PCM Selector: Penalized Covariate-Mediator Selection Operator for Evaluating Linear Causal Effects
- Technical Appendix-

Hisayoshi Nanmo[1, 2], Manabu Kuroki [2]
[1]Chugai Pharmaceutical Co., Ltd., Nihonbashi Muromachi, Chuo-ku, Tokyo, Japan
[2]Yokohama National University, Tokiwadai, Hodogaya-ku, Yokohama, Japan
nanmohisayoshi@gmail.com, kuroki-manabu-zm@ynu.ac.jp

A Derivation of PCM estimator

Note that PCM estimator can be derived by repeated application of the blockwise inversion formula of the invertible matrix (Bernstein, 2009): For the invertible matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}, \tag{A.1}$$

we have

$$\begin{aligned}\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} &= \begin{pmatrix} A^{-1}+A^{-1}B\left(D-CA^{-1}B\right)^{-1}CA^{-1} & -A^{-1}B\left(D-CA^{-1}B\right)^{-1} \\ -\left(D-CA^{-1}B\right)^{-1}CA^{-1} & \left(D-CA^{-1}B\right)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \left(A-BD^{-1}C\right)^{-1} & -\left(A-BD^{-1}C\right)^{-1}BD^{-1} \\ -D^{-1}C\left(A-BD^{-1}C\right)^{-1} & D^{-1}+D^{-1}C\left(A-BD^{-1}C\right)^{-1}BD^{-1} \end{pmatrix},\end{aligned} \tag{A.2}$$

where A, D, $D-CA^{-1}B$ and $A-BD^{-1}C$ are invertible square submatrices. Then, the derivation of PCM estimators is based on the following steps:

Step 1: The derivation of $\check{B}^{\dagger}_{mx.c}$

Step 2: The derivation of $\check{\beta}^{\dagger}_{yx.cm}$, $\check{B}^{\dagger}_{ys.xc\overline{s}}$ and $\check{B}^{\dagger}_{y\overline{s}.xcs}$

Step 1: The derivation of $\check{B}^{\dagger}_{mx.c}$

When the sum-of-squares matrix of $\{X\}\cup \boldsymbol{C}$ is invertible, by using the idea of the subderivative, for $p=1$, we find that the values of $B_{mx.c}$, $B_{mz.x\overline{z}}$ and $B_{m\overline{z}.xz}$ that minimize equation (8) are given by

$$\begin{pmatrix} \check{B}^{\dagger}_{mx.c} \\ \check{B}^{\dagger}_{mz.x\overline{z}} \\ \check{B}^{\dagger}_{m\overline{z}.xz} \end{pmatrix} = \begin{pmatrix} \hat{B}_{mx.c} \\ \hat{B}_{mz.x\overline{z}} \\ \hat{B}_{m\overline{z}.xz} \end{pmatrix} - n\rho_1 \begin{pmatrix} S_{xx} & S_{xz} & S_{x\overline{z}} \\ S_{zx} & S_{zz} & S_{z\overline{z}} \\ S_{\overline{z}x} & S_{\overline{z}z} & S_{\overline{z}\overline{z}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0}^T_{q_m} \\ \mathbf{0}_{q_z,q_m} \\ \boldsymbol{\gamma}_{m\overline{z}.xz}\odot \mathrm{sign}(\check{B}^{\dagger}_{m\overline{z}.xz}) \end{pmatrix}. \tag{A.3}$$

Then, letting

$$\begin{pmatrix} S_{xx} & S_{xz} & S_{x\overline{z}} \\ S_{zx} & S_{zz} & S_{z\overline{z}} \\ S_{\overline{z}x} & S_{\overline{z}z} & S_{\overline{z}\overline{z}} \end{pmatrix}^{-1} = \begin{pmatrix} S^{xx} & S^{xz} & S^{x\overline{z}} \\ S^{zx} & S^{zz} & S^{z\overline{z}} \\ S^{\overline{z}x} & S^{\overline{z}z} & S^{\overline{z}\overline{z}} \end{pmatrix}, \tag{A.4}$$

since we have

$$\begin{pmatrix} S^{x\overline{z}} \\ S^{z\overline{z}} \end{pmatrix} = \begin{pmatrix} -\hat{B}_{\overline{z}x.z}S^{-1}_{\overline{z}\overline{z}.xz} \\ -\hat{B}_{\overline{z}z.x}S^{-1}_{\overline{z}\overline{z}.xz} \end{pmatrix}, \tag{A.5}$$

from the blockwise inversion formula of the invertible matrix (Bernstein, 2009), we derive

$$\begin{aligned}\check{B}^{\dagger}_{mx.c} &= \hat{B}_{mx.c} - n\rho_1 S^{x\overline{z}}\boldsymbol{\gamma}_{m\overline{z}.xz}\odot \mathrm{sign}(\check{B}^{\dagger}_{m\overline{z}.xz}) \\ &= \hat{B}_{mx.c} + n\rho_1 \hat{B}_{\overline{z}x.z}S^{-1}_{\overline{z}\overline{z}.xz}\boldsymbol{\gamma}_{m\overline{z}.xz}\odot \mathrm{sign}(\check{B}^{\dagger}_{m\overline{z}.xz}).\end{aligned} \tag{A.6}$$

Step 2: The derivation of $\check{\beta}^{\dagger}_{yx.cm}$, $\check{B}^{\dagger}_{ys.xc\bar{s}}$ and $\check{B}^{\dagger}_{y\bar{s}.xcs}$

Similarly, when the sum-of-squares matrix of $\{X\} \cup \boldsymbol{C} \cup \boldsymbol{M}$ is invertible, by using the idea of the subderivative, for $p = 1$, we find that the values of $\beta_{yx.cm}$, $B_{ys.xc\bar{s}}$, $B_{yz.xm\bar{z}}$, $B_{y\bar{s}.xcs}$ and $B_{y\bar{z}.xmz}$ that minimize equation (7) are given by

$$\begin{pmatrix} \check{\beta}^{\dagger}_{yx.cm} \\ \check{B}^{\dagger}_{ys.xc\bar{s}} \\ \check{B}^{\dagger}_{yz.xm\bar{z}} \\ \check{B}^{\dagger}_{y\bar{s}.xcs} \\ \check{B}^{\dagger}_{y\bar{z}.xmz} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{yx.cm} \\ \hat{B}_{ys.xc\bar{s}} \\ \hat{B}_{yz.xm\bar{z}} \\ \hat{B}_{y\bar{s}.xcs} \\ \hat{B}_{y\bar{z}.xmz} \end{pmatrix} - n\lambda_1 \begin{pmatrix} s_{xx} & S_{xs} & S_{xz} & S_{x\bar{s}} & S_{x\bar{z}} \\ S_{sx} & S_{ss} & S_{sz} & S_{s\bar{s}} & S_{s\bar{z}} \\ S_{zx} & S_{zs} & S_{zz} & S_{z\bar{s}} & S_{z\bar{z}} \\ S_{\bar{s}x} & S_{\bar{s}s} & S_{\bar{s}z} & S_{\bar{s}\bar{s}} & S_{\bar{s}\bar{z}} \\ S_{\bar{z}x} & S_{\bar{z}s} & S_{\bar{z}z} & S_{\bar{z}\bar{s}} & S_{\bar{z}\bar{z}} \end{pmatrix}^{-1}$$

$$\times \begin{pmatrix} \zeta_1 \mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm}) \\ \mathbf{0}_{q_s} \\ \mathbf{0}_{q_z} \\ \xi_1 \boldsymbol{\gamma}_{\bar{s}x.c} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\ (1 - \zeta_1 - \xi_1)\boldsymbol{\gamma}_{y\bar{z}.xmz} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz}) \end{pmatrix}. \tag{A.7}$$

Then, letting

$$\begin{pmatrix} s_{xx} & S_{xs} & S_{xz} & S_{x\bar{s}} & S_{x\bar{z}} \\ S_{sx} & S_{ss} & S_{sz} & S_{s\bar{s}} & S_{s\bar{z}} \\ S_{zx} & S_{zs} & S_{zz} & S_{z\bar{s}} & S_{z\bar{z}} \\ S_{\bar{s}x} & S_{\bar{s}s} & S_{\bar{s}z} & S_{\bar{s}\bar{s}} & S_{\bar{s}\bar{z}} \\ S_{\bar{z}x} & S_{\bar{z}s} & S_{\bar{z}z} & S_{\bar{z}\bar{s}} & S_{\bar{z}\bar{z}} \end{pmatrix}^{-1} = \begin{pmatrix} s^{xx} & S^{xs} & S^{xz} & S^{x\bar{s}} & S^{x\bar{z}} \\ S^{sx} & S^{ss} & S^{sz} & S^{s\bar{s}} & S^{s\bar{z}} \\ S^{zx} & S^{zs} & S^{zz} & S^{z\bar{s}} & S^{z\bar{z}} \\ S^{\bar{s}x} & S^{\bar{s}s} & S^{\bar{s}z} & S^{\bar{s}\bar{s}} & S^{\bar{s}\bar{z}} \\ S^{\bar{z}x} & S^{\bar{z}s} & S^{\bar{z}z} & S^{\bar{z}\bar{s}} & S^{\bar{z}\bar{z}} \end{pmatrix}, \tag{A.8}$$

we have

$$\begin{aligned} \begin{pmatrix} s^{xx} & S^{xs} \\ S^{sx} & S^{ss} \end{pmatrix} &= \begin{pmatrix} s^{-1}_{xx.cm} & -s^{-1}_{xx.c\bar{s}} S_{xs.c\bar{s}} S^{-1}_{ss.xc\bar{s}} \\ -S^{-1}_{ss.c\bar{s}} S_{sx.c\bar{s}} s^{-1}_{xx.cm} & S^{-1}_{ss.xc\bar{s}} \end{pmatrix} \\ &= \begin{pmatrix} s^{-1}_{xx.cm} & -\hat{B}_{sx.c\bar{s}} S^{-1}_{ss.xc\bar{s}} \\ -\hat{B}_{xs.c\bar{s}} s^{-1}_{xx.cm} & S^{-1}_{ss.xc\bar{s}} \end{pmatrix}, \end{aligned} \tag{A.9}$$

$$\begin{aligned} \begin{pmatrix} S^{\bar{s}\bar{s}} & S^{\bar{s}\bar{z}} \\ S^{\bar{z}\bar{s}} & S^{\bar{z}\bar{z}} \end{pmatrix} &= \begin{pmatrix} S^{-1}_{\bar{s}\bar{s}.xcs} & -S^{-1}_{\bar{s}\bar{s}.xsz} S_{\bar{s}\bar{z}.xsz} S^{-1}_{\bar{z}\bar{z}.xmz} \\ -S^{-1}_{\bar{z}\bar{z}.xsz} S_{\bar{z}\bar{s}.xsz} S^{-1}_{\bar{s}\bar{s}.xcs} & S^{-1}_{\bar{z}\bar{z}.xmz} \end{pmatrix} \\ &= \begin{pmatrix} S^{-1}_{\bar{s}\bar{s}.xcs} & -\hat{B}_{\bar{z}\bar{s}.xsz} S^{-1}_{\bar{z}\bar{z}.xmz} \\ -\hat{B}_{\bar{s}\bar{z}.xsz} S^{-1}_{\bar{s}\bar{s}.xcs} & S^{-1}_{\bar{z}\bar{z}.xmz} \end{pmatrix}, \end{aligned} \tag{A.10}$$

$$\begin{pmatrix} S^{zx} & S^{zs} \\ S^{\bar{s}x} & S^{\bar{s}s} \\ S^{\bar{z}x} & S^{\bar{z}s} \end{pmatrix} = -\begin{pmatrix} \hat{B}_{xz.\bar{s}\bar{z}} & \hat{B}_{sz.\bar{s}\bar{z}} \\ \hat{B}_{x\bar{s}.c} & \hat{B}_{s\bar{s}.c} \\ \hat{B}_{x\bar{z}.z\bar{s}} & \hat{B}_{s\bar{z}.z\bar{s}} \end{pmatrix} \begin{pmatrix} s^{-1}_{xx.cm} & -\hat{B}_{sx.c\bar{s}} S^{-1}_{ss.xc\bar{s}} \\ -\hat{B}_{xs.c\bar{s}} s^{-1}_{xx.cm} & S^{-1}_{ss.xc\bar{s}} \end{pmatrix} \tag{A.11}$$

$$= \begin{pmatrix} \hat{B}_{sz.\bar{s}\bar{z}} \hat{B}_{xs.c\bar{s}} s^{-1}_{xx.cm} - \hat{B}_{xz.\bar{s}\bar{z}} s^{-1}_{xx.cm} & \hat{B}_{xz.\bar{s}\bar{z}} \hat{B}_{sx.c\bar{s}} S^{-1}_{ss.xc\bar{s}} - \hat{B}_{sz.\bar{s}\bar{z}} S^{-1}_{ss.xc\bar{s}} \\ \hat{B}_{s\bar{s}.c} \hat{B}_{xs.c\bar{s}} s^{-1}_{xx.cm} - \hat{B}_{x\bar{s}.c} s^{-1}_{xx.cm} & \hat{B}_{x\bar{s}.c} \hat{B}_{sx.c\bar{s}} S^{-1}_{ss.xc\bar{s}} - \hat{B}_{s\bar{s}.c} S^{-1}_{ss.xc\bar{s}} \\ \hat{B}_{s\bar{z}.z\bar{s}} \hat{B}_{xs.c\bar{s}} s^{-1}_{xx.cm} - \hat{B}_{x\bar{z}.z\bar{s}} s^{-1}_{xx.cm} & \hat{B}_{x\bar{z}.z\bar{s}} \hat{B}_{sx.c\bar{s}} S^{-1}_{ss.xc\bar{s}} - \hat{B}_{s\bar{z}.z\bar{s}} S^{-1}_{ss.xc\bar{s}} \end{pmatrix},$$

$$\begin{pmatrix} S^{x\bar{s}} & S^{x\bar{z}} \\ S^{s\bar{s}} & S^{s\bar{z}} \\ S^{z\bar{s}} & S^{z\bar{z}} \end{pmatrix} = -\begin{pmatrix} \hat{B}_{\bar{s}x.sz} & \hat{B}_{\bar{z}x.sz} \\ \hat{B}_{\bar{s}s.xz} & \hat{B}_{\bar{z}s.xz} \\ \hat{B}_{\bar{s}z.xs} & \hat{B}_{\bar{z}z.xs} \end{pmatrix} \begin{pmatrix} S^{-1}_{\bar{s}\bar{s}.xcs} & -\hat{B}_{\bar{z}\bar{s}.xsz} S^{-1}_{\bar{z}\bar{z}.xmz} \\ -\hat{B}_{\bar{s}\bar{z}.xsz} S^{-1}_{\bar{s}\bar{s}.xcs} & S^{-1}_{\bar{z}\bar{z}.xmz} \end{pmatrix} \tag{A.12}$$

$$= \begin{pmatrix} \hat{B}_{\bar{z}x.sz} \hat{B}_{\bar{s}\bar{z}.xsz} S^{-1}_{\bar{s}\bar{s}.xcs} - \hat{B}_{\bar{s}x.sz} S^{-1}_{\bar{s}\bar{s}.xcs} & \hat{B}_{\bar{s}x.sz} \hat{B}_{\bar{z}\bar{s}.xsz} S^{-1}_{\bar{z}\bar{z}.xmz} - \hat{B}_{\bar{z}x.sz} S^{-1}_{\bar{z}\bar{z}.xmz} \\ \hat{B}_{\bar{z}s.xz} \hat{B}_{\bar{s}\bar{z}.xsz} S^{-1}_{\bar{s}\bar{s}.xcs} - \hat{B}_{\bar{s}s.xz} S^{-1}_{\bar{s}\bar{s}.xcs} & \hat{B}_{\bar{s}s.xz} \hat{B}_{\bar{z}\bar{s}.xsz} S^{-1}_{\bar{z}\bar{z}.xmz} - \hat{B}_{\bar{z}s.xz} S^{-1}_{\bar{z}\bar{z}.xmz} \\ \hat{B}_{\bar{z}z.xs} \hat{B}_{\bar{s}\bar{z}.xsz} S^{-1}_{\bar{s}\bar{s}.xcs} - \hat{B}_{\bar{s}z.xs} S^{-1}_{\bar{s}\bar{s}.xcs} & \hat{B}_{\bar{s}z.xs} \hat{B}_{\bar{z}\bar{s}.xsz} S^{-1}_{\bar{z}\bar{z}.xmz} - \hat{B}_{\bar{z}z.xs} S^{-1}_{\bar{z}\bar{z}.xmz} \end{pmatrix}$$

from the blockwise inversion formula of the invertible matrix (Bernstein, 2009). Thus, we derive

$$\check{\beta}^{\dagger}_{yx.cm} = \hat{\beta}_{yx.cm} - n\lambda_1 \Big\{ \zeta_1 s^{xx} \mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm}) + \xi_1 S^{x\bar{s}} \boldsymbol{\gamma}_{\bar{s}x.c} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs})$$

$$
\begin{aligned}
&\quad +(1-\zeta_1-\xi_1)S^{x\bar{z}}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz})\Big\} \\
&= \hat{\beta}_{yx.cm}-n\lambda_1\Big\{\zeta_1 s^{-1}_{xx.cm}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm})-\xi_1\hat{B}_{\bar{s}x.sc}S^{-1}_{\bar{s}\bar{s}.xcs}\boldsymbol{\gamma}_{\bar{s}x.c}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\
&\quad -(1-\zeta_1-\xi_1)\hat{B}_{\bar{z}x.zm}S^{-1}_{\bar{z}\bar{z}.xmz}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz})\Big\}, \qquad \text{(A.13)}
\end{aligned}
$$

$$
\begin{aligned}
\check{B}^{\dagger}_{ys.xc\bar{s}} &= \hat{B}_{ys.xc\bar{s}}-n\lambda_1\Big\{\zeta_1 S^{sx}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm})+\xi_1 S^{s\bar{s}}\boldsymbol{\gamma}_{\bar{s}x.c}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\
&\quad +(1-\zeta_1-\xi_1)S^{s\bar{z}}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz})\Big\} \\
&= \hat{B}_{ys.xc\bar{s}}-n\lambda_1\Big\{-\zeta_1\hat{B}_{xs.c\bar{s}}s^{-1}_{xx.cm}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm}) \\
&\quad -\xi_1\hat{B}_{\bar{s}s.xc}S^{-1}_{\bar{s}\bar{s}.xcs}\boldsymbol{\gamma}_{\bar{s}x.c}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\
&\quad -(1-\zeta_1-\xi_1)\hat{B}_{\bar{z}s.xz\bar{s}}S^{-1}_{\bar{z}\bar{z}.xmz}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz})\Big\}, \qquad \text{(A.14)}
\end{aligned}
$$

$$
\begin{aligned}
\check{B}^{\dagger}_{y\bar{s}.xcs} &= \hat{B}_{y\bar{s}.xcs}-n\lambda_1\Big\{\zeta_1 S^{\bar{s}x}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm})+\xi_1 S^{\bar{s}\bar{s}}\boldsymbol{\gamma}_{\bar{s}x.c}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\
&\quad +(1-\zeta_1-\xi_1)S^{\bar{s}\bar{z}}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz})\Big\} \\
&= \hat{B}_{y\bar{s}.xcs}-n\lambda_1\Big\{-\zeta_1\hat{B}_{x\bar{s}.cs}s^{-1}_{xx.cm}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm})+\xi_1 S^{-1}_{\bar{s}\bar{s}.xcs}\boldsymbol{\gamma}_{\bar{s}x.c}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\
&\quad -(1-\zeta_1-\xi_1)\hat{B}_{\bar{z}\bar{s}.xsz}S^{-1}_{\bar{z}\bar{z}.xmz}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz})\Big\}. \qquad \text{(A.15)}
\end{aligned}
$$

By combining the above equations, we derive

$$
\begin{aligned}
\begin{pmatrix} \check{\beta}^{\dagger}_{yx.cm} \\ \check{B}^{\dagger}_{ys.xc\bar{s}} \\ \check{B}^{\dagger}_{y\bar{s}.xcs} \end{pmatrix} &= \begin{pmatrix} \hat{\beta}_{yx.cm} \\ \hat{B}_{ys.xc\bar{s}} \\ \hat{B}_{y\bar{s}.xcs} \end{pmatrix} + n\lambda_1 \begin{pmatrix} -1 & \hat{B}_{\bar{s}x.sc} & \hat{B}_{\bar{z}x.zm} \\ \hat{B}_{xs.c\bar{s}} & \hat{B}_{\bar{s}s.xc} & \hat{B}_{\bar{z}s.xz\bar{s}} \\ \hat{B}_{x\bar{s}.cs} & -I_{q_{\bar{s}}} & \hat{B}_{\bar{z}\bar{s}.xsz} \end{pmatrix} \\
&\quad \times \begin{pmatrix} \zeta_1 s^{-1}_{xx.cm}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm}) \\ \xi_1 S^{-1}_{\bar{s}\bar{s}.xcs}\boldsymbol{\gamma}_{\bar{s}x.c}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\ (1-\zeta_1-\xi_1)S^{-1}_{\bar{z}\bar{z}.xmz}\boldsymbol{\gamma}_{y\bar{z}.xmz}\odot\mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz}) \end{pmatrix}. \qquad \text{(A.16)}
\end{aligned}
$$

B Proof of Theorem 1

First, letting $\boldsymbol{v}$ be an active set from $\boldsymbol{x}\cup\boldsymbol{m}\cup\boldsymbol{c}$, for penalized estimators, such as ridge-type estimators and LASSO-type estimators, $\check{B}_{yv}$ and the ordinary least-squares estimator $\hat{B}_{yv}$ in the regression model of Y on $\boldsymbol{V}$, from Zou (2006), we have

$$\check{B}_{yv}=(S_{vv}+\Gamma)^{-1}S_{vy}=(I_{vv}+S^{-1}_{vv}\Gamma)^{-1}\hat{B}_{yv}, \qquad \text{(B.17)}$$

$$\mathrm{var}(\check{B}_{yv})=\sigma_{yy.v}(S_{vv}+\Gamma)^{-1}S_{vv}(S_{vv}+\Gamma)^{-1} \qquad \text{(B.18)}$$

approximately[1], where $\sigma_{yy.v}$, S_{vv} and Γ are the conditional variance of Y given $\boldsymbol{V}$, the sum-of-products matrix of $\boldsymbol{v}$ and a semi-positive diagonal matrix determined by the penalty parameter in the regression model of Y on $\boldsymbol{v}$, respectively.

Then, we prove Theorem 1 by the following steps:

Step 1: Compare $\mathrm{var}(\check{B}_{yv})$ and $\mathrm{var}(\hat{B}_{yv})$

Step 2: Compare $\mathrm{var}(\check{B}^{*}_{mx.c}\hat{B}_{ym.c})$ and $\mathrm{var}(\hat{B}_{mx.c}\hat{B}_{ym.c})$

Step 3: Compare $\mathrm{var}(\check{B}^{*}_{mx.c}\hat{B}_{ym.c})$ and $\mathrm{var}(\check{B}^{*}_{mx.c}\check{B}^{*}_{ym.c})$

[1]Given an active set, LASSO-type estimators can be replaced by ridge-type estimators through the local quadratic approximation (Zou, 2006).

Step 4: Compare $\mathrm{var}(\check{B}^*_{mx.c}\check{B}^*_{ym.c})$ and $\mathrm{var}(\check{\beta}^*_{yx.c})$

Step 1: Comparison between var$(\check{B}_{yv})$ and var$(\hat{B}_{yv})$

We have

$$\begin{aligned}\mathrm{var}(\check{B}_{yv}) - \mathrm{var}(\hat{B}_{yv}) &= \sigma_{yy.v}(S_{vv}+\Gamma)^{-1}S_{vv}(S_{vv}+\Gamma)^{-1} - \sigma_{yy.v}S_{vv}^{-1} \\ &= \sigma_{yy.v}(S_{vv}+\Gamma)^{-1}\left\{S_{vv} - (S_{vv}+\Gamma)S_{vv}^{-1}(S_{vv}+\Gamma)\right\}(S_{vv}+\Gamma)^{-1} \\ &= -\sigma_{yy.v}(S_{vv}+\Gamma)^{-1}\left\{2\Gamma + \Gamma S_{vv}^{-1}\Gamma\right\}(S_{vv}+\Gamma)^{-1}. \end{aligned} \tag{B.19}$$

Since $2\Gamma + \Gamma S_{vv}^{-1}\Gamma$ is a semi-positive definite matrix, if $\boldsymbol{v} = \boldsymbol{x}\cup\boldsymbol{c}$, then we derive

$$\boldsymbol{\omega}^T\mathrm{var}(\check{B}_{yv})\boldsymbol{\omega} \leq \boldsymbol{\omega}^T\mathrm{var}(\hat{B}_{yv})\boldsymbol{\omega}$$

for any q_v-dimensional nonzero vector $\boldsymbol{\omega}$, which leads to

$$\mathrm{var}(\check{\beta}_{yx.c}) \leq \mathrm{var}(\hat{\beta}_{yx.c}). \tag{B.20}$$

Here,

$$\mathrm{var}(\hat{B}_{mx.c}\hat{B}_{ym.c}) \leq \mathrm{var}(\hat{\beta}_{yx.c}) \tag{B.21}$$

was derived in Kuroki and Hayashi (2014, 2016).

Step 2: Comparison between var$(\check{B}^*_{mx.c}\hat{B}_{ym.c})$ and var$(\hat{B}_{mx.c}\hat{B}_{ym.c})$

By applying the result of Step 1 to the relationship between $\check{B}^*_{mx.c}$ and $\hat{B}_{mx.c}$, we have

$$\begin{aligned}&\mathrm{var}(\check{B}^*_{mx.c}\hat{B}_{ym.c}) - \mathrm{var}(\hat{B}_{mx.c}\hat{B}_{ym.c}) \\ =\;& \mathrm{var}(E(\check{B}^*_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) + E(\mathrm{var}(\check{B}^*_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) \\ & -\mathrm{var}(E(\hat{B}_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) - E(\mathrm{var}(\hat{B}_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) \\ =\;& \mathrm{var}(\check{B}^*_{mx.c}B_{ym.c}) + \sigma_{yy.mc}E(\check{B}^*_{mx.c}S_{mm.c}^{-1}\check{B}^{*T}_{mx.c}) \\ & -\mathrm{var}(\hat{B}_{mx.c}B_{ym.c}) - \sigma_{yy.mc}E(\hat{B}_{mx.c}S_{mm.c}^{-1}\hat{B}^{T}_{mx.c}) \\ \leq\;& \sigma_{yy.mc}\left\{E(\check{B}^*_{mx.c}S_{mm.c}^{-1}\check{B}^{*T}_{mx.c}) - E(\hat{B}_{mx.c}S_{mm.c}^{-1}\hat{B}^{T}_{mx.c})\right\}. \end{aligned} \tag{B.22}$$

Here, we have

$$\check{B}^*_{mx.c}S_{mm.c}^{-1}\check{B}^{*T}_{mx.c} - \hat{B}_{mx.c}S_{mm.c}^{-1}\hat{B}^{T}_{mx.c} = S_{xm.c}S_{mm.c}^{-1}S_{mx.c}\left((s_{xx.c}-\gamma)^{-2} - s_{xx.c}^{-2}\right) \leq 0.$$

Referring to equation (B.17), γ is given by

$$\gamma = S_{xc}(S_{cc}+\Lambda_{cc})^{-1}S_{cx}, \tag{B.23}$$

where S_{cc}, $S_{cx} = S_{xc}^T$ and Λ_{cc} are the sum-of-products matrix of $\boldsymbol{c}$, the sum-of-cross-products matrix between $\boldsymbol{c}$ and $\boldsymbol{x}$, and the positive diagonal matrix determined by the penalty parameter in the regression model of $\boldsymbol{M}$ on X and $\boldsymbol{C}$, respectively. This shows that

$$\mathrm{var}(\check{B}^*_{mx.c}\hat{B}_{ym.c}) \leq \mathrm{var}(\hat{B}_{mx.c}\hat{B}_{ym.c}).$$

Step 3: Comparison between var$(\check{B}^*_{mx.c}\hat{B}_{ym.c})$ and var$(\check{B}^*_{mx.c}\check{B}^*_{ym.c})$

By applying the result of Step 1 to the relationship between $\check{B}^*_{ym.c}$ and $\hat{B}_{ym.c}$, we have

$$\begin{aligned}&\mathrm{var}(\check{B}^*_{mx.c}\check{B}^*_{ym.c}) - \mathrm{var}(\check{B}^*_{mx.c}\hat{B}_{ym.c}) \\ =\;& \mathrm{var}(E(\check{B}^*_{mx.c}\check{B}^*_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) + E(\mathrm{var}(\check{B}^*_{mx.c}\check{B}^*_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) \\ & -\mathrm{var}(E(\check{B}^*_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) - E(\mathrm{var}(\check{B}^*_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) \\ \leq\;& \mathrm{var}(E(\check{B}^*_{mx.c}\check{B}^*_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) - \mathrm{var}(E(\check{B}^*_{mx.c}\hat{B}_{ym.c}|\boldsymbol{x},\boldsymbol{c},\boldsymbol{m})) \\ =\;& B_{ym.c}^T\mathrm{var}(\check{B}^*_{mx.c}(I_{q_m,q_m}+S_{mm.c}^{-1}\Gamma)^{-1})B_{ym.c} - B_{ym.c}^T\mathrm{var}(\check{B}^*_{mx.c})B_{ym.c}. \end{aligned} \tag{B.24}$$

Here, $I_{q_m,q_m} + S_{mm.c}^{-1}\Gamma$ is a semipositive definite matrix and

$$\boldsymbol{\omega}^T (I_{q_m,q_m} + S_{mm.c}^{-1}\Gamma)\boldsymbol{\omega} - \boldsymbol{\omega}^T\boldsymbol{\omega} \geq 0 \tag{B.25}$$

holds for any q_m-dimensional nonzero vector $\boldsymbol{\omega}$, which leads to

$$\mathrm{var}(\check{B}_{mx.c}^{*}\check{B}_{ym.c}^{*}) - \mathrm{var}(\check{B}_{mx.c}^{*}\hat{B}_{ym.c}) \leq 0. \tag{B.26}$$

Thus, Steps 1$\sim$3 show that

$$\mathrm{var}(\check{B}_{mx.c}^{*}\check{B}_{ym.c}^{*}) \leq \mathrm{var}(\hat{B}_{mx.c}\hat{B}_{ym.c}) \leq \mathrm{var}(\hat{\beta}_{yx.c})$$

holds approximately.

Step 4: Comparison between var$(\check{B}_{mx.c}^{*}\check{B}_{ym.c}^{*})$ and var$(\check{\beta}_{yx.c}^{*})$

For the optimal semi-positive diagonal matrix Γ to yield $\check{\beta}_{yx.c}^{*}$, which may not be optimal for $\check{B}_{mx.c}^{*}\hat{B}_{ym.c}$, we have

$$\sigma_{yy.xc} = \sigma_{yy.xmc} + B_{ym.xc}^T\Sigma_{mm.xc}B_{ym.xc} = \sigma_{yy.mc} + B_{ym.c}^T\Sigma_{mm.xc}B_{ym.c} \tag{B.27}$$

since X is conditionally independent of Y given $\boldsymbol{M} \cup \boldsymbol{C}$. Thus, from equation (B.23), we have

$$\begin{aligned}
\mathrm{var}(\check{\beta}_{yx.c}^{*}) - \mathrm{var}(\check{B}_{mx.c}^{*}\hat{B}_{ym.c}) &= \sigma_{yy.xc}E\left(\frac{s_{xx.c}}{(s_{xx.c}-\gamma)^2}\right) \\
&\quad - B_{ym.c}^T\mathrm{var}(\check{B}_{mx.c}^{*})B_{ym.c} - \sigma_{yy.mc}E(\check{B}_{mx.c}^{*}S_{mm.c}^{-1}\check{B}_{mx.c}^{*T}) \\
&\geq \sigma_{yy.mc}E\left(\frac{s_{xx.c} - S_{xm.c}S_{mm.c}^{-1}S_{mx.c}}{(s_{xx.c}-\gamma)^2}\right) = \sigma_{yy.mc}E\left(\frac{s_{xx.mc}}{(s_{xx.c}-\gamma)^2}\right) \geq 0.
\end{aligned} \tag{B.28}$$

Thus, together with the results of Step 3, we have

$$\mathrm{var}(\check{\beta}_{yx.c}^{*}) \geq \mathrm{var}(\check{B}_{mx.c}^{*}\hat{B}_{ym.c}) \geq \mathrm{var}(\check{B}_{mx.c}^{*}\check{B}_{ym.c}^{*}) \tag{B.29}$$

approximately.

C Numerical Experiments

In this section, we conduct numerical experiments to compare the performances of LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA, least squares methods, and PCM Selector.

C.1 Loss Functions

Traditional Penalized Regression Analysis

For an q_c-dimensional regression vector $B_{yc.xm}$ and a q_m-dimensional regression vector $B_{ym.xc}$, let $B_y = (\beta_{yx.cm}, B_{yc.xm}^T, B_{ym.xc}^T)^T = (\beta_1, \beta_2, ..., \beta_{q_c+q_m+1})^T$ and $\lambda, \lambda' > 0$. First, the L_1-penalized loss function of adaptive LASSO (Zou, 2006) is defined as

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{x}\beta_{yx.cm} - \boldsymbol{c}B_{yc.xm} - \boldsymbol{m}B_{ym.xc}\|_2^2 + \lambda||\boldsymbol{\gamma} \odot B_y||_1, \tag{C.30}$$

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

$$\boldsymbol{\gamma} = \left(\frac{1}{|\tilde{\beta}_1|^\eta}, \frac{1}{|\tilde{\beta}_2|^\eta}, \ldots, \frac{1}{|\tilde{\beta}_{q_c+q_m+1}|^\eta}\right)^T \tag{C.31}$$

with tuning parameter $\eta \geq 0$, where $\tilde{\beta}_i$, $i = 1, 2, \ldots, q_c + q_m + 1$, is the standard ridge estimator of B_y given a penalty parameter λ' (Hoerl and Kennard, 1970). In particular, equation (C.30) is the L_p-penalized loss function of the standard LASSO (Tibshirani, 1996) when $\eta = 0$.

For $0 \leq \phi \leq 1$, the L_p-penalized loss function of the Elastic Net (Zou, 2006) is given by

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{x}\beta_{yx.cm} - \boldsymbol{c}B_{yc.xm} - \boldsymbol{m}B_{ym.xc}\|_2^2 + \lambda\left((1-\phi)||B_y||_2^2 + \phi||B_y||_1^1\right). \tag{C.32}$$

Partially Adaptive $\mathbf{L}_p$-Regularization Multiple Regression Analysis (PAL$_p$MA)

For an $n \times q_{\bar{z}}$ observation matrix $\bar{\boldsymbol{z}}$ and penalty parameter $\lambda_p \geq 0$, the L_p-penalized loss function of the original PAL$_p$MA (Nanmo and Kuroki, 2022) is given by

$$\frac{1}{2n}||\boldsymbol{y} - \boldsymbol{x}\beta_{yx.c} - \boldsymbol{z}B_{yz.x\bar{z}} - \bar{\boldsymbol{z}}B_{y\bar{z}.xz}||_2^2 + \lambda_p||\boldsymbol{\gamma} \odot B_{y\bar{z}.xz}||_p^p, \quad p = 1, 2 \tag{C.33}$$

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

$$\boldsymbol{\gamma} = \left(\frac{1}{|\tilde{\beta}_{y\bar{z}_1.xc}|^{\eta_p}}, \frac{1}{|\tilde{\beta}_{y\bar{z}_2.xc}|^{\eta_p}}, \ldots, \frac{1}{|\tilde{\beta}_{y\bar{z}_{q_{\bar{z}}}.xc}|^{\eta_p}}\right)^T \tag{C.34}$$

with tuning parameter $\eta_p \geq 0$, where $\tilde{B}_{y\bar{z}.xz}^T = (\tilde{\beta}_{y\bar{z}_1.xc}, \ldots, \tilde{\beta}_{y\bar{z}_{q_{\bar{z}}}.xc})^T$ is derived from

$$\begin{pmatrix} \tilde{\beta}_{yx.z\bar{z}} \\ \tilde{B}_{yz.x\bar{z}} \\ \tilde{B}_{y\bar{z}.xz} \end{pmatrix} = \begin{pmatrix} s_{xx} & S_{xz} & S_{x\bar{z}} \\ S_{zx} & S_{zz} & S_{z\bar{z}} \\ S_{\bar{z}x} & S_{\bar{z}z} & n\lambda I_{q_{\bar{z}}} + S_{\bar{z}\bar{z}} \end{pmatrix}^{-1} \begin{pmatrix} s_{xy} \\ S_{zy} \\ S_{\bar{z}y} \end{pmatrix} \tag{C.35}$$

given a penalty parameter $\lambda > 0$. Letting $\check{B}_y^{\dagger T} = (\check{\beta}_{yx.c}^{\dagger}, \check{B}_{yz.x\bar{z}}^{\dagger T}, \check{B}_{y\bar{z}.xz}^{\dagger T})^T$ be the estimator that minimizes the L_p-penalized loss function (C.33) for $p = 1$, the estimators of the total effect of PAL$_1$MA are defined by correcting the bias term of $\check{\beta}_{yx.c}^{\dagger}$ using $\boldsymbol{\gamma}$. Here, the stability of the estimated $\boldsymbol{\gamma}$ may have an effect on the bias correction. To avoid this difficulty, in the numerical experiments, we apply the following standardized weight vector to $\boldsymbol{\gamma}$ in equation (C.33):

$$\boldsymbol{\gamma}' = \left(\sum_{i=1}^{q_{\bar{z}}} \frac{1}{|\tilde{\beta}_{y\bar{z}_i.xc}|^{\eta_p}}\right)^{-1} \left(\frac{1}{|\tilde{\beta}_{y\bar{z}_1.xc}|^{\eta_p}}, \frac{1}{|\tilde{\beta}_{y\bar{z}_2.xc}|^{\eta_p}}, \ldots, \frac{1}{|\tilde{\beta}_{y\bar{z}_{q_{\bar{z}}}.xc}|^{\eta_p}}\right)^T. \tag{C.36}$$

Then, in the framework of PAL$_1$MA, the total effect is estimated by

$$\check{\beta}_{yx.c}^{*} = \check{\beta}_{yx.c}^{\dagger} - \frac{n\lambda_1}{\tilde{s}_{xx.c}^{\dagger}} \tilde{B}_{x\bar{z}.z}^{\dagger T} \boldsymbol{\gamma} \odot \mathrm{sign}(\check{B}_{y\bar{z}.xz}^{\dagger}), \tag{C.37}$$

$$\tilde{s}_{xx.c}^{\dagger} = ||\boldsymbol{x} - \boldsymbol{z}\tilde{B}_{xz.\bar{z}}^{\dagger} - \bar{\boldsymbol{z}}\tilde{B}_{x\bar{z}.z}^{\dagger}||_2^2 \tag{C.38}$$

where $\tilde{B}_{x\bar{z}.z}^{\dagger}$ and $\tilde{B}_{xz.\bar{z}}^{\dagger}$ are PAL$_2$MA estimators derived from the L_p-penalized loss function with a standardized weight vector $\boldsymbol{\gamma}''$, a penalty parameter $\lambda' \geq 0$ and a tuning parameter $\eta_2 > 0$ such that $\boldsymbol{x}$ and $\boldsymbol{y}$ are replaced by an empty set and $\boldsymbol{x}$, respectively, in equation (C.33) for $p = 2$.

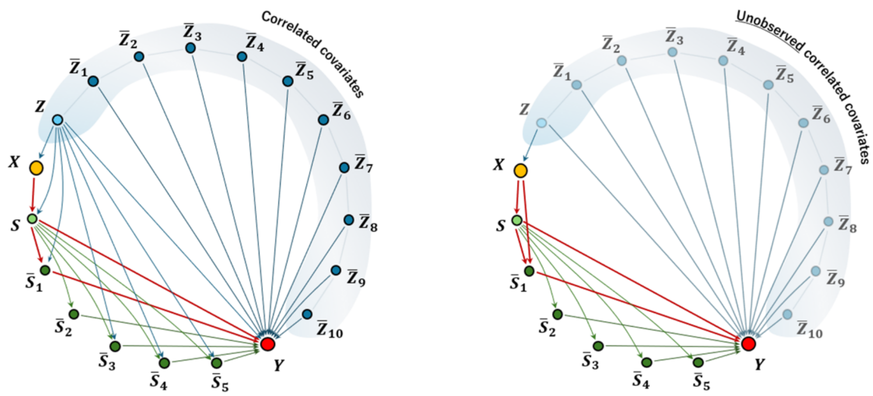
Note that the R package "`glmnet`" (version 4.1.8) (Friedman et al., 2023) is utilized to perform LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA and PCM Selector. All experiments were carried out on an Intel Core i7-1360P CPU running at 2.20 GHz.

C.2 Parameter settings

For simplicity, letting X and Y be the treatment variable and the response variable, respectively, consider linear SCMs with 18 explanatory variables for Y in the form

$$\left.\begin{array}{l} Y = \alpha_{ys}S + \alpha_{yz}Z + \overline{\boldsymbol{S}}A_{y\overline{s}} + \overline{\boldsymbol{Z}}A_{y\overline{z}} + \epsilon_y \\ \overline{\boldsymbol{S}} = XA_{\overline{s}x} + SA_{\overline{s}s} + ZA_{\overline{s}z} + \boldsymbol{\epsilon}_{\overline{s}} \\ S = \alpha_{sx}X + \alpha_{sz}Z + \epsilon_s \\ X = \alpha_{xz}Z + \epsilon_x \end{array}\right\} \tag{C.39}$$

where $\overline{\boldsymbol{S}}$ and $\overline{\boldsymbol{Z}}$ include 5 and 10 variables, respectively. In addition, $A_{y\overline{s}} = (\alpha_{y\overline{s}_1}, \alpha_{y\overline{s}_2}, \ldots, \alpha_{y\overline{s}_5})^T$, $A_{y\overline{z}} = (\alpha_{y\overline{z}_1}, \alpha_{y\overline{z}_2}, \ldots, \alpha_{y\overline{z}_{10}})^T$, $A_{\overline{s}x} = (\alpha_{\overline{s}_1x}, \alpha_{\overline{s}_2x}, \ldots, \alpha_{\overline{s}_5x})$, $A_{\overline{s}s} =$



(a) **(b)**

Fig. A. Causal diagram. The red arrows show the total effect of interest. X: treatment variable; Y: response variable; S: intermediate variable that can be selected using prior causal knowledge; $\overline{\boldsymbol{S}} = \{\overline{S}_1, \overline{S}_2, \ldots, \overline{S}_5\}$: a set of intermediate variables for which it is uncertain which element should be added to evaluate the total effects; Z: covariate that can be selected using prior causal knowledge; $\overline{\boldsymbol{Z}} = \{\overline{Z}_1, \overline{Z}_2, \ldots, \overline{Z}_{10}\}$: a set of covariates for which it is uncertain which element should be added to evaluate the total effects.

$(\alpha_{\overline{s}_1 s}, \alpha_{\overline{s}_2 s}, \ldots, \alpha_{\overline{s}_5 s})$ and $A_{\overline{s}z} = (\alpha_{\overline{s}_1 z}, \alpha_{\overline{s}_2 z}, \ldots, \alpha_{\overline{s}_5 z})$. Fig. A (a) shows that (i) S satisfies the front-door-like criterion relative to (X, Y) with Z and (ii) Z satisfies the back-door criterion relative to (X, Y). Fig. A (b) shows that (i) $\{S, \overline{S}_1\}$ satisfies the front-door criterion relative to (X, Y) and (ii) $\boldsymbol{C} = \{Z, \overline{\boldsymbol{Z}}\}$ satisfies the back-door criterion relative to (X, Y) but is unobserved. Here, S and $\{S, \overline{S}_1\}$ are the minimally sufficient sets of intermediate variables that satisfies the front-door-like criterion for Fig. A (a), and satisfies the front-door criterion for Fig. A (b), respectively.

To construct the population variance-covariance matrix with the linear SCMs (C.39), we first assigned one of 0.1 and 0.8 to α_{xz} and α_{sx} depending on Fig. A (a) and to α_{sx} depending on Fig. A (b). Here, multicollinearity may occur between X and the covariates satisfying the back-door criterion or intermediate variables satisfying the front-door-like criterion when we assign 0.8 to the direct effects on X but may not occur when we assign 0.1 to the direct effects on X. The direct effect α_{ys} was set to 0.4, the direct effects $\alpha_{\overline{s}_1 s}, \alpha_{\overline{s}_2 s}, \ldots, \alpha_{\overline{s}_5 s} (\in A_{\overline{s}s})$ were all set to 0.2, and the direct effects $\alpha_{\overline{s}_2 x}, \alpha_{\overline{s}_3 x}, \ldots, \alpha_{\overline{s}_5 x} (\in A_{\overline{s}x})$ were all set to 0 in the both settings Fig. A (a) and (b). The direct effects $\alpha_{\overline{s}_1 z}, \alpha_{\overline{s}_2 z}, \ldots, \alpha_{\overline{s}_5 z} (\in A_{\overline{s}z})$ were all set to 0.2 in the settings Fig. A (a) and were all set to 0 in the settings Fig. A (b). In addition, the direct effects $\alpha_{y\overline{z}_1}, \alpha_{y\overline{z}_2}, \ldots, \alpha_{y\overline{z}_{10}} (\in A_{y\overline{z}})$ and $\alpha_{y\overline{s}_2}, \alpha_{y\overline{s}_3}, \alpha_{y\overline{s}_4}, \alpha_{y\overline{s}_5} (\in A_{y\overline{s}})$ were randomly and independently generated according to a uniform distribution on the interval $[-0.2, 0.2]$. The other direct effects are given in Table A. In addition, the population variance-covariance

Table A. Direct effects

(a) S satisfies the front-door-like criterion relative to (X, Y) with Z
Z satisfies the back-door criterion relative to (X, Y)

Fig. A (a)	α_{xz}	α_{sx}	$\alpha_{\overline{s}_1 x}$	α_{yz}	α_{sz}	$\alpha_{y\overline{s}_1}$
(a_1)	0.1	0.1	0.0	0.2	0.2	$U([-0.2, 0.2])$
(a_2)	0.1	0.8	0.0	0.2	0.2	$U([-0.2, 0.2])$
(a_3)	0.8	0.1	0.0	0.2	0.2	$U([-0.2, 0.2])$
(a_4)	0.8	0.8	0.0	0.2	0.2	$U([-0.2, 0.2])$

(b) $\{S, \overline{S}_1\}$ satisfies the front-door criterion relative to (X, Y)

Fig. A (b)	α_{xz}	α_{sx}	$\alpha_{\overline{s}_1 x}$	α_{yz}	α_{sz}	$\alpha_{y\overline{s}_1}$
(b_1)	0.2	0.1	0.2	0.0	0.0	0.2
(b_2)	0.2	0.8	0.2	0.0	0.0	0.2

$U([-0.2, 0.2])$: direct effects determined by random numbers from the uniform distribution on the interval $[-0.2, 0.2]$.

matrices of the covariates $\boldsymbol{C}$ were randomly generated using the "`randcorr`" package (available from `https://www.rdocumentation.org/packages/randcorr/versions/1.0/topics/randcor r-package`) according to Pourahmadi and Wang (2015). In addition, we assume that (i) the random disturbances ϵ_x, ϵ_s, $\boldsymbol{\epsilon}_{\overline{s}}$ and ϵ_y independently follow normal distributions with mean 0 and variance or variance-covariance matrix $\sigma_{xx.c}$, $\sigma_{ss.cz}$, $\Sigma_{\overline{ss}.xsz}$, and $\sigma_{yy.cm}$, respectively, and (ii) the random disturbances are independent of their non-descendants. Here, the variances and variance-covariance matrices $\sigma_{xx.c}$, $\sigma_{ss.cz}$, $\Sigma_{\overline{ss}.xsz}$, and $\sigma_{yy.cm}$ are determined to satisfy the criterion that the variance of each variable in the corresponding linear SCM equals one.

Regarding the tuning penalty parameters for LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA, and PCM Selector, the "`glmnet`" package is utilized in this paper. For tuning the penalty parameter in the L_p-penalized loss function, we referred to the search range $(0, \sqrt{\log(q')/n}]$ proposed by Bühlmann and van de Geer (2011), where q' is the number of variables corresponding to the penalized part of regression coefficients. In addition, the search ranges of the other tuning parameters were set to $\zeta_p, \xi_p \in \{0.01, 0.02, \ldots, 0.99\}$ for PCM Selector ($\zeta_p + \xi_p \in [0, 1]$ for $p = 1, 2$) and $\eta, \eta_1, \eta_2 \in \{0.1, 0.2, 0.3, ..., 2.9, 3.0\}$ for adaptive LASSO and PAL$_p$MA. The mixing parameter ϕ of Elastic Net was set to $\phi \in \{0.01, 0.02, 0.03, ..., 0.98, 0.99\}$.

Based on the abovementioned parameter ranges, we conducted all possible selections based on threefold cross-validation to determine the combination of parameters based on the mean squared error. The results of the parameter tuning are shown in Table B. Note that the parameter settings of PAL$_1$MA and PCM Selector in this paper are somewhat empirical; i.e., they may not be determined as optimally as in other penalized regression analyses. The development of optimal parameter tuning for PAL$_1$MA and PCM Selector is left for future work.

Table B. Parameter settings based on threefold cross-validation

Fig. A (a)	LASSO	adaptive LASSO			Elastic Net		PAL$_1$MA			PCM Selector						
	λ	λ	η	λ'	λ	ϕ	λ_1	η_1	λ	λ_1	ζ_1	ξ_1	ρ_1	λ	ρ	τ_{yx}
(a_1)	0.407	0.269	0.100	451.940	0.407	0.910	0.392	1.300	249.858	0.009	0.400	0.030	0.213	3.377	52.653	0.045
(a_2)	0.142	0.151	0.100	5.242	0.142	0.920	0.306	1.200	4.516	0.013	0.610	0.010	0.213	3.107	64.798	0.362
(a_3)	0.407	0.407	0.100	395.298	0.399	0.900	0.294	1.200	308.926	0.017	0.390	0.050	0.213	3.157	69.484	0.045
(a_4)	0.028	0.073	1.900	4.561	0.033	0.930	0.020	0.900	2.070	0.013	0.270	0.190	0.213	3.565	44.102	0.362

Fig. A (b)	PCM Selector						
	λ_1	ζ_1	ξ_1	ρ_1	λ	ρ	τ_{yx}
(b_1)	0.076	0.000	1.000	-	253.515	-	0.085
(b_2)	0.346	0.000	1.000	-	3.726	-	0.402

ρ, ρ_1, λ, λ_1, λ': penalty parameters; η, η_1, ζ_1, ξ_1, ξ_2: tuning parameters; ϕ: mixing parameter; τ_{yx}: total effect of X on Y.

Table C. Parameter settings in replications

Fig. A (a)	adaptive LASSO	PAL$_1$MA				PCM Selector					
	λ'	λ	λ_2	η_2	λ'	λ	ρ	λ_2	ξ_2	ρ_2	ρ_2'
(a_1)	167.620	131.372	0.000	0.000	0.000	128.569	34.981	0.000	0.000	0.000	0.000
(a_2)	127.820	114.470	0.000	0.000	0.000	126.195	33.734	0.000	0.000	0.000	0.000
(a_3)	150.285	130.230	0.000	0.000	0.000	124.013	37.411	0.000	0.000	0.000	0.000
(a_4)	54.846	109.683	0.000	0.000	0.000	116.405	35.220	0.000	0.000	0.030	0.000

Fig. A (b)	PCM Selector					
	λ	ρ	λ_2	ξ_2	ρ_2	ρ_2'
(b_1)	69.487	-	0.000	1.000	0.000	-
(b_2)	70.590	-	0.000	1.000	0.000	-

ρ, ρ_2, ρ_2', λ, λ', λ_2: penalty parameters; η_2, ξ_2: tuning parameters. All parameter values in this table are shown as means in 5000 replications.

C.3 Analysis

For 5000 replications, we generated 15 random samples of 18 variables from a multivariate normal distribution with a zero mean vector and the population variance-covariance matrix generated by the above procedure. Tables D and E show the numerical results by LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA, the OLS method, the TSLS method, and PCM Selector based on Table A. Here, the TSLS methods are based on front-door-like criterion in Setting (a) and based on front-door criterion in Setting (b). In addition, for the OLS and TSLS methods, we select a set of covariates $\boldsymbol{C}$ in Fig. A (a). In Fig. A (b), it is assumed that a set of covariates is not observed, and thus the total effect can not be estimated by using the back-door criterion.

From Figs. B and C and Tables D and E, 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 LASSO, adaptive LASSO, and Elastic Net. This would be serious because it provides such a misleading interpretation that the external intervention of the treatment variable X does not have no effect on the change of Y. In contrast, the coincidence rates of PAL$_1$MA and PCM Selector are still higher than those of the other regression analyses. Here, when the true total effect is far from zero, the coincidence rates are high for each regression analysis.

2. The OLS method provides an unbiased estimator of the total effect through a whole set of covariates that satisfy the back-door criterion, and the TSLS method also provides an unbiased estimator of the total effect through a minimally sufficient set of intermediate variables that satisfy the front-door-like criterion with a whole set of covariates. Given this finding, the estimators from the penalized regression analysis are expected to be close to both the OLS estimate and the TSLS estimate. However, from Tables D and E, the estimates from PAL$_1$MA are close to the OLS estimates, but the estimates from PCM Selector are close to the TSLS estimates (not including x). The difference between the OLS (PAL$_1$MA) estimate and the TSLS (PCM Selector) estimate may be due to the small sample size problem. Here, note that both PCM Selector and PAL$_1$MA provide better estimation accuracy than the OLS and the TSLS methods in most cases.

3. The variances of the estimated total effects from PCM Selector are larger than those from the other traditional penalized regression analyses but smaller than those from OLS and TSLS methods in most cases. In addition, it seems that PAL$_1$MA provides better estimation accuracy than PCM Selector in some cases. This seems to contradict Theorem 1, but it is not, because Theorem 1 is derived under the assumption that PAL$_1$MA and PCM Selector utilize the same set of covariates and the same weight matrix.

4. PCM Selector provides consistent or less biased estimators of the total effect than other regression analyses.

Overall, the coincidence rates between the signs of the estimated total effects and the true total effect from PCM Selector seem equal to or higher than those from the other regression analyses. In addition, in some cases of Fig. A, PCM Selector may not select a set of covariates/intermediate variables that satisfies the front-door-like criterion, and such a missing covariate/intermediate variable may provide biased estimates of the total effects. Then, PCM Selector may reverse the direction of the regression coefficient in such situations. However, regarding PCM Selector, such a drawback is eliminated by selecting smaller values of the penalized parameters based on the whole set of covariates and intermediate variables. That is to say, since a set of covariates and intermediate variables is selected based on the sign of the estimated total effect of X on Y by PCM Selector with the smaller penalized parameter values, we can verify that the lack of sufficient confounders and intermediate variables does not interfere with the qualitative interpretation of the total effects. Thus, PCM Selector and PAL$_1$MA can provide less biased estimators than the other penalized regression analyses in most cases. This indicates that the estimation of the total effect by PCM Selector does not lead to a misleading qualitative interpretation compared to the standard penalized regression analysis.

Finally, we would like to emphasize that most of the current penalized regression analyses, such as LASSO, adaptive LASSO, Elastic Net, and PAL$_p$MA, can not be applied to evaluate the total effects when a set of covariates satisfying the back-door criterion cannot be observed. In contrast, although we discussed the performances of LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA and PCM Selector separately, PCM Selector provides a wider class, including LASSO, adaptive LASSO, and PAL$_p$MA. In addition, PCM Selector is also applicable to some situations where a set of covariates satisfying the back-door criterion cannot be observed.

Table D. Results based on cross-validation.

(a) S satisfies the front-door-like criterion relative to (X, Y) with Z
Z satisfies the back-door criterion relative to (X, Y)

(a_1) $\tau_{yx} = 0.045$

	Mean	SD	Bias	RMSE	Sign
LASSO	0.005	0.034	-0.040	0.053	0.068
adaptive LASSO	0.018	0.087	-0.028	0.091	0.192
Elastic Net	0.006	0.041	-0.039	0.056	0.093
$\mathrm{PAL_1MA}$	0.058	0.306	0.013	0.306	0.578
PCM Selector	0.047	0.338	0.001	0.338	0.566
Front-door-like (including x)	0.029	0.597	-0.016	0.598	0.543
Front-door-like (not including x)	0.042	0.418	-0.003	0.418	0.559
Back-door	0.052	0.633	0.007	0.633	0.551

(a_2) $\tau_{yx} = 0.362$

	Mean	SD	Bias	RMSE	Sign
LASSO	0.250	0.191	-0.112	0.221	0.821
adaptive LASSO	0.256	0.196	-0.106	0.223	0.817
Elastic Net	0.256	0.191	-0.106	0.219	0.833
$\mathrm{PAL_1MA}$	0.417	0.263	0.055	0.269	0.933
PCM Selector	0.409	0.500	0.047	0.503	0.833
Front-door-like (including x)	0.425	2.258	0.063	2.258	0.614
Front-door-like (not including x)	0.419	0.479	0.057	0.483	0.834
Back-door	0.419	0.547	0.057	0.550	0.817

(a_3) $\tau_{yx} = 0.045$

	Mean	SD	Bias	RMSE	Sign
LASSO	0.013	0.045	-0.033	0.056	0.117
adaptive LASSO	0.017	0.057	-0.028	0.063	0.138
Elastic Net	0.017	0.054	-0.028	0.060	0.156
$\mathrm{PAL_1MA}$	0.054	0.792	0.009	0.792	0.528
PCM Selector	0.036	0.718	-0.010	0.718	0.526
Front-door-like (including x)	-0.008	1.577	-0.053	1.578	0.515
Front-door-like (not including x)	0.030	1.051	-0.015	1.051	0.524
Back-door	0.054	1.591	0.009	1.591	0.532

(a_4) $\tau_{yx} = 0.362$

	Mean	SD	Bias	RMSE	Sign
LASSO	0.306	0.350	-0.056	0.355	0.681
adaptive LASSO	0.351	0.719	-0.011	0.719	0.683
Elastic Net	0.301	0.331	-0.061	0.337	0.689
$\mathrm{PAL_1MA}$	0.375	0.796	0.013	0.796	0.703
PCM Selector	0.370	0.952	0.008	0.952	0.677
Front-door-like (including x)	0.390	8.281	0.027	8.281	0.535
Front-door-like (not including x)	0.380	1.187	0.018	1.187	0.660
Back-door	0.380	1.312	0.018	1.312	0.658

Mean: sample mean; SD: standard deviation; Bias: bias between the true value and the sample mean; RMSE: root mean squared error: Sign: coincidence rate between the signs of the true value and the estimates; Front-door-like (including x): the treatment variable X, the intermediate variable S and the set of covariates $\boldsymbol{C}$ are used for the front-door-like criterion; Front-door-like (not including x): intermediate variable S and the set of covariates $\boldsymbol{C}$ are used for the front-door-like criterion; Back-door: the set of covariates $\boldsymbol{C}$ are used for the back-door criterion. τ_{yx} shows true value of total effect.

Table E. Results based on cross-validation.

(b) $\{S, \overline{S}_1\}$ satisfies the front-door criterion relative to (X, Y)
$\boldsymbol{C}$ satisfies the back-door criterion relative to (X, Y)

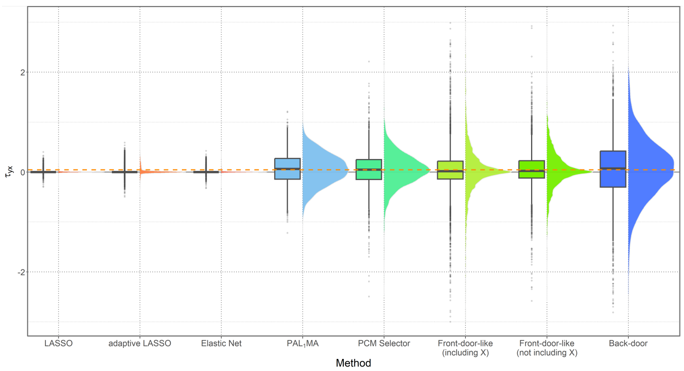
(b_1) $\tau_{yx} = 0.085$

	Mean	SD	Bias	RMSE	Sign
PCM Selector	0.083	0.204	-0.002	0.204	0.666
Front-door (minimal)	0.091	0.172	0.006	0.173	0.713
Front-door (whole)	0.090	0.244	0.004	0.244	0.654

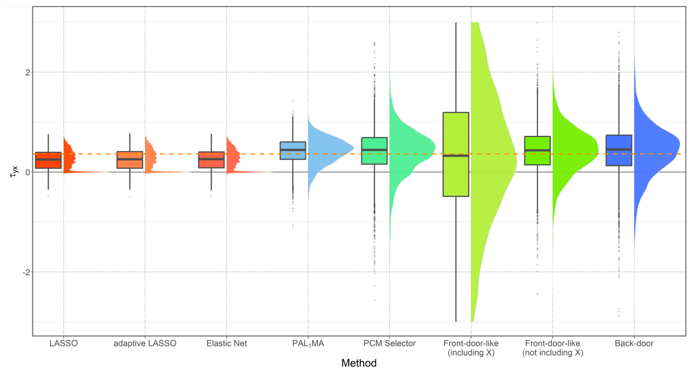
(b_2) $\tau_{yx} = 0.402$

	Mean	SD	Bias	RMSE	Sign
PCM Selector	0.448	0.549	0.046	0.551	0.808
Front-door (minimal)	0.468	0.552	0.066	0.556	0.818
Front-door (whole)	0.462	0.692	0.060	0.694	0.770

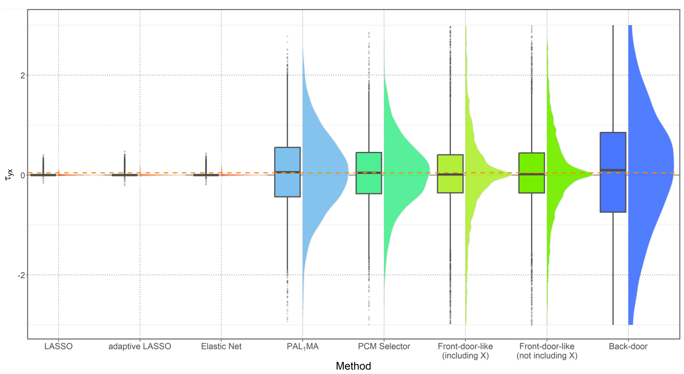
Mean: sample mean; SD: standard deviation; bias: bias between the true value and the sample mean; RMSE: root mean squared error: Sign: coincidence rate between the signs of the true value and the estimates; Front-door (minimal): the minimal subset of intermediate variables $\{S, \overline{S}_1\}$ is used for the front-door criterion. Front-door (whole): the set of intermediate variables $\boldsymbol{M}$ is used for the front-door criterion. τ_{yx} shows the true value of the total effect.



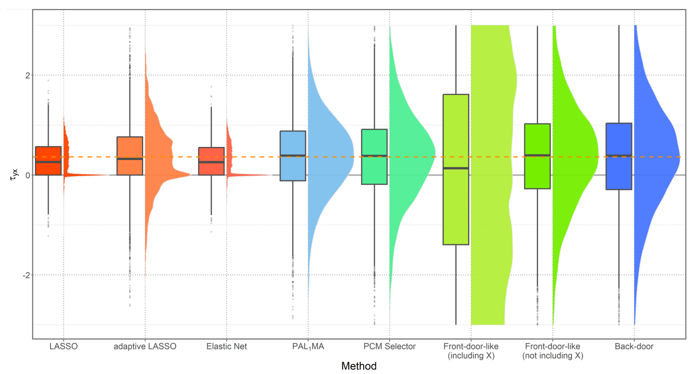
(a_1)



(a_2)



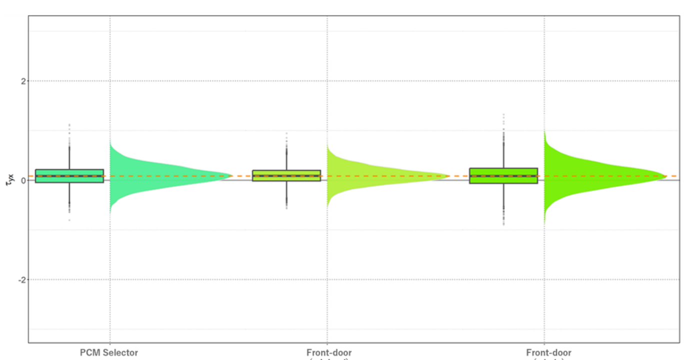
(a_3)



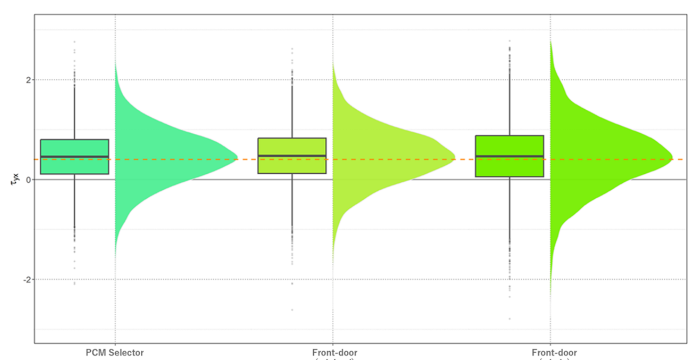
(a_4)

(a) S satisfies the front-door-like criterion relative to (X, Y) with Z
Z satisfies the back-door criterion relative to (X, Y)

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



(b_1)



(b_2)

(b) $\{S, \overline{S}_1\}$ satisfies the front-door criterion relative to (X, Y)
$\boldsymbol{C}$ satisfies the back-door criterion relative to (X, Y)

Fig. C. Violin plots of the estimated total effects based on 5000 replications from the numerical experiments. The dashed lines show the true total effects.

D Application to a Real-World Dataset

D.1 Problem Setting

In this section, we apply LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA, PCM Selector, the OLS method, and the TSLS method to a case study of setting up coating conditions for car bodies, as reported by Okuno et al. (1986) and reanalyzed by Kuroki (2012) and Nanmo and Kuroki (2021).

According to Kuroki (2012), car bodies are coated to improve 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 lead to 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 was 38, and the dataset is available from Okuno et al. (1986). In addition, the observed variables of interest are as follows:

Process conditions

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), paint temperature (X_8), temperature (X_9), and degree of moisture (X_{10})

Response

The transfer efficiency (Y).

Table F shows the randomly selected data from the whole dataset given by Okuno et al. (1986). Note that our discussion is based on Table F and considers a situation where the OLS method with the all-variable selection procedure cannot be applied.

According to Kuroki (2012), there are some differences among these variables in terms of the controllability level: X_1, X_2, ..., X_6 can be controlled (i.e., treatment variables); 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. Here, we assume that the cause-effect relationships in the coating process are as shown in Fig. D. From Fig. D, sets of covariates, including X_{10}, satisfy the back-door criterion relative to (X_1, Y). In addition, X_2, X_7 and $\{X_2, X_7\}$ satisfy the front-door-like criterion relative to (X_1, Y). For details on this case study, refer to Kuroki (2012).

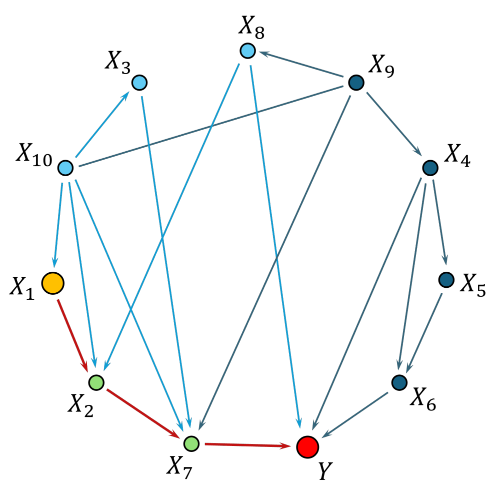


Fig. D. Graphical representation of the case study of setting up coating conditions for car bodies. The red-directed path shows the total effect of interest.

Table F. Randomly selected data from 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	16.7	35.0	4.9	40.0	5.0	3.9	168.0	25.0	20.0	25.0	28.7
2	16.7	28.0	8.3	40.0	2.8	5.0	112.0	32.0	22.0	29.0	19.6
3	33.0	25.5	6.5	40.0	4.0	4.0	276.0	20.0	22.5	25.0	17.8
4	44.0	29.5	6.5	30.0	2.1	5.0	120.0	6.7	7.0	30.0	21.7
5	33.0	28.3	8.3	40.0	2.0	3.0	318.0	20.0	19.0	30.0	22.8
6	44.0	29.5	6.5	30.0	4.9	5.0	180.0	6.7	7.0	30.0	54.8
7	16.7	28.0	8.3	40.0	4.5	1.0	128.0	33.0	10.5	39.0	19.5
8	44.0	24.2	5.0	30.0	2.0	5.0	108.0	28.0	22.5	25.0	19.8
9	16.7	50.0	5.0	40.0	3.0	2.0	112.0	10.0	10.5	39.0	40.2
10	33.0	28.3	6.7	40.0	3.0	5.0	208.0	20.0	19.0	30.0	19.3
11	44.0	25.8	6.7	40.0	4.1	5.0	132.0	22.0	8.2	46.0	13.4
12	16.7	50.0	8.3	40.0	5.0	5.1	112.0	10.0	10.5	39.0	24.0

D.2 Analysis

In this section, we evaluate the total effect of X_1 on Y because similar observations can be derived regarding other treatment variables. In this case study, we assume that $\{X_2, X_7\}$ is a subset of intermediate variables selected according to prior causal knowledge and that $\{X_3, X_8, X_{10}\}$ is a subset of covariates selected according to prior causal knowledge. In contrast, $\{X_4, X_5, X_6, X_9\}$ is a subset for which it is uncertain which element should be selected to evaluate the total effects.

For LASSO, adaptive LASSO, Elastic Net and PAL$_1$MA, $\{X_1, X_3, X_4, X_5, X_6, X_8, X_9, X_{10}\}$ were included as explanatory variables. In particular, the regression coefficients of $\{X_4, X_5, X_6, X_9\}$ are penalized in PAL$_1$MA. In contrast, regarding PCM Selector, $\{X_1, X_3, X_4, X_5, X_6, X_8, X_9, X_{10}\}$ were included as explanatory variables in the L_p-penalized loss function (8) with the response variables $\{X_2, X_7\}$, and the regression coefficients of $\{X_4, X_5, X_6, X_9\}$ were penalized. In addition, $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}\}$ were included as explanatory variables in the L_p-penalized loss function (7) and the regression coefficients of $\{X_1, X_4, X_5, X_6, X_9\}$ were penalized. With respect to the TSLS method based on the front-door-like criterion, we used all intermediate variables and covariates to evaluate the total effect of X_1 on Y. For the OLS method based on the back-door criterion, we also used all covariates to evaluate the total effect of X_1 on Y. Furthermore, to characterize the estimation accuracy, the standard deviations were calculated based on the leave-one-out method.

Table G shows the results obtained by each regression analysis. Here, parameter tuning was conducted by the same procedure as in Section C.2. We also provide the violin plots of the estimated total effect by each regression analysis shown in Fig. E and the solution paths with the selected variables shown in Figs. F and G.

First, according to Okuno et al. (1986), the dilution ratio (X_1) is an important factor that increases both rust protection quality and visual appearance. However, from Table G and Fig. E, the total effect of X_1 on Y is estimated as zero by LASSO, adaptive LASSO, and Elastic Net, which is problematic because it provides such a misleading interpretation that it is not useful to control X_1 to achieve the aim. In contrast, PAL$_1$MA, PCM Selector, the OLS method (with a back-door criterion based on all covariates), and the TSLS method (with the front-door-like criterion (not including X_1) based on all covariates) estimate the total effect of X_1 on Y as a negative value. Here, since X_1 is highly correlated with X_2 and X_4 $\left(\frac{\sigma_{x_1x_2}}{\sqrt{\sigma_{x_1x_1}\sigma_{x_2x_2}}} = -0.593, \frac{\sigma_{x_1x_4}}{\sqrt{\sigma_{x_1x_1}\sigma_{x_4x_4}}} = -0.686\right)$, compared to other correlation relationships between variables, the total effect of X_1 on Y is estimated as a positive value and the estimation accuracy is worse when using the TSLS method with the front-door-like criterion including X_1.

Second, the OLS method provides an unbiased estimator of the total effect through a whole set of covariates that satisfy the back-door criterion, and the TSLS method also provides an unbiased estimator of the total effect through a whole set of intermediate variables that satisfy the front-door-like criterion with a whole set of covariates that satisfy the back-door criterion. Given this finding, it is desirable for the estimators from the penalized regression analysis to be close to both the OLS estimate and the TSLS estimate. From this observation, from Table G and Fig. E, the estimates from PAL$_1$MA are close to the OLS estimates, but the estimates from PCM Selector are close to the TSLS estimates. The difference between the OLS (PAL$_1$MA) estimate and the TSLS (PCM Selector) estimate may be due to the small sample size problem or the model misspecification problem. In fact, Kuroki (2012) applied graphical modeling (Whittaker, 2009) based on some prior causal knowledge to the sample correlation matrix given by Okuno et al. (1986), and selected Fig. D by considering the simplicity ($dev = 34.28$, $df = 36$, p-value$= 0.55$). Here, note that both PCM Selector and PAL$_1$MA provide better estimation accuracy than the OLS and the TSLS methods. The standard deviation from PAL$_1$MA is lower than that from PCM Selector, but this difference seems not to be significant.

Third, from Figs. F and G, LASSO, adaptive LASSO, and Elastic Net select X_8, $\{X_5, X_6, X_8\}$, and $\{X_3, X_4, X_5, X_6, X_8\}$, respectively, which may be difficult to interpret the results from the viewpoint of causal inference because these sets of covariates do not satisfy the back-door criterion. In contrast, PAL$_1$MA selects $\{X_1, X_3, X_8, X_{10}\}$, which satisfies the back-door criterion. PCM Selector also selects $\{X_3, X_8, X_{10}\}$ regarding $\{X_2, X_7\}$; $\{X_2, X_7\}$ satisfies the front-door-like criterion relative to (X_1, Y) with $\{X_3, X_8, X_{10}\}$. This implies that PAL$_1$MA and PCM Selector could help us to interpret the results from the viewpoint of causal inference.

Fourth, as shown in Figs. F and G, LASSO, adaptive LASSO, and Elastic Net estimate the total effect of X_1 on Y as zero with zero standard deviation because X_1 is judged not to be active by these penalized regression analyses. In contrast, the estimated 95% confidence intervals from the OLS method, PAL$_1$MA, and PCM Selector do not include zero. From this observation, it is judged that X_1 has a negative effect on Y by the OLS method, PAL$_1$MA, and PCM Selector, but the hypothesis that X_1 has no effect on Y may not be rejected by LASSO, adaptive LASSO, Elastic Net, or the TSLS method. Here, it seems that PAL$_1$MA provides better estimation accuracy than PCM Selector. This seems to contradict Theorem 1, but it is not, because Theorem 1 is derived under the assumption that PAL$_1$MA and PCM Selector utilize the same set of covariates and the same weight matrix. In the case study, PAL$_1$MA selects the different sets of covariates and different weight matrices from the PCM Selector.

Figs. F and G show the solution paths for the penalty parameter when the other parameters are fixed at the values in Table G. PCM Selector and PAL$_1$MA automatically excluded X_4, X_5, X_6, and X_9, and it is uncertain whether they should be included given the value of the penalty parameter based on cross-validation. However, since cross-validation with datasets split into training and test datasets aims to achieve better prediction accuracy for the response variable and not proper qualitative variable selection, if we prefer to achieve proper qualitative variable selection from a causal inference perspective, the value of the penalty parameter can be selected according to the importance levels of the variables presented by the solution paths. Therefore, the development of optimal parameter tuning for PCM Selector is left for future work.

Table G. Results.

Method	Estimate	SD	Parameters					
			λ_1	ρ_1	η_1	ζ_1	ξ_1	ϕ
LASSO	0.000	0.000	0.416	-	-	-	-	-
adaptive LASSO	0.000	0.000	0.308	-	0.500	-	-	-
Elastic Net	0.000	0.030	0.416	-	-	-	-	0.340
$\mathrm{PAL_1MA}$	-0.250	0.089	0.252	-	0.900	-	-	-
PCM Selector	-0.160	0.117	0.366	0.062	-	0.430	0.000	
Front-door-like (including x)	8.124	0.827	-	-	-	-	-	-
Front-door-like (not including x)	-0.167	0.253	-	-	-	-	-	-
Back-door	-0.268	0.249	-	-	-	-	-	-

Estimate: estimates of the total effect with $n = 12$; SD: standard deviation based on the leave-one-out method; λ_1: penalty parameter for L_1 penalization for the response model; ρ_1: penalty parameter for L_1 penalization for the mediator model; η_1, ζ_1, ξ_1: tuning parameters; ϕ: mixing parameter. The tuning parameters for weight vectors are adaptive LASSO: $\lambda' = 13.960$; $\mathrm{PAL_1MA}$: $\lambda = 99.800$; PCM Selector: $\lambda = 120.105$, $\rho = 0.165$. All tuning parameters for bias correction were set to 0.

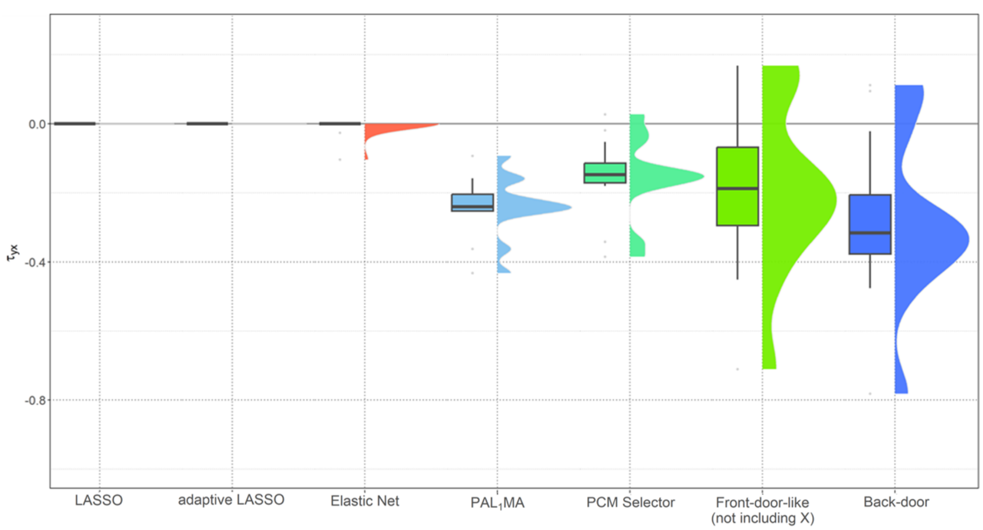


Fig. E. Violin plots of the case study for setting up the coating conditions for car bodies.

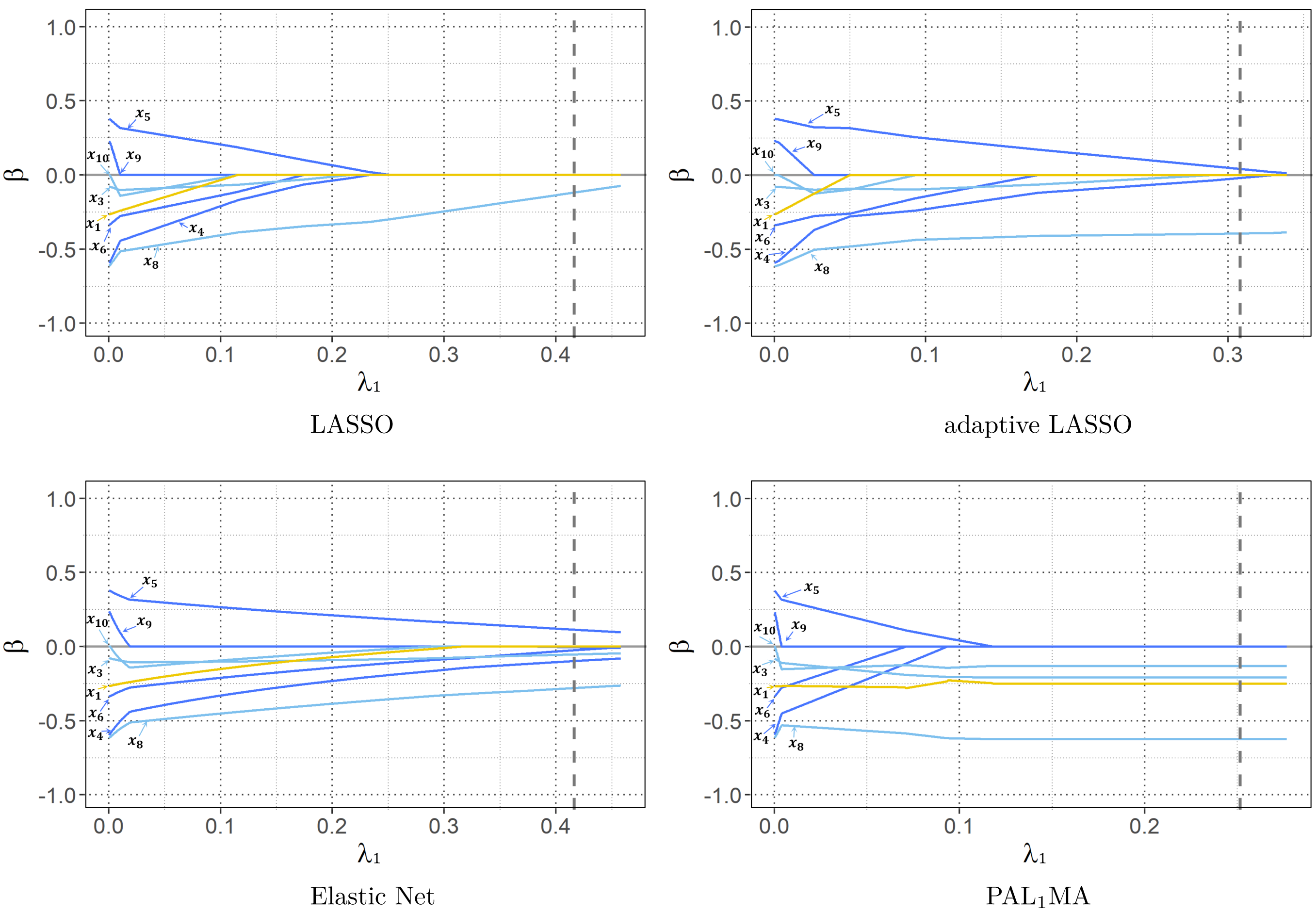
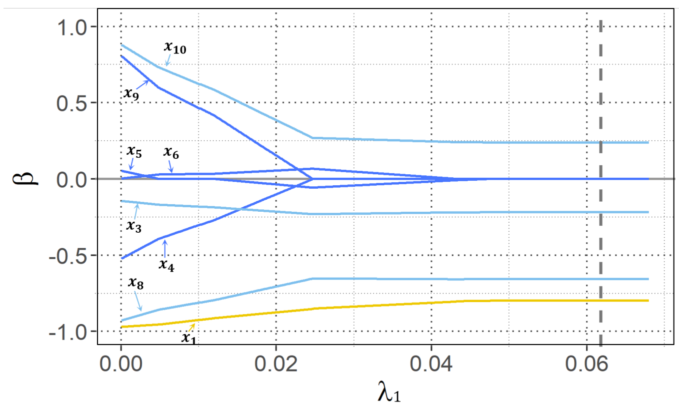
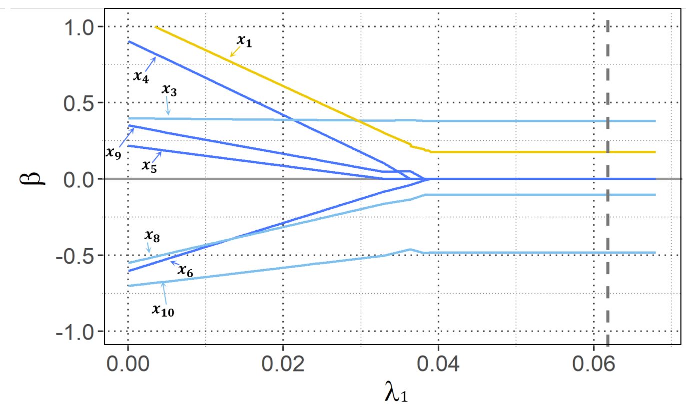


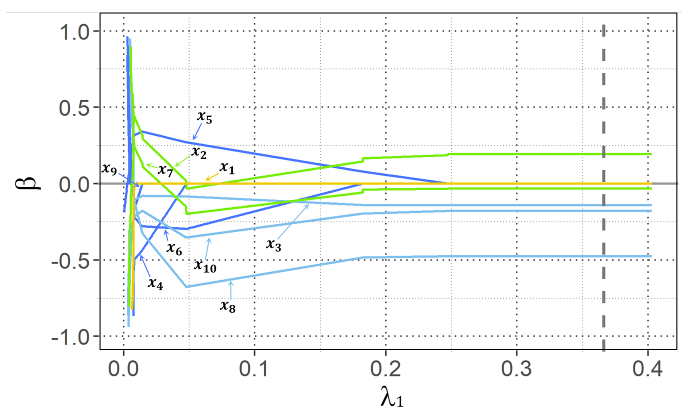
Fig. F. Solution paths for the penalty parameter λ_1 when the other parameters are fixed to the values in Table G. The dashed vertical line represents the values of λ_1 from Table G. The yellow line indicates the regression coefficient of X_1; the light blue line indicates the regression coefficients of covariates $\{X_3, X_8, X_{10}\}$; and the blue line indicates the regression coefficients of covariates $\{X_4, X_5, X_6, X_9\}$.



PCM Selector: X_2



PCM Selector: X_7



PCM Selector: Y

Fig. G. Solution paths for the penalty parameter λ_1 when the other parameters are fixed to the values in Table G. "PCM Selector: X_2" shows the regression coefficients with the response variable X_2 at the first stage, "PCM Selector: X_7" shows the regression coefficients with the response variable X_7 at the first stage, and "PCM Selector: Y" shows the regression coefficients with the response variable Y at the second stage. The dashed vertical line represents the value of λ_1 from Table G. The yellow line indicates the regression coefficient of X_1; the light blue line indicates the regression coefficients of covariates $\{X_3, X_8, X_{10}\}$; the blue line indicates the regression coefficients of covariates $\{X_4, X_5, X_6, X_9\}$; and the green line indicates the regression coefficients of intermediate variables $\{X_2, X_7\}$.

References

Bernstein, D. S. (2009). *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press.

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

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2023). Package 'glmnet'. *CRAN R Repositary*, 595.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

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

Kuroki, M. and Hayashi, T. (2014). On the estimation accuracy of the total effect using intermediate characteristics. *Journal of The Japanese Society for Quality Control*, 44(4):429–440.

Kuroki, M. and Hayashi, T. (2016). On the estimation accuracy of causal effects using supplementary variables. *Scandinavian Journal of Statistics*, 43(2):505–519.

Nanmo, H. and Kuroki, M. (2021). Exact variance formula for the estimated mean outcome with external intervention based on the front-door criterion in gaussian linear structural equation models. *Journal of Multivariate Analysis*, 185:104766.

Nanmo, H. and Kuroki, M. (2022). Partially adaptive regularized multiple regression analysis for estimating linear causal effects. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, pages 1456–1465.

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

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

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

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

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