot xz}[k]\right) \\ & + \left\langle \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B^\sharp_{yw\cdot xz}[k]}, B - B^\sharp_{yw\cdot xz}[k] \right\rangle + g(B) + \frac{l_f}{2}||B - B^\sharp_{yw\cdot xz}[k]||_2^2 \Big), \end{aligned} \tag{A.11}$$

where l_f is a Lipschitz constant with respect to the partial derivative function

$$\frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B). \tag{A.12}$$

Here, through the partial derivative of convex function (A.11) with respect to B, equation (A.11) can also be rewritten as

$$B^{\sharp}_{yw\cdot xz}[k+1] = \mathrm{prox}_{\frac{1}{l_f}g}\left(B^{\sharp}_{yw\cdot xz}[k] - \frac{1}{l_f}\frac{\partial}{\partial B}f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B^{\sharp}_{yw\cdot xz}[k]}\right)\left(= T_{l_f}\left(B^{\sharp}_{yw\cdot xz}[k]\right)\right), \quad \text{(A.13)}$$

where

$$\mathrm{prox}_a(b) = \begin{cases} b-a & : \ b \geq a \\ 0 & : \ -a < b < a \\ b+a & : \ b \leq -a \end{cases} \quad \text{(A.14)}$$

and we define

$$T_{l_f}(B') = \mathrm{prox}_{\frac{1}{l_f}g}\left(B' - \frac{1}{l_f}\frac{\partial}{\partial B}f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B'}\right) \quad \text{(A.15)}$$

for any fixed $\beta_{yx\cdot zw}$ and $B_{yz\cdot xw}$. Then, we obtain equation (25) by replacing $\beta_{yx\cdot zw}$ and $B_{yz\cdot xw}$ with $\beta_{yx\cdot zw}[k]$ and $B_{yz\cdot xw}[k]$, respectively.

Second, since the sum of squares matrix of X and $\boldsymbol{Z}$ is invertible in the paper, clearly, the solution of $(\beta_{yx\cdot zw}[k+1],\ B^T_{yz\cdot xw}[k+1])^T$ given $B^{\sharp}_{yw\cdot xz}[k+1]$ can be derived as the least squares estimators of $(b, B)^T$:

$$\begin{pmatrix} \beta_{yx\cdot zw}[k+1] \\ B_{yz\cdot xw}[k+1] \end{pmatrix} = \begin{pmatrix} s_{xx} & S_{xz} \\ S^T_{xz} & S_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{x}^T \\ \boldsymbol{z}^T \end{pmatrix} (\boldsymbol{y} - \boldsymbol{w}B_{yw\cdot xz}[k+1]). \quad \text{(A.16)}$$

Thus, letting $\{(\beta_{yx\cdot zw}[l], B^T_{yz\cdot xw}[l])^T\}_{l\geq 0}$ and $\{B^{\sharp}_{yw\cdot xz}[k]\}_{k\geq 0}$ be the sequence generated by procedure (A.10) for solving the minimization problem with respect to loss function (A.7), $L^{\sharp}_1(\beta_{yx\cdot zw}[l], B_{yz\cdot xw}[l], B^{\sharp}_{yw\cdot xz}[k])$ is a monotonically decreasing function of l and k.

A.3 Proof

Under the preparation in Section A.2, we prove the following lemmas to prove Theorem 3.

Lemma 1 *For a given $\beta_{yx\cdot zw}$ and $B_{yz\cdot xw}$ in equation (A.7), we have*

$$\begin{aligned} & L^{\sharp}_1\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_1\right) - L^{\sharp}_1\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)\right) \\ \geq\ & \frac{l_f}{2}||B_1 - T_{l_f}(B_2)||^2_2 - \frac{l_f}{2}||B_1 - B_2||^2_2 + d_f(B_1, B_2) \end{aligned} \quad \text{(A.17)}$$

for any B_1 and B_2, where $d_f(B_1, B_2)$ is the Bregman distance with $f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)$ between B_1 and B_2; i.e.,

$$\begin{aligned} d_f(B_1, B_2) = & f^{\sharp}\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_1\right) - f^{\sharp}\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) \\ & - \left\langle \frac{\partial}{\partial B}f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, B_1 - B_2 \right\rangle. \end{aligned} \quad \text{(A.18)}$$

and $\langle \boldsymbol{a}, \boldsymbol{b}\rangle$ is an inner product between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, i.e., $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \boldsymbol{a}^T\boldsymbol{b}$.

Proof of Lemma 1: Letting

$$\psi(\boldsymbol{b}) = f^{\sharp}\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) + \left\langle \frac{\partial}{\partial B}f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, \boldsymbol{b} - B_2 \right\rangle + g(\boldsymbol{b}) + \frac{l_f}{2}||\boldsymbol{b} - B_2||^2_2, \quad \text{(A.19)}$$

ψ is an l_f-strongly convex function from Proposition 3. Referring to equations (A.11) and (A.13), we have

$$\mathrm{prox}_{\frac{1}{l_f}g}\left(B_2 - \frac{1}{l_f}\frac{\partial}{\partial B}f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}\right) \quad \text{(A.20)}$$

$$= \underset{b}{\operatorname{argmin}} \left(f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) + \left\langle \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, \boldsymbol{b} - B_2 \right\rangle + g(\boldsymbol{b}) + \frac{l_f}{2}||\boldsymbol{b} - B_2||_2^2 \right)$$

and $T_{l_f}(B_2) = \underset{b}{\operatorname{argmin}} \psi(\boldsymbol{b})$. Thus, from Proposition 2, we have

$$\psi(B_1) - \psi(T_{l_f}(B_2)) \geq \frac{l_f}{2}||B_1 - T_{l_f}(B_2)||_2^2. \tag{A.21}$$

Here, letting $\lambda_{\max}(A)$ be the maximum eigenvalue of a $p \times p$ symmetric matrix A, when we define

$$\|A\|_{op} = \sup_{x \neq 0} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \lambda_{\max}(A),$$

for all B' and B'', we have

$$\begin{aligned} ||\frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B'} - \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B''}||_2 &= ||(\boldsymbol{w}^\sharp)^T(\boldsymbol{w}^\sharp B'' - \boldsymbol{w}^\sharp B')||_2^2 \\ &\leq ||(\boldsymbol{w}^\sharp)^T \boldsymbol{w}^\sharp||_{op}||B'' - B'||_2^2 \leq \lambda_{\max}(S_{ww}^\sharp)||B'' - B'||_2^2, \end{aligned} \tag{A.22}$$

then $f(\boldsymbol{x})$ is a $\lambda_{\max}(S_{ww}^\sharp)(= l_f)$-smooth function. In addition, from the Cauchy–Schwarz inequality and equation (A.22),

$$\begin{aligned} &\left\langle \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B'} - \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B''}, B' - B'' \right\rangle \\ &\quad \leq ||\frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B'} - \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B''}||_2||B' - B''||_2 \\ &\quad \leq l_f||B' - B''||_2^2. \end{aligned} \tag{A.23}$$

Letting

$$h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B) = \frac{l_f}{2}||B||_2^2 - f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B), \tag{A.24}$$

from equation (A.24), we have

$$\begin{aligned} &\left\langle \frac{\partial}{\partial B} h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B'} - \frac{\partial}{\partial B} h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B''}, B' - B'' \right\rangle \\ &\quad = \left\langle l_f(B' - B'') - \left(\frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B'} - \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B''} \right), B' - B'' \right\rangle \\ &\quad \geq 0 \end{aligned} \tag{A.25}$$

for any B' and B''. From equation (A.25), since $h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)$ is a convex function with respect to B and satisfies the first-order condition, we obtain

$$\begin{aligned} & h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)) \geq h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2) + \left\langle \frac{\partial}{\partial B} h(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, T_{l_f}(B_2) - B_2 \right\rangle \\ \Leftrightarrow \quad & \frac{l_f}{2}||T_{l_f}(B_2)||_2^2 - f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)) \geq \frac{l_f}{2}||B_2||_2^2 - f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2) \\ & + \left\langle l_f B_2 - \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, T_{l_f}(B_2) - B_2 \right\rangle \\ \Leftrightarrow \quad & f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)\right) \leq f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) \\ & + \left\langle \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, T_{l_f}(B_2) - B_2 \right\rangle + \frac{l_f}{2}||T_{l_f}(B_2) - B_2||_2^2 \\ \Leftrightarrow \quad & L_1^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)) \leq \psi(T_{l_f}(B_2)). \end{aligned} \tag{A.26}$$

From equation (A.26) together with equation (A.21), we derive

$$\psi(B_1) - L_1^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)) \geq \frac{l_f}{2}||B_1 - T_{l_f}(B_2)||_2^2, \tag{A.27}$$

for any fixed $\beta_{yx\cdot zw}$ and $B_{yz\cdot xw}$ and any B_1.

From equations (A.19) and (A.27), we obtain

$$\begin{aligned} & f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) + \left\langle \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, B_1 - B_2 \right\rangle + g(B_1) \\ & + \frac{l_f}{2}||B_1 - B_2|| - L_1^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)) \geq \frac{l_f}{2}||B_1 - T_{l_f}(B_2)||_2^2, \end{aligned} \tag{A.28}$$

and adding $f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_1\right)$ to the right-hand side of the above function for rearrangement, we obtain

$$\begin{aligned} & L_1^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_1\right) - L_1^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, T_{l_f}(B_2)\right) \\ \geq \; & \frac{l_f}{2}||B_1 - T_{l_f}(B_2)||_2^2 - \frac{l_f}{2}||B_1 - B_2||_2^2 + f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_1\right) - f^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) \\ & - \left\langle \frac{\partial}{\partial B} f^\sharp(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, B_1 - B_2 \right\rangle \end{aligned} \tag{A.29}$$

which completes the proof. □

Lemma 2 *Let $\{B_{yw\cdot xz}^\sharp[k]\}_{k\geq 0}$ be the sequence generated by the sequential minimization of equation (A.10) given $\beta_{yx\cdot zw}$ and $B_{yz\cdot xw}$. Then, for optimal solution $B_{yw\cdot xz}^{\sharp *}$, there exists some natural number K for any $\epsilon > 0$ such that*

$$||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k+1]||_2^2 < \epsilon \tag{A.30}$$

for all $k \geq K$.

Proof of Lemma 2: Letting $B_1 = B_{yw\cdot xz}^{\sharp *}$ and $B_2 = B_{yw\cdot xz}^\sharp[k]$ in Lemma 1, from $B_{yw\cdot xz}^\sharp[k+1] = T_{l_f}(B_{yw\cdot xz}^\sharp[k])$ and the nonnegativity of the Bregman distance regarding the convex functions, we have

$$\begin{aligned} & \frac{2}{l_f}(L_1^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_{yw\cdot xz}^{\sharp *}\right) - L_1^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_{yw\cdot xz}^\sharp[k+1]\right)) \\ & \geq ||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k+1]||_2^2 - ||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k]||_2^2 + \frac{2}{l_f} d_f(B_{yw\cdot xz}^{\sharp *}, B_{yw\cdot xz}^\sharp[k]) \\ & \geq ||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k+1]||_2^2 - ||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k]||_2^2. \end{aligned} \tag{A.31}$$

Thus, noting that $\left\{L_1^\sharp\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_{yw\cdot xz}^\sharp[k+1]\right)\right\}_{k\geq 0}$ is a monotonically decreasing sequence with respect to k, we have

$$0 \leq ||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k+1]||_2^2 \leq ||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k]||_2^2, \tag{A.32}$$

i.e., $\{B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k]\}_{k\geq 0}$ is also a monotonically decreasing sequence with respect to k. Noting that the formulation of $B_{yw\cdot xz}^\sharp[k]$ is the paraphrase of equation (A.3) in Proposition 1 given $\beta_{yx\cdot zw}$ and $B_{yw\cdot xz}$, $B_{yw\cdot xz}^\sharp[k]$ converges to $B_{yw\cdot xz}^{\sharp *}$ for $k \to \infty$. In other words, there exists some natural number K for any $\epsilon > 0$ such that

$$||B_{yw\cdot xz}^{\sharp *} - B_{yw\cdot xz}^\sharp[k+1]||_2^2 < \epsilon \tag{A.33}$$

for all $k \geq K$. □

Lemma 3 *Let $\{B^{\sharp}_{yw\cdot xz}[k]\}_{k\geq 0}$ be the sequence generated by the sequential minimization of equation (A.10) given an optimal solution $\beta^*_{yx\cdot zw}$ and $B^*_{yz\cdot xw}$. Then, for optimal solution $B^{\sharp *}_{yw\cdot xz}$,*

$$\begin{aligned} & L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[k+1]\right) - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right) \\ & \leq \frac{\lambda_{\max}(S^{\sharp}_{ww})}{2k}||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[0]||_2^2 \end{aligned} \tag{A.34}$$

holds for all $k \geq 0$.

Proof of Lemma 3: For any $i \geq 0$, letting $B_1 = B^{\sharp *}_{yw\cdot xz}$ and $B_2 = B^{\sharp}_{yw\cdot xz}[i]$ in Lemma 1, since we have

$$\begin{aligned} & \frac{2}{l_f}(L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right) - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[i+1]\right)) \\ & \geq ||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[i+1]||_2^2 - ||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[i]||_2^2 + \frac{2}{l_f} d_f(B^{\sharp *}_{yw\cdot xz}, B^{\sharp}_{yw\cdot xz}[i]) \\ & \geq ||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[i+1]||_2^2 - ||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[i]||_2^2, \end{aligned} \tag{A.35}$$

we obtain

$$\begin{aligned} & \frac{2}{l_f}\sum_{i=0}^{k-1}(L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right) - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[i+1]\right)) \\ & \geq ||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[k]||_2^2 - ||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[0]||_2^2 \geq -||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[0]||_2^2. \end{aligned} \tag{A.36}$$

Here, noting that $\left\{L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[i+1]\right)\right\}_{i\geq 0}$ is a monotonically decreasing sequence with respect to $i \geq 0$, we derive

$$\begin{aligned} & k(L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[k]\right) - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right)) \\ & \leq \sum_{i=0}^{k-1}(L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[i+1]\right) - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right)) \\ & \leq \frac{l_f}{2}||B^{\sharp *}_{yw\cdot xz} - B^{\sharp}_{yw\cdot xz}[0]||_2^2. \end{aligned} \tag{A.37}$$

Finally, noting that $l_f = \lambda_{\max}(S^{\sharp}_{ww})$, we derive Lemma 3. □

From Lemma 3, according to equation (A.16), we provide the optimal solution $\beta^*_{yx\cdot zw}$ and $B^*_{yz\cdot xw}$ given $B^*_{yw\cdot xz}$ as

$$\begin{pmatrix} \beta^*_{yx\cdot zw} \\ B^*_{yz\cdot xw} \end{pmatrix} = \begin{pmatrix} s_{xx} & S_{xz} \\ S_{xz}^T & S_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{x}^T \\ \boldsymbol{z}^T \end{pmatrix} \left(\boldsymbol{y} - \boldsymbol{w}B^*_{yw\cdot xz}\right). \tag{A.38}$$

Then, the following lemma is obtained.

Lemma 4 *Let $\{\beta_{yx\cdot zw}[k]\}_{k\geq 0}$, $\{B_{yz\cdot xw}[k]\}_{k\geq 0}$ and $\{B^{\sharp}_{yw\cdot xz}[k]\}_{k\geq 0}$ be the sequences generated by i-PROGLES and $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{z})$. Then, for optimal solution $\beta^*_{yx\cdot zw}$, $B^*_{yz\cdot xw}$, there exists some natural number K for any $\epsilon \geq 0$ such that*

$$\begin{aligned} & L_1^{\sharp}\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B^{\sharp}_{yw\cdot xz}[k+1]\right) \\ & \quad - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[k+1]\right) \leq \frac{\lambda_{\max}(S_{uu})}{2}\lambda_{\max}(S^{\sharp}_{wu}S^{-2}_{uu}S^{\sharp}_{uw})\epsilon \end{aligned} \tag{A.39}$$

for all $k \geq K$.

Proof of Lemma 4: For all $k \geq 0$, we obtain

$$L_1^{\sharp}\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B^{\sharp}_{yw\cdot xz}[k+1]\right) - L_1^{\sharp}\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp}_{yw\cdot xz}[k+1]\right)$$

$$
\begin{aligned}
&\leq \frac{1}{2}||\boldsymbol{u}((\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1])^T - (\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw})^T)||_2^2 \\
&\leq \frac{1}{2}||\boldsymbol{u}||_{op}^2||(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1])^T - (\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw})^T||_2^2 \\
&\leq \frac{\lambda_{\max}(S_{uu})}{2}||(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1])^T - (\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw})^T||_2^2. \qquad \text{(A.40)}
\end{aligned}
$$

From equations (A.16) and (A.38), we obtain

$$
\begin{aligned}
&||(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1])^T - (\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw})^T||_2^2 \\
&\quad = || \left(\boldsymbol{u}^T\boldsymbol{u}\right)^{-1}\boldsymbol{u}^T\boldsymbol{w}^\sharp \left(B^{\sharp *}_{yw\cdot xz} - B^\sharp_{yw\cdot xz}[k+1]\right) ||_2^2 \\
&\quad \leq || \left(\boldsymbol{u}^T\boldsymbol{u}\right)^{-1}\boldsymbol{u}^T\boldsymbol{w}^\sharp||_{op}^2|| \left(B^{\sharp *}_{yw\cdot xz} - B^\sharp_{yw\cdot xz}[k+1]\right) ||_2^2 \qquad \text{(A.41)}
\end{aligned}
$$

Here, there exists a maximum eigenvalue of $S^\sharp_{wu}S^{-2}_{uu}S^\sharp_{uw}$ because the sum of squares matrix of $\boldsymbol{x}$ and $\boldsymbol{z}$ is invertible. Thus, from Lemma 2, there exists some natural number K for any $\epsilon > 0$ such that

$$
||(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1])^T - (\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw})^T||_2^2 \leq \lambda_{\max}(S^\sharp_{wu}S^{-2}_{uu}S^\sharp_{uw})\epsilon \qquad \text{(A.42)}
$$

for all $k \geq K$. From equations (A.40) and equation (A.42), we obtain Lemma 4. □

Theorem 3 *Let $\{\beta_{yx\cdot zw}[k]\}_{k\geq 0}$, $\{B_{yz\cdot xw}[k]\}_{k\geq 0}$ and $\{B_{yw\cdot xz}[k]\}_{k\geq 0}$ be the sequences of $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$, respectively, generated by i-PROGLES, and let $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{z})$. When $\beta^*_{yx\cdot zw}$, $B^*_{yz\cdot xw}$ and $B^*_{yw\cdot xz}$ minimize equation (19) regarding $\beta_{yx\cdot zw}$, $B_{yz\cdot xw}$ and $B_{yw\cdot xz}$, respectively, there exists a natural number K for any $\epsilon > 0$ such that*

$$
\begin{aligned}
&L_1\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^*_{yw\cdot xz}\right) - L_1\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B_{yw\cdot xz}[k+1]\right) \\
&\quad \leq \frac{\lambda_{\max}(S^\sharp_{ww})}{2k}||B^\sharp_{yw\cdot xz}[0] - B^{\sharp *}_{yw\cdot xz}||_2^2 + \frac{\lambda_{\max}(S_{uu})}{2}\lambda_{\max}(S^\sharp_{wu}S^{-2}_{uu}S^\sharp_{uw})\epsilon. \qquad \text{(A.43)}
\end{aligned}
$$

holds for any $k \geq K$, where $B^\sharp_{yw\cdot xz}[k] = \boldsymbol{\gamma} \odot B_{yw\cdot xz}[k]$ and $B^{\sharp *}_{yw\cdot xz} = \boldsymbol{\gamma} \odot B^*_{yw\cdot xz}$.

Proof of Theorem 3: Noting that

$$
\begin{aligned}
&L_1\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^*_{yw\cdot xz}\right) - L_1\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B_{yw\cdot xz}[k+1]\right) \\
&\quad = L_1^\sharp\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right) - L_1^\sharp\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B^\sharp_{yw\cdot xz}[k+1]\right) \\
&\quad = L_1^\sharp\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^{\sharp *}_{yw\cdot xz}\right) - L_1^\sharp\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^\sharp_{yw\cdot xz}[k+1]\right) \\
&\quad + L_1^\sharp\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^\sharp_{yw\cdot xz}[k+1]\right) - L_1^\sharp\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B^\sharp_{yw\cdot xz}[k+1]\right), \qquad \text{(A.44)}
\end{aligned}
$$

from Lemmas 3 and 4, we have

$$
\begin{aligned}
&L_1\left(\beta^*_{yx\cdot zw}, B^*_{yz\cdot xw}, B^*_{yw\cdot xz}\right) - L_1\left(\beta_{yx\cdot zw}[k+1], B_{yz\cdot xw}[k+1], B_{yw\cdot xz}[k+1]\right) \\
&\quad \leq \frac{\lambda_{\max}(S^\sharp_{ww})}{2k}||B^\sharp_{yw\cdot xz}[0] - B^{\sharp *}_{yw\cdot xz}||_2^2 + \frac{\lambda_{\max}(S_{uu})}{2}\lambda_{\max}(S^\sharp_{wu}S^{-2}_{uu}S^\sharp_{uw})\epsilon. \qquad \text{(A.45)}
\end{aligned}
$$

□.

B Numerical Experiments

In this section, we conduct numerical experiments to compare the performance of LASSO, adaptive LASSO, elastic net, SCAD, MCP, OLS and PAL$_1$MA.

B.1 Loss Functions

For an r-dimensional regression vector $B_{yz\cdot xw}$ and a q-dimensional regression vector $B_{yw\cdot xz}$, let $B_y = (\beta_{yx\cdot zw}, B_{yz\cdot xw}^T, B_{yw\cdot xz}^T)^T = (\beta_1, \beta_2, ..., \beta_{q+r+1})^T$ and $\lambda, \lambda_1, \lambda_2 \geq 0$. First, the loss function of adaptive LASSO [Zou, 2006] is defined as

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}B_{yw\cdot xz}||_2^2 + \lambda||\boldsymbol{\gamma} \odot B_y||_1^1, \tag{B.1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_{q+r+1})^T$ is a weight vector such that

$$\boldsymbol{\gamma} = \left(\frac{1}{|\tilde{\beta}_1|^\xi}, \frac{1}{|\tilde{\beta}_2|^\xi}, \ldots, \frac{1}{|\tilde{\beta}_{q+r+1}|^\xi}\right)^T \tag{B.2}$$

for the non-invertible sum of squares matrix of the explanatory variables with tuning parameter $\xi \geq 0$ and

$$\boldsymbol{\gamma} = \left(\frac{1}{|\hat{\beta}_1|^\xi}, \frac{1}{|\hat{\beta}_2|^\xi}, \ldots, \frac{1}{|\hat{\beta}_{q+r+1}|^\xi}\right)^T \tag{B.3}$$

for the invertible sum of squares matrix of the explanatory variables with tuning parameter $\xi \geq 0$. In particular, equation (B.1) is the loss function of the standard LASSO [Tibshirani, 1996] when $\xi = 0$ and the loss function of OLS regression when $\lambda = 0$.

Second, for $0 \leq \phi \leq 1$, the loss function of the elastic net [Zou and Hastie, 2005] is given by

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}B_{yw\cdot xz}||_2^2 + \lambda\left((1-\phi)||B_y||_2^2 + \phi||B_y||_1^1\right). \tag{B.4}$$

Third, consider the following type of loss function:

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx\cdot zw} - \boldsymbol{z}B_{yz\cdot xw} - \boldsymbol{w}B_{yw\cdot xz}||_2^2 + \sum_{j=1}^{q+r+1} p_{\lambda,\xi}(\beta_j). \tag{B.5}$$

Then, for $\xi > 1$, the loss function of MCP [Zhang, 2010] is given by defining the function $p_{\lambda,\xi}$ in equation (B.5) as follows:

$$p_{\lambda,\xi}(x) = \begin{cases} \lambda|x| - \dfrac{|x|^2}{2\xi} & : \ |x| \leq \xi\lambda \\ \dfrac{1}{2}\xi\lambda^2 & : \ |x| > \xi\lambda \end{cases}. \tag{B.6}$$

In addition, for $\xi > 2$, the loss function of SCAD [Fan and Li, 2001] is given by defining the function $p_{\lambda,\xi}$ in equation (B.5) as follows:

$$p_{\lambda,\xi}(x) = \begin{cases} \lambda|x| & : \ |x| \leq \lambda \\ \dfrac{\xi\lambda|x| - 0.5(|x|^2 + \lambda^2)}{\xi - 1} & : \ \lambda < |x| < \xi\lambda \\ \dfrac{\lambda^2(\xi^2 - 1)}{2(\xi - 1)} & : \ |x| > \xi\lambda \end{cases}. \tag{B.7}$$

In this paper, we use the "glmnet" package (version 4.0.2) [Friedman, Hastie and Tibshirani, 2010] to perform LASSO, adaptive LASSO and elastic net, and the "ncvreg" package [Breheny and Huang, 2011] to conduct SCAD and MCP. The "glmnet" and "ncvreg" packages are available from `https://glmnet.stanford.edu/` and `http://pbreheny.github.io/ncvreg/`, respectively.

Table A. Path coefficients

(a) Z satisfies the back-door criterion

Fig. A (a)	α_{yz}	α_{xz}	α_{yx}	A_{yw}
(a_1)	0.5	0.5	$U([-3,3])$	$U([-3,3])$
(a_2)	0.5	2.5	$U([-3,3])$	$U([-3,3])$
(a_3)	2.5	0.5	$U([-3,3])$	$U([-3,3])$
(a_4)	2.5	2.5	$U([-3,3])$	$U([-3,3])$

(b) $\{Z_1, Z_2\}$ satisfies the back-door criterion

Fig. A (b)	α_{yz_1}	α_{xz_1}	α_{yz_2}	α_{xz_2}	α_{yx}	A_{yw}
(b_1)	0.5	0.5	0.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_2)	0.5	0.5	0.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_3)	0.5	0.5	2.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_4)	0.5	0.5	2.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_5)	0.5	2.5	0.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_6)	0.5	2.5	0.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_7)	0.5	2.5	2.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_8)	0.5	2.5	2.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_9)	2.5	0.5	0.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_{10})	2.5	0.5	0.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_{11})	2.5	0.5	2.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_{12})	2.5	0.5	2.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_{13})	2.5	2.5	0.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_{14})	2.5	2.5	0.5	2.5	$U([-3,3])$	$U([-3,3])$
(b_{15})	2.5	2.5	2.5	0.5	$U([-3,3])$	$U([-3,3])$
(b_{16})	2.5	2.5	2.5	2.5	$U([-3,3])$	$U([-3,3])$

$U([-3,3])$: path coefficients that have been determined by the random number from the uniform distribution on the interval $[-3,3]$.

B.2 Parameter settings

For simplicity, letting X and Y be the treatment variable and the response variable, respectively, consider the linear SCMs with 42 explanatory variables for Y in the form of

$$\left.\begin{array}{l} Y = \alpha_{yx}X + \alpha_{yz}Z + A_{yw}\boldsymbol{W} + \epsilon_y \\ X = \alpha_{xz}Z + \epsilon_x \end{array}\right\} \tag{B.8}$$

for Fig. A (a) ($\boldsymbol{W}$ includes 40 variables), and

$$\left.\begin{array}{l} Y = \alpha_{yx}X + \alpha_{yz_1}Z_1 + \alpha_{yz_2}Z_2 + A_{yw}\boldsymbol{W} + \epsilon_y \\ X = \alpha_{xz_1}Z_1 + \alpha_{xz_2}Z_2 + \epsilon_x \end{array}\right\} \tag{B.9}$$

for Fig. A (b) ($\boldsymbol{W}$ includes 39 variables). Fig. A (a) shows that (i) Z satisfies the back-door criterion relative to (X, Y), and (ii) the path coefficients of $\boldsymbol{W}$ on Y are regularized, but Z is not. Fig. A (b) shows that (i) $\{Z_1, Z_2\}$ satisfies the back-door criterion relative to (X, Y), and (ii) the path coefficients of $\{Z_2\} \cup \boldsymbol{W}$ on Y are regularized, but Z_1 is not. Theorem 1 holds in Fig. A (a), so $\boldsymbol{W}$ is collapsible. However, $\{Z_2\} \cup \boldsymbol{W}$ is not in Fig. A (b); thus, the estimated total effect may be biased.

To construct the population variance-covariance matrix, first, we assigned one of 0.5 and 2.5 to α_{yz} and α_{xz}, depending on Fig. A (a), and α_{yz_1}, α_{yz_2}, α_{xz_1} and α_{xz_2}, depending on Fig. A (b). Multicollinearity may occur between X and the covariates satisfying the back-door criterion when we assign 2.5 to the path coefficients on X but may not occur when we assign 0.5 to the path coefficients on X. Other path coefficients were randomly and independently generated according to the uniform distribution on the interval $[-3, 3]$. These parameter settings are shown in Table A. In addition, the population variance-covariance matrices of the covariates $\{Z\} \cup \boldsymbol{W}$ in Fig. A (a) and $\{Z_1, Z_2\} \cup \boldsymbol{W}$ in Fig. A (b) are also randomly generated using the

"randcorr" package (available from `https://www.rdocumentation.org/packages/randcorr/versions/1.0/topics/randcorr-package`) according to Pourahmadi and Wang [2015]. Furthermore, we assume that (i) the random disturbances ϵ_x and ϵ_y independently follow normal distributions with mean zero and variance one, and (ii) the random disturbances are also independent of their non-descendants.

Regarding tuning the regularization parameter λ, the "glmnet" package was utilized for LASSO, adaptive LASSO and elastic net. Here, the search ranges were set to $\xi \in \{0.1, 0.2, 0.3, ..., 2.9, 3.0\}$ for the tuning parameter ξ of adaptive LASSO and $\phi \in \{0.01, 0.02, 0.03, ..., 0.98, 0.99\}$ for the mixing parameter ϕ of elastic net. For MCP and SCAD, the "ncvreg" package was applied to determine the regularized parameter λ. Here, the search ranges were set to $\xi \in \{1.5, 2.0, 2.5, ..., 19.5, 20.0\}$ for the tuning parameter ξ of MCP and $\xi \in \{2.5, ..., 19.5, 20.0\}$ for the tuning parameter ξ of SCAD. In contrast, in PAL$_1$MA, we conducted all possible selection based on three fold cross-validation to determine the regularization parameter λ_1 from the search range $\lambda_1 \in \{0.01, 0.011, ..., 0.049, 0.050\}$ and the tuning parameter ξ_1 from the search range $\xi_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Similarly, bias correction was also conducted through all possible selections to determine the regularization parameter λ_2 from the search range $\lambda_2 \in \{0.00, 0.01, 0.02, 0.03\}$ and the tuning parameter ξ_2 from the search range $\xi_2 \in \{0.00, 0.01, 0.02, 0.03\}$. Note that such parameter settings of PAL$_1$MA in this paper are somewhat empirical; i.e., they may not be optimally determined compared to other regularized regression analyses. The development of optimal parameter tuning for PAL$_1$MA is saved for future work. The parameter tuning results are shown in Table B.

B.3 Analysis

For 5000 replications, we generated 30 random samples of 42 variables from a multivariate normal distribution with a zero mean vector and the population variance-covariance matrix generated by the above procedure. Tables C and C' show the numerical results by LASSO, adaptive LASSO, elastic net, SCAD, MCP, PAL$_1$MA and OLS based on Table B. Here, for OLS, we select a set of covariates based on prior causal knowledge; i.e., Z and $\{Z_1, Z_2\}$ are selected in Figs. A (a) and (b), respectively.

From Figs. B and B' and Tables C and C', we make the following observations:

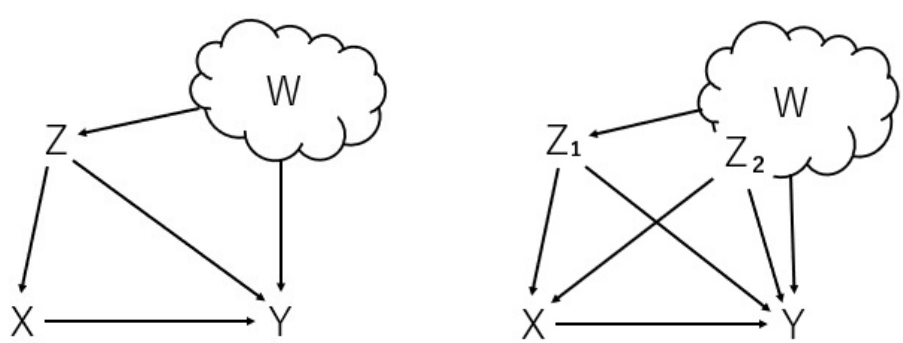
1. When the total effect is close to zero, the coincidence rates between the signs of the estimated total effects and the true total effects are low for each regression analysis, but those of PAL$_1$MA are still higher than those of the other regression analyses.

2. When the true total effect is far from zero, the coincidence rates are high for each regression analysis.

3. When there is high spurious correlation, the coincidence rates for PAL$_1$MA are lower than those for elastic net, but the differences are not significant. This situation may occur because the variances of the estimated total effects are larger than those of the other regression analyses.

4. Except for Case (b_8), PAL$_1$MA provides fewer bias estimates than the other regularized regression analyses. In Case (b_8), PAL$_1$MA provides more biased estimates than SCAD and MCP but higher coincidence rates than these regularized regression analyses.

.



(a) (b)

Fig. A. Causal diagram

5. The variance of the estimated total effects from $\mathrm{PAL}_1\mathrm{MA}$ are larger than those from the other regularized regression analyses but smaller than those from OLS regression for most cases.

6. From Figs. B and B', the interquartile ranges of $\mathrm{PAL}_1\mathrm{MA}$ include the true value of the total effects in all cases, but the other regularized regression analyses do not include this value in most cases.

Overall, the coincidence rates between the signs of the estimated total effects and the true total effect from $\mathrm{PAL}_1\mathrm{MA}$ seem equal to or higher than those from the other regression analyses. In addition, $\mathrm{PAL}_1\mathrm{MA}$ can provide less biased estimators than the other regularized regression analyses in most cases. In some cases of Figs. A (b), $\mathrm{PAL}_1\mathrm{MA}$ does not select a set $\{Z_1, Z_2\}$ of covariates satisfies the back-door criterion, and such a missing covariate (Z_2) provides biased estimates of the total effects. However, since the regression coefficient of Z_2 takes a small value in such cases, $\mathrm{PAL}_1\mathrm{MA}$ seems not reverse the direction of the regression coefficient in most case. Here, as seen from the following section, note that such a drawback can be eliminated by selecting smaller values of the regularization parameters based on the whole set of covariates. These results imply that the estimation of the total effect by $\mathrm{PAL}_1\mathrm{MA}$ does not lead to the misleading qualitative interpretation compared to the standard regularized regression analysis.

Table B. Parameter settings

(a) Z satisfies the back-door criterion

Fig. A (a)	LASSO	adaptive LASSO		Elastic net		MCP		SCAD		PAL$_1$MA				Total effect
	λ	ξ	λ	ϕ	λ	ξ	λ	ξ	λ	ξ_1	λ_1	ξ_2	λ_2	τ_{yx}
(a_1)	0.0080	1.8000	1.1510	0.2500	0.0250	11.5000	0.0470	16.0000	0.0380	0.3000	0.0100	0.0000	0.0000	-0.1520
(a_2)	0.0240	0.9000	1.9240	0.6600	0.0640	5.0000	0.0340	16.5000	0.0720	0.2000	0.0100	0.0001	0.0001	0.1290
(a_3)	0.0070	0.7000	0.0460	0.0400	0.1680	15.0000	0.0350	12.0000	0.0900	0.1000	0.0100	0.0013	0.0014	-0.0200
(a_4)	0.0150	0.8000	0.0440	0.1100	0.0500	19.0000	0.0320	12.5000	0.0810	0.3000	0.0100	0.0000	0.0000	0.3770

(b) $\{Z_1, Z_2\}$ satisfies the back-door criterion

Fig. A (b)	LASSO	adaptive LASSO		Elastic net		MCP		SCAD		PAL$_1$MA				Total effect
	λ	ξ	λ	ϕ	λ	ξ	λ	ξ	λ	ξ_1	λ_1	ξ_2	λ_2	τ_{yx}
(b_1)	0.0110	1.2000	0.1340	0.2400	0.0290	20.0000	0.0370	16.5000	0.0440	0.2000	0.0100	0.0005	0.0005	-0.1520
(b_2)	0.0560	1.2000	0.4380	0.9700	0.0550	20.0000	0.0880	13.0000	0.0960	0.1000	0.0100	0.0005	0.0006	-0.6700
(b_3)	0.0070	1.3000	0.8200	0.7500	0.0360	13.0000	0.0400	5.0000	0.0450	0.1000	0.0100	0.0024	0.0026	-0.1840
(b_4)	0.0850	1.1000	0.7270	0.2400	0.0320	14.0000	0.1510	16.5000	0.0650	0.1000	0.0100	0.0011	0.0012	-0.7780
(b_5)	0.0630	0.8000	0.0570	0.7900	0.0100	16.5000	0.0730	13.0000	0.0750	0.4000	0.0100	0.0000	0.0000	0.3520
(b_6)	0.0090	0.3000	0.0520	0.7500	0.0320	12.5000	0.1170	15.5000	0.0550	0.3000	0.0100	0.0000	0.0000	-0.2630
(b_7)	0.0060	1.0000	0.0840	0.2800	0.0230	17.0000	0.0320	20.0000	0.1410	0.5000	0.0100	0.0000	0.0000	0.2980
(b_8)	0.1060	1.2000	0.3590	0.8400	0.0150	10.0000	0.0620	9.5000	0.0720	0.1000	0.0100	0.0005	0.0006	-0.8240
(b_9)	0.0360	0.8000	0.0910	0.7900	0.0250	16.5000	0.0770	3.5000	0.1510	0.4000	0.0100	0.0000	0.0000	0.1580
(b_{10})	0.0070	1.2000	0.1280	0.5500	0.0240	13.5000	0.1240	15.0000	0.0760	0.3000	0.0100	0.0000	0.0000	-0.1850
(b_{11})	0.0750	1.5000	0.6240	0.5700	0.0490	4.0000	0.0960	17.0000	0.1260	0.3000	0.0100	0.0000	0.0000	-0.4260
(b_{12})	0.0330	1.5000	1.5240	0.4300	0.0890	11.0000	0.1020	16.0000	0.1150	0.1000	0.0300	0.0000	0.0000	-0.0840
(b_{13})	0.0980	2.3000	2.0180	0.7100	0.0570	5.0000	0.1580	3.0000	0.1320	0.2000	0.0100	0.0001	0.0001	-0.8140
(b_{14})	0.0630	1.2000	0.2350	0.5500	0.0250	19.0000	0.0610	15.0000	0.0710	0.3000	0.0100	0.0000	0.0000	-0.2580
(b_{15})	0.0330	1.4000	0.6870	0.9900	0.0500	12.0000	0.0840	10.5000	0.0540	0.3000	0.0100	0.0000	0.0000	-0.1950
(b_{16})	0.0860	1.2000	0.8960	0.8400	0.0180	6.0000	0.0490	17.0000	0.0420	0.1000	0.0100	0.0008	0.0009	-0.9100

λ, λ_1, λ_2: regularization parameters; ξ, ξ_1, ξ_2: tuning parameters; ϕ: mixing parameter; τ_{yx}: total effect of X on Y.

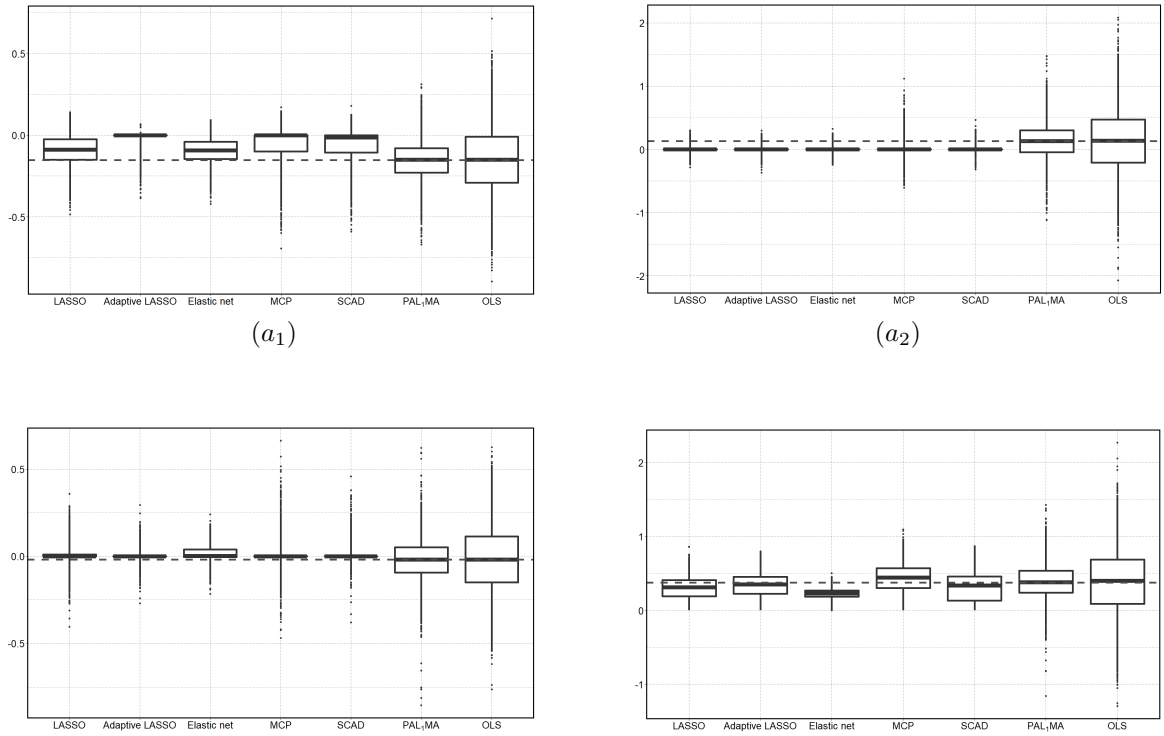
Table C. Results based on cross-validation.

(a) Z satisfies the back-door criterion

	(a_1)					(a_2)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.0971	0.0550	0.0098	0.0824	0.8282	0.0079	0.1216	0.0163	0.0385	0.1482
adaptive LASSO	-0.0185	0.1336	0.0200	0.0463	0.2332	-0.0024	0.1319	0.0187	0.0366	0.0708
Elastic net	-0.0983	0.0538	0.0081	0.0720	0.8904	0.0054	0.1241	0.0163	0.0304	0.1174
MCP	-0.0619	0.0902	0.0176	0.0971	0.4794	0.0275	0.1020	0.0278	0.1320	0.1642
SCAD	-0.0635	0.0886	0.0163	0.0919	0.5440	0.0034	0.1261	0.0168	0.0309	0.0626
$\mathrm{PAL_1MA}$	-0.1552	0.0031	0.0136	0.1165	0.9190	0.1333	0.0038	0.0787	0.2804	0.6908
OLS	-0.1480	0.0041	0.0426	0.2063	0.7630	0.1293	0.0002	0.2708	0.5204	0.6048

	(a_3)					(a_4)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	0.0036	0.0232	0.0038	0.0572	0.2456	0.2959	0.0808	0.0309	0.1562	0.9390
adaptive LASSO	0.0022	0.0218	0.0012	0.0277	0.0720	0.3328	0.0439	0.0285	0.1632	0.9554
Elastic net	0.0137	0.0333	0.0032	0.0459	0.2820	0.2297	0.1469	0.0254	0.0617	0.9998
MCP	0.0052	0.0248	0.0044	0.0616	0.1130	0.4180	0.0413	0.0474	0.2137	0.8942
SCAD	0.0083	0.0279	0.0023	0.0385	0.0214	0.3059	0.0708	0.0469	0.2047	0.8070
$\mathrm{PAL_1MA}$	-0.0210	0.0013	0.0140	0.1184	0.5688	0.3897	0.0131	0.0572	0.2389	0.9564
OLS	-0.0189	0.0007	0.0372	0.1929	0.5398	0.3840	0.0073	0.1974	0.4442	0.8100

mean: sample mean; bias: bias between the true value and the sample mean; mse: mean squared error: sd: standard deviation; sign: coincidence rate between the signs of the true value and the estimates.



(a_1) (a_2) (a_3) (a_4)

(a) Z satisfies the back-door criterion.

Fig. B. Boxplots of the estimated total effects based on 5000 replications from the numerical experiments. The dashed lines show the true total effects.

Table C'. Results based on cross-validation.

(b) $\{Z_1, Z_2\}$ satisfies the back-door criterion

	(b_1)					(b_2)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.0756	0.0761	0.0129	0.0844	0.6876	-0.5308	0.1390	0.0451	0.1606	0.9920
adaptive LASSO	-0.0001	0.1517	0.0230	0.0021	0.0032	-0.6033	0.0664	0.0355	0.1763	0.9948
Elastic net	-0.0816	0.0702	0.0103	0.0732	0.8074	-0.5291	0.1407	0.0448	0.1583	0.9936
MCP	-0.0522	0.0995	0.0178	0.0888	0.4406	-0.5979	0.0719	0.0427	0.1936	0.9458
SCAD	-0.0465	0.1052	0.0185	0.0861	0.4128	-0.6378	0.0319	0.0328	0.1783	0.9602
$\mathrm{PAL_1MA}$	-0.1485	0.0032	0.0192	0.1386	0.8712	-0.6984	0.0287	0.0273	0.1628	0.9990
OLS	-0.1528	0.0010	0.0542	0.2329	0.7470	-0.6697	0.0001	0.1126	0.3356	0.9718

	(b_3)					(b_4)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.0957	0.0879	0.0166	0.0941	0.7560	-0.4448	0.3337	0.1313	0.1414	0.9904
adaptive LASSO	0.0000	0.1835	0.0337	0.0000	0.0000	-0.5685	0.2099	0.0672	0.1521	0.9964
Elastic net	-0.0560	0.1276	0.0222	0.0772	0.5674	-0.4267	0.3517	0.1349	0.1055	1.0000
MCP	-0.0568	0.1267	0.0277	0.1081	0.4088	-0.4490	0.3294	0.1342	0.1603	0.9626
SCAD	-0.0766	0.1069	0.0325	0.1451	0.3810	-0.5625	0.2159	0.0759	0.1711	0.9824
$\mathrm{PAL_1MA}$	-0.1686	0.0149	0.0221	0.1480	0.8948	-0.6766	0.1018	0.0601	0.2231	0.9976
OLS	-0.1860	0.0025	0.0522	0.2284	0.7936	-0.7760	0.0025	0.1495	0.3867	0.9718

	(b_5)					(b_6)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	0.1219	0.2297	0.0646	0.1087	0.7640	-0.1216	0.1412	0.0311	0.1058	0.7900
adaptive LASSO	0.2227	0.1290	0.0312	0.1205	0.9406	-0.1227	0.1401	0.0306	0.1049	0.7932
Elastic net	0.2277	0.1239	0.0300	0.1211	0.9482	-0.1020	0.1608	0.0346	0.0937	0.7646
MCP	0.1233	0.2283	0.0689	0.1294	0.6744	-0.0515	0.2114	0.0522	0.0867	0.3956
SCAD	0.1167	0.2350	0.0726	0.1317	0.6618	-0.0793	0.1835	0.0455	0.1085	0.5452
$\mathrm{PAL_1MA}$	0.3811	0.0295	0.0552	0.2331	0.9586	-0.2976	0.0348	0.0493	0.2193	0.9320
OLS	0.3498	0.0018	0.2565	0.5064	0.7704	-0.2560	0.0068	0.5220	0.7225	0.6364

	(b_7)					(b_8)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	0.1124	0.2296	0.0644	0.1078	0.7414	-0.5468	0.2768	0.0887	0.1099	0.9998
adaptive LASSO	0.1189	0.2232	0.0615	0.1082	0.7696	-0.6804	0.1432	0.0334	0.1137	1.0000
Elastic net	0.1487	0.1934	0.0463	0.0942	0.9154	-0.6009	0.2227	0.0656	0.1263	1.0000
MCP	0.1530	0.1891	0.0653	0.1718	0.6476	-0.7764	0.0472	0.0196	0.1319	0.9976
SCAD	0.0993	0.2428	0.0773	0.1353	0.5856	-0.7703	0.0533	0.0196	0.1296	0.9974
$\mathrm{PAL_1MA}$	0.4603	0.1183	0.0829	0.2625	0.9686	-0.7105	0.1131	0.0577	0.2120	0.9994
OLS	0.3408	0.0012	0.2443	0.4943	0.7704	-0.8224	0.0012	0.1699	0.4122	0.9718

mean: sample mean; bias: bias between the true value and the sample mean; mse: mean squared error: sd: standard deviation; sign: coincidence rate between the signs of the true value and the estimates.

Table C'. Results based on cross-validation.

(b) $\{Z_1, Z_2\}$ satisfies the back-door criterion

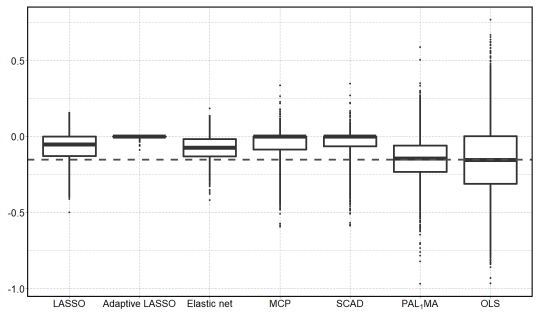
	(b_9)					(b_{10})				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	0.1000	0.0583	0.0115	0.0898	0.7886	-0.0626	0.1219	0.0223	0.0863	0.5282
adaptive LASSO	0.0762	0.0822	0.0131	0.0797	0.7044	-0.0445	0.1400	0.0254	0.0758	0.4102
Elastic net	0.1136	0.0448	0.0098	0.0881	0.8576	-0.0493	0.1352	0.0232	0.0703	0.5136
MCP	0.0776	0.0807	0.0156	0.0954	0.6076	-0.0124	0.1721	0.0316	0.0446	0.1178
SCAD	0.0442	0.1142	0.0229	0.0994	0.3606	-0.0201	0.1644	0.0302	0.0568	0.1910
PAL_1MA	0.1706	0.0122	0.0106	0.1024	0.9592	-0.1923	0.0078	0.0225	0.1499	0.9208
OLS	0.1573	0.0010	0.0496	0.2227	0.7704	-0.1805	0.0040	0.2569	0.5068	0.6364

	(b_{11})					(b_{12})				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.1674	0.2586	0.0837	0.1296	0.8384	0.0063	0.0901	0.0107	0.0504	0.1188
adaptive LASSO	-0.3059	0.1201	0.0365	0.1486	0.9726	0.0000	0.0838	0.0070	0.0000	0.0000
Elastic net	-0.2310	0.1949	0.0546	0.1287	0.9476	0.0100	0.0938	0.0109	0.0461	0.1050
MCP	-0.2607	0.1653	0.0827	0.2353	0.7000	0.0098	0.0936	0.0109	0.0463	0.0162
SCAD	-0.1207	0.3053	0.1089	0.1251	0.7076	0.0098	0.0937	0.0104	0.0407	0.0140
PAL_1MA	-0.4059	0.0201	0.0294	0.1703	0.9926	-0.0396	0.0443	0.0223	0.1427	0.6152
OLS	-0.4230	0.0030	0.0424	0.2060	0.9718	-0.0821	0.0017	0.0509	0.2256	0.6364

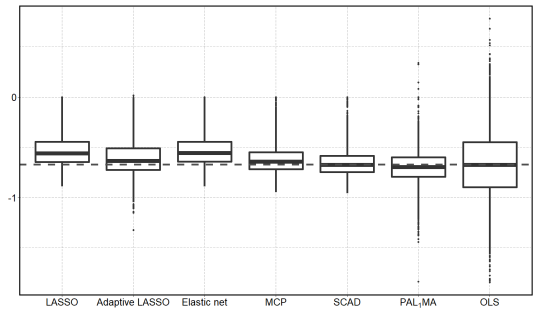
	(b_{13})					(b_{14})				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.3901	0.4240	0.1979	0.1347	0.9904	-0.0045	0.2535	0.0648	0.0235	0.0820
adaptive LASSO	-0.6497	0.1643	0.0631	0.1899	0.9996	-0.0004	0.2577	0.0664	0.0061	0.0072
Elastic net	-0.4351	0.3789	0.1597	0.1269	0.9980	-0.0154	0.2426	0.0606	0.0415	0.2374
MCP	-0.4831	0.3310	0.1517	0.2054	0.9704	-0.0047	0.2534	0.0650	0.0276	0.0620
SCAD	-0.5735	0.2405	0.1021	0.2103	0.9662	-0.0039	0.2542	0.0652	0.0250	0.0588
PAL_1MA	-0.8383	0.0243	0.1000	0.3153	0.9968	-0.2747	0.0167	0.0478	0.2180	0.9206
OLS	-0.8114	0.0026	0.1640	0.4050	0.9718	-0.2532	0.0049	0.5067	0.7118	0.6364

	(b_{15})					(b_{16})				
LASSO	0.0193	0.2141	0.0486	0.0524	0.0164	-0.5053	0.4043	0.1776	0.1191	0.9992
adaptive LASSO	0.0000	0.1948	0.0380	0.0000	0.0000	-0.6185	0.2910	0.0992	0.1205	0.9998
Elastic net	0.0190	0.2138	0.0484	0.0513	0.0104	-0.5709	0.3387	0.1305	0.1259	1.0000
MCP	0.0220	0.2168	0.0513	0.0651	0.0030	-0.7268	0.1827	0.0632	0.1725	0.9976
SCAD	0.0313	0.2261	0.0591	0.0895	0.0076	-0.7094	0.2001	0.0673	0.1650	0.9966
PAL_1MA	-0.1051	0.0898	0.1163	0.3291	0.6364	-0.7830	0.1266	0.0778	0.2485	0.9990
OLS	-0.1905	0.0043	0.2857	0.5345	0.6364	-0.9069	0.0027	0.2047	0.4525	0.9718

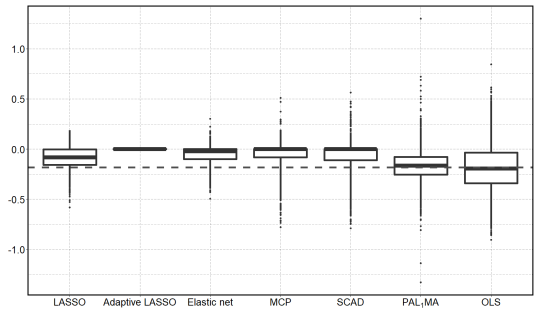
mean: sample mean; bias: bias between the true value and the sample mean; mse: mean squared error: sd: standard deviation; sign: coincidence rate between the signs of the true value and the estimates.



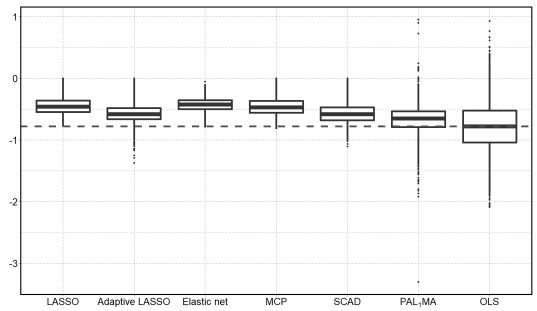
(b_1)



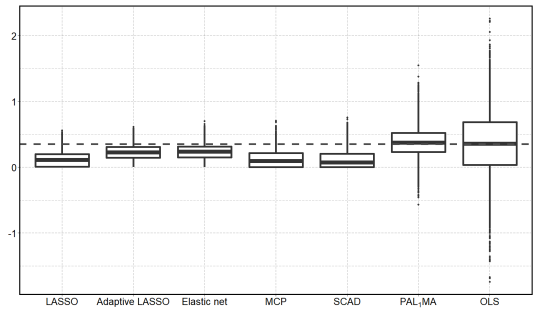
(b_2)



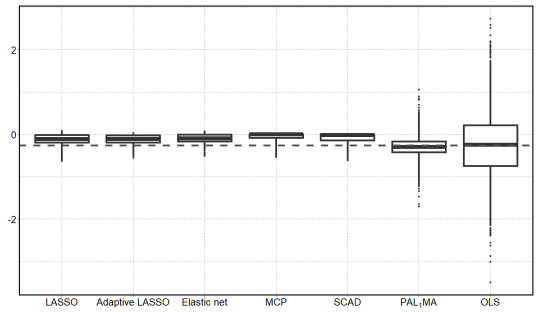
(b_3)



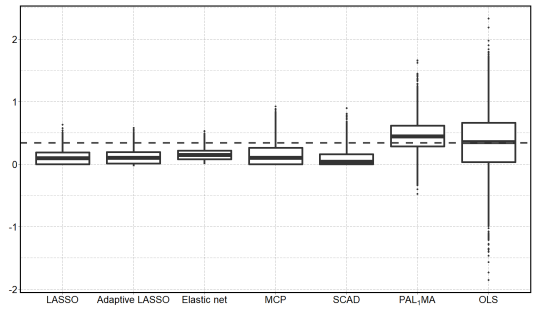
(b_4)



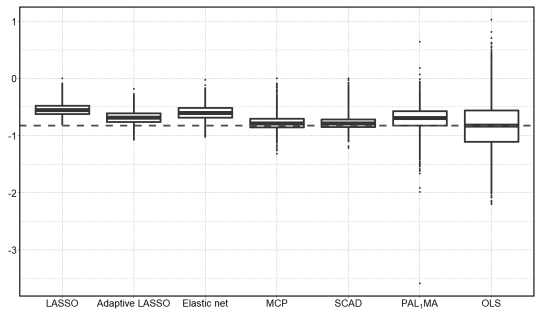
(b_5)



(b_6)



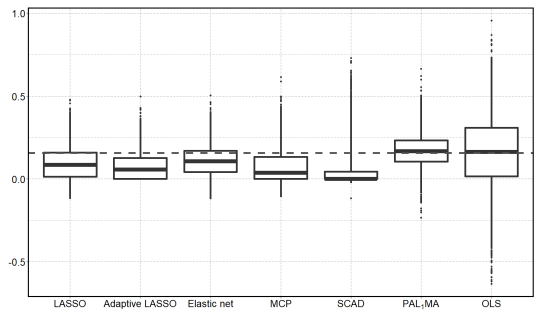
(b_7)



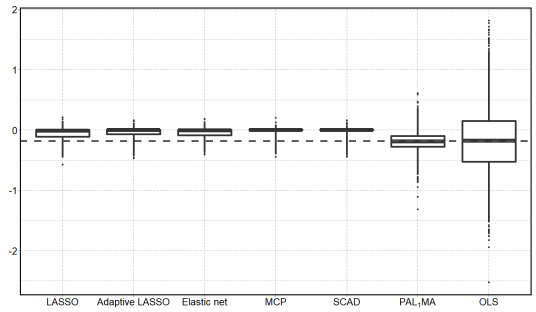
(b_8)

(b) $\{Z_1, Z_2\}$ satisfies the back-door criterion.

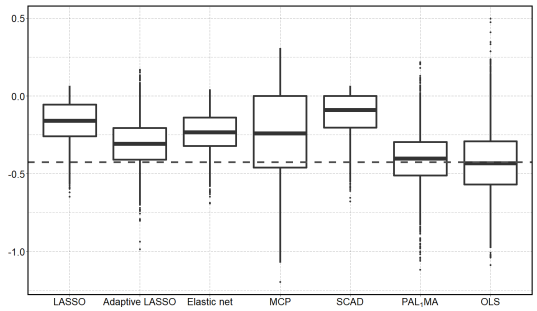
Fig. B'. Boxplots of the estimated total effects based on 5000 replications from the numerical experiments. The dashed lines show the true total effects.



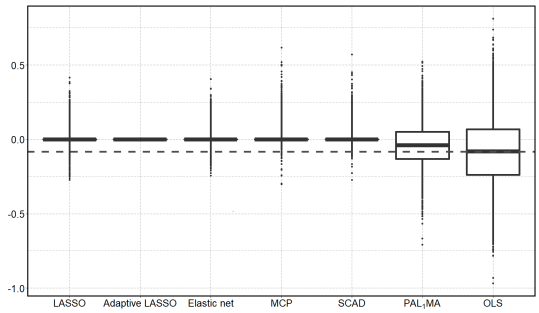
(b_9)



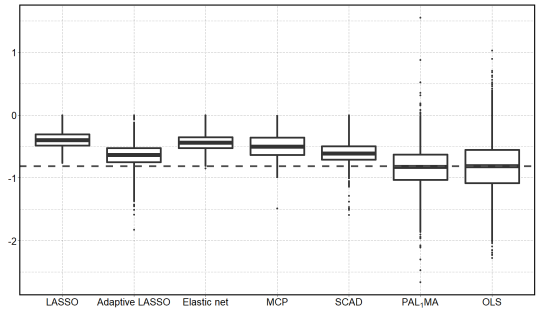
(b_{10})



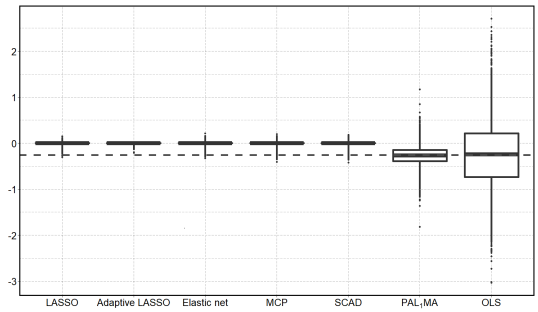
(b_{11})



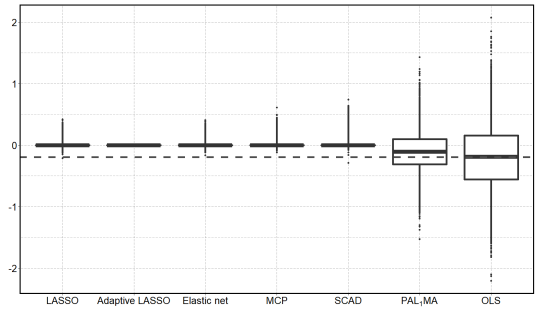
(b_{12})



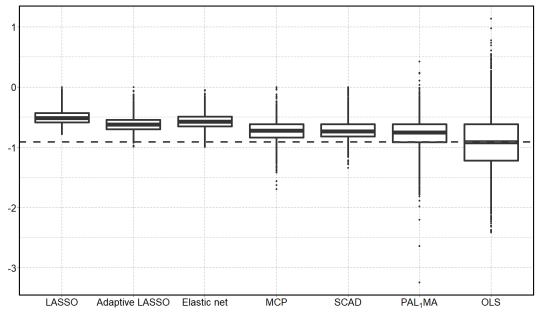
(b_{13})



(b_{14})



(b_{15})



(b_{16})

(b) $\{Z_1, Z_2\}$ satisfies the back-door criterion.

Fig. B'. Boxplots of the estimated total effects based on 5000 replications from the numerical experiments. The dashed lines show the true total effects.

C Case Study

C.1 Background

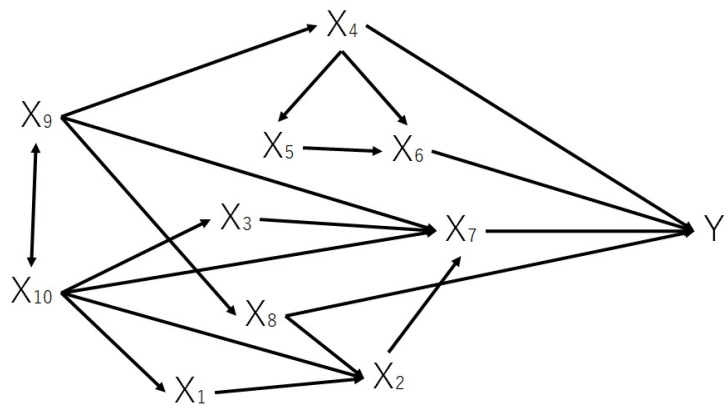


Fig. C. Causal diagram [Kuroki, 2012]

In this section, we apply LASSO, adaptive LASSO, elastic net, SCAD, MCP, PAL$_1$MA and OLS to a case study of setting up coating conditions for car bodies, reported by Okuno et al. [1986] and reanalyzed by Kuroki [2012].

According to Okuno et al. [1986], car bodies are coated to increase both the rust protection quality and the visual appearance. A certain coating thickness must be ensured in the coating process. At the time of the study, this process was conducted by operators who sprayed the car bodies with paint, which depended on the operators' skills and could cause low transfer efficiency. Okuno et al. [1986] collected nonexperimental data on the coating process to examine the process conditions and to increase the transfer efficiency. The sample size is 38, and the dataset is available from Okuno et al. [1986]. In addition, the observed variables of interest are as follows:

Process condition

The dilution ratio (X_1), degree of viscosity (X_2), gun speed (X_3), spray distance (X_4), air pressure (X_5), pattern width (X_6), fluid output (X_7), temperature of the paint (X_8), temperature (X_9), and degree of moisture (X_{10})

Response

The transfer efficiency (Y).

Table D shows the randomly selected data from the whole dataset given by Okuno et al. [1986]. Here, note that our discussion is based on Table D to consider a situation where OLS and the all-variable selection procedure cannot be applied.

Table D. Randomly selected data from the paper by Okuno et al [1986].

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
1	33	28.3	6.7	40.0	3.0	5.0	208.0	20.0	19.0	30.0	19.3
2	16.7	35.0	5.0	40.0	5.0	5.5	108.0	25.0	10.5	39.0	7.3
3	16.7	35.0	8.3	30.0	2.1	3.0	112.0	25.0	20.0	25.0	35.2
4	33	25.0	8.3	40.0	4.0	4.1	240.0	34.0	22.5	25.0	18.4
5	44	29.5	6.5	30.0	2.1	5.0	120.0	6.7	7.0	30.0	21.7
6	16.7	35.0	4.9	40.0	5.0	3.9	168.0	25.0	20.0	25.0	28.7
7	44	29.5	8.3	40.0	2.1	2.2	200.0	7.0	7.0	30.0	37.8
8	44	25.8	6.7	40.0	4.1	5.0	132.0	22.0	8.2	46.0	13.4
9	33	25.5	6.5	40.0	4.0	4.0	276.0	20.0	22.5	25.0	17.8

Table E. Results based on cross-validation.

Method	non-regulaized variables	estimate	sd	selected variables	parameters λ	ξ	ϕ
LASSO	–	0.0470	0.0914	X_2,X_5,X_6	0.1640	–	–
adaptive LASSO	–	0.0759	0.0862	X_2,X_5,X_6,X_{10}	0.2160	0.5000	–
Elastic net	–	0.1395	0.0963	X_2,X_5,X_6,X_8,X_{10}	0.1350	–	0.5500
MCP	–	0.0000	0.0621	X_6	0.3020	4.0000	–
SCAD	–	0.1357	0.1124	X_2,X_4,X_5,X_6,X_{10}	0.0820	17.5000	–
PAL$_1$MA	X_8,X_{10}	0.2221	0.1111	X_2,X_5,X_6,X_8,X_{10}	0.00005	0.1000	–
PAL$_1$MA	X_8	0.2242	0.0950	X_2,X_5,X_6,X_8,X_{10}	0.00003	0.5000	–
PAL$_1$MA	X_{10}	0.2251	0.0854	X_2,X_5,X_6,X_{10}	0.00003	0.5000	–
PAL$_1$MA	–	0.2297	0.0795	X_2,X_5,X_6,X_{10}	0.00002	0.5000	–
OLS	–	0.2455	0.1314	X_2,X_8,X_{10}	–	–	–

estimate: estimates of the total effect with $n = 9$; sd: standard deviation based on leave-one-out method; selected variables: selected explanatory variables by variable selection; parameter: regularized, tuning and mixing parameters. Here, λ_2 and ξ_2 of PAL$_1$MA were selected as zero by leave-one-out method.

According to Okuno et al. [1986], there is some difference among these variables in terms of the controllability level: X_1, X_2, ..., X_6 can be controlled; X_7 and X_8 result from other factors and are difficult to control; and X_9 and X_{10} are environmental conditions that cannot be controlled. In addition, Kuroki [2012] assumed that the cause-effect relationships in the coating process are as shown in Fig. C. From Fig. C, $\{X_8, X_{10}\}$ satisfies the back-door criterion relative to (X_2, Y). For details on this case study, refer to Okuno et al. [1986] and Kuroki [2012].

C.2 Analysis

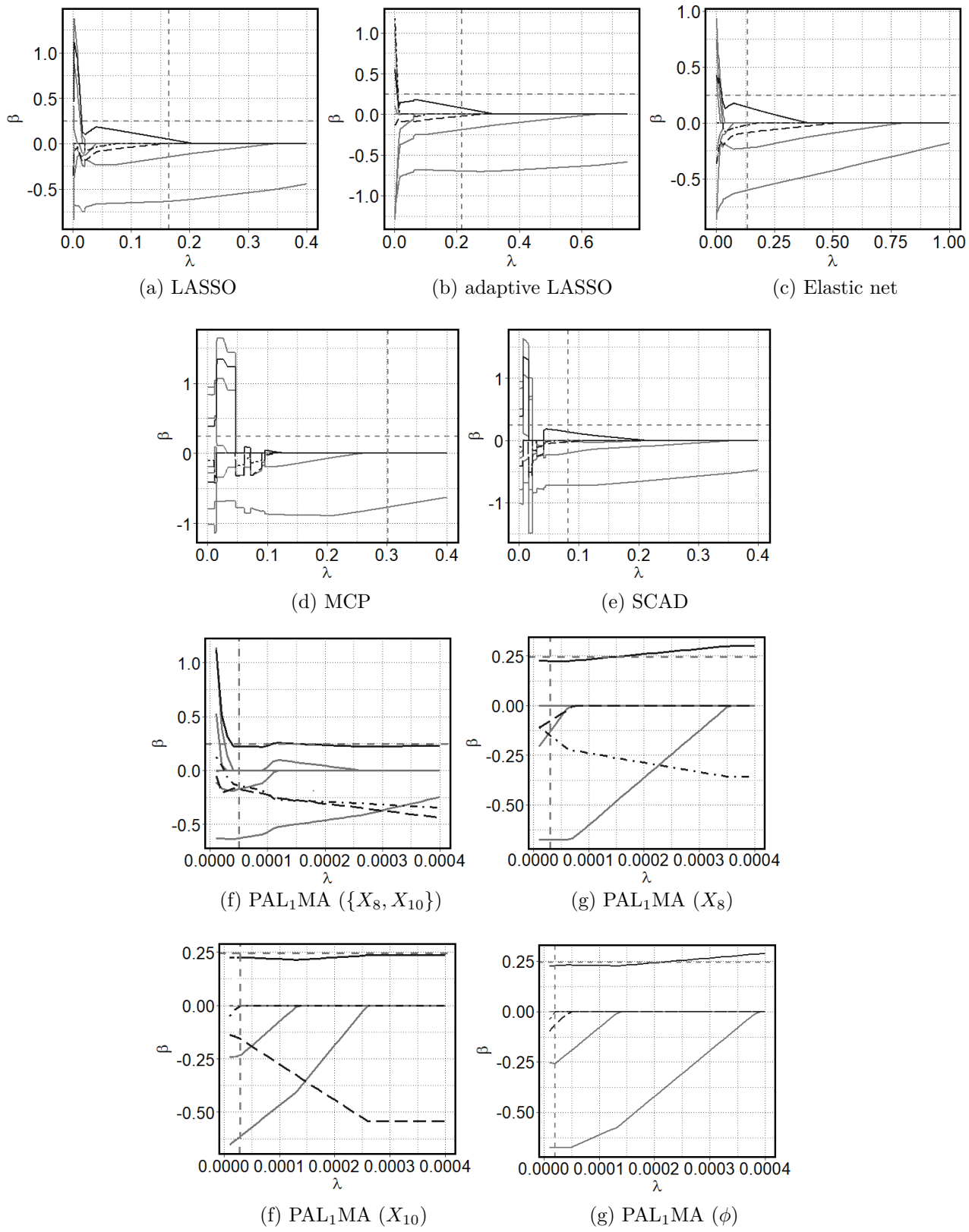
In this section, we are concerned with the evaluation of the total effect of X_2 on Y because similar observations can be derived regarding other controllable variables. Table E shows the results obtained by each regression analysis. Here, parameter tuning was conducted by the same procedure as in Section B.

First, according to Okuno et al. [1986], it is well known that the viscosity (X_2) is an important factor that increases both the rust protection quality and visual appearance. However, from Table E, the total effect of X_2 on Y is estimated as zero by MCP, which is problematic because it provides such a misleading interpretation that it is no use to control X_2 to achieve the aim.

Second, OLS regression provides the unbiased estimator of the total effect through a set $\{X_8, X_9\}$ that satisfies the back-door criterion. Given this finding, it is desirable that the estimators from the regularized regression analysis be close to the OLS estimate. From the viewpoint of this observation, the estimates from PAL$_1$MA are close to the OLS estimates for each selected variable, but those from the other regularized regression analyses are not close to these estimates.

Third, when the regression coefficient of X_8 is regularized, for PAL$_1$MA, X_8 is not selected, but $\{X_5, X_6\}$ is selected. This phenomenon may occur because the OLS estimate of the regression coefficient of X_8 is very small (-0.083) in the regression model of Y on X_2, X_5, X_6, X_8 and X_{10}. However, even if a set of sufficient confounders is not available by PAL$_1$MA, by checking the solution paths shown in Fig. D, we can verify that missing sufficient confounders do not interfere with the qualitative interpretation of the total effects by PAL$_1$MA for any λ.

Fourth, from Fig. E, the sample ranges of elastic net, PAL$_1$MA and OLS do not include zero, but those of the other regression analyses include zero. From this observation, it is judged that X_2 would have a positive effect on Y from elastic net, PAL$_1$MA and OLS, but the other regression analyses may not result in the rejection of the hypothesis that X_2 has no effect on Y.



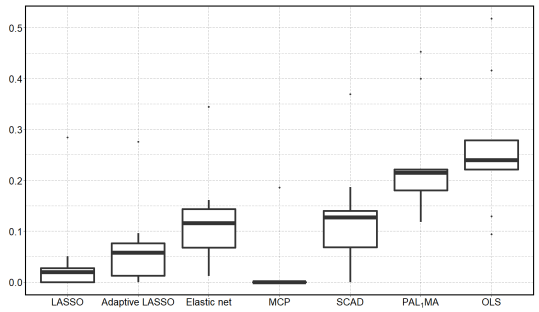
(a) LASSO (b) adaptive LASSO (c) Elastic net

(d) MCP (e) SCAD

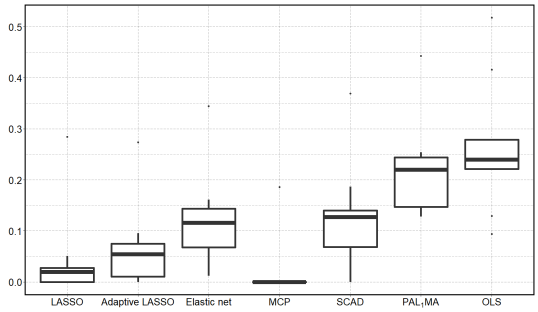
(f) $\mathrm{PAL_1MA}$ ($\{X_8, X_{10}\}$) (g) $\mathrm{PAL_1MA}$ (X_8)

(f) $\mathrm{PAL_1MA}$ (X_{10}) (g) $\mathrm{PAL_1MA}$ (ϕ)

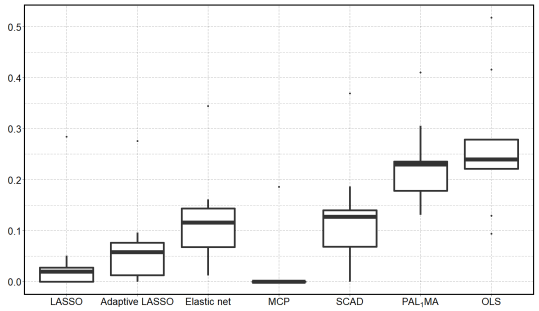
Fig. D. Solution paths of the regularization parameter λ when both ξ and ϕ are fixed to the value in Table E. Here, the dashed horizontal lines and the dashed vertical lines show the value of λ from Table E. The bold solid line: the regression coefficient of X_2; the dot-dashed line: the regression coefficient of X_8; the dashed line: the regression coefficient of X_{10}; the thin solid line: the regression coefficients of the other covariates.



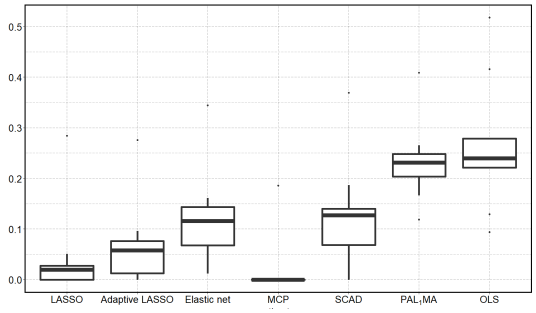
(a) $\mathrm{PAL}_1\mathrm{MA}$ ($\{X_8, X_{10}\}$)



(b) $\mathrm{PAL}_1\mathrm{MA}$ (X_8)

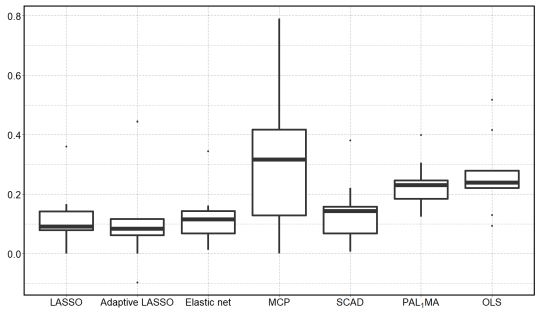


(c) $\mathrm{PAL}_1\mathrm{MA}$ (X_{10})

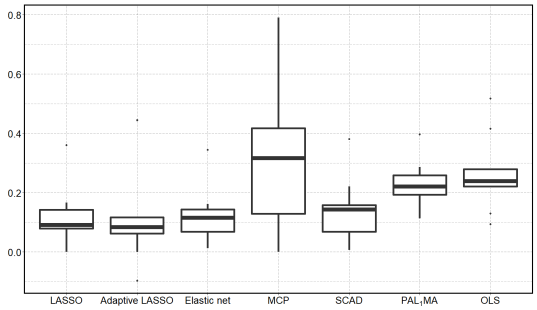


(d) $\mathrm{PAL}_1\mathrm{MA}$ (ϕ)

Fig. E. Boxplots of the case study for setting up the coating conditions for car bodies



(a) $\mathrm{PAL}_1\mathrm{MA}$ (X_{10})



(b) $\mathrm{PAL}_1\mathrm{MA}$ (ϕ)

Fig. F. Boxplots of the case study for setting up the coating conditions for car bodies

Here, Table F also shows the results obtained by conducting parameter tuning to select $\{X_8, X_{10}\}$ satisfying the back-door criterion relative to (X_2, Y) with the best prediction accuracy possible. To select $\{X_8, X_{10}\}$, in Table E, the regularization parameters have been set to smaller values than those in Table E. First, both the estimates and the standard deviations of LASSO, adaptive LASSO, MCP and SCAD in Table F are larger than those in Table E, but there seems to be no significant change in those of $\mathrm{PAL}_1\mathrm{MA}$ between Tables E and F. Second, compared to Table E, covariates other than X_8 and X_{10} are selected in Table F. Especially for MCP, an intermediate variable X_7 is also selected against the back-door criterion to select X_8 and X_{10}, which may be problematic in the context of statistical causal inference. Third, from Figs. F (a) and (b), although the sample ranges of LASSO, adaptive LASSO, MCP and SCAD include zero, OLS or $\mathrm{PAL}_1\mathrm{MA}$ does not include zero. From this observation, it is judged that X_2 would have a positive effect on Y from elastic net, the $\mathrm{PAL}_1\mathrm{MA}$ and OLS, but the other regression analyses may not result in the rejection of the hypothesis that X_2 has no effect on Y.

Table F. Results

Method	non-regulaized variables	estimate	sd	selected variables	parameters		
					λ	ξ	ϕ
LASSO	–	0.1453	0.1070	X_2,X_5,X_6,X_8,X_{10}	0.0752	–	–
adaptive LASSO	–	0.1438	0.1856	X_2,X_3,X_4,X_5,X_6,X_8,X_{10}	0.0170	0.5000	–
Elastic net	–	0.1395	0.0963	X_2,X_5,X_6,X_8,X_{10}	0.1350	–	0.5500
MCP	–	0.4254	0.2473	X_1,X_2,X_4,X_5,X_6,X_7,X_8,X_9,X_{10}	0.0140	4.0000	–
SCAD	–	0.1568	0.1158	X_2,X_4,X_5,X_6,X_8,X_{10}	0.0680	17.5000	–
PAL_1MA	X_{10}	0.2254	0.0835	X_2,X_5,X_6,X_8,X_{10}	0.00002	0.5000	–
PAL_1MA	–	0.2276	0.0816	X_2,X_5,X_6,X_8,X_{10}	0.00001	0.5000	–
OLS	–	0.2455	0.1314	X_2,X_8,X_{10}	–	–	–

estimate: estimates of the total effect with $n = 9$; sd: standard deviation based on method; selected variables: selected explanatory variables by the variable selection; parameter: regularized, tuning and mixing parameters. Here, λ_2 and ξ_2 of PAL_1MA were selected as zero by three fold cross-validation.

References

Beck, A. *First-order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

Breheny, P. and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, **5**:232-253, 2011.

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

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**:1–22, 2010.

Kuroki, M. Optimizing an external intervention using a structural equation model with an application to statistical process analysis. *Journal of Applied Statistics*, **39**:673-694, 2012.

Okuno, T., Katayama, Z., Kamigori, N., Itoh, T., Irikura, N., and Fujiwara, N. *Multivariate Data Analysis in Industry*, JUSE Press, 1986.

Pourahmadi, M. and Wang, X, Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor, *Statistics and Probability Letters*, **106**:5-12, 2015.

Sardy, S., Bruce, A. G., and Tseng, P. Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics*, **9**:361-379, 2000.

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

Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**:894-942, 2010.

Zou, H. The adaptive LASSO and its oracle properties. *Journal of the American statistical association*, **101**:1418-1429, 2006.

Zou, H. and Hastie, T. Regularization and variable selection via the Elastic net. *Journal of the Royal Statistical Society: Series B*, **67**:301-320, 2005.