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

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

Abstract

For a data-generating process for random variables that can be described with a linear structural equation model, we consider a situation in which (i) a set of covariates satisfying the back-door criterion cannot be observed or (ii) such a set can be observed, but standard statistical estimation methods cannot be applied to estimate causal effects because of multicollinearity/high-dimensional data problems. We propose a novel two-stage penalized regression approach, the penalized covariate-mediator selection operator (PCM Selector), to estimate the causal effects in such scenarios. Unlike existing penalized regression analyses, when a set of intermediate variables is available, PCM Selector provides a consistent or less biased estimator of the causal effect. In addition, PCM Selector provides a variable selection procedure for intermediate variables to obtain better estimation accuracy of the causal effects than does the back-door criterion.

Technical Appendix —
https://doi.org/10.48550/arXiv.2412.18180

Introduction

Background

Auxiliary variables are those that are not considered to be of interest in themselves but help us to evaluate causal effects and/or understand the data-generating process in practical studies. For example, an intermediate variable is often considered an auxiliary variable because it is used to evaluate causal effects (Pearl 2001, 2009), to understand the data-generating process in the context of mediation analysis (Baron and Kenny 1986; Imai et al. 2011; Mackinnon 2008) and to improve the estimation accuracy of causal effects (Cox 1960; Hayashi and Kuroki 2014).

In the context of linear structural equation models, this paper focuses on estimating causal effects using intermediate variables. For cases in which the data-generating process for random variables can be described by nonparametric structural equation models and the corresponding directed acyclic graph, Pearl (2009) provided the front-door criterion as the identification condition for causal effects based on intermediate variables. In addition, in the framework of linear structural equation models, Kuroki (2000), Nanmo and Kuroki (2021), and Kuroki and Tezuka (2023) formulated the exact variance of causal effects based on the front-door criterion. Furthermore, Kuroki and Cai (2004), Hui and Zhongguo (2008), and Ramsahai (2012) compared some identification conditions in terms of the asymptotic estimation accuracy of causal effects. On the other hand, under the assumption that a treatment variable is associated with a response variable through a univariate intermediate variable, from the viewpoint of the asymptotic estimation accuracy, Cox (1960) showed that the estimation accuracy of the regression coefficient of the treatment variable on the response variable in the single linear regression model can be improved by using a joint linear regression model based on the response variable and the intermediate variable. In addition, Kuroki and Hayashi (2014) and Hayashi and Kuroki (2014) derived the same results as Cox (1960) in terms of the exact variance of causal effects. Gupta, Lipton, and Childers (2021) derived the same results as Kuroki and Hayashi (2014) and Hayashi and Kuroki (2014) for cases in which a multivariate intermediate variable is available.

In existing studies, it is noted that causal effects can be estimated by standard statistical estimation methods, e.g., the maximum likelihood estimation (MLE) method and the ordinary least squares (OLS) method. Thus, many covariates affect both the treatment variable and the response variable and are highly correlated with each other in reality. This situation leads to a multicollinearity problem, which decreases the estimation accuracy of the causal effects and leads to the formulation of an unreliable plan that prevents us from conducting appropriate policy decision-making. In addition, when the sample size is smaller than the number of explanatory variables in the regression analysis, high-dimensional data analysis also suffers from multicollinearity problems, which cause overfitting and interfere with obtaining admissible solutions for regression coefficients. Recently, due to the development of technological advances in collecting data with many variables to better understand a given phenomenon of interest, the multicollinearity problem has become serious in many domains. To overcome this difficulty, numerous kinds of variable selection techniques based on penalized regression analysis, e.g., the least absolute shrinkage and selection operator (LASSO), adaptive LASSO, and Elastic Net, have been proposed by many statistical and AI researchers and practitioners (Bühlmann and van de Geer 2011; Efron et al. 2004; Tibshirani 1996; Van et al. 2014; Zou 2006; Zou and Hastie 2005). However, the present countermeasures against the multicollinearity problem are formulated independently of the problem of identifying causal effects. Thus, although stable results of regression analysis may be derived by these countermeasures from the viewpoint of prediction, they may yield a seriously biased estimate of the causal effect. Nanmo and Kuroki (2022) proposed partially adaptive L_p-penalized multiple regression analysis (PAL$_p$MA) based on the back-door criterion to overcome these drawbacks. However, because of the formulation of PAL$_p$MA, this method is not applicable to situations where a sufficient set of confounders is not available. In addition, PAL$_p$MA selects a set of covariates to derive a consistent or less biased estimator of causal effects but does not consider the estimation accuracy of the causal effects.

Contributions

For cases in which the data-generating process for random variables can be described with a linear structural equation model, we consider a situation where (i) a set of covariates satisfying the back-door criterion cannot be observed or (ii) such a set can be observed, but standard statistical estimation methods cannot be applied to estimate causal effects because of the multicollinearity/high-dimensional data problem. Then, we propose a novel two-stage penalized regression approach, the penalized covariate-mediator selection operator (PCM Selector), to estimate causal effects. In addition to the desirable properties of PAL$_p$MA, PCM Selector also has the following properties:

(i) Cox (1960) noted that introducing intermediate variables enables us to improve the estimation accuracy of the regression coefficients in some situations. However, Cox's consideration was not used in formulating PAL$_p$MA, LASSO, and other penalized regression analyses. In contrast, based on Cox's consideration, PCM Selector selects covariates and intermediate variables to evaluate the causal effects with better estimation accuracy than PAL$_p$MA and other penalized regression analyses.

(ii) PCM Selector without intermediate variables is consistent with PAL$_p$MA. In this sense, PCM Selector is considered a generalization of PAL$_p$MA, and thus provides a wider class including LASSO and adaptive LASSO. In addition, to our knowledge, there has been much less discussion of the selection problem for intermediate variables in the context of penalized regression analysis. In contrast, PCM Selector selects intermediate variables in the context of penalized regression analysis.

From these properties, PCM Selector contributes to solving the multicollinearity/high-dimensional data problems of evaluating causal effects in statistical causal inference. Given the space constraints, the proofs, several numerical experiments, and a case study are provided in the Technical Appendix.

Linear Structural Causal Model

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

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

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

where $pa(V_i)$ denotes a set of parents of V_i in DAG G and random disturbances $\epsilon_{v_1}, \epsilon_{v_2}, \ldots, \epsilon_{v_{q_v}}$ are assumed to be independently distributed with mean 0 and constant variance. In addition, μ_{v_i} is an intercept, and $\alpha_{v_i v_j}(\neq 0)$ is called a direct effect of V_j on V_i $(i, j = 1, 2, \ldots, q_v\,;\, i \neq j)$. Then, equation (1) is called a linear structural causal model (linear SCM) in this paper.

The linear SCM is a parametric version of Pearl's nonparametric structural causal model (PCM).

To proceed with our discussion, we define some notation. For univariate variables X and Y and a set of variables $\boldsymbol{Z}$, let $\sigma_{xy.z}$ and $\sigma_{xx.z}$ be the conditional covariance between X and Y given $\boldsymbol{Z} = \boldsymbol{z}$ and the conditional variance of X given $\boldsymbol{Z} = \boldsymbol{z}$, respectively. Then, the regression coefficient of X in the single linear regression model of Y on X and $\boldsymbol{Z}$ is denoted by $\beta_{yx.z} = \sigma_{xy.z}/\sigma_{xx.z}$. For sets of variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ ($\boldsymbol{Y}$ can be univariate), let $\Sigma_{xy.z}$ and $\Sigma_{xx.z}$ be the conditional cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z} = \boldsymbol{z}$ and the conditional variance-covariance matrix of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$, respectively. Then, the regression coefficient vector of $\boldsymbol{X}$ in the (single/joint) linear regression model of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$ is denoted by $B_{yx.z} = \Sigma_{xx.z}^{-1}\Sigma_{xy.z}$. In particular, for univariate Y and $\boldsymbol{X} = \{X_1, X_2, ..., X_{q_x}\}$, the i-th element of $B_{yx.z}$ is denoted by $\beta_{yx_i.xz}$ for $i = 1, 2, \cdots, q_x$. For univariate X and $\boldsymbol{Y} = \{Y_1, Y_2, ..., Y_{q_y}\}$, the i-th element of $B_{yx.z}$ is denoted by $\beta_{y_i x.z}$ for $i = 1, 2, \cdots, q_y$. The set of variables $\boldsymbol{Z}$ is omitted from these arguments if it is an empty set. A similar notation is used for the remaining statistical parameters.

The main purpose of this paper is to estimate the total effects from observed data in the context of linear SCMs. The total effect τ_{yx} of X on Y is defined as the total sum of the products of the direct effects on the sequence of directed arrows along all the directed paths from X to Y. To achieve our aim, we introduce the back-door and front-door-like criteria (Pearl 2009) as the representative identification conditions for the total effects. Here, when causal effects, such as direct, indirect, and total effects, can be determined uniquely from the variance/covariance parameters of observed variables, they are said to be identifiable; that is, they can be estimated consistently. Note that direct and indirect effects are also known as representative causal effects in the context of the linear SCM. However, we are concerned with the evaluation of the total effects using intermediate variables because (i) the direct effect can be discussed in the framework of PAL$_p$MA (Nanmo and Kuroki 2021) through the "single-door criterion" (Pearl 2009), and PCM Selector is a generalization of PAL$_p$MA, and (ii) the problem of evaluating the indirect effects is within the scope of PCM Selector in some situations. Here, the indirect effect of X on Y is defined as the sum of the products of the direct effects on the sequence of directed arrows along the directed paths of interest from X to Y, excluding the direct effect of X on Y.

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

(i) No vertex in $\boldsymbol{Z}$ is a descendant of X; and

(ii) $\boldsymbol{Z}$ d-separates X from Y in the DAG obtained by deleting all the directed arrows emerging from X from the DAG G.

If a set $\boldsymbol{Z}$ of observed variables satisfies the back-door criterion relative to (X, Y) in a causal diagram G, then the total effect τ_{yx} is identifiable and is given by the formula $\beta_{yx.z}$ (Pearl 2009). As seen from Rule 2 (Action/observation exchange) of do-calculus (Pearl 2009), note that X and Y of Definition 2 can be generalized to sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Here, a covariate is defined as an element of the nondescendants of X and Y. In addition, a set of covariates is called a sufficient set of confounders if it satisfies the back-door criterion; otherwise, it is called an insufficient set of confounders.

Definition 3 *(Front-Door-Like Criterion) Let $\{X, Y\}$, $\boldsymbol{S}$, $\boldsymbol{Z}_1 \cup \boldsymbol{Z}_2$ be disjoint subsets of $\boldsymbol{V}$ in the DAG G, where X is a nondescendant of Y. If a set $\boldsymbol{S}$ of vertices satisfies the following conditions relative to an ordered pair (X, Y) together with $\boldsymbol{Z}_1 \cup \boldsymbol{Z}_2$, then $\boldsymbol{S}$ is said to satisfy the front-door-like criterion relative to (X, Y) with $\boldsymbol{Z}_1 \cup \boldsymbol{Z}_2$.*

(i) $\boldsymbol{S}$ intercepts all the directed paths from X to Y;

(ii) $\boldsymbol{Z}_1$ satisfies the back-door criterion relative to $(X, \boldsymbol{S})$; and

(iii) $\boldsymbol{Z}_2 \cup \{X\}$ satisfies the back-door criterion relative to $(\boldsymbol{S}, Y)$.

If a set $\boldsymbol{S}$ of observed variables satisfies the front-door-like criterion relative to (X, Y) with $\boldsymbol{Z}_1 \cup \boldsymbol{Z}_2$ in a causal diagram G, then the total effect τ_{yx} is identifiable and is given by the formula $B_{sx.z_1} B_{ys.xz_2}$. The front-door-like criterion is considered an extended version of the front-door criterion (Pearl 2009) since it is consistent with the front-door criterion when $\boldsymbol{Z}_1 \cup \boldsymbol{Z}_2$ is empty.

Here, an intermediate variable relative to (X, Y) is defined as one that is a descendant of X and an ancestor of Y simultaneously. In addition, a set of intermediate variables is called a sufficient set if it satisfies the front-door-like criterion; otherwise, it is called an insufficient set of intermediate variables.

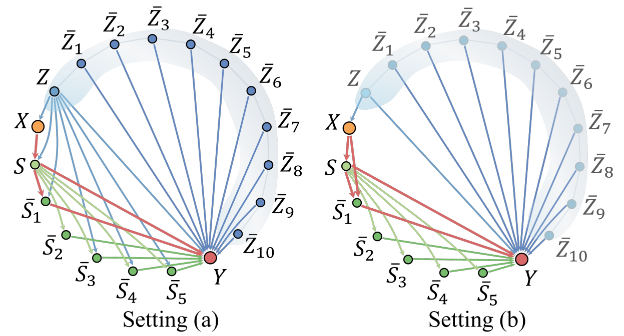

Figure 1: Causal diagram. The thick red arrows show the total effect of interest. X: treatment variable; Y: response variable; S: intermediate variable that can be selected using prior causal knowledge; $\overline{\boldsymbol{S}} = \{\overline{S}_1, \ldots, \overline{S}_5\}$: a set of intermediate variables 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; $\overline{\boldsymbol{Z}} = \{\overline{Z}_1, \ldots, \overline{Z}_{10}\}$: a set of covariates for which it is uncertain which element should be added to evaluate the total effects.

PCM Selector

Problem Setting

In this paper, we partition a set of observed variables into the following three disjoint sets:

(i) $\{X, Y\}$: X and Y are the treatment and response variables, respectively.

(ii) $\boldsymbol{C} = \boldsymbol{Z} \cup \overline{\boldsymbol{Z}}$ ('C' for covariates): a set of covariates satisfying the back-door criterion relative to (X, Y) ($\boldsymbol{Z} \cap \overline{\boldsymbol{Z}}$ is empty), where $\boldsymbol{Z}$ and $\overline{\boldsymbol{Z}}$ are the first q_z components and the next $q_{\overline{z}}$ components of $\boldsymbol{C}$, respectively. Here, $\boldsymbol{Z}$ is a subset including some covariates selected using prior causal knowledge ($\boldsymbol{Z}$ may be an empty set, a sufficient set of confounders, or an insufficient set of confounders), but $\overline{\boldsymbol{Z}}$ is a subset of covariates for which it is uncertain which element of $\overline{\boldsymbol{Z}}$ should be added to evaluate the total effects.

(iii) $\boldsymbol{M} = \boldsymbol{S} \cup \overline{\boldsymbol{S}}$ ('M' for intermediate variables): a set of intermediate variables satisfying the front-door-like criterion relative to (X, Y) with $\boldsymbol{C}$ ($\boldsymbol{S} \cap \overline{\boldsymbol{S}}$ is empty), where $\boldsymbol{S}$ and $\overline{\boldsymbol{S}}$ are the first q_s components and the next $q_{\overline{s}}$ components of $\boldsymbol{M}$, respectively. Here, $\boldsymbol{S}$ is a subset including some intermediate variables selected using prior causal knowledge ($\boldsymbol{S}$ may be an empty set, a sufficient set of intermediate variables, or an insufficient set of intermediate variables), but $\overline{\boldsymbol{S}}$ is a subset for which it is uncertain which element of $\overline{\boldsymbol{S}}$ should be added to evaluate the total effects.

Then, for sample size n, consider the following joint linear regression model of $\{Y\} \cup \boldsymbol{M}$:

$$\boldsymbol{y} = \boldsymbol{x}\beta_{yx.cm} + \boldsymbol{c}B_{yc.xm} + \boldsymbol{m}B_{ym.xc} + \boldsymbol{\epsilon}_{y.xcm}, \quad (2)$$

$$\boldsymbol{m} = \boldsymbol{x}B_{mx.c} + \boldsymbol{c}B_{mc.x} + \boldsymbol{\epsilon}_{m.xc}, \quad (3)$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ represent n-dimensional observation vectors of X and Y, respectively. $\boldsymbol{c}$ and $\boldsymbol{m}$ are an $n \times (q_z + q_{\overline{z}})$ observation matrix of $\boldsymbol{C}$ and an $n \times (q_s + q_{\overline{s}})$ observation matrix of $\boldsymbol{M}$, respectively. Here, $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{c}$ and $\boldsymbol{m}$ are standardized to sample mean 0 and sample variance 1 in advance. In addition, we assume that the elements of the random error vector $\boldsymbol{\epsilon}_{y.xcm}$ are independent and identically distributed with mean 0 and finite variance $\sigma_{yy.xcm}$. Furthermore, the column vectors of the random error matrix $\boldsymbol{\epsilon}_{m.xc}$ are independent and identically distributed with zero mean vector and variance-covariance matrix $\Sigma_{mm.xc}$ for $M \in \boldsymbol{M}$ and are also independent of the elements of $\boldsymbol{\epsilon}_{y.xcm}$.

Under the above setting, this paper focuses on situations where the sum-of-squares matrix of $\{X\} \cup \boldsymbol{S} \cup \boldsymbol{Z}$ is invertible but that of $\{X\} \cup \boldsymbol{C} \cup \boldsymbol{M}$ is not; this is because if it is invertible, then the total effect is estimable by the OLS method (Pearl 2009).

Estimator

For univariates X and Y and a set of variables $\boldsymbol{Z}$, let $s_{xx.z}$ and $s_{xy.z}$ be the sum-of-squares of X given $\boldsymbol{Z}$ and the sum of cross-products between X and Y given $\boldsymbol{Z}$, respectively. In addition, for sets of variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ ($\boldsymbol{Y}$ can be univariate), let $S_{xx.z}$ and $S_{xy.z}$ be the sum-of-squares matrix of $\boldsymbol{X}$ given $\boldsymbol{Z}$ and the sum-of-cross-products matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$, respectively. Here, the set of variables $\boldsymbol{Z}$ is omitted from these arguments if it is an empty set. A similar notation is used for the remaining sums of squares/cross-products. Furthermore, $\boldsymbol{0}_q$, $\boldsymbol{0}_{q,r}$, $\boldsymbol{1}_q$ and I_q are a q-dimensional zero vector, a $q \times r$ zero matrix, a q-dimensional one vector, and a $q \times q$ identity matrix, respectively.

Then, the proposed penalized regression approach, PCM Selector, is formulated as follows:

First, when the sum-of-squares matrix of $\{X\} \cup \boldsymbol{C} \cup \boldsymbol{M}$ is invertible, let

$$\hat{\beta}_{yx.cm} = s_{xy.cm}/s_{xx.cm}, \quad \hat{B}_{yc.xm} = S^{-1}_{cc.xm}S_{cy.xm}, \quad \hat{B}_{ym.xc} = S^{-1}_{mm.xc}S_{my.xc}, \quad (4)$$

and when the sum-of-squares matrix of $\{X\} \cup \boldsymbol{C} \cup \boldsymbol{M}$ is not invertible, let

$$\left(\tilde{\beta}_{yx.cm}, \tilde{B}_{ys.xc\overline{s}}, \tilde{B}_{yz.xm\overline{z}}, \tilde{B}_{y\overline{s}.xcs}, \tilde{B}_{y\overline{z}.xmz}\right)^T = \begin{pmatrix} n\lambda + s_{xx} & S_{xs} & S_{xz} & S_{x\overline{s}} & S_{x\overline{z}} \\ S_{sx} & S_{ss} & S_{sz} & S_{s\overline{s}} & S_{s\overline{z}} \\ S_{zx} & S_{zs} & S_{zz} & S_{z\overline{s}} & S_{z\overline{z}} \\ S_{\overline{s}x} & S_{\overline{s}s} & S_{\overline{s}z} & n\lambda I_{q_{\overline{s}}} + S_{\overline{s}\overline{s}} & S_{\overline{s}\overline{z}} \\ S_{\overline{z}x} & S_{\overline{z}s} & S_{\overline{z}z} & S_{\overline{z}\overline{s}} & n\lambda I_{q_{\overline{z}}} + S_{\overline{z}\overline{z}} \end{pmatrix}^{-1} \times (s_{xy}, S_{sy}, S_{zy}, S_{\overline{s}y}, S_{\overline{z}y})^T \quad (5)$$

for the penalty parameter $\lambda > 0$, where $S_{yx.z} = S^T_{xy.z}$ and the superscript "T" represents the transposed vector/matrix. Here, equation (5) is consistent with equation (4) for $\lambda = 0$. In addition, when the sum-of-squares matrix of $\{X\} \cup \boldsymbol{C}$ is not invertible, let

$$\begin{pmatrix} \tilde{B}_{mx.c} \\ \tilde{B}_{mz.x\overline{z}} \\ \tilde{B}_{m\overline{z}.xz} \end{pmatrix} = \begin{pmatrix} s_{xx} & S_{xz} & S_{x\overline{z}} \\ S_{zx} & S_{zz} & S_{z\overline{z}} \\ S_{\overline{z}x} & S_{\overline{z}z} & n\rho I_{q_{\overline{z}}} + S_{\overline{z}\overline{z}} \end{pmatrix}^{-1} \begin{pmatrix} S_{xm} \\ S_{zm} \\ S_{\overline{z}m} \end{pmatrix} \quad (6)$$

for the penalty parameter $\rho > 0$. For $p = 1, 2$, consider the L_p-penalized loss function

$$\begin{aligned} &L_p(\beta_{yx.cm}, B_{yc.xm}, B_{ym.xc}) \\ &= \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{x}\beta_{yx.cm} - \boldsymbol{c}B_{yc.xm} - \boldsymbol{m}B_{ym.xc}\|_2^2 \\ &\quad + \lambda_p \left(\zeta_p\|\beta_{yx.cm}\|_p^p + \xi_p\|\boldsymbol{\gamma}_{\overline{s}x.c} \odot B_{y\overline{s}.xcs}\|_p^p \right. \\ &\quad \left. + (1 - \zeta_p - \xi_p)\|\boldsymbol{\gamma}_{y\overline{z}.xmz} \odot B_{y\overline{z}.xmz}\|_p^p\right) \quad (7) \end{aligned}$$

for the tuning parameters $\zeta_p \geq 0$ and $\xi_p \geq 0$ such that $\zeta_p + \xi_p \in [0, 1]$, the penalty parameter λ_p corresponding to the L_p norm ($\lambda_p \geq 0$), and the multivariate response type L_p-penalized loss function

$$\begin{aligned} L_p(B_{mx.c}, B_{mc.x}) &= \frac{1}{2n}\|\boldsymbol{m} - \boldsymbol{x}B_{mx.c} - \boldsymbol{c}B_{mc.x}\|_F^2 \\ &\quad + \rho_p\|\mathrm{vec}\left(\boldsymbol{\gamma}_{m\overline{z}.xz} \odot B_{m\overline{z}.xz}\right)\|_p^p \quad (8) \end{aligned}$$

for the penalty parameter ρ_p corresponding to the L_p norm ($\rho_p \geq 0$). Here, $\odot$, $\|\cdot\|_p^p$, and $\|\cdot\|_F$ refer to the Hadamard product, the L_p norm, and the Frobenius norm, respectively. In addition, for $\overline{s}_i \in \overline{\boldsymbol{S}}\,(i = 1, 2, ..., q_{\overline{s}})$, $\overline{z}_i \in \overline{\boldsymbol{Z}}\,(i = 1, 2, ..., q_{\overline{z}})$ and $m_i \in \boldsymbol{M}\,(i = 1, 2, ..., q_m)$, the standardized weight vectors $\boldsymbol{\gamma}_{\overline{s}x.c}$ and $\boldsymbol{\gamma}_{y\overline{z}.xmz}$ and the standardized weight matrix $\boldsymbol{\gamma}_{m\overline{z}.xz}$ are given by

$$\boldsymbol{\gamma}_{\overline{s}x.c} = \left(\sum_{i=1}^{q_{\overline{s}}} \frac{1}{|\tilde{\beta}_{\overline{s}_i x.c}|}\right)^{-1} \times \left(\frac{1}{|\tilde{\beta}_{\overline{s}_1 x.c}|}, \frac{1}{|\tilde{\beta}_{\overline{s}_2 x.c}|}, \ldots, \frac{1}{|\tilde{\beta}_{\overline{s}_{q_{\overline{s}}} x.c}|}\right)^T, \quad (9)$$

$$\boldsymbol{\gamma}_{y\overline{z}.xmz} = \left(\sum_{i=1}^{q_{\overline{z}}} \frac{1}{|\tilde{\beta}_{y\overline{z}_i.xmc}|}\right)^{-1} \times \left(\frac{1}{|\tilde{\beta}_{y\overline{z}_1.xmc}|}, \frac{1}{|\tilde{\beta}_{y\overline{z}_2.xmc}|}, \ldots, \frac{1}{|\tilde{\beta}_{y\overline{z}_{q_{\overline{z}}}.xmc}|}\right)^T, \quad (10)$$

$$\boldsymbol{\gamma}_{m\overline{z}.xz} = \left[\left(\sum_{k=1}^{q_{\overline{z}}}\sum_{\ell=1}^{q_m} \frac{1}{|\tilde{\beta}_{m_\ell \overline{z}_k.xc}|}\right)^{-1} \frac{1}{|\tilde{\beta}_{m_j \overline{z}_i.xc}|}\right]_{1 \leq i \leq q_{\overline{z}}, 1 \leq j \leq q_m} \quad (11)$$

respectively, where $|\cdot|$ refers to the absolute value, and the vec operator, $\mathrm{vec}(A)$, denotes the vectorization of an $q \times r$ matrix A, which is the $q \times r$-dimensional vector obtained by stacking the columns of matrix A on top of one another. Equation (7) is different from the standard penalized loss function in the following ways:

(i) The penalty parameter λ_p is not assigned to $B_{yz.xm\overline{z}}$ and $B_{ys.xc\overline{s}}$ in equation (7) in order not to remove covariates ($\boldsymbol{Z}$) and intermediate variables ($\boldsymbol{S}$) selected using prior causal knowledge.

(ii) The weight vector constructed by $\tilde{B}_{\bar{s}x.c}$ of $\tilde{B}_{mx.c} = (\tilde{B}_{sx.c}, \tilde{B}_{\bar{s}x.c})$, but not that constructed by $\tilde{B}_{y\bar{s}.xcs}$, is assigned to $B_{y\bar{s}.xcs}$. Equation (9) shows that the indirect effect of X on Y decreases via $\overline{S}_i \in \overline{\boldsymbol{S}}$ to zero when $B_{\bar{s}_i x.c}$ approaches zero.

(iii) Standardizing each weight vector enable us to fairly select covariates and intermediate variables in order of priority.

For $p = 1$, $\beta_{yx.cm}$, $B_{yc.xm}$ and $B_{ym.xc}$, which minimize equation (7), and $B_{mx.c}$ and $B_{mc.x}$, which minimize equation (8), are called PCM estimators, denoted by $\check{\beta}^{\dagger}_{yx.cm}$, $\check{B}^{\dagger}_{yc.xm}$, $\check{B}^{\dagger}_{ym.xc}$, $\check{B}^{\dagger}_{mx.c}$, and $\check{B}^{\dagger}_{mc.x}$, respectively. Since equation (7) is consistent with the partially adaptive L_p-penalized loss function given by Nanmo and Kuroki (2022) when ζ_p and ξ_p respectively are zero and $\boldsymbol{M}$ is an empty set, PCM Selector is considered a generalization of PAL$_p$MA. Under the assumption that the sum-of-squares matrix of $\{X\} \cup \boldsymbol{C} \cup \boldsymbol{M}$ is invertible, letting $\lambda_p = 0$, $\beta_{yx.cm}$, $B_{yc.xm}$ and $B_{ym.xc}$, which minimize equation (7), are given by the OLS estimators, i.e., equation (4). In addition, Let $p = 2$, $\lambda_2 = 3\lambda > 0$, $\zeta_p = 1/3$, $\xi_p = 1/3$, $\boldsymbol{\gamma}_{\bar{s}x.c} = \mathbf{1}_{q_{\bar{s}}}$ and $\boldsymbol{\gamma}_{y\bar{z}.xmz} = \mathbf{1}_{q_{\bar{z}}}$. Then, $\beta_{yx.cm}$, $B_{ym.xc}$ and $B_{yc.xm}$, which minimize equation (7), are given by the ridge-type estimators in equation (5).

Here, in order to avoid confusion by the notation in the following discussion, regarding equations (7) and (8) for $p = 1$, let $\{X\}$, $\overline{\boldsymbol{S}}$ and $\overline{\boldsymbol{Z}}$ be active sets for a given $\lambda_1, \rho_1 > 0$, which is a subset of variables with nonzero regression coefficients that do not include any elements of $\boldsymbol{Z} \cup \boldsymbol{S}$. In addition, let $q_{\bar{s}}$ and $q_{\bar{z}}$ be the numbers of variables in the active sets $\overline{\boldsymbol{S}}$ and $\overline{\boldsymbol{Z}}$, respectively. Then, under the assumption that the sum-of-squares matrix of explanatory variables $\{X\} \cup \boldsymbol{C} \cup \boldsymbol{M}$ is invertible, when X is active, $\check{\beta}^{\dagger}_{yx.cm}$, $\check{B}^{\dagger}_{ys.xc\bar{s}}$, $\check{B}^{\dagger}_{y\bar{s}.xcs}$ and $\check{B}^{\dagger}_{mx.c}$ are given by

$$\left(\check{\beta}^{\dagger}_{yx.cm}, \check{B}^{\dagger}_{ys.xc\bar{s}}, \check{B}^{\dagger}_{y\bar{s}.xcs}\right)^T = \left(\hat{\beta}_{yx.cm}, \hat{B}_{ys.xc\bar{s}}, \hat{B}_{y\bar{s}.xcs}\right)^T + n\lambda_1 \begin{pmatrix} -1 & \hat{B}_{\bar{s}x.sc} & \hat{B}_{\bar{z}x.zm} \\ \hat{B}_{xs.c\bar{s}} & \hat{B}_{\bar{s}s.xc} & \hat{B}_{\bar{z}s.xz\bar{s}} \\ \hat{B}_{x\bar{s}.cs} & -I_{q_{\bar{s}}} & \hat{B}_{\bar{z}\bar{s}.xsz} \end{pmatrix} \times \begin{pmatrix} \zeta_1 s^{-1}_{xx.cm}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm}) \\ \xi_1 S^{-1}_{\bar{s}\bar{s}.xcs}\boldsymbol{\gamma}_{\bar{s}x.c} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\ (1-\zeta_1-\xi_1) S^{-1}_{\bar{z}\bar{z}.xmz}\boldsymbol{\gamma}_{y\bar{z}.xmz} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz}) \end{pmatrix}, \quad (12)$$

$$\check{B}^{\dagger}_{mx.c} = \hat{B}_{mx.c} + n\rho_1 \hat{B}_{\bar{z}x.z} S^{-1}_{\bar{z}\bar{z}.xz} \boldsymbol{\gamma}_{m\bar{z}.xz} \odot \mathrm{sign}(\check{B}^{\dagger}_{m\bar{z}.xz}), \quad (13)$$

where

$$\begin{aligned} &\hat{B}_{\bar{s}x.sc} = s^{-1}_{xx.sc} S_{x\bar{s}.sc}, \quad \hat{B}_{\bar{z}x.zm} = s^{-1}_{xx.zm} S_{x\bar{z}.zm}, \\ &\hat{B}_{xs.c\bar{s}} = S^{-1}_{ss.c\bar{s}} S_{sx.x\bar{s}}, \quad \hat{B}_{\bar{s}s.xc} = S^{-1}_{ss.xc} S_{s\bar{s}.xc}, \\ &\hat{B}_{\bar{z}s.xz\bar{s}} = S^{-1}_{ss.xz\bar{s}} S_{s\bar{z}.xz\bar{s}}, \quad \hat{B}_{x\bar{s}.cs} = S^{-1}_{\bar{s}\bar{s}.cs} S_{\bar{s}x.cs}, \\ &\hat{B}_{\bar{z}\bar{s}.xsz} = S^{-1}_{\bar{s}\bar{s}.xsz} S_{\bar{s}\bar{z}.xsz}, \quad \hat{B}_{ys.xc\bar{s}} = S^{-1}_{ss.xz\bar{s}} S_{ys.xz\bar{s}}, \\ &\hat{B}_{y\bar{s}.xcs} = S^{-1}_{\bar{s}\bar{s}.xzs} S_{y\bar{s}.xzs} \end{aligned} \quad (14)$$

In addition, for a $q \times r$ matrix $A = (a_{ij})_{1 \le i \le q, 1 \le j \le r}$, $\mathrm{sign}(A) = (\mathrm{sign}(a_{ij}))_{1 \le i \le q, 1 \le j \le r}$, where

$$\mathrm{sign}(a_{ij}) = \begin{cases} 1 & a_{ij} > 0 \\ 0 & a_{ij} = 0 \\ -1 & a_{ij} < 0 \end{cases} \quad (15)$$

for $i = 1, 2, ..., q$, $j = 1, 2, ..., r$. When X is not active, $\check{\beta}^{\dagger}_{yx.cm}$ is evaluated as zero. In addition, $\check{B}^{\dagger}_{ys.c\bar{s}}$ and $\check{B}^{\dagger}_{y\bar{s}.cs}$ are obtained by omitting the subscript x in equation (12) except for $\boldsymbol{\gamma}_{\bar{s}x.c}$ and replacing $\hat{B}_{\bar{s}x.sc}$, $\hat{B}_{\bar{z}x.zm}$, $\hat{B}_{xs.c\bar{s}}$, $\hat{B}_{x\bar{s}.cs}$ and $s^{-1}_{xx.cm}$ with zeros in equation (12). Note that $\boldsymbol{\gamma}_{\bar{s}x.c}$ is given by equation (9) regardless of whether X is active or not.

Here, for $\lambda_2, \rho_2, \rho_2' \ge 0$ and $\xi_2 \in [0, 1]$, to reduce the bias, based on the derived active sets, the following estimators are considered:

(a) $\tilde{B}^{\dagger}_{xc.m}$ **and** $\tilde{B}^{\dagger}_{xm.c}$**:** $B_{xc.m}$ and $B_{xm.c}$ that minimize

$$\begin{aligned} &L_2(B_{xc.m}, B_{xm.c}) \\ &= \frac{1}{2n}\|\boldsymbol{x} - \boldsymbol{z}B_{xz.\bar{z}m} - \bar{\boldsymbol{z}}B_{x\bar{z}.zm} - \boldsymbol{s}B_{xs.c\bar{s}} - \bar{\boldsymbol{s}}B_{x\bar{s}.cs}\|_2^2 \\ &\quad + \lambda_2\left\{\xi_2\|B_{x\bar{s}.cs}\|_2^2 + (1-\xi_2)\|B_{x\bar{z}.zm}\|_2^2\right\}, \end{aligned} \quad (16)$$

(b) $\tilde{B}^{\dagger}_{\bar{s}x.cs}$**,** $\tilde{B}^{\dagger}_{\bar{s}s.xc}$ **and** $\tilde{B}^{\dagger}_{\bar{s}c.xs}$**:** $B_{\bar{s}x.cs}$, $B_{\bar{s}s.xc}$ and $B_{\bar{s}c.xs}$ that minimize

$$\begin{aligned} &L_2(B_{\bar{s}x.cs}, B_{\bar{s}s.xc}, B_{\bar{s}c.xs}) \\ &= \frac{1}{2n}\|\bar{s} - \boldsymbol{x}B_{\bar{s}x.sc} - \boldsymbol{s}B_{\bar{s}s.xc} - \boldsymbol{z}B_{\bar{s}z.xs\bar{z}} - \bar{\boldsymbol{z}}B_{\bar{s}\bar{z}.xsz}\|_F^2 \\ &\quad + \rho_2\|\mathrm{vec}\left(B_{\bar{s}\bar{z}.xsz}\right)\|_2^2, \end{aligned} \quad (17)$$

(c) $\tilde{B}^{\dagger}_{\bar{z}x.zm}$**,** $\tilde{B}^{\dagger}_{\bar{z}z.xm}$ **and** $\tilde{B}^{\dagger}_{\bar{z}m.xz}$**:** $B_{\bar{z}x.zm}$, $B_{\bar{z}z.xm}$ and $B_{\bar{z}m.xz}$ that minimize

$$\begin{aligned} &L_2(B_{\bar{z}x.zm}, B_{\bar{z}z.xm}, B_{\bar{z}m.xz}) \\ &= \frac{1}{2n}\|\bar{\boldsymbol{z}} - \boldsymbol{x}B_{\bar{z}x.zm} - \boldsymbol{z}B_{\bar{z}z.xm} - \boldsymbol{s}B_{\bar{z}s.xz\bar{s}} - \bar{\boldsymbol{s}}B_{\bar{z}\bar{s}.xzs}\|_F^2 \\ &\quad + \rho_2'\|\mathrm{vec}\left(B_{\bar{z}\bar{s}.xsz}\right)\|_2^2. \end{aligned} \quad (18)$$

Then, based on equations (12) and (13), when X is active, consider

$$\left(\check{\beta}^{*}_{yx.cm}, \check{B}^{*}_{ys.xc\bar{s}}, \check{B}^{*}_{y\bar{s}.xcs}\right)^T = \left(\check{\beta}^{\dagger}_{yx.cm}, \check{B}^{\dagger}_{ys.xc\bar{s}}, \check{B}^{\dagger}_{y\bar{s}.xcs}\right)^T - n\lambda_1 \begin{pmatrix} -1 & \tilde{B}^{\dagger}_{\bar{s}x.sc} & \tilde{B}^{\dagger}_{\bar{z}x.zm} \\ \tilde{B}^{\dagger}_{xs.c\bar{s}} & \tilde{B}^{\dagger}_{\bar{s}s.xc} & \tilde{B}^{\dagger}_{\bar{z}s.xz\bar{s}} \\ \tilde{B}^{\dagger}_{x\bar{s}.cs} & -I_{q_{\bar{s}}} & \tilde{B}^{\dagger}_{\bar{z}\bar{s}.xsz} \end{pmatrix} \times \begin{pmatrix} \zeta_1 \tilde{s}^{\dagger -1}_{xx.cm}\mathrm{sign}(\check{\beta}^{\dagger}_{yx.cm}) \\ \xi_1 \tilde{S}^{\dagger +}_{\bar{s}\bar{s}.xcs}\boldsymbol{\gamma}_{\bar{s}x.c} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{s}.xcs}) \\ (1-\zeta_1-\xi_1) \tilde{S}^{\dagger +}_{\bar{z}\bar{z}.xmz}\boldsymbol{\gamma}_{y\bar{z}.xmz} \odot \mathrm{sign}(\check{B}^{\dagger}_{y\bar{z}.xmz}) \end{pmatrix}, \quad (19)$$

$$\check{B}^{*}_{mx.c} = \check{B}^{\dagger}_{mx.c} - n\rho_1 \hat{B}_{\bar{z}x.z} \hat{S}^{+}_{\bar{z}\bar{z}.xz} \boldsymbol{\gamma}_{m\bar{z}.xz} \odot \mathrm{sign}(\check{B}^{\dagger}_{m\bar{z}.xz}), \quad (20)$$

where $\boldsymbol{m}$ and $\boldsymbol{c}$ of $\check{B}^*_{mx.c}$ are constructed by both $\boldsymbol{S} \cup \boldsymbol{Z}$ and a subset of $\overline{\boldsymbol{S}} \cup \overline{\boldsymbol{Z}}$ corresponding to the active sets of $\check{B}^{\dagger}_{y\overline{s}.xcs}$ and $\check{B}^{\dagger}_{y\overline{c}.xmz}$,

$$\tilde{s}^{\dagger}_{xx.cm} = \|\boldsymbol{x} - \boldsymbol{c}\tilde{B}^{\dagger}_{xc.m} - \boldsymbol{m}\tilde{B}^{\dagger}_{xm.c}\|_2^2, \tag{21}$$

$$\tilde{S}^{\dagger}_{\overline{ss}.xcs} = ||\overline{\boldsymbol{s}} - \boldsymbol{x}\tilde{B}^{\dagger}_{\overline{s}x.cs} - \boldsymbol{s}\tilde{B}^{\dagger}_{\overline{s}s.xc} - \boldsymbol{c}\tilde{B}^{\dagger}_{\overline{s}c.xs}||_G, \tag{22}$$

$$\tilde{S}^{\dagger}_{\overline{zz}.xmz} = ||\overline{\boldsymbol{z}} - \boldsymbol{x}\tilde{B}^{\dagger}_{\overline{z}x.cs} - \boldsymbol{m}\tilde{B}^{\dagger}_{\overline{z}m.xz} - \boldsymbol{z}\tilde{B}^{\dagger}_{\overline{z}z.xm}||_G, \tag{23}$$

$$\hat{S}_{\overline{zz}.xz} = ||\overline{\boldsymbol{z}} - \boldsymbol{x}\hat{B}_{\overline{z}x.z} - \boldsymbol{z}\hat{B}_{\overline{z}z.x}||_G, \tag{24}$$

and $\|A\|_G$ and A^+ denote the gram matrix A^TA and the generalized inverse of a matrix A (Bernstein 2009), respectively. When X is not active, $\check{\beta}^*_{yx.cm}$ is evaluated as zero. In addition, $\check{B}^*_{ys.c\overline{s}}$ and $\check{B}^*_{y\overline{s}.cs}$ are obtained by omitting the subscript x from equation (19) except for $\boldsymbol{\gamma}_{\overline{s}x.c}$ and replacing $\tilde{B}^{\dagger}_{\overline{s}x.sc}$, $\tilde{B}^{\dagger}_{\overline{z}x.zm}$, $\tilde{B}^{\dagger}_{xs.c\overline{s}}$, $\tilde{B}^{\dagger}_{x\overline{s}.cs}$ and $\tilde{s}^{\dagger -1}_{xx.cm}$ with zeros in equation (19). Note that $\boldsymbol{\gamma}_{\overline{s}x.c}$ is given by equation (9) regardless of whether X is active or not.

Then, we formulate the modified PCM estimator of the total effect τ_{yx} as

$$\check{\tau}^*_{yx} = \check{\beta}^*_{yx.cm} + \check{B}^*_{mx.c}\check{B}^*_{ym.xc}$$

when X is active according to equation (7) and

$$\check{\tau}^*_{yx} = \check{B}^*_{mx.c}\check{B}^*_{ym.c}$$

when X is not active according to equation (7). Hereafter, the modified PCM estimator is called the PCM estimator.

Regarding PCM estimators, the following theorems hold:

Theorem 1 *For an active set $\boldsymbol{M} \cup \boldsymbol{C}$, when the OLS estimators are available, if X is conditionally independent of Y given $\boldsymbol{M} \cup \boldsymbol{C}$, then the following inequalities approximately hold under the normality:*

$$var(\check{B}^*_{mx.c}\check{B}^*_{ym.c}) \leq var(\hat{B}_{mx\cdot c}\hat{B}_{ym\cdot c}) \leq var(\hat{\beta}_{yx\cdot c}) \tag{25}$$

$$var(\check{B}^*_{mx.c}\check{B}^*_{ym.c}) \leq var(\check{\beta}^*_{yx.c}) \tag{26}$$

for the optimal tuning and penalty parameters.

The first inequality is given in the Technical Appendix. The second inequality is shown in Kuroki and Hayashi (2014, 2016). Theorem 1 shows that the estimation accuracy of the total effect can be improved compared to that of the OLS method through PCM Selector based on a set of variables that make X and Y conditionally independent.

Theorem 2 *For an active set $\boldsymbol{M} \cup \boldsymbol{C}$, when the OLS estimators are available, if X is conditionally independent of Y given $\boldsymbol{M} \cup \boldsymbol{C}$ and $\boldsymbol{M}' \cup \boldsymbol{C}$, the following inequalities approximately hold under the normality:*

$$var(\check{B}^*_{m'x.c}\check{B}^*_{ym'.c}) \leq var(\hat{B}_{m'x.c}\hat{B}_{ym'.c}) \leq var(\hat{B}_{mx.c}\hat{B}_{ym.c}) \tag{27}$$

for $\boldsymbol{M}' \subset \boldsymbol{M}$.

The first inequality is simply obtained from Theorem 1, and the second inequality is shown in Kuroki and Hayashi (2014, 2016). Theorem 2 provides a statistical guideline for selecting a set of intermediate variables to derive a more efficient estimator of the total effects.

Numerical Experiment

In this section, we present a numerical experiment to compare the performances of LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA, the OLS method, the two-stage least squares (TSLS) method and PCM Selector. For brevity, consider the linear SCM

$$\left.\begin{array}{l} Y = \alpha_{ys}S + \alpha_{yz}Z + \overline{\boldsymbol{S}}A_{y\overline{s}} + \overline{\boldsymbol{Z}}A_{y\overline{z}} + \epsilon_y \\ \overline{\boldsymbol{S}} = XA_{\overline{s}x} + SA_{\overline{s}s} + ZA_{\overline{s}z} + \boldsymbol{\epsilon}_{\overline{s}} \\ S = \alpha_{sx}X + \alpha_{sz}Z + \epsilon_s \\ X = \alpha_{xz}Z + \epsilon_x \end{array}\right\} \tag{28}$$

for Figure 1, where $\overline{\boldsymbol{Z}}$ and $\overline{\boldsymbol{S}}$ include 10 covariates and 5 intermediate variables ($\boldsymbol{M} = \{S\} \cup \overline{\boldsymbol{S}}$), respectively. In Figure 1, Setting (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) and Setting (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 Setting (a), and satisfies the front-door criterion for Setting (b), respectively.

To set up the numerical experiment, we first construct the population variance-covariance matrix. To eliminate arbitrariness, the true values of the direct effects are $\alpha_{ys} = 0.4$, $\alpha_{\overline{s}s} = 0.2$ $(\in A_{\overline{s}s})$, $\alpha_{\overline{s}_2x}, \alpha_{\overline{s}_3x}, \alpha_{\overline{s}_4x}, \alpha_{\overline{s}_5x}$ $(\in A_{\overline{s}x})$ are set to 0 and $\alpha_{y\overline{z}_1}, \alpha_{y\overline{z}_2}, \ldots, \alpha_{y\overline{z}_{10}} (\in A_{y\overline{z}})$, $\alpha_{y\overline{s}_2}, \alpha_{y\overline{s}_3}, \alpha_{y\overline{s}_4}, \alpha_{y\overline{s}_5} (\in A_{y\overline{s}})$ are randomly and independently generated according to a uniform distribution on the interval $[-0.2, 0.2]$ in the both settings (a) and (b). The other direct effects are given as follows: Setting (a) $\alpha_{xz} = 0.8$, $\alpha_{\overline{s}_1x} = 0.0$, $\alpha_{sx} = 0.1$, $\alpha_{yz} = \alpha_{sz} = \alpha_{\overline{s}z} = 0.2$ $(\alpha_{\overline{s}z} \in A_{\overline{s}z})$, $\alpha_{y\overline{s}_1}$ is randomly generated according to a uniform distribution on the interval $[-0.2, 0.2]$; Setting (b) $\alpha_{xz} = \alpha_{\overline{s}_1x} = \alpha_{y\overline{s}_1} = 0.2$, $\alpha_{sx} = 0.8$, $\alpha_{yz} = \alpha_{sz} = \alpha_{\overline{s}z} = 0.0$ $(\alpha_{\overline{s}z} \in A_{\overline{s}z})$.

In addition, we assume that the random disturbances ϵ_x, ϵ_y, ϵ_s and $\boldsymbol{\epsilon}_{\overline{s}}$ independently follow a normal distribution in which X, Y, S, $\overline{\boldsymbol{S}}$ and $\boldsymbol{C}$ are standardized to mean 0 and the unit variance. Furthermore, the population variance-covariance matrix of $\boldsymbol{C}$ is randomly determined according to Pourahmadi and Wang (2015).

We generated 15 random samples of 18 variables from a multivariate normal distribution with a zero mean vector and the above variance-covariance matrix for 5000 replications. Table 1 shows the basic statistics of the total effects estimated by LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA, the OLS method, the TSLS methods, and PCM Selector based on the given penalty and tuning parameters. 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 Setting (a). In Setting (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. Regarding the parameter tuning for LASSO, adaptive LASSO, Elastic Net, PAL$_1$MA and PCM Selector, see Section C in the Technical Appendix.

Setting (a)	$\tau_{yx} = 0.045$				parameter settings						
	Mean	SD	Bias	Sign	λ	η	ϕ	λ_1	ρ_1	ζ_1	ξ_1
LASSO	0.013	0.045	-0.033	0.117	0.407	-	-	-	-	-	-
adaptive LASSO	0.017	0.057	-0.028	0.138	0.407	0.100	-	-	-	-	-
Elastic Net	0.017	0.054	-0.028	0.156	0.399	-	0.910	-	-	-	-
PAL_1MA	0.054	0.792	0.009	0.528	0.294	1.200	-	-	-	-	-
PCM Selector	0.036	0.718	-0.010	0.526	-	-	-	0.017	0.213	0.270	0.190
Front-door-like (including x)	-0.008	1.577	-0.053	0.515	-	-	-	-	-	-	-
Front-door-like (not including x)	0.030	1.051	-0.015	0.524	-	-	-	-	-	-	-
Back-door	0.054	1.591	0.009	0.532	-	-	-	-	-	-	-

Setting (b)	$\tau_{yx} = 0.402$				parameter settings			
	Mean	SD	Bias	Sign	λ_1	ρ_1	ζ_1	ξ_1
PCM Selector	0.448	0.549	0.046	0.808	0.346	-	0.000	1.000
Front-door (minimal)	0.468	0.552	0.066	0.818	-	-	-	-
Front-door (whole)	0.462	0.692	0.060	0.770	-	-	-	-

Table 1: Results based on cross-validation. Mean: sample mean; SD: standard deviation; Bias: bias between the true value and the sample mean; 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): the intermediate variable S and the set of covariates $\boldsymbol{C}$ are used for the front-door-like criterion; Back-door: the set of covariates $\boldsymbol{C}$ is used for the back-door criterion; Front-door (minimal): a minimally sufficient set of intermediate variables is used for the front-door criterion. Front-door (whole): the set of intermediate variables $\boldsymbol{M}$ is used for the front-door criterion.$\lambda, \lambda_1, \rho_1$: penalty parameters; η: tuning parameter for the adaptive weights (Zou 2006); ϕ: tuning parameter for the elastic net penalty (Zou and Hastie 2005); ζ_1, ξ_1: tuning parameters; $\lambda = 3.157$, $\rho = 69.484$ for equations (5) and (6) in Setting (a) and $\lambda = 3.726$ for equation (6) in Setting (b); τ_{yx}: true value of total effect.

According to Table 1, PCM Selector provides better estimation accuracy than PAL_1MA and the least squares methods. In addition, Table 1 shows that PCM Selector generally provides an estimation that is biased but less biased than the TSLS methods in the present parameter setting. Furthermore, 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 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 PCM Selector and PAL_1MA are not low.

The Technical Appendix provides further discussion.

Conclusion

In current situations where advanced artificial intelligence technology enables us to collect large datasets, it is not difficult to observe many covariates and intermediate variables. In such situations, it would be reasonable to consider such sets of variables to evaluate total effects. However, it is difficult to evaluate the total effects reliably when multicollinearity/high-dimensional data problems occur in this situation. To solve this problem, we establish PCM Selector, which is considered as a wider class, including adaptive LASSO and PAL_pMA, to provide a less biased estimator of total effects with better estimation accuracy. In addition, through numerical experiments and a case study in the Technical Appendix, we confirmed that PCM Selector is superior to other methods. Interestingly, there are some situations where the total effect is not identifiable, but the indirect effects are identifiable (Inoue, Ritz, and Arah 2022). Although the current penalized regression analyses are not applicable to such situations, PCM Selector is applicable for evaluating the indirect effect.

Finally, although PCM Selector is formulated based on single/joint linear regression models, it would be interesting to extend our approach to a wide variety of statistical models, including generalized linear models. Such an extension would be straightforward - the loss function would be replaced with a more general form. This extension will be left for future work.

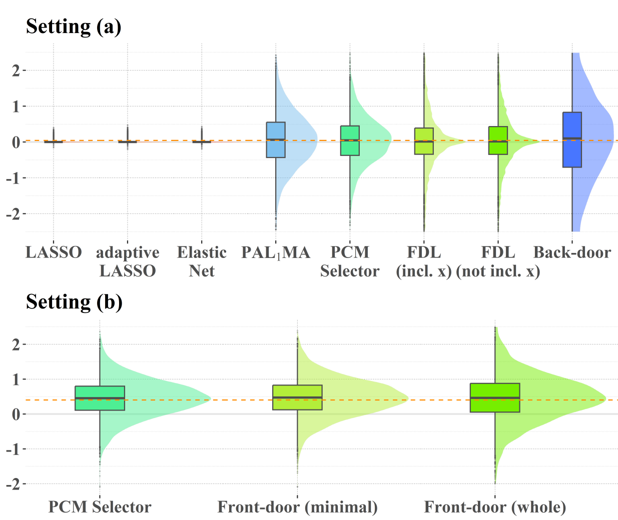

Figure 2: Violin plots of estimated total effects. The dashed lines show the true total effects. FDL: Front-door-like.

Acknowledgements

We would like to acknowledge the helpful comments of the five anonymous reviewers. This research was partially funded by the JSPS Grant-in-Aids for Scientific Research 23K21700 and 24K14851.

References

Baron, R.; and Kenny, D. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6): 1173–1182.

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.

Cox, D. R. 1960. Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1): 172–176.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. LEAST ANGLE REGRESSION. *Annals of Statistics*, 32(2): 407–499.

Gupta, S.; Lipton, Z. C.; and Childers, D. 2021. Estimating treatment effects with observed confounders and mediators. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, 982–991.

Hayashi, T.; and Kuroki, M. 2014. On Estimating causal effects based on supplemental variables. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, 312–319.

Hui, H.; and Zhongguo, Z. 2008. Comparing identifiability criteria for causal effects in Gaussian causal models (In Chinese). *Acta Mathematica Scientia, Series A*, 28(5): 808–817.

Imai, K.; Keele, L. J.; Tingley, D.; and Yamamoto, T. 2011. Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*, 105(4): 765–789.

Inoue, K.; Ritz, B.; and Arah, O. A. 2022. Causal effect of chronic pain on mortality through opioid prescriptions: Application of the front-door formula. *Epidemiology*, 33(4): 572–580.

Kuroki, M. 2000. Selection of post-treatment variables for estimating total effect from empirical research. *Journal of the Japan Statistical Society*, 30(2): 115–128.

Kuroki, M.; and Cai, Z. 2004. Selection of identifiability criteria for total effects by using path diagrams. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 333–340.

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.

Kuroki, M.; and Tezuka, T. 2023. The estimated causal effect on the variance based on the front-door criterion in Gaussian linear structural equation models: an unbiased estimator with the exact variance. *Statistical Papers*, 1–24.

Mackinnon, D. 2008. *Introduction to Statistical Mediation Analysis*. Erlbaum.

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*, 1456–1465.

Pearl, J. 2001. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 411–420.

Pearl, J. 2009. *Causality: Models of Reasoning and Inference, The 2nd Edition*. Cambridge University 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.

Ramsahai, R. R. 2012. Supplementary variables for causal estimation. *Causality: Statistical Perspectives and Applications*, 218–233.

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

Van, S.; Geer, P. b.; Ritov, Y.; and Dezeure, R. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3): 1166–1202.

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

Zou, H.; and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2): 301–320.