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

Hisayoshi Nanmo[1] **Manabu Kuroki**[2]

[1]Chugai Pharmaceutical Co., Ltd., Nihonbashi Muromachi, Chuo-ku, Tokyo, Japan
[2]Yokohama National University , Tokiwadai, Hodogaya-ku, Yokohama, Japan

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 $\operatorname*{argmin}_x (f(\boldsymbol{x}) + g(\boldsymbol{x}))$ 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*

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

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

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:

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

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

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

where

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

and we define

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

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 $(b, B)^T$:

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

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*

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

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.,

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

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

$$\begin{aligned}
&\mathrm{prox}_{\frac{1}{l_f} g}\left(B_2 - \frac{1}{l_f}\frac{\partial}{\partial B} f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}\right) \quad \text{(A.20)}\\
&= \underset{b}{\mathrm{argmin}}\left(f^{\sharp}\left(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B_2\right) + \left\langle \frac{\partial}{\partial B} f^{\sharp}(\beta_{yx\cdot zw}, B_{yz\cdot xw}, B)_{B=B_2}, \boldsymbol{b} - B_2 \right\rangle + g(\boldsymbol{b}) + \frac{l_f}{2}||\boldsymbol{b} - B_2||_2^2\right)
\end{aligned}$$

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

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

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

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

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

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

Letting

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

from equation (A.24), we have

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

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

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

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

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

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

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

we obtain

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

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

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

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

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

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 et al., 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_1MA, 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_1MA 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_1MA is saved for future work. The parameter tuning results are shown in Table B.

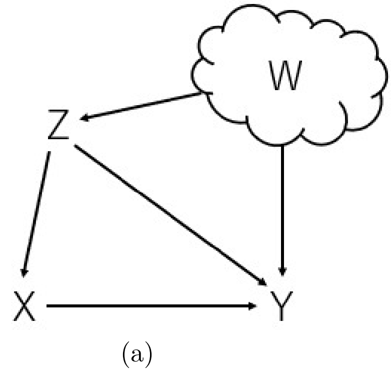
B.3 ANALYSIS

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_1MA 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.

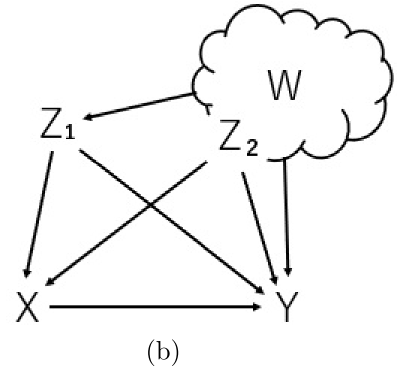
From Figs. B and B’ and Tables C and C’, we make the following observations:

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_1MA 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_1MA 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_1MA provides fewer bias estimates than the other regularized regression analyses. In Case (b_8), PAL_1MA provides more biased estimates than SCAD and MCP but higher coincidence rates than these regularized regression analyses.
5. The variance of the estimated total effects from PAL_1MA are larger than those from the other regularized regression analyses but smaller than those from OLS regression for most cases.

.



(a)



(b)

Fig. A. Causal diagram

6. From Figs. B and B', the interquartile ranges of $\mathrm{PAL_1MA}$ 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.
7. The running time of the i-PROGLES is slightly longer than those of other regularized regression analyses.

Overall, the coincidence rates between the signs of the estimated total effects and the true total effect from $\mathrm{PAL_1MA}$ seem equal to or higher than those from the other regression analyses. In addition, $\mathrm{PAL_1MA}$ can provide less biased estimators than the other regularized regression analyses in most cases. In some cases of Figs. A (b), $\mathrm{PAL_1MA}$ 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, $\mathrm{PAL_1MA}$ 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, although an estimated total effects may not be stable in some situations. These results imply that the estimation of the total effect by $\mathrm{PAL_1MA}$ 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		$\mathrm{PAL}_1\mathrm{MA}$				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		$\mathrm{PAL}_1\mathrm{MA}$				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.

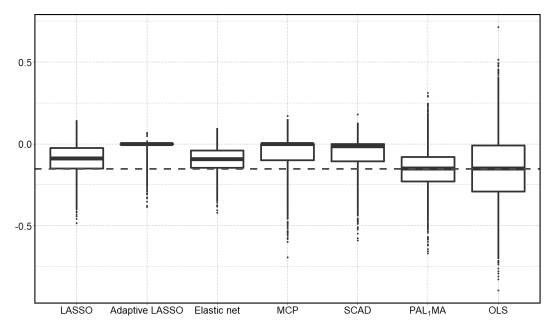
Table C. Results based on cross-validation.

(a) Z satisfies the back-door criterion

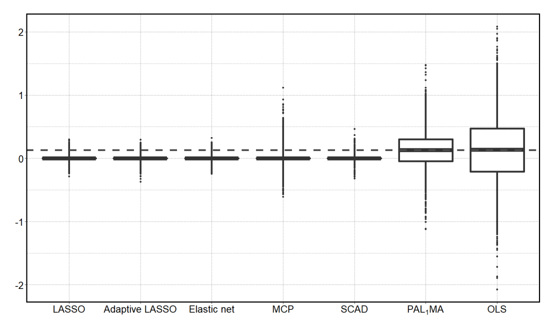
	(a_1)					(a_2)				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.0971	0.0550	0.0098	0.0824	0.8282	0.0079	0.1216	**0.0163**	0.0385	0.1482
adaptive LASSO	-0.0185	0.1336	0.0200	**0.0463**	0.2332	-0.0024	0.1319	0.0187	0.0366	0.0708
Elastic net	-0.0983	0.0538	**0.0081**	0.0720	0.8904	0.0054	0.1241	**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
PAL$_1$MA	-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
PAL$_1$MA	-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

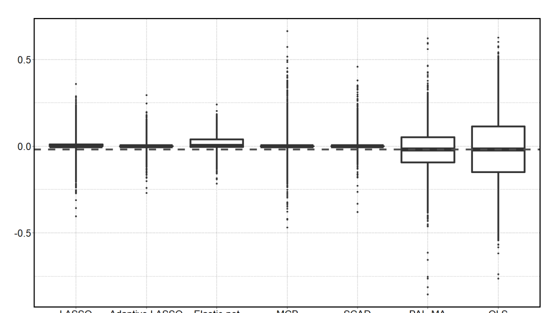
mean: sample mean; bias: bias between the true value and the sample mean; mse: mean squared error: sd: standard deviation; sign: coincidence rate between the signs of the true value and the estimates The best results for each columns are highlighted in boldface.



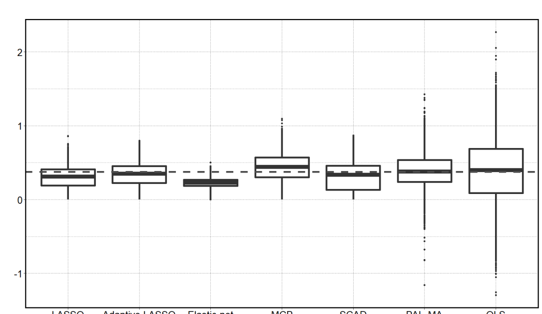
(a_1)



(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. The best results for each columns are highlighted in boldface.

Table C'. Results based on cross-validation.

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

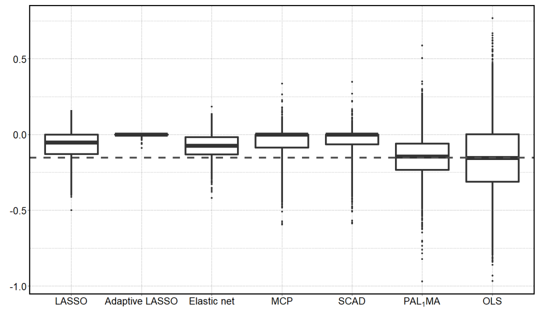
	(b_9)					(b_{10})				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	0.1000	0.0583	0.0115	0.0898	0.7886	-0.0626	0.1219	**0.0223**	0.0863	0.5282
adaptive LASSO	0.0762	0.0822	0.0131	**0.0797**	0.7044	-0.0445	0.1400	0.0254	0.0758	0.4102
Elastic net	0.1136	0.0448	**0.0098**	0.0881	0.8576	-0.0493	0.1352	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$_1$MA	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$_1$MA	-0.4059	**0.0201**	**0.029**	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

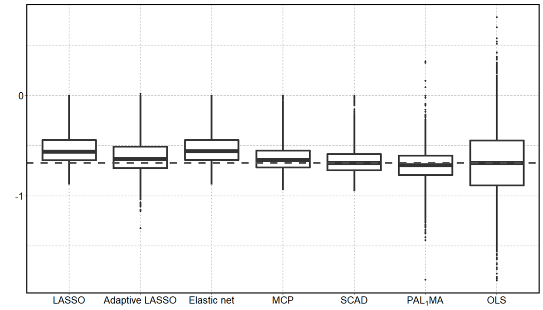
	(b_{13})					(b_{14})				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
LASSO	-0.3901	0.4240	0.1979	0.1347	0.9904	-0.0045	0.2535	0.0648	0.0235	0.0820
adaptive LASSO	-0.6497	0.1643	**0.0631**	0.1899	**0.9996**	-0.0004	0.2577	0.0664	**0.0061**	0.0072
Elastic net	-0.4351	0.3789	0.1597	**0.1269**	0.9980	-0.0154	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$_1$MA	-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})				
	mean	bias	mse	sd	sign	mean	bias	mse	sd	sign
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$_1$MA	-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

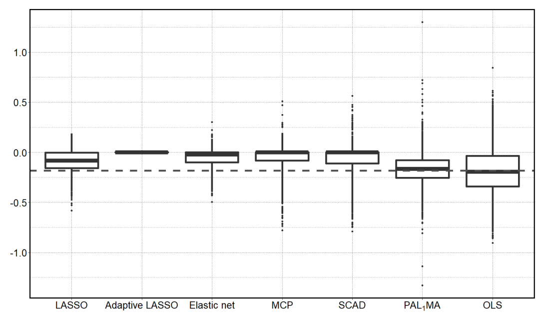
mean: sample mean; bias: bias between the true value and the sample mean; mse: mean squared error: sd: standard deviation; sign: coincidence rate between the signs of the true value and the estimates The best results for each columns are highlighted in boldface.



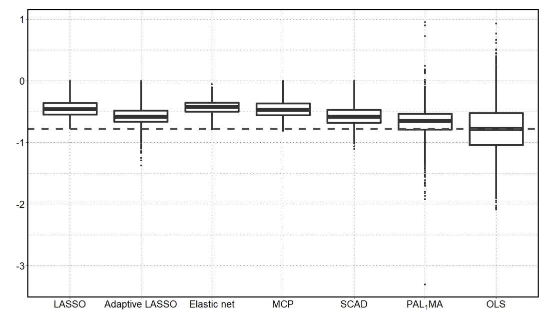
(b_1)



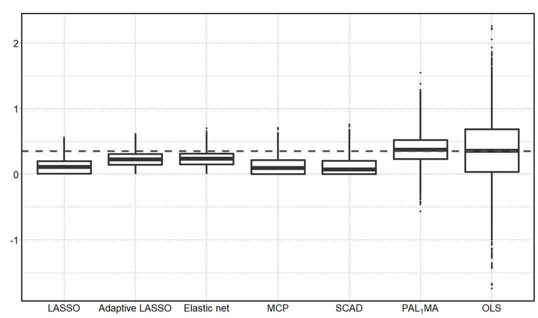
(b_2)



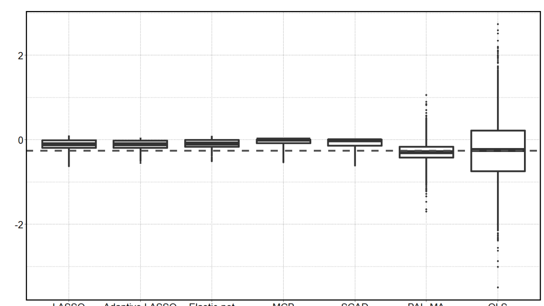
(b_3)



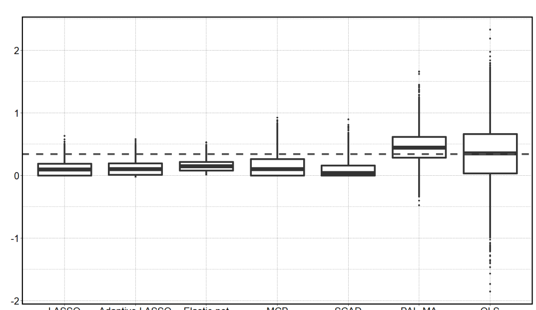
(b_4)



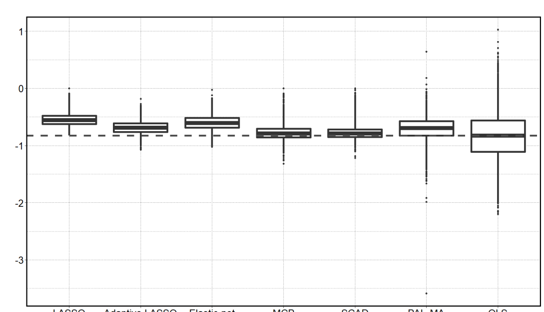
(b_5)



(b_6)



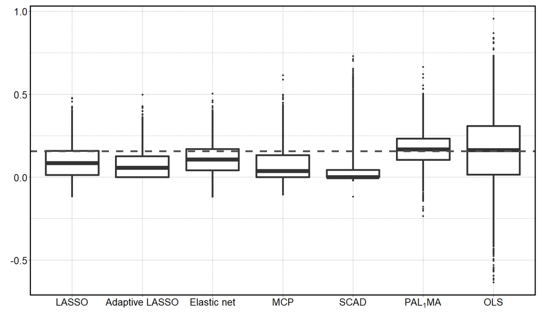
(b_7)



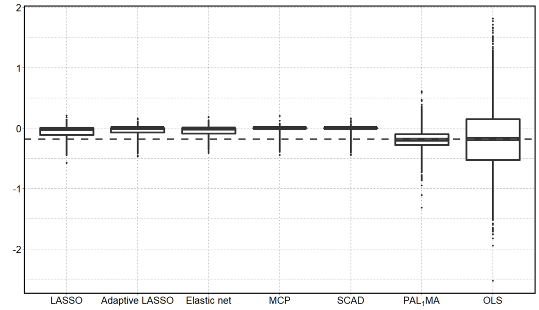
(b_8)

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

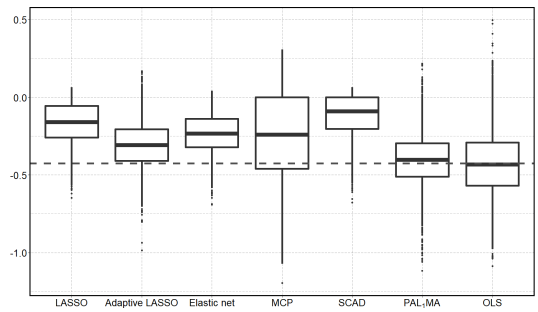
Fig. B'. Boxplots of the estimated total effects based on 5000 replications from the numerical experiments. The dashed lines show the true total effects.



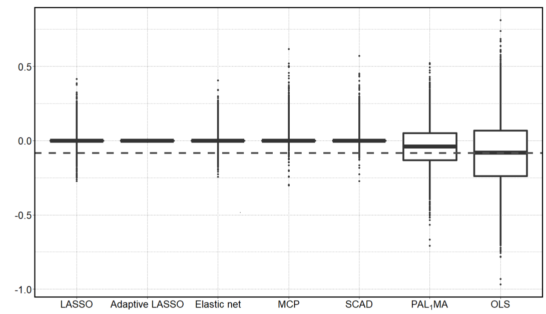
(b_9)



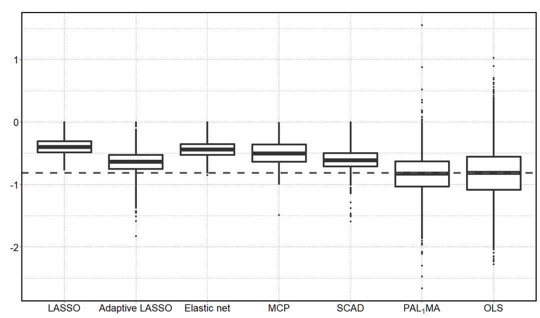
(b_{10})



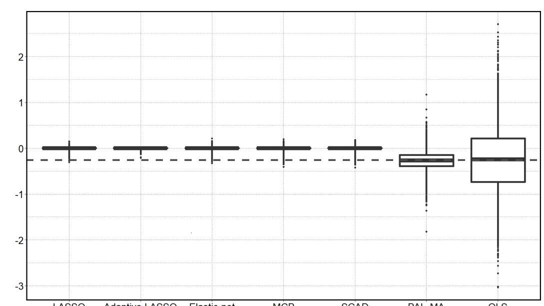
(b_{11})



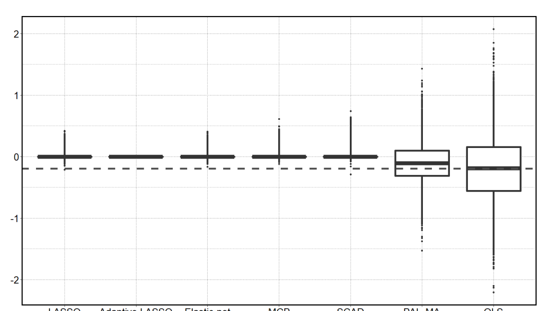
(b_{12})



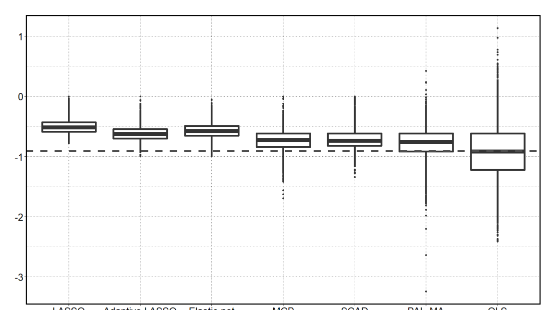
(b_{13})



(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

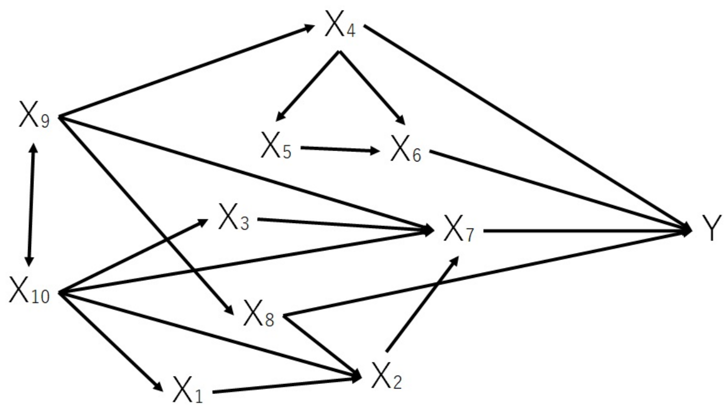


Fig. C. Causal diagram [Kuroki, 2012]

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.

According to Okuno et al. [1986], there is some difference among these variables in terms of the controllability

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.

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

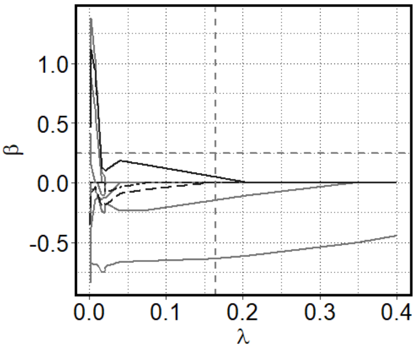
In this section, we are concerned with the evaluation of the total effect of X_2 on Y because similar observations can be derived regarding other controllable variables. Table E shows the results obtained by each regression analysis. Here, parameter tuning was conducted by the same procedure as in Section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

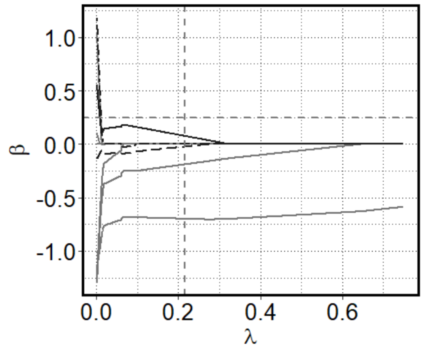
Second, OLS regression provides the unbiased estimator of the total effect through a set $\{X_8, X_9\}$ that satisfies the back-door criterion. Given this finding, it is desirable that the estimators from the regularized regression analysis be close to the OLS estimate. From the viewpoint of this observation, 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.

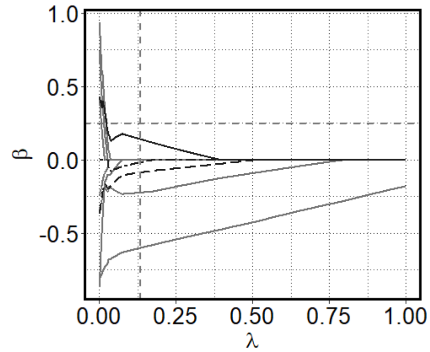
Fourth, from Fig. E, the sample ranges of elastic net, PAL$_1$MA and OLS do not include zero, but those of the other regression analyses include zero. From this observation, it is judged that X_2 would have a positive effect on Y from elastic net, PAL$_1$MA and OLS, but the other regression analyses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



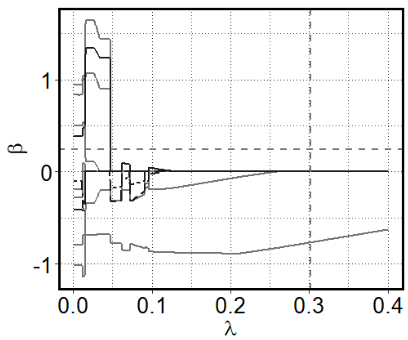
(a) LASSO



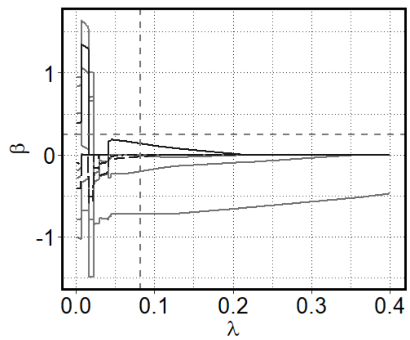
(b) adaptive LASSO



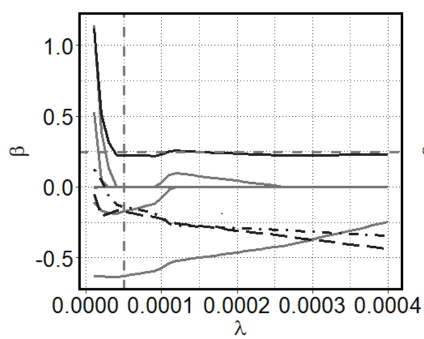
(c) Elastic net



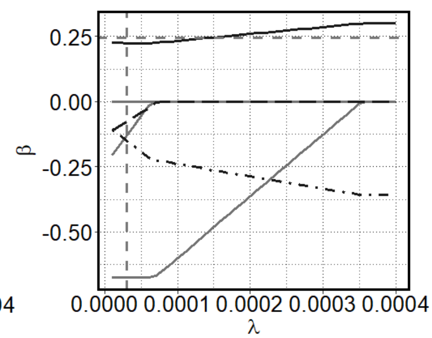
(d) MCP



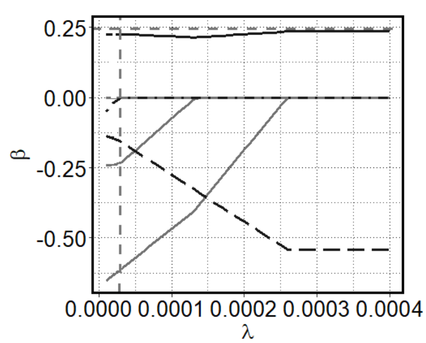
(e) SCAD



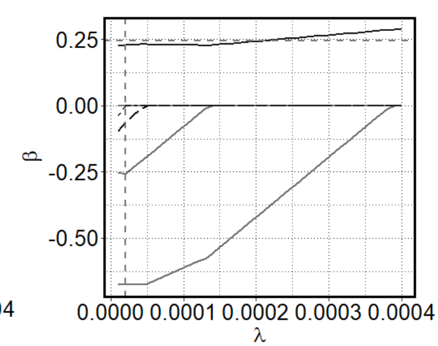
(f) $\mathrm{PAL_1MA}$ ($\{X_8, X_{10}\}$)



(g) $\mathrm{PAL_1MA}$ (X_8)

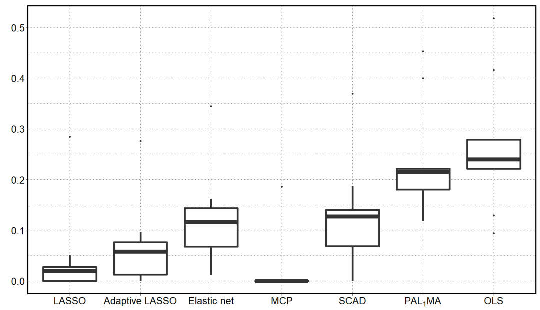


(f) $\mathrm{PAL_1MA}$ (X_{10})

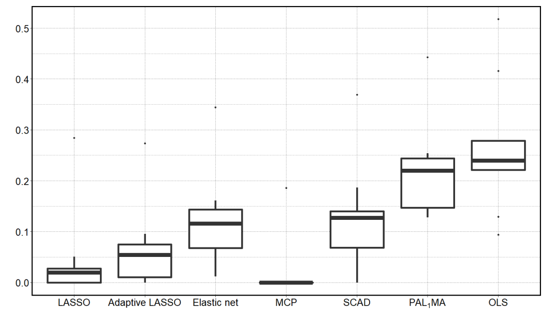


(g) $\mathrm{PAL_1MA}$ (ϕ)

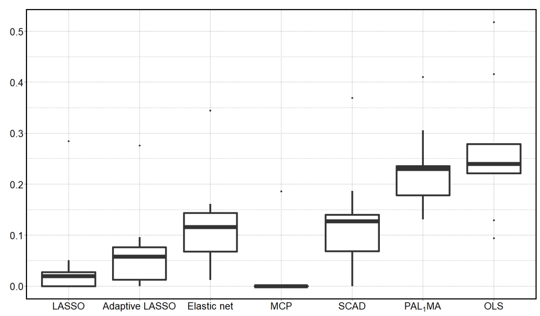
Fig. D. Solution paths of the regularization parameter λ when both ξ and ϕ are fixed to the value in Table E. Here, the dashed horizontal lines and the dashed vertical lines show the value of λ from Table E. The bold solid line: the regression coefficient of X_2; the dot-dashed line: the regression coefficient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



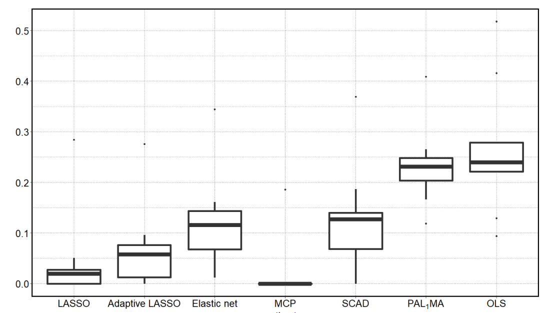
(a) PAL$_1$MA ($\{X_8, X_{10}\}$)



(b) PAL$_1$MA (X_8)

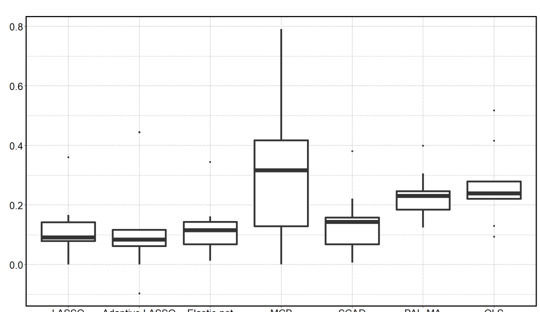


(c) PAL$_1$MA (X_{10})

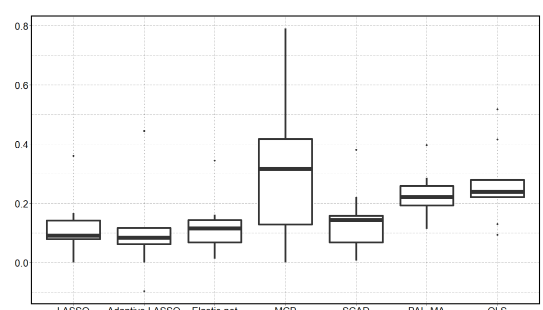


(d) PAL$_1$MA (ϕ)

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



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



(b) PAL$_1$MA (ϕ)

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

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$_1$MA	X_{10}	0.2254	0.0835	X_2,X_5,X_6,X_8,X_{10}	0.00002	0.5000	–
PAL$_1$MA	–	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$_1$MA were selected as zero by three fold cross-validation.

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 PAL$_1$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 PAL$_1$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 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.

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.