Joint Adaptive Penalty for Unbalanced Mediation Pathways

Hanying Jiang[*1], Kris Sankaran[†1], and Yinqiu He[‡1]

[1]Department of Statistics, University of Wisconsin-Madison, USA

Abstract

Mediation analysis has been widely used to investigate how a treatment influences an outcome through intermediate variables, known as mediators. Analyzing a mediation mechanism typically requires assessing multiple model parameters that characterize distinct pathwise effects. Classical methods that estimate these parameters individually can be inefficient, particularly when the underlying pathwise effects exhibit substantial imbalance. To address this challenge, this work proposes a new joint adaptive penalty that integrates information across entire mediation mechanisms, thereby enhancing both parameter estimation and pathway selection. We establish theoretical guarantees for the proposed method under an asymptotic framework and conduct extensive numerical studies to demonstrate its superior performance in scenarios with unbalanced mediation pathways.

Keywords: Mediation analysis, adaptive penalty, structural equation models.

1 Introduction

Mediation analysis explores whether and how a treatment influences an outcome through intermediate mediator variables (MacKinnon, 2012; Tingley et al., 2014). It decomposes the treatment effect on the outcome into the indirect/mediation effect through the mediators and the direct effect through other causal mechanisms. This decomposition can improve our understanding of complex mechanisms and inform the design of effective interventions. In biomedical research, for example, mediation analysis has been used to examine how exposure to fine particulate matter increases mortality through metabolic and cardiovascular diseases (Bai et al., 2022), and how antibiotic treatments influence asthma development through changes in the microbiome (Toivonen et al., 2021). Across various scientific fields, modern technological advances enable the simultaneous measurement of numerous candidate mediators, such as genomic features (Abrishamcar et al., 2022; Yang et al., 2024), neural signatures

[*]hjiang252@wisc.edu

[†]ksankaran@wisc.edu

[‡]yinqiu.he@wisc.edu

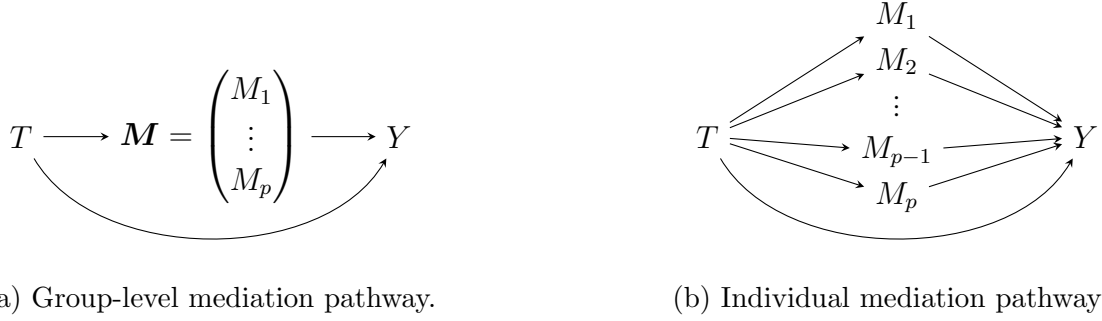(a) Group-level mediation pathway.   (b) Individual mediation pathways.

Figure 1: Directed acyclic graphs for mediation pathways.

(Chén et al., 2021; Zhao et al., 2022), and metabolic traits (Ko et al., 2023; Lu et al., 2023). Developing efficient analysis for multiple mediators can help uncover active pathways for more targeted interventions and deepen our understanding of complicated systems.

In this work, we focus on the multi-mediator framework where the goal is to understand how an exposure/treatment T influences an outcome Y through p potential mediators $\boldsymbol{M} = [M_1, \ldots, M_p]^\top$. To achieve this, classical mediation analyses model the relationship between exposure, potential mediators, and the outcome through directed acyclic graphs. One class of existing studies investigates multiple mediators as a group, which can be illustrated as in Figure 1a. This class includes methods examining the overall group-level mediation effect (VanderWeele and Vansteelandt, 2014; Zhou et al., 2020; Hao and Song, 2023) or jointly transforming multiple mediators, such as to principal components (Huang and Pan, 2016; Chén et al., 2018; Bellavia et al., 2019; Zhao et al., 2020).

To further reveal the detailed causal mechanisms through the observed mediators, various efforts have also been made to study pathways via each individual mediator (Imai and Yamamoto, 2013; Vansteelandt and Daniel, 2017; Jérolon et al., 2021), as illustrated in Figure 1b. In this case, a central question is to estimate the mediation effect through each individual mediator M_j and assess whether it is significantly different from zero, indicating whether M_j plays an active mediating role or not. Estimating mediation effects and identifying active pathways through individual mediators can uncover complicated causal mechanisms and guide the development of effective interventions. To address the above question, researchers have proposed statistical methods from both frequentist and Bayesian perspectives. From the frequentist perspective, pathwise effects are often modeled as fixed parameters, commonly as coefficients in structural equation models. Examining mediation effects is then formulated as estimating these coefficients and determining whether or not the corresponding estimands are zero. Within this framework, one research line focuses on fitting a joint model of $(T, \boldsymbol{M}, Y)$ based on observed data, often incorporating regularizations to handle multiple mediators. Examples include the minimax concave penalty (Zhang et al., 2016), de-biased LASSO (Gao et al., 2019), adaptive LASSO (Zhang, 2022) and pathway-specific penalty (Zhao and Luo, 2022). Another research line considers scenarios where estimators for pathwise effects along $T \to M_j$ and $M_j \to Y$ have been obtained along with their asymptotic distributions. Then these studies primarily focus on correcting for multiple comparisons across multiple mediators (Dai et al., 2022; Liu et al., 2022; Du et al., 2023). From the Bayesian perspective, pathwise effects are modeled as random coefficients in structural equation models. Estimation and

variable selection can be achieved through Bayesian shrinkage estimation with appropriate priors (Song et al., 2020, 2021).

This work focuses on fitting the joint model of $(T, \boldsymbol{M}, Y)$ with regularization from the frequentist perspective. Within this paradigm, most existing methods apply separate regularizations for the exposure-to-mediator $(T \to \boldsymbol{M})$ and mediator-to-outcome $(\boldsymbol{M} \to Y)$ paths (Gao et al., 2019; Zhang, 2022). However, such a separate-fitting strategy does not effectively combine joint information across the two sets of paths $T \to \boldsymbol{M}$ and $\boldsymbol{M} \to Y$ in mediation analysis. This limitation becomes especially problematic if the pathwise effects are *unbalanced*. For example, when the pathwise effect along $T \to M_j$ is strong but that along $M_j \to Y$ is weak, examining the two paths separately may result in overlooking this active mediation pathway $T \to M_j \to Y$ as a whole. A more detailed illustration under the classical parallel model is provided in Section 3.1. In the existing literature, Zhao and Luo (2022) introduces a pathway-specific LASSO method and jointly fits the model over $(T, \boldsymbol{M}, Y)$. However, it can be computationally expensive, and its accuracy of identifying active pathways can be low, as further demonstrated in the Supporting Information.

This work focuses on fitting the joint model of $(T, \boldsymbol{M}, Y)$ with regularization from the frequentist perspective. Within this paradigm, most existing methods apply separate regularizations for the exposure-to-mediator $(T \to \boldsymbol{M})$ and mediator-to-outcome $(\boldsymbol{M} \to Y)$ paths (Gao et al., 2019; Zhang, 2022). However, such a separate-fitting strategy does not effectively combine joint information across the two sets of paths $T \to \boldsymbol{M}$ and $\boldsymbol{M} \to Y$ in mediation analysis. This limitation becomes especially problematic if the pathwise effects are *unbalanced*. For example, when the pathwise effect along $T \to M_j$ is strong but that along $M_j \to Y$ is weak, examining the two paths separately may result in overlooking this active mediation pathway $T \to M_j \to Y$ as a whole. A more detailed illustration under the classical parallel model is provided in Section 3.1. In the existing literature, Zhao and Luo (2022) introduces a pathway-specific LASSO method and jointly fits the model over $(T, \boldsymbol{M}, Y)$. However, it can be computationally expensive, and its accuracy of identifying active pathways can be low, as further demonstrated in Section F of the appendix.

To overcome the above challenges, this work proposes a new joint adaptive penalty that combines information across distinct and potentially unbalanced pathwise effects while maintaining low computational cost. The penalty is constructed by incorporating adaptive weights informed by the significance of target mediation effects. Theoretically, we establish asymptotic guarantees showing that the proposed penalty controls estimation errors and achieves consistent selection of active mediation pathways. Through extensive numerical studies, we demonstrate our method is scalable and yields superior performance across various scenarios. The new penalization framework will advance current mediation analysis with multiple mediators, facilitating more important scientific discoveries.

The rest of the paper is organized as follows. Section 2 introduces the framework under which our analysis is conducted. Section 3 introduces the new proposed penalty, including its construction and asymptotic theory. Section 5 reviews the comparable methods in the existing literature. Section 6 conducts numerical experiments to compare the proposed method and the existing methods under finite samples. In Section 7, we demonstrate our method by investigating the effect of gastrectomy on total cholesterol level mediated by the gut microbiome. We conclude this paper with discussions in Section 8. All the proofs are deferred to the appendix.

We will use the following notations throughout the paper. For two sequences of real numbers (a_n) and (b_n), we let $a_n \gg b_n$ denote $\lim_{n \to \infty} b_n/a_n = 0$, let $a_n \ll b_n$ denote $\lim_{n \to \infty} a_n/b_n = 0$, and let $a_n \lesssim b_n$ denote that there exists a constant $C > 0$ such that $|a_n| \leq C|b_n|$ for all n. For a sequence of random variables (X_n) and a sequence of real numbers (a_n), we let $X_n = O_p(a_n)$ represent that for any $\epsilon > 0$, there is a positive constant C_ϵ such that $\sup_n \Pr(|X_n| \geq C_\epsilon |a_n|) < \epsilon$. We use $\to_p$ to denote convergence in probability. For a vector $\boldsymbol{x} = [x_1, \ldots, x_p]^\top \in \mathbb{R}^p$ and a positive integer q, let $\|\boldsymbol{x}\|_q = (\sum_{j=1}^p |x_j|^q)^{1/q}$ represent the ℓ_q norm of $\boldsymbol{x}$. For a matrix $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{m \times n}$, let $\boldsymbol{A}^\top$ represent its transpose, and let $\|\boldsymbol{A}\|_F = (\text{tr}(\boldsymbol{A}^\top \boldsymbol{A}))^{1/2}$ represent its Frobenius norm. Let $A \perp\!\!\!\perp B \mid \mathcal{E}$ represent the independence of random variables A and B conditional on an event $\mathcal{E}$.

2   Model and Setup

We consider that the exposure T, p potential mediators $\boldsymbol{M}$, and outcome Y follow the canonical linear structural equation model (MacKinnon, 2012):

$$\boldsymbol{M} = \boldsymbol{\alpha}^* T + \boldsymbol{\zeta}_M^{*\top} \boldsymbol{X} + \boldsymbol{E}, \qquad Y = \eta^* T + \boldsymbol{\beta}^{*\top} \boldsymbol{M} + \boldsymbol{\zeta}_Y^{*\top} \boldsymbol{X} + \epsilon, \qquad (1)$$

where $\boldsymbol{\alpha}^* = [\alpha_1^*, \ldots, \alpha_p^*]^\top \in \mathbb{R}^p$, $\boldsymbol{\beta}^* = [\beta_1^*, \ldots, \beta_p^*]^\top \in \mathbb{R}^p$, $\boldsymbol{\zeta}_M^* = [\zeta_{M,ij}^*]_{1 \leq i \leq q, 1 \leq j \leq p} \in \mathbb{R}^{q \times p}$, and $\boldsymbol{\zeta}_Y^* = [\zeta_{Y,1}^*, \ldots, \zeta_{Y,q}^*]^\top \in \mathbb{R}^q$. Additionally, $\boldsymbol{X}$ represents a q-dimensional observed pre-treatment confounding variable, and we assume its first element is set to be one to allow for an intercept. The random errors $\boldsymbol{E}$ and ϵ are independent with zero mean, and $(\boldsymbol{E}, \epsilon)$ are independent of $(T, \boldsymbol{X})$. We emphasize that the model (1) is considered for the simplicity of illustration and interpretation, whereas our Joint Adaptive Penalty proposed in Section 3 is general and could potentially be extended under other models for mediation pathway analysis.

Coefficients in the model (1) can be connected with causal estimands, particularly individual mediation/indirect effects under the counterfactual framework (Imai and Yamamoto, 2013; Loh et al., 2022). In the existing literature, one class of works examines classical natural indirect effects through individual mediators, and identification typically requires that the causal relationships between multiple mediators be either absent or known (Daniel et al., 2015; Taguri et al., 2018). Another class of studies examines interventional indirect effects, which are defined by setting the mediator to a random draw from the distribution of the counterfactual mediator, and often do not require knowledge of the causal structure among mediators (Vansteelandt and Daniel, 2017). Under the canonical parallel path model (1), the analytical forms of the natural and interventional indirect effects coincide under suitable assumptions (Jérolon et al., 2021; Loh et al., 2022). For simplicity, we only review the standard natural indirect effect and present its analytical form under (1). Notably, the conditions for identifying interventional indirect effects are generally weaker, with further discussions available in Loh et al. (2022) and Miles (2023).

Let $\boldsymbol{M}(t) = (M_1(t), \ldots, M_p(t))^\top$ represent the potential value of $\boldsymbol{M}$ under the treatment status t. For each $j = 1, \ldots, p$, let $\boldsymbol{M}_{-j}$ denote entries in $\boldsymbol{M}$ excluding M_j, and similarly define $\boldsymbol{M}_{-j}(t)$. Let $Y(t, \boldsymbol{m})$ represent the potential outcome of Y if T and $\boldsymbol{M}$

were set to be t and $\boldsymbol{m}$, respectively. To define the natural indirect effect through a mediator, we follow the framework in Imai and Yamamoto (2013) which assumes no causal ordering between mediators. In this case, we can simultaneously let $(M_j(t'), \boldsymbol{M}_{-j}(t''))$ denote potential outcomes of M_j and $\boldsymbol{M}_{-j}$ if T were set to be t' and t'', respectively, for $j = 1, \ldots, p$. We then consider the natural indirect effect through M_j defined as $\delta_j(t'; t) = \mathbb{E}\left[Y(t, M_j(t'), \boldsymbol{M}_{-j}(t))\right] - \mathbb{E}\left[Y(t, M_j(t), \boldsymbol{M}_{-j}(t))\right]$, where t' and t are the treatment statuses being compared. To identify $\delta_j(t'; t)$, we assume the standard consistency condition (Condition 1) and the sequential ignorability condition for multiple causally unrelated mediators (Condition 2) first introduced in Jérolon et al. (2021).

Condition 1. *For all possible values of t and $\boldsymbol{m}$, $\boldsymbol{M} = \boldsymbol{M}(t)$ and $Y = Y(t, \boldsymbol{M}(t))$ if $T = t$, and $Y = Y(t, \boldsymbol{m})$ if $T = t$ and $\boldsymbol{M} = \boldsymbol{m}$.*

Condition 2. *For $j = 1, \ldots, p$ and all possible values of $t, t', t'', m,$ and $\boldsymbol{w}$, assume (i) $\{Y(t, m, \boldsymbol{w}), M_j(t'), \boldsymbol{M}_{-j}(t'')\} \perp\!\!\!\perp T \mid \{\boldsymbol{X} = \boldsymbol{x}\}$, (ii) $Y(t', m, \boldsymbol{w}) \perp\!\!\!\perp (M_j(t), \boldsymbol{M}_{-j}(t)) \mid \{T = t, \boldsymbol{X} = \boldsymbol{x}\}$, and (iii) $Y(t, m, \boldsymbol{w}) \perp\!\!\!\perp (M_j(t'), \boldsymbol{M}_{-j}(t)) \mid \{T = t, \boldsymbol{X} = \boldsymbol{x}\}$.*

Detailed interpretations of Condition 2 can be found in Jérolon et al. (2021). Notably, Condition 2 allows the mediators to be uncausally correlated after conditioning on the treatment and observed pretreatment confounders, e.g., due to unmeasured pretreatment confounders. Such flexibility can accommodate mediators that covary together in real-world applications. We next derive an analytical formula of $\delta(t, t')$ in Lemma 1 below. It generalizes Corollary 3.2 in Jérolon et al. (2021) by relaxing their assumptions on the Gaussianity and constant correlations of noise terms.

Lemma 1. *Under the model (1) and Conditions 1 and 2, $\delta_j(t'; t) = \alpha_j^* \beta_j^* (t' - t)$.*

Lemma 1 shows that $\delta_j(t'; t)$ is proportional to the product of coefficients $\alpha_j^* \beta_j^*$ under the model (1). Therefore, a mediator M_j and its corresponding individual pathway $T \to M_j \to Y$ are referred to as active if $\alpha_j^* \beta_j^* \neq 0$.

Remark 1. *In scenarios with multiple mediators, various definitions of indirect effects have been proposed, and the associated identification conditions can differ from and even relax those discussed above, including allowing known and unknown causal orderings between mediators (Daniel et al., 2015; Loh et al., 2022). Despite that, the same product-of-coefficients form have been consistently observed (Imai and Yamamoto, 2013; Daniel et al., 2015; Huang and Pan, 2016; Loh et al., 2022). Our proposed method will be based on these products $\alpha_j^* \beta_j^*$ and thus is inherently generalizable beyond the particular set of definitions and assumptions reviewed above.*

3 Joint Adaptive Penalty

3.1 Unbalanced Mediation Pathways and Challenges

We aim to fit the model (1) while performing efficient identification on the set of mediators M_j with active mediation effects. Under our studied model (1), an individual pathway $T \to M_j \to Y$ consists of two paths $T \to M_j$ and $M_j \to Y$ with effects characterized by the

two coefficients α_j^* and β_j^*, respectively. Based on the relative magnitudes between α_j^* and β_j^*, we can divide active pathways into balanced and unbalanced cases. Specifically, there can exist three scenarios: (a) balanced pathway where the absolute values of α_j^* and β_j^* are similar, (b) unbalanced pathway with the scale of α_j^* being much larger than that of β_j^*, and (c) unbalanced pathway with the scale of α_j^* being much smaller than that of β_j^*. These scenarios are visualized as in Figure 2 (a)–(c). As mentioned in Section 1, existing methods that apply separate regularizations on the effects α_j^* and β_j^* may lack power for identifying a significant mediation effect $\alpha_j^* \beta_j^*$ under unbalanced scenarios, as one of the coefficients can be too small to detect its significance.

$$T \xrightarrow{\alpha_j^*} M_j \xrightarrow{\beta_j^*} Y \qquad\qquad T \xrightarrow{\alpha_j^*} M_j \xrightarrow{\beta_j^*} Y \qquad\qquad T \xrightarrow{\alpha_j^*} M_j \xrightarrow{\beta_j^*} Y$$

$$\text{(a) Balanced.} \qquad\qquad \text{(b) Unbalanced } |\alpha_j^*| > |\beta_j^*|. \qquad\qquad \text{(c) Unbalanced } |\alpha_j^*| < |\beta_j^*|.$$
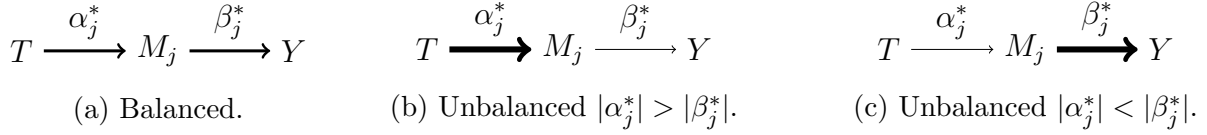
Figure 2: Visualization of balanced and unbalanced mediation pathways. Line widths of the solid links show relative magnitudes of the corresponding pathwise effects.

3.2 Procedure

To address the above inherent challenge of identifying effects of unbalanced pathways, we propose a new penalty that can be adaptive to the joint pathway effect of interest, referred to as Joint Adaptive Penalty (JAP) below. The overall idea is that less regularization shall be used to estimate α_j^* and β_j^* if a targeted mediation effect $\alpha_j^* \beta_j^*$ is significant, and vice versa. As true coefficients are unknown in practice, our proposal proceeds in two stages: first, obtain suitable initial estimates of the mediation effects $\alpha_j^* \beta_j^*$'s, and second, refit the model with regularization adjusted according to the initial estimates. For concreteness, we next describe our proposed method based on the ℓ_1-norm LASSO penalty of coefficients (Tibshirani, 1996; Zou, 2006). But we emphasize the idea is general and could potentially be extended under other regularizations.

In particular, consider a dataset with n independently and identically distributed observations $\{T_i, \boldsymbol{M}_i, Y_i, \boldsymbol{X}_i\}_{i=1}^n$. Let $\mathbf{T}_n = [T_1, \ldots, T_n]^\top \in \mathbb{R}^n$, $\mathbf{M}_n = [M_{ij}]_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p} \in \mathbb{R}^{n \times p}$, $\mathbf{Y}_n = [Y_1, \ldots, Y_n]^\top \in \mathbb{R}^n$, and $\mathbf{X}_n = [X_{ij}]_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant q} \in \mathbb{R}^{n \times q}$. We assume the data following the model (1), i.e.,

$$\begin{aligned} \mathbf{M}_n &= \mathbf{T}_n \boldsymbol{\alpha}^{*\top} + \mathbf{X}_n \boldsymbol{\zeta}_M^* + \mathbf{E}_n, \\ \mathbf{Y}_n &= \mathbf{T}_n \eta^* + \mathbf{M}_n \boldsymbol{\beta}^* + \mathbf{X}_n \boldsymbol{\zeta}_Y^* + \boldsymbol{\epsilon}_n, \end{aligned} \tag{2}$$

where $\mathbf{E}_n = [E_{ij}]_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$ and $\boldsymbol{\epsilon}_n = [\epsilon_1, \ldots, \epsilon_n]^\top$ are the error matrix/vector of the exposure-to-mediator and mediator-to-outcome models, whose rows/entries are independently and identically distributed with zero mean. For the simplicity of presentation, We let

$$\begin{aligned} \mathbf{D}_M &= (\mathbf{T}_n, \mathbf{X}_n) \in \mathbb{R}^{n \times (1+q)}, & \boldsymbol{\theta}_M &= (\boldsymbol{\alpha}, \boldsymbol{\zeta}_M^\top)^\top \in \mathbb{R}^{(1+q) \times p}, \\ \mathbf{D}_Y &= (\mathbf{T}_n, \mathbf{X}_n, \mathbf{M}_n) \in \mathbb{R}^{n \times (1+q+p)}, & \boldsymbol{\theta}_Y &= (\eta, \boldsymbol{\zeta}_Y^\top, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{1+q+p} \end{aligned}$$

6

represent design matrices and coefficients in the exposure-to-mediator and mediator-to-outcome models, respectively. We propose the following procedure that can estimate model coefficients and identify active mediation pathways.

Step 1: Initialization. Construct initial estimates $\hat{\boldsymbol{\alpha}}_n^0 = [\hat{\alpha}_{n1}^0, \ldots, \hat{\alpha}_{np}^0]^\top$ and $\hat{\boldsymbol{\beta}}_n^0 = [\hat{\beta}_{n1}^0, \ldots, \hat{\beta}_{np}^0]^\top$ for $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$, and let $\hat{\alpha}_{nj}^0 \hat{\beta}_{nj}^0$ be initial estimates for $\alpha_{nj}^* \beta_j^*$ across $j = 1, \ldots, p$.

Step 2: Joint Adaptive Penalized Regression. Estimate coefficients in the model (1) by solving

$$\hat{\boldsymbol{\theta}}_{M,n} = \underset{\boldsymbol{\theta}_M \in \mathbb{R}^{(1+q) \times p}}{\operatorname{argmin}} \ \ell_M(\boldsymbol{\theta}_M; \mathbf{D}_M, \mathbf{M}_n) + \lambda_{n\alpha} \sum_{j=1}^p \frac{|\alpha_j|}{\hat{w}_{nj,\alpha}}, \tag{3}$$

$$\hat{\boldsymbol{\theta}}_{Y,n} = \underset{\boldsymbol{\theta}_Y \in \mathbb{R}^{1+q+p}}{\operatorname{argmin}} \ \ell_Y(\boldsymbol{\theta}_Y; \mathbf{D}_Y, \mathbf{Y}_n) + \lambda_{n\beta} \sum_{j=1}^p \frac{|\beta_j|}{\hat{w}_{nj,\beta}}, \tag{4}$$

where $\ell_M(\cdot)$ and $\ell_Y(\cdot)$ represent the loss functions that fit data without regularizations imposed on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $\lambda_{n\alpha} \geqslant 0$ and $\lambda_{n\beta} \geqslant 0$ are regularization parameters for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, and we construct pathway adaptive weights

$$\hat{w}_{nj,\alpha} = |\hat{\alpha}_{nj}^0 \hat{\beta}_{nj}^0|^{\gamma_\alpha} + |\hat{\alpha}_{nj}^0|^{2\eta_\alpha} \qquad \text{and} \qquad \hat{w}_{nj,\beta} = |\hat{\alpha}_{nj}^0 \hat{\beta}_{nj}^0|^{\gamma_\beta} + |\hat{\beta}_{nj}^0|^{2\eta_\beta}, \tag{5}$$

with prespecified constants $\gamma_\alpha > 2\eta_\alpha > 0$ and $\gamma_\beta > 2\eta_\beta > 0$. The proposed joint adaptive penalized estimates $\hat{\boldsymbol{\alpha}}_n = (\hat{\alpha}_{n1}, \ldots, \hat{\alpha}_{np})^\top$ and $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{n1}, \ldots, \hat{\beta}_{np})^\top$ for $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ are constructed as taking the corresponding components in $\hat{\boldsymbol{\theta}}_{M,n}$ and $\hat{\boldsymbol{\theta}}_{Y,n}$.

Step 3: Identification of Active Mediation Pathways. The proposed penalized estimates can be used to identify active mediation pathways with selection efficiency improved from combining two pathway effects. In particular, we construct the set

$$\hat{\mathcal{A}}_n = \left\{ j : \ \hat{\alpha}_{nj} \hat{\beta}_{nj} \neq 0, j = 1, \ldots, p \right\}. \tag{6}$$

The proposed penalties in (3) and (4) extend the classical LASSO penalty and yield tailored estimation for mediation pathway analysis. When $\eta_\alpha = \gamma_\alpha = \eta_\beta = \gamma_\beta = 0$, the weights in (5) reduce to constants, and (3) and (4) become equivalent to fitting LASSO to the exposure-to-mediator and mediator-to-outcome models, respectively. Our proposed penalties achieve estimation that is adaptive to mediation pathway properties by incorporating the weights in (5). In particular, each weight in (5) consists of two parts: one proportional to the exponential of $|\hat{\alpha}_{nj}^0 \hat{\beta}_{nj}^0|$, the absolute value of the initial estimate of the mediation effect, and the other proportional to the exponential of the magnitude of the initial estimate of a single coefficient, i.e., $|\hat{\alpha}_{nj}^0|$ or $|\hat{\beta}_{nj}^0|$. Without the former, the proposed penalty reduces to the adaptive LASSO (Zou, 2006), which adaptively assigns a smaller penalty to a coefficient if its initial estimate is significant. Our proposal generalizes the idea by assigning a smaller penalty to a coefficient if either its own initial estimate is significant or the initial estimate of its corresponding mediation effect is significant. In this way, if a single coefficient is weak but its corresponding mediation effect is significant, it is less likely to be missed using the proposed penalties compared to using the LASSO or adaptive LASSO. This can be especially helpful under the types of unbalanced mediation pathways illustrated in Figure 2.

The proposed initialization+refitting strategy is advantageous for efficient computation and flexible implementation. In particular, the adaptive pathway information $|\hat{\alpha}_{nj}^0 \hat{\beta}_{nj}^0|$ from initialization is fixed during the refitting stage, and the refitting optimization in (3) and (4) are convex with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, allowing for the application of a wide range of standard solvers and flexible choice of loss functions. On the other hand, although it may be tempting to consider a penalty that is a function of $|\alpha_j \beta_j|$, i.e., directly examining targeted mediation effects without initialization, we note that $|\alpha_j \beta_j|$ is not a convex function with respect to (α_j, β_j) as pointed out by Zhao and Luo (2022). As a result, extra adjustments may be needed and the computation can be burdensome.

More generally, the proposed adaptive weights may be combined not only with LASSO but also with any other penalties that one prefers. As an example, for another penalty $\mathcal{P}(\alpha_j)$ used in fitting the $T \to M_j$ model, we can similarly reweight the penalty as $\mathcal{P}(\alpha_j)/\hat{w}_{nj,\alpha}$ to achieve adaptive properties under unbalanced mediation pathways. In the above discussion, we illustrate the proposed idea using $\mathcal{P}(\alpha_j) = |\alpha_j|$ which generalizes LASSO for its popularity and simplicity of implementation. We will stick to this choice in the remainder of this paper for ease of presentation, but we expect that similar theoretical and numerical properties can be achieved using other forms of penalties under suitable conditions.

Remark 2. *The proposed adaptive weighting strategy can be readily extended beyond the model discussed above, making it a versatile tool under a range of problems with potential unbalanced parameters. A key strength of our approach is its ability to borrow statistical efficiency across models through initialization, while preserving low computational cost in the refitting phase. This flexibility makes it particularly well-suited for examining target causal effects that depend on multiple model parameters, a common characteristic in mediation pathway models across diverse data types, including compositional data (Sohn and Li, 2019; Sohn et al., 2022; Jiang et al., 2024), categorical and count data (Hao et al., 2025), and survival outcome (Tchetgen Tchetgen, 2011). Furthermore, while the adaptive strategy is demonstrated under two-step pathways as illustrated in Figure 2, it can be similarly generalized to multi-step causal chains involving multiple parameters (Shi and Li, 2022), underscoring its potential as a comprehensive and scalable method for analyzing complex causal effects.*

3.3 Implementation

We next discuss the implementation of Steps 1 and 2 in the proposed procedure.

Step 1: To obtain reliable numerical results, the initial estimates $(\hat{\alpha}_{nj}^0, \hat{\beta}_{nj}^0)$ should adequately approximate the true values (α_j^*, β_j^*), while ensuring the stability of inverse weights in (3) and (4). To this end, we will combine ordinary least squares (OLS) estimates with appropriate lower bounds. We find this estimate exhibits both efficient and stable performance through the extensive numerical studies in Section 6. In particular, let $\hat{\alpha}_{nj,o}$ and $\hat{\beta}_{nj,o}$ denote the OLS estimates of α_j^* and β_j^* under the model (1). Then we define $\mathcal{T}(x, l) = \text{sign}(x)(\max\{|x| - l, 0\} + l)$ and construct

$$\hat{\alpha}_{nj}^0 = \mathcal{T}(\hat{\alpha}_{nj,o}, l_{nj,\alpha}), \qquad \text{and} \qquad \hat{\beta}_{nj}^0 = \mathcal{T}(\hat{\beta}_{nj,o}, l_{nj,\beta}), \tag{7}$$

where we set lower bounds $l_{nj,\alpha} = l_0 \cdot \hat{se}(\hat{\alpha}_{nj,o})$ and $l_{nj,\beta} = l_0 \cdot \hat{se}(\hat{\beta}_{nj,o})$ with $\hat{se}(\hat{\alpha}_{nj,o})$ and $\hat{se}(\hat{\beta}_{nj,o})$ denoting the estimated standard error of $\hat{\alpha}_{nj,o}$ and $\hat{\beta}_{nj,o}$ from the OLS regression. This construction ensures that $|\hat{\alpha}_{nj}^0| \geq l_{nj,\alpha}$ and $|\hat{\beta}_{nj}^0| \geq l_{nj,\beta}$, preventing the penalty weights from diverging too fast when OLS estimates $\hat{\alpha}_{nj,o}$ and $\hat{\beta}_{nj,o}$ approach zero, thereby enhancing robustness to poor OLS estimates. A detailed theoretical investigation of the asymptotic properties will be provided in Section 4. Beyond this specific construction (7), other initializations with similar properties could also be used.

Step 2: In the adaptive penalized regression, loss functions $\ell_M(\cdot)$ and $\ell_Y(\cdot)$ can be specified by users and take general forms. One simple yet effective choice is the quadratic loss function. When the dimension of $\boldsymbol{X}$ becomes higher, regularization of coefficients $\boldsymbol{\zeta}_M$ and $\boldsymbol{\zeta}_Y$ may also be added to improve estimation efficiency. For instance, we may choose

$$
\begin{aligned}
\ell_M(\boldsymbol{\theta}_M; \mathbf{D}_M, \mathbf{M}_n) &= \|\mathbf{M}_n - \mathbf{D}_M\boldsymbol{\theta}_M\|_{\mathrm{F}}^2 + \mathcal{P}_M(\boldsymbol{\zeta}_M), \\
\ell_Y(\boldsymbol{\theta}_Y; \mathbf{D}_Y, \mathbf{Y}_n) &= \|\mathbf{Y}_n - \mathbf{D}_Y\boldsymbol{\theta}_Y\|_{\mathrm{F}}^2 + \mathcal{P}_Y(\boldsymbol{\zeta}_Y, \eta),
\end{aligned}
\tag{8}
$$

where for $A \in \{M, Y\}$, $\mathcal{P}_A(\cdot)$ represents the penalty of the corresponding input coefficients. In (8), setting $\mathcal{P}_A(\cdot) = 0$ gives vanilla quadratic loss functions, and $\mathcal{P}_A(\cdot)$ can also be a function with respect to only a subset of input parameters, e.g., the coefficient of the intercept is often not penalized in practice. Numerically, our extensive experiments suggest that (8) could yield sufficiently good and stable empirical results, as will be detailed in Section 6. Computationally, under (8), optimizations (3) and (4) may be transformed to examining penalized coefficients only, allowing the potential to further reduce computational complexity. This is shown by Proposition 1 below. To facilitate the subsequent presentation, we define $\mathbf{R}_M = \mathbf{M}_n$ and $\mathbf{R}_Y = \mathbf{Y}_n$. Moreover, for $A \in \{M, Y\}$, let $\boldsymbol{\theta}_{AP}$ and $\boldsymbol{\theta}_{AU}$ represent penalized and unpenalized coefficients in $\boldsymbol{\theta}_A$, respectively, and let $\mathbf{D}_{AP}$ and $\mathbf{D}_{AU}$ be corresponding columns in the design matrix $\mathbf{D}_A$, respectively.

Proposition 1. *For $A \in \{M, Y\}$, assume $\mathbf{D}_A$ has full column rank, and $\mathcal{P}_A(\cdot)$ is convex. Solving $(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_Y)$ by (3) and (4) is equivalent to solving that for $A \in \{M, Y\}$,*

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{AP} &= \underset{\boldsymbol{\theta}_{AP}}{\arg\min} \|\mathbf{P}_{AU}^{\perp}(\mathbf{R}_A - \mathbf{D}_{AP}\boldsymbol{\theta}_{AP})\|_F^2 + \bar{\mathcal{P}}_A(\boldsymbol{\theta}_{AP}), \\
\hat{\boldsymbol{\theta}}_{AU} &= \mathbf{D}_{AU}^{\dagger}(\mathbf{R}_A\mathbf{D}_{AP}\hat{\boldsymbol{\theta}}_{AP}),
\end{aligned}
\tag{9}
$$

where $\mathbf{D}_{AU}^{\dagger} = (\mathbf{D}_{AU}^{\top}\mathbf{D}_{AU})^{-1}\mathbf{D}_{AU}^{\top}$, $\mathbf{P}_{AU}^{\perp} = \mathbf{I}_{n \times n} - \mathbf{D}_{AU}\mathbf{D}_{AU}^{\dagger}$, $\mathbf{I}_{n \times n}$ denotes an $n \times n$ identity matrix, and $\bar{\mathcal{P}}_M(\boldsymbol{\theta}_{MP}) = \lambda_{n\alpha} \sum_{j=1}^{p} |\alpha_j|/\hat{w}_{nj,\alpha} + \mathcal{P}_M(\boldsymbol{\zeta}_M)$ and $\bar{\mathcal{P}}_Y(\boldsymbol{\theta}_{YP}) = \lambda_{n\beta} \sum_{j=1}^{p} |\beta_j|/\hat{w}_{nj,\beta} + \mathcal{P}_Y(\boldsymbol{\zeta}_Y, \eta)$ represent two augmented penalties.

Proposition 1 shows that optimizing penalized losses only requires examining $\boldsymbol{\theta}_{AP}$, which typically has a lower dimension than $\boldsymbol{\theta}_A$. This reduction in dimensionality might help decrease computational complexity when iterative optimization is needed. Moreover, once the penalized coefficient estimate $\hat{\boldsymbol{\theta}}_{AP}$ is obtained, unpenalized coefficient estimate $\hat{\boldsymbol{\theta}}_{AU}$ can be computed through closed-form formulae. When $\mathcal{P}_A(\cdot)$ takes ℓ_1-norm based penalty, $\hat{\boldsymbol{\theta}}_{AP}$ can be easily obtained by a standard LASSO-based solver, such as the R package `glmnet` (Friedman et al., 2010). When $\mathcal{P}_M(\cdot) = 0$, a closed-form formula for penalized coefficients $\hat{\boldsymbol{\alpha}}$ can

in fact be obtained; see Remark 4 in Appendix. While we discuss the implementation under (8), we emphasize that the proposed adaptive strategy in (3) and (4) is general and can be used with any other loss functions.

4 Asymptotic Theory

This section establishes asymptotic guarantees for the proposed procedure in Section 3, giving insights into the advantage of the joint adaptive penalty for unbalanced pathways. We begin by examining properties of the initialization and then analyze the joint adaptive estimators.

4.1 Initialization

As discussed in Section 3.3, ideal initial estimates should accurately approximate the true coefficients with stability near zero. This is formalized as Condition 3 below.

Condition 3 (Initialization). *The constructed initial estimates $\hat{\boldsymbol{\alpha}}_n^0$ and $\hat{\boldsymbol{\beta}}_n^0$ in (5) satisfy*

(i) *$\sqrt{n}(\hat{\alpha}_{nj}^0 - \alpha_j^*) = O_p(1)$ and $\sqrt{n}(\hat{\beta}_{nj}^0 - \beta_j^*) = O_p(1)$;*

(ii) *$1/\hat{\alpha}_{nj}^0 = O_p(\sqrt{n})$ and $1/\hat{\beta}_{nj}^0 = O_p(\sqrt{n})$.*

We will show that the initial estimates in Section 3.3 satisfy Condition 3. We require the following condition.

Condition 4 (Moments). *Assume that*

(i) *$\operatorname{cov}(\boldsymbol{E}) = \boldsymbol{\Sigma}$ and $\operatorname{cov}(\epsilon) = \sigma^2$, where $\boldsymbol{\Sigma}$ and σ^2 are fixed and positive (definite);*

(ii) *as $n \to \infty$, $\mathbf{X}_n^\top \mathbf{X}_n / n \to \boldsymbol{\Sigma}_X$, $\|\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n\|_2^2 / n \to \sigma_T^2$, and $\max_{1 \le i \le n} \left(\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n \right)_i^2 / n \to 0$, where $\boldsymbol{\Sigma}_X$ and σ_T^2 are fixed and positive (definite), $\mathbf{P}_{\mathbf{X}_n}^\perp = \mathrm{I}_{n \times n} - \mathbf{X}_n (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n$, and $\left(\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n \right)_i$ represents the ith element of $\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n$.*

Condition 4 imposes regularity conditions on the second-order moments of the error terms and covariates in the model (2), which are common in the regression analysis (Seber and Lee, 2003).

Proposition 2. *Under the model (2) and Condition 4, the initial estimates $\hat{\boldsymbol{\alpha}}_n^0$ and $\hat{\boldsymbol{\beta}}_n^0$ constructed in (7) satisfy Condition 3.*

4.2 Estimation and Selection Consistency

Under the above framework, we define the targeted set of active mediators as the indices in the following set:
$$\mathcal{A}^* = \left\{ j : \alpha_j^* \beta_j^* \neq 0, \quad j = 1, \ldots, p \right\}, \tag{10}$$

which can be interpreted as the mediators for which pathways $T \to M_j$ and $M_j \to Y$ have nonzero signals. We now study the asymptotics of our method. For ease of illustration, we consider that quadratic loss functions are used in (3) and (4), i.e.,

$$
\begin{aligned}
\ell_M(\boldsymbol{\theta}_M; \mathbf{D}_M, \mathbf{M}_n) &= \|\mathbf{M}_n - \mathbf{D}_M \boldsymbol{\theta}_M\|_{\mathrm{F}}^2, \\
\ell_Y(\boldsymbol{\theta}_Y; \mathbf{D}_Y, \mathbf{Y}_n) &= \|\mathbf{Y}_n - \mathbf{D}_Y \boldsymbol{\theta}_Y\|_{\mathrm{F}}^2.
\end{aligned}
\tag{11}
$$

These loss functions are common in practice and clarify the essence of signal adaptation achieved by the proposed method. More generally, we expect that conclusions for other loss functions can be similarly established given suitable assumptions.

Theorem 1. *Assume Conditions 3-4 under the model* (2). *Suppose the tuning parameters satisfy $n^{1/2-\eta_\alpha} \ll \lambda_{n\alpha} \ll n^{1/2}$ and $n^{1/2-\eta_\beta} \ll \lambda_{n\beta} \ll n^{1/2}$, where $0 < 2\eta_\alpha < \gamma_\alpha$ and $0 < 2\eta_\beta < \gamma_\beta$ are specified in* (5). *Then with the use of the quadratic loss functions in* (11), *the proposed estimates in* (3) *and* (4) *satisfy $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_F^2 + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\|_F^2 \to_p 0$. Moreover, $\hat{\mathcal{A}}_n$ constructed in* (6) *satisfies*

$$
\lim_{n \to \infty} \Pr(\hat{\mathcal{A}}_n = \mathcal{A}^*) = 1.
\tag{12}
$$

Theorem 1 provides asymptotic guarantees for the joint adaptive penalty in both parameter estimation and active pathway selection, laying the theoretical foundation for its practical utility.

To further gain insights into how the proposed method adapts to unbalanced pathways, we compare the proposed joint adaptive penalty weights with those from a standard adaptive LASSO penalty, i.e., comparing weights $\hat{w}_{nj,\alpha}$ and $\hat{w}_{nj,\beta}$ in (3) and (4) with the weights in adaptive LASSO: $\hat{w}_{\mathrm{AL},nj,\alpha} = |\hat{\alpha}_{nj}^0|^{2\eta_\alpha}$ and $\hat{w}_{\mathrm{AL},nj,\beta} = |\hat{\beta}_{nj}^0|^{2\eta_\beta}$, respectively.

Proposition 3. *Assume Conditions 3 and 4, $0 < 2\eta_\alpha < \gamma_\alpha$ and $0 < 2\eta_\beta < \gamma_\beta$. Then*

$$
\begin{aligned}
\frac{\hat{w}_{nj,\alpha}}{\hat{w}_{\mathrm{AL},nj,\alpha}} &\to_p 1 + |\alpha_j^*|^{\gamma_\alpha - 2\eta_\alpha} |\beta_j^*|^{2\gamma_\alpha}, \\
\frac{\hat{w}_{nj,\beta}}{\hat{w}_{\mathrm{AL},nj,\beta}} &\to_p 1 + |\alpha_j^*|^{2\gamma_\beta} |\beta_j^*|^{\gamma_\beta - 2\eta_\beta}.
\end{aligned}
\tag{13}
$$

Proposition 3 shows that for an active pathway $T \to M_j \to Y$ where α_j^* and β_j^* are nonzero, the two ratios in (13) are greater than 1. This implies that the proposed joint adaptive penalty is of smaller order compared to the adaptive LASSO penalty. In contrast, for a nonactive pathway where at least one of α_j^* and β_j^* is zero, both ratios in (13) equal 1. This implies that the joint adaptive penalty and the adaptive LASSO penalty would remain at the same order. Therefore, it is easier to distinguish the active pathways from the remaining ones with the joint adaptive penalty.

5   Related Methods

In this section, we review related methods in the literature to set the stage for the numerical comparisons in Section 6. Since our proposed method emphasizes regularized estimation of

the model (1) along with the identification of active pathways through individual mediators, we focus on methods with comparable setups and objectives, specifically those addressing the dual goals of estimation and identification. It is worth noting that the proposed method is flexible and can incorporate other methodological advances; see Remark 3. For a broader overview on modern mediation analysis with multiple mediators, we refer readers to comprehensive reviews (Clark-Boucher et al., 2023; Blum et al., 2020; Zeng et al., 2021). In the following, we organize our discussion by separately examining frequentist and Bayesian approaches.

Under the frequentist perspective, mediation effects, i.e., $\alpha_j^* \beta_j^*$ under the model (1), are typically treated as fixed parameters. One research line utilizes a fitting-and-testing strategy: first fitting the T-$\boldsymbol{M}$ and $\boldsymbol{M}$-Y models in (1) separately and obtaining asymptotic distributions of the coefficient estimates; second, identifying active mediation pathways by testing hypotheses $H_{0,j} : \alpha_j^* \beta_j^* = 0$ for $j = 1, \ldots, p$ with appropriate adjustments for multiple comparisons. For example, Zhang et al. (2016) first perform an additional initial screening to reduce the number of mediators by applying sure independence screening (Fan and Lv, 2008) to the $\boldsymbol{M}$-Y model. They then fit T-$\boldsymbol{M}$ and $\boldsymbol{M}$-Y models and obtain asymptotic distributions of the coefficient estimates by the ordinary least squares regression and minimax concave penalty regularized regression (Zhang, 2010), respectively. Gao et al. (2019) propose a similar approach but obtain $\hat{\beta}_j$ estimates and their asymptotic distributions through debiased LASSO (Zhang and Zhang, 2014; van de Geer et al., 2014) to mitigate biases. Zhang (2022) applies the adaptive LASSO to both T-$\boldsymbol{M}$ and $\boldsymbol{M}$-Y models with asymptotic distributions derived in Zou (2006). As an alternative to the above methods based on separate model fitting, Zhao and Luo (2022) examines the joint model (1) with a penalty involving $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously. In particular, it proposes to penalize mediation effects by $\sum_{j=1}^p |\alpha_j \beta_j|$, alongside elastic net regularizations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ individually (Zou and Hastie, 2005). The mediators with $\hat{\alpha}_j \hat{\beta}_j \neq 0$ are identified as active.

Remark 3. *While Section 3 focuses on model selection through regularization, our proposed method can be seamlessly combined with other widely used strategies, such as sure independence screening and multiple hypothesis testing, similar to the methods reviewed above. Notably, our theoretical analysis has derived asymptotic distributions of coefficient estimates, providing the foundation needed to apply established testing procedures for mediation pathways (Dai et al., 2022; Liu et al., 2022; He et al., 2024). To improve the accuracy of inference under finite sample sizes, future research could explore high-order refinements for uncertainty quantification (Chatterjee and Lahiri, 2013). In current simulations and data analysis, our proposed method in Section 3 demonstrates sufficiently good performance. For clarity and conciseness, we do not pursue post-selection inference in this paper and leave these intriguing directions for future investigation.*

Under the Bayesian perspective, estimating mediation effects and identifying active mediation pathways can be framed as a Bayesian shrinkage estimation problem for the coefficients under (1). In this vein, Song et al. (2020) specifies separate shrinkage priors for α_j and β_j, followed by model selection based on posterior inclusion probabilities. Alternatively, Song et al. (2021) proposes joint prior distributions on (α_j, β_j), utilizing hard thresholding to specifically target non-zero mediation effects.

6 Simulation Studies

We next evaluate the finite-sample performance of the proposed method and compare with the methods reviewed in Section 5 through comprehensive simulations. As highlighted above, the proposed penalty can be advantageous for identifying active yet unbalanced mediation pathways. In the following, Section 6.1 presents the simulation settings, encompassing both balanced and unbalanced pathways, and provides implementation details for all methods under comparison. Section 6.2 reports numerical results focusing on the accuracy of selecting active mediation pathways, highlighting the unique advantages of our proposed penalty.

6.1 Setup

Data Generation. We simulate data under the model (1), where T is randomly assigned according to a Bernoulli distribution with success probability 0.5, the mediator-to-outcome-model noise ϵ is independently drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and pretreatment covariates $\boldsymbol{X}$ are set to be empty for simplicity. We consider two cases of exposure-to-mediator-model noises $\boldsymbol{E}$ below to examine the impact of different dependence patterns.

(I) Draw $\boldsymbol{E}$ independently from a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = (\rho^{|i-j|})$, where $\rho \in \{0, 0.4, 0.8\}$, corresponding to uncorrelated, moderate-correlation, and high-correlation settings.

(II) After drawing n independent copies of $\boldsymbol{E}$ as in Case (I), randomly permute p dimensions of the n copies together to allow for correlations between mediators that are non-adjacent in terms of dimension indices. Note $\rho = 0$ is excluded in this case as permutation would not change the distribution.

In each case above, we set $p = 150$, $\eta^* = 1$, and $\sigma^2 = 1$. We vary the sample size $n \in \{500, 1000, 1500, 2000\}$ to understand its effect.

For the pairwise coefficients $\{(\alpha_j^*, \beta_j^*) : j = 1, \ldots, p\}$, we consider six groups of patterns: for each group $k = 1, \ldots, 6$, we set

$$(\alpha_j^*, \beta_j^*) = C \cdot (a_k, b_k), \quad \text{for } j \in \mathcal{G}_k := \{(k-1)p/6 + 1, \ldots, kp/6\}, \tag{14}$$

where values of (a_k, b_k) are defined in Table 1, and C^2 represents the effect magnitude.

	Active			Inactive		
Group k	1	2	3	4	5	6
a_k	1	$1/\delta$	δ	0	1	0
b_k	1	δ	$1/\delta$	1	0	0

Table 1: Patterns of six groups of coefficient pairs $(\alpha_j^*, \beta_j^*) = C \cdot (a_k, b_k)$ with $|\delta| < 1$.

We aim to select the mediators M_j with $j \in \cup_{k=1}^3 \mathcal{G}_k$. These three active groups satisfy $\alpha_j^* \beta_j^* = C^2 \neq 0$, and correspond to three types of relationships between treatment-mediator and mediator-outcome effects, i.e.,

$$\frac{\alpha_j^*}{\beta_j^*} = 1 \ \text{ for } j \in \mathcal{G}_1, \qquad \frac{\alpha_j^*}{\beta_j^*} = \frac{1}{\delta^2} > 1 \ \text{ for } j \in \mathcal{G}_2, \quad \text{and} \quad \frac{\alpha_j^*}{\beta_j^*} = \delta^2 < 1 \ \text{ for } j \in \mathcal{G}_3$$

when $|\delta| < 1$. When fixing the direct effect, the above relationships can give three types of directed acyclic graphs visualized as in Figure 3. As $|\delta|$ becomes smaller, the discrepancy

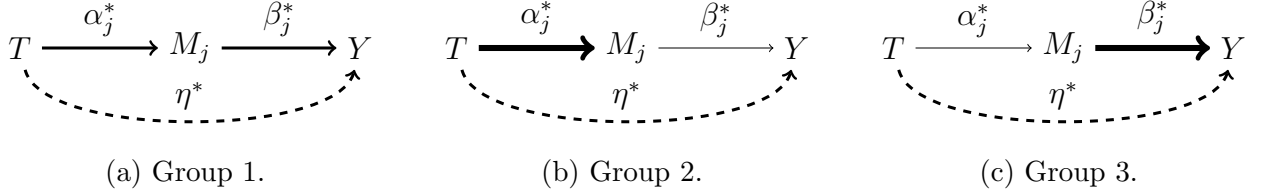(a) Group 1.   (b) Group 2.   (c) Group 3.

Figure 3: Directed acyclic graphs of three active groups. Line widths of the solid links indicate relative magnitudes of the corresponding pathwise effects. Dashed line represents the direct effect that is fixed in the analysis.

between α_j^* and β_j^* becomes larger, which could make it more challenging for methods that examine pathwise effects α_j^* and β_j^* separately. We aim to exclude the mediators M_j with $j \in \cup_{k=4}^6 \mathcal{G}_k$. The three inactive groups satisfy $\alpha_j^* \beta_j^* = 0$ and correspond to three directed acyclic graphs visualized as in Figure 4. In our simulations below, we set $C = 1$.

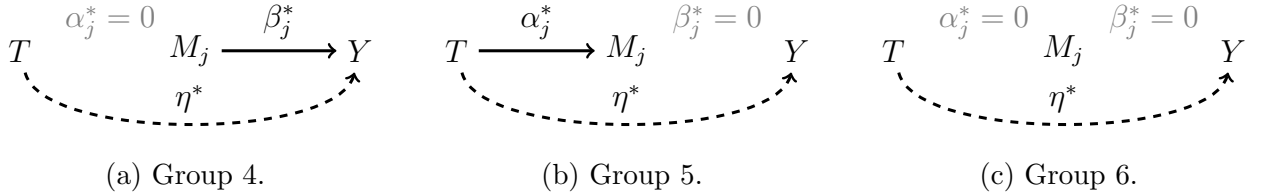(a) Group 4.   (b) Group 5.   (c) Group 6.

Figure 4: Directed acyclic graphs of three inactive groups. The absence of an edge indicates no causal effects.

Implementation Details. We implement the proposed method as discussed in Section 3.3. For l_0 used in the initialization, our experiments suggest that the results are not sensitive to the choice of l_0, and we fix $l_0 = 5$ below. Moreover, we use vanilla quadratic loss functions in (3) and (4). For the hyperparameters in (5), we consider $\gamma_\alpha, \gamma_\beta \in \{0.75, 1, \ldots, 3\}$ and $\eta_\alpha, \eta_\beta \in \{0.25, 0.5, 0.75, 1, 1.25\}$, subject to the constraints $\gamma_\alpha > 2\eta_\alpha$ and $\gamma_\beta > 2\eta_\beta$. For $\lambda_{n\alpha}$ in (3) and $\lambda_{n\beta}$ in (4), we consider exponentially spaced values ranging from e^0 to e^5 and from e^3 to e^8, respectively, using a step size of 0.1 in the exponent. We select the hyperparameters $(\gamma_\alpha, \eta_\alpha, \lambda_{n\alpha})$ and $(\gamma_\beta, \eta_\beta, \lambda_{n\beta})$ by balancing variable selection stability (VSS, Sun et al. (2013)) and mean squared error (MSE) in the following sense. Given each candidate pair $\boldsymbol{\gamma}' = (\gamma_\alpha, \eta_\alpha)$, we first identify the smallest $\lambda_{n\alpha}$ value that achieves the highest VSS computed via 5-fold cross-validation and denote it as $\lambda_{n\alpha}(\boldsymbol{\gamma}')$. We then fit the model for

each $(\boldsymbol{\gamma}', \lambda_{n\alpha}(\boldsymbol{\gamma}'))$ and compute MSE on the full data. The final hyperparameters are chosen to minimize MSE. This two-stage procedure aims to achieve good model fit and robust variable selection simultaneously. More details on the implementation can be found in Jiang (2025).

We compare with the methods reviewed in Section 5. For the ease of reference, we adopt abbreviations of methods as used in Clark-Boucher et al. (2023). In particular, HIMA (Zhang et al., 2016), HDMA (Gao et al., 2019), and MedFix (Zhang, 2022) are conducted using codes by Clark-Boucher et al. (2023). BSLMM (Song et al., 2020) and PTG (Song et al., 2021) are conducted by R package bama. For the frequentist methods involving multiple testing, we choose 0.05 as the significance level and apply Bonferroni correction. For Bayesian methods, we use 0.5 as the cutoff of posterior inclusion probability. All other hyperparameters remain at their default values. As Pathway LASSO incurs prohibitive computational costs and demonstrates low selection accuracy in our settings, we present a separate analysis of it in Appendix F.

In addition, to demonstrate the unique feature of the proposed penalty for mediation pathway selection, we also compare with two established regularized regression methods for model selection: LASSO (Tibshirani, 1996) and adaptive LASSO (Zou, 2006, abbreviated as AL). Under (3) and (4), LASSO corresponds to $\hat{w}_{nj,\alpha} = 1$ and $\hat{w}_{nj,\beta} = 1$, and AL corresponds to $\hat{w}_{nj,\alpha} = |\hat{\alpha}_{nj}^0|^{2\eta_\alpha}$, $\hat{w}_{nj,\beta} = |\hat{\beta}_{nj}^0|^{2\eta_\beta}$. For a fair comparison, we tune the hyperparameters of LASSO and AL using similar strategy to that above for our proposed method and construct selection set same as in (6).

6.2   Numerical Results on Selection Accuracy

Figures 5 and 6 display estimated probabilities of correctly recovering the active set $\mathcal{A}^*$ over 100 Monte Carlo replicates under Cases (I) and (II) of mediator-noise simulations, respectively. For clear presentation and discussion, we group methods into three classes: penalty-based (LASSO, AL, JAP), testing-based (HIMA, HDMA, MedFix), and Bayesian methods (BSLMM, PTG).

The empirical results show that JAP achieves the highest accuracy across most regimes. In fact, it was only outperformed by HIMA, LASSO, and AL when $n = 500$, $\rho = 0.8$, and δ is close to zero. For a fixed n, the accuracy of JAP increases as ρ becomes smaller. For a fixed ρ, the accuracy of JAP increases as n increases. These observations suggest that JAP might be advantageous under scenarios with small correlations and large sample sizes. Fixing (n, ρ), the selection accuracies of all methods, except PTG, decrease as δ becomes smaller. This is reasonable since Table 1 implies that as δ decreases, β_j^* in Group 2 and α_j^* in Group 3 would become smaller, making it more challenging to identify these small nonzero α_j^* and β_j^* separately. Indeed, all the methods tend to have more false negatives under smaller δ in our numerical studies. Our proposed JAP leverages the information on the product effect $\alpha_j^* \beta_j^*$ and thus can achieve higher accuracy than the other methods do in the challenging small-δ scenarios.

For the other two penalty-based methods, LASSO and AL, their relative performances vary across scenarios. In Case (I), LASSO tends to outperform AL under a large n and a larger ρ, whereas the relationship is at times reversed in Case (II). The results suggest that LASSO might gain from larger within-group correlations but not between-group correlations.
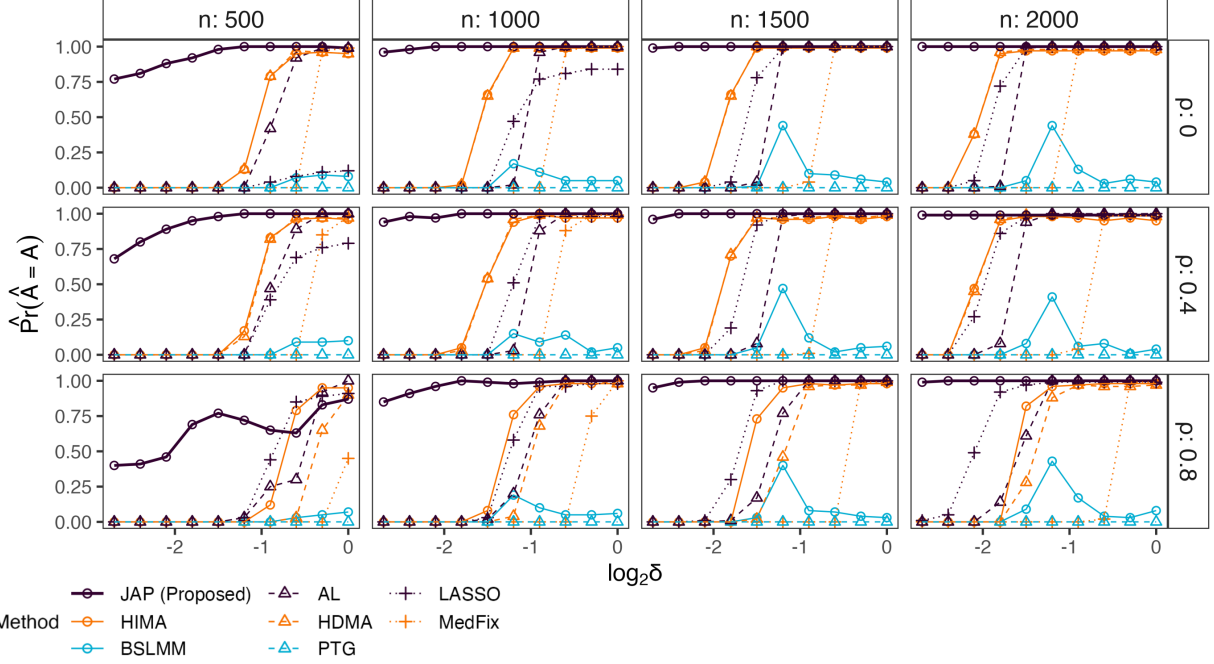
Figure 5: The empirical probability of selecting the active set $\mathcal{A}^*$ correctly. The rows correspond to different correlations ρ and the columns correspond to different sample sizes n. We use three colors to represent three method classes (penalty-based, testing-based, and Bayesian), while employing distinct linetypes and point shapes to differentiate within each method class. JAP is highlighted with a wider line.

For the three testing-based methods, we observe that HIMA consistently performs the best, while HDMA is close to HIMA under no-correlation and low-correlation regimes but worse than HIMA in high-correlation regimes. MedFix has the lowest selection accuracy among the three in all scenarios. Both Bayesian methods, PTG and BSLMM, exhibit relatively low selection accuracies and hardly improve as n or δ increases. The underperformance of PTG may be attributed to its design for sparse settings. The accuracy of BSLMM, unlike the other methods, shows non-monotonicity with respect to δ, which may arise from our fixed choice of hyperparameters; it may perform better if hyperparameters can be chosen in an appropriate data-driven way.

7   Data Analysis

We demonstrate the use of the proposed method by analyzing a gastrectomy dataset from the Curated Gut Microbiome-Metabolome Data Resource (Muller et al., 2022). Our goal is to investigate how the relationships between gastrectomy and the total cholesterol (TC) levels in patients may be mediated through gut microbiome. Gastrectomy, the surgical removal of all or part of the stomach, is commonly performed to treat conditions such as gastric cancer (Penna and Allum, 2013), peptic ulcer (Maki et al., 1967), and morbid obesity (Bennett et al., 2007). TC is a crucial health indicator and is known to be associated with cardiovascular disease risks, including acute myocardial infarction and stroke (Jeong et al., 2018). Previous
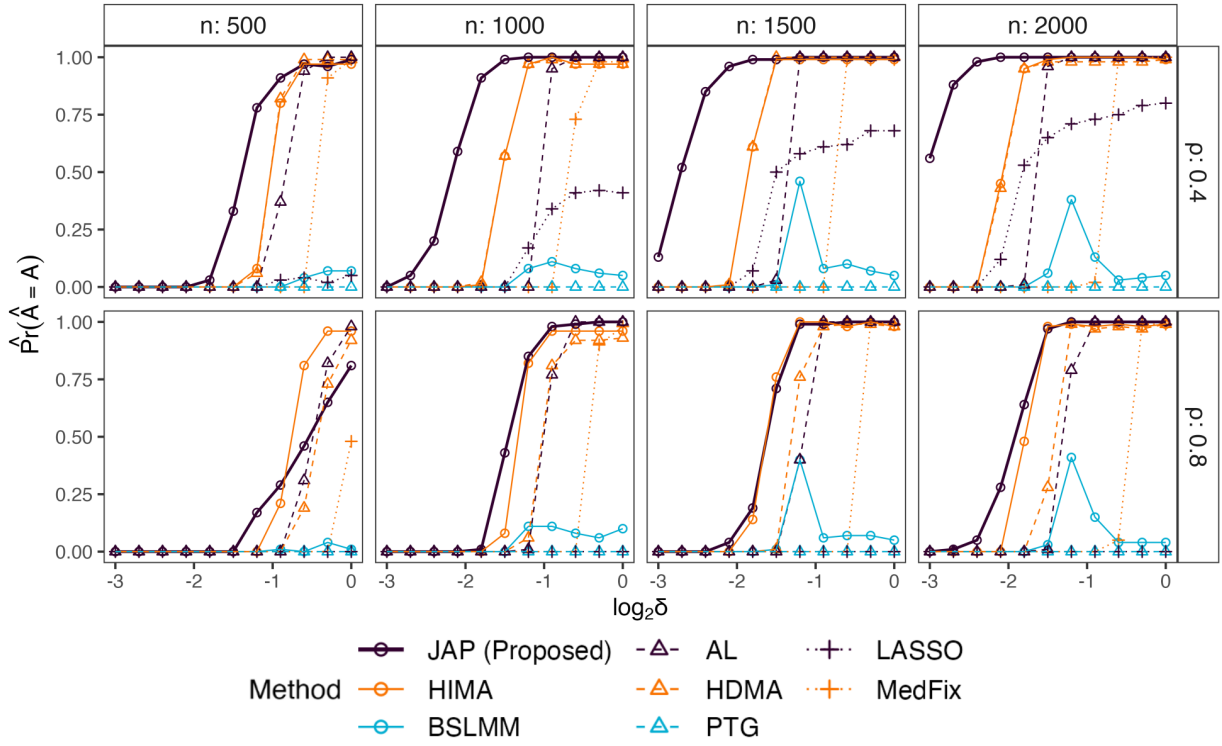
Figure 6: The probability of selecting the active set $\mathcal{A}^*$ correctly with randomly reordered mediator model noises $\mathbf{E}_n$.
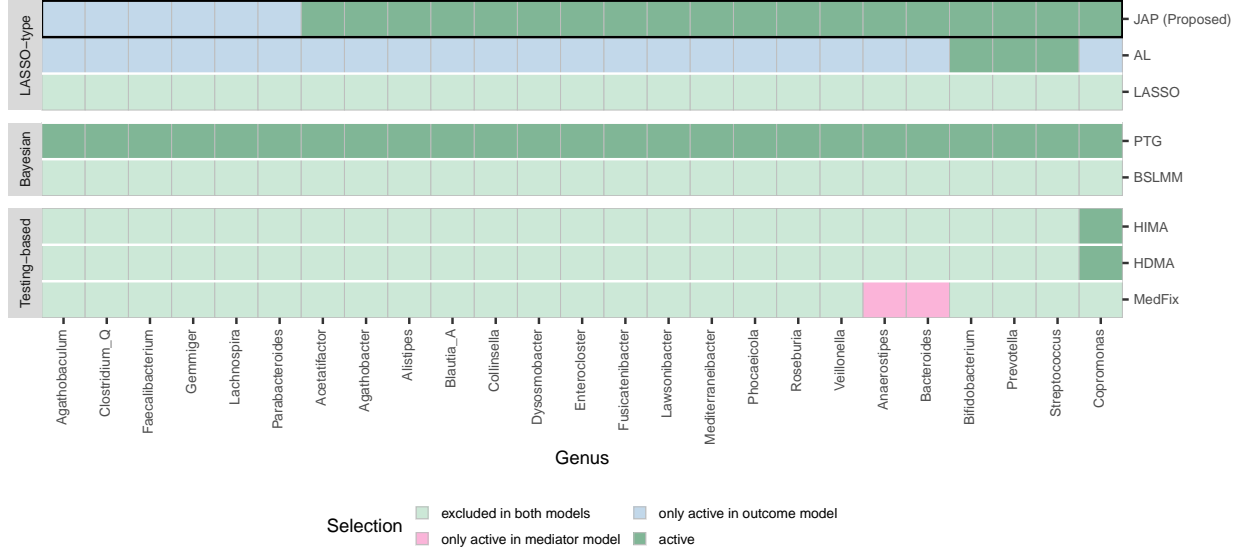
Figure 7: The genera selected by different methods. We use different colors to indicate the four selection results: excluded by both the mediator and outcome models, only included by the mediator model, only included by the outcome model, and active mediators that are included in both models.

studies indicate that TC levels tend to reduce after gastrectomy (Tromba et al., 2017; Lee et al., 2015), and both gastrectomy and TC are related to the gut microbiome (Deng et al., 2023; Vourakis et al., 2021; Erawijantari et al., 2020). Mediation analysis can further reveal the role the gut microbiome plays in the process of gastrectomy influencing TC, enhancing our understanding towards the fundamental biological mechanisms underlying the observed effect. Specifically, we model gastrectomy as the treatment, TC levels as the outcome, and observed gut microbiome as potential mediators and aim to identify microbiome genera with significant mediation effects.

In particular, the studied dataset involves patients undergoing total colonoscopy at the National Cancer Center Hospital, Tokyo, Japan. Their demographic information, clinical parameters, and faecal samples have been collected. More detailed information of data collection can be found in Erawijantari et al. (2020). After eliminating the records with missing data, our sample contains 82 subjects, consisting of 42 participants who have undergone gastrectomy for gastric cancer and 40 controls. For the gut microbiome, we first filter out genera with a prevalence lower than 0.9 to concentrate on the core microbiome shared across study subjects. To accommodate the compositional nature of the count data (Gloor et al., 2017), we apply a centered log-ratio (CLR) transformation with a pseudocount of 1. We then remove the genera with average transformed abundance smaller than 5, resulting in a total of 25 genera. Throughout the analysis, we adjust for covariates age and gender with ℓ_1-penalty for their coefficients in the outcome model when fitting the three penalty-based methods.

Figure 7 displays the results of the eight methods, implemented and organized in three classes as in Section 6. For each method (rows) and each microbiome genus j (columns), we use four colors to represent four types of results: (a) both treatment-to-mediator effect

α_j and mediator-to-outcome effect β_j are significant, indicating the existence of mediation effect $\alpha_j\beta_j$ through the corresponding microbiome genera, (b) only α_j is significant, (c) only β_j is significant, and (d) both α_j and β_j are not significant.

In the results, MedFix, BSLMM and LASSO do not identify any significant effects. On the other hand, PTG suggests all genera have significant mediation effects, which may be overly optimistic as it tends to give more false positives in our simulations. HIMA and HDMA are similar and only identify one genus, *Copromonas*, with significant mediation effect. AL suggests all mediator-to-outcome effects β_j's are significant, whereas only three genera have significant mediation effects $\alpha_j\beta_j$. For those three genera, *Bifidobacterium* and *Prevotella* have been discovered to be associated with the gastrectomy and TC (Lin et al., 2018; Sánchez-Alcoholado et al., 2019; Jia et al., 2023; Deng et al., 2023; Vourakis et al., 2021; Yu et al., 2020), and the fluctuation of *Streptococcus* abundance after gastrectomy has also been observed before (Yu et al., 2020).

The proposed method JAP identifies 19 genera with significant mediation effects, including *Roseburia* and *Bacteroides*, whose relationships with gastrectomy and TC have been reported in the literature (Yu et al., 2020; Kissmann et al., 2024; Deng et al., 2023; Wu et al., 2022; Jia et al., 2023). Compared to the other methods, it can yield more discoveries while maintaining some discernment. Also, JAP is similar to AL in terms of identifying all mediator-to-outcome effects β_j's, which is understandable by the connections between their constructions. But JAP identifies more treatment-to-mediator effects α_j's than AL does, which could be attributed to the effective use of joint pathwise information in JAP. For example, existing studies have shown that the abundances of *Bacteroides* and *Veillonella* changed after gastrectomy (Yu et al., 2020; Kissmann et al., 2024; Deng et al., 2023; Wu et al., 2022). Their treatment-to-mediator effects α_j's have been identified by JAP but not AL. The results suggest that the proposed JAP can be a powerful method in practice.

8 Discussion

In this work, we propose a novel joint adaptive penalty for regularized mediation analysis. Our approach incorporates adaptive weights informed by the significance of mediation effects to improve statistical efficiency in estimating and identifying unbalanced mediation pathways. Theoretically, we establish rigorous asymptotic guarantee for controlling the estimation error and consistent selection of active mediation pathways. Numerically, we demonstrate the adaptability and scalability of the proposed method across diverse scenarios. Our proposed strategy provides a flexible and powerful framework for analyzing mediation effects, opening several new avenues for future research.

First, as noted in Remark 2, the adaptive weights can be extended to mediation analysis models for diverse data types and causal chains (Jiang et al., 2024; Hao et al., 2025; Tchetgen Tchetgen, 2011). While our core idea of incorporating the significance of target causal effects remains applicable across different models and targets, the constructions and performance of extensions would require case-by-case investigation, presenting important questions and opportunities for future exploration.

Second, although this paper considers scenarios with fixed-dimensional mediators for the ease of understanding and illustration, we anticipate that our theoretical results can

be extended to accommodate high-dimensional mediators under appropriate assumptions (Huang et al., 2008). Methodologically, the proposed method could potentially be enhanced by combining with other widely used strategies, such as sure independence screening (Fan and Lv, 2008). Developing appropriate refinements and comprehensive understanding for the proposed adaptive weights in high-dimensional scenarios would be important areas for future research.

Third, this paper focuses on model estimation and consistent selection, where achieving the latter typically requires strong assumptions on signal strengths. More generally, alternative measures of variable selection performance, such as false discovery rate, could also be considered. This may be achieved by developing post-selection inference tools for the proposed adaptive penalty and combining them with multiple testing methods (Liu et al., 2022), as discussed in Remark 3. Understanding how the improved estimation efficiency influences the final performance of pathway identification would be an interesting future research direction.

Funding

Acknowledgement

References

Abrishamcar, S., Chen, J., Feil, D., Kilanowski, A., Koen, N., Vanker, A., Wedderburn, C. J., Donald, K. A., Zar, H. J., Stein, D. J., et al. (2022). Dna methylation as a potential mediator of the association between prenatal tobacco and alcohol exposure and child neurodevelopment in a south african birth cohort. *Translational psychiatry*, 12(1):418.

Bai, L., Benmarhnia, T., Chen, C., Kwong, J. C., Burnett, R. T., van Donkelaar, A., Martin, R. V., Kim, J., Kaufman, J. S., and Chen, H. (2022). Chronic exposure to fine particulate matter increases mortality through pathways of metabolic and cardiovascular disease: insights from a large mediation analysis. *Journal of the American Heart Association*, 11(22):e026660.

Bellavia, A., James-Todd, T., and Williams, P. L. (2019). Approaches for incorporating environmental mixtures as mediators in mediation analysis. *Environment international*, 123:368–374.

Bennett, J. M., Mehta, S., and Rhodes, M. (2007). Surgery for morbid obesity. *Postgraduate medical journal*, 83(975):8–15.

Blum, M. G., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., and Slama, R. (2020). Challenges raised by mediation analysis in a high-dimension setting. *Environmental health perspectives*, 128(5):055001.

Center for High Throughput Computing (2006). Center for high throughput computing.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232 – 1259.

Chén, O. Y., Cao, H., Phan, H., Nagels, G., Reinen, J. M., Gou, J., Qian, T., Di, J., Prince, J., Cannon, T. D., et al. (2021). Identifying neural signatures mediating behavioral symptoms and psychosis onset: High-dimensional whole brain functional mediation analysis. *NeuroImage*, 226:117508.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136.

Clark-Boucher, D., Zhou, X., Du, J., Liu, Y., Needham, B. L., Smith, J. A., and Mukherjee, B. (2023). Methods for mediation analysis with high-dimensional dna methylation data: Possible choices and comparisons. *Plos Genetics*, 19(11):e1011022.

Dai, J. Y., Stanford, J. L., and LeBlanc, M. (2022). A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117(537):198–213.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14.

Deng, C., Pan, J., Zhu, H., and Chen, Z.-Y. (2023). Effect of gut microbiota on blood cholesterol: A review on mechanisms. *Foods*, 12(23):4308.

Du, J., Zhou, X., Clark-Boucher, D., Hao, W., Liu, Y., Smith, J. A., and Mukherjee, B. (2023). Methods for large-scale single mediator hypothesis testing: Possible choices and comparisons. *Genetic epidemiology*, 47(2):167–184.

Erawijantari, P. P., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Saito, Y., Fukuda, S., Yachida, S., and Yamada, T. (2020). Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut*, 69(8):1404–1415.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911.

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

Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E. L., and Cui, Y. (2019). Testing mediation effects in high-dimensional epigenetic studies. *Frontiers in genetics*, 10:1195.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of statistics*, pages 1993–2010.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224.

Hao, W., Chen, C., and Song, P. X.-K. (2025). A class of directed acyclic graphs with mixed data types in mediation analysis. *Canadian Journal of Statistics*, page e70016.

Hao, W. and Song, P. X. (2023). A simultaneous likelihood test for joint mediation effects of multiple mediators. *Statistica Sinica*, 33(4):2305–2326.

He, Y., Song, P. X., and Xu, G. (2024). Adaptive bootstrap tests for composite null hypotheses in the mediation pathway analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):411–434.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171.

Jeong, S.-M., Choi, S., Kim, K., Kim, S. M., Lee, G., Park, S. Y., Kim, Y.-Y., Son, J. S., Yun, J.-M., and Park, S. M. (2018). Effect of change in total cholesterol levels on cardiovascular disease among young adults. *Journal of the American Heart Association*, 7(12):e008819.

Jérolon, A., Baglietto, L., Birmelé, E., Alarcon, F., and Perduca, V. (2021). Causal mediation analysis in presence of multiple mediators uncausally related. *The International Journal of Biostatistics*, 17(2):191–221.

Jia, B., Zou, Y., Han, X., Bae, J.-W., and Jeon, C. O. (2023). Gut microbiome-mediated mechanisms for reducing cholesterol levels: implications for ameliorating cardiovascular disease. *Trends in Microbiology*, 31(1):76–91.

Jiang, H. (2025). https://github.com/hhhanying/AdaptMLASSO.

Jiang, H., Miao, X., Thairu, M. W., Beebe, M., Grupe, D. W., Davidson, R. J., Handelsman, J., and Sankaran, K. (2024). Multimedia: multimodal mediation analysis of microbiome data. *Microbiology Spectrum*, 0(0):e01131–24.

Kissmann, A.-K., Paß, F., Ruzicka, H.-M., Dorst, I., Stieger, K. R., Weil, T., Gihring, A., Elad, L., Knippschild, U., and Rosenau, F. (2024). An increase in prominent probiotics represents the major change in the gut microbiota in morbidly obese female patients upon bariatric surgery. *Women*, 4(1):86–104.

Ko, J., Sequeira, I. R., Skudder-Hill, L., Cho, J., Poppitt, S. D., and Petrov, M. S. (2023). Metabolic traits affecting the relationship between liver fat and intrapancreatic fat: a mediation analysis. *Diabetologia*, 66(1):190–200.

Lee, J. W., Kim, E. Y., Yoo, H. M., Park, C. H., and Song, K. Y. (2015). Changes of lipid profiles after radical gastrectomy in patients with gastric cancer. *Lipids in Health and Disease*, 14:1–9.

Lin, X.-H., Huang, K.-H., Chuang, W.-H., Luo, J.-C., Lin, C.-C., Ting, P.-H., Young, S.-H., Fang, W.-L., Hou, M.-C., and Lee, F.-Y. (2018). The long term effect of metabolic profile and microbiota status in early gastric cancer patients after subtotal gastrectomy. *PLoS One*, 13(11):e0206930.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A. A., and Lin, X. (2022). Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, 117(537):67–81.

Loh, W. W., Moerkerke, B., Loeys, T., and Vansteelandt, S. (2022). Disentangling indirect effects through multiple mediators without assuming any causal structure among the mediators. *Psychological Methods*, 27(6):982.

Lu, S., Xie, Q., Kuang, M., Hu, C., Li, X., Yang, H., Sheng, G., Xie, G., and Zou, Y. (2023). Lipid metabolism, bmi and the risk of nonalcoholic fatty liver disease in the general population: evidence from a mediation analysis. *Journal of Translational Medicine*, 21(1):192.

MacKinnon, D. (2012). *Introduction to statistical mediation analysis*. Routledge.

Maki, T., Shiratori, T., Hatafuku, T., and Sugawara, K. (1967). Pylorus-preserving gastrectomy as an improved operation for gastric ulcer. *Surgery*, 61(6):838–845.

Miles, C. H. (2023). On the causal interpretation of randomised interventional indirect effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1154–1172.

Muller, E., Algavi, Y. M., and Borenstein, E. (2022). The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj Biofilms and Microbiomes*, 8(1):79.

Penna, M. and Allum, W. (2013). New treatments for gastric cancer: are they changing clinical practice? *Clinical Practice*, 10(5):649.

Sánchez-Alcoholado, L., Gutiérrez-Repiso, C., Gómez-Pérez, A. M., García-Fuentes, E., Tinahones, F. J., and Moreno-Indias, I. (2019). Gut microbiota adaptation after weight loss by roux-en-y gastric bypass or sleeve gastrectomy bariatric surgeries. *Surgery for Obesity and Related Diseases*, 15(11):1888–1895.

Seber, G. A. and Lee, A. J. (2003). *Linear regression analysis*. John Wiley & Sons.

Shi, C. and Li, L. (2022). Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, 117(540):2014–2027.

Sohn, M. B. and Li, H. (2019). Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics*, 13(1):661–681.

Sohn, M. B., Lu, J., and Li, H. (2022). A compositional mediation model for a binary outcome: application to microbiome studies. *Bioinformatics*, 38(1):16–21.

Song, Y., Zhou, X., Kang, J., Aung, M. T., Zhang, M., Zhao, W., Needham, B. L., Kardia, S. L., Liu, Y., Meeker, J. D., et al. (2021). Bayesian sparse mediation analysis with targeted penalization of natural indirect effects. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(5):1391–1412.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L., Roux, A. V. D., Needham, B. L., Smith, J. A., and Mukherjee, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76(3):700–710.

Sun, W., Wang, J., and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research*, 14(1):3419–3440.

Taguri, M., Featherstone, J., and Cheng, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Statistical methods in medical research*, 27(1):3–19.

Tchetgen Tchetgen, E. J. (2011). On causal mediation analysis with a survival outcome. *The international journal of biostatistics*, 7(1):00001022021557467 91351.

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

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*.

Toivonen, L., Schuez-Havupalo, L., Karppinen, S., Waris, M., Hoffman, K. L., Camargo Jr, C. A., Hasegawa, K., and Peltola, V. (2021). Antibiotic treatments during infancy, changes in nasal microbiota, and asthma development: population-based cohort study. *Clinical Infectious Diseases*, 72(9):1546–1554.

Tromba, L., Tartaglia, F., Carbotta, S., Sforza, N., Pelle, F., Colagiovanni, V., Carbotta, G., Cavaiola, S., and Casella, G. (2017). The role of sleeve gastrectomy in reducing cardiovascular risk. *Obesity surgery*, 27:1145–1151.

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

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

Vansteelandt, S. and Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258–265.

Vourakis, M., Mayer, G., and Rousseau, G. (2021). The role of gut microbiota on cholesterol metabolism in atherosclerosis. *International journal of molecular sciences*, 22(15):8074.

Wu, C., Zhao, Y., Zhang, Y., Yang, Y., Su, W., Yang, Y., Sun, L., Zhang, F., Yu, J., Wang, Y., et al. (2022). Gut microbiota specifically mediates the anti-hypercholesterolemic effect of berberine (bbr) and facilitates to predict bbr's cholesterol-decreasing efficacy in patients. *Journal of advanced research*, 37:197–208.

Yang, H., Liu, Z., Wang, R., Lai, E.-Y., Schwartz, J., Baccarelli, A. A., Huang, Y.-T., and Lin, X. (2024). Causal mediation analysis for integrating exposure, genomic, and phenotype data. *Annual Review of Statistics and Its Application*, 12.

Yu, D., Shu, X.-O., Howard, E. F., Long, J., English, W. J., and Flynn, C. R. (2020). Fecal metagenomics and metabolomics reveal gut microbial changes after bariatric surgery. *Surgery for Obesity and Related Diseases*, 16(11):1772–1782.

Zeng, P., Shao, Z., and Zhou, X. (2021). Statistical methods for mediation analysis in the era of high-throughput genomics: current successes and future challenges. *Computational and structural biotechnology journal*, 19:3209–3224.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154.

Zhang, Q. (2022). High-dimensional mediation analysis with applications to causal gene identification. *Statistics in biosciences*, 14(3):432–451.

Zhao, Y., Chen, T., Cai, J., Lichenstein, S., Potenza, M. N., and Yip, S. W. (2022). Bayesian network mediation analysis with application to the brain functional connectome. *Statistics in medicine*, 41(20):3991–4005.

Zhao, Y., Lindquist, M. A., and Caffo, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. *Computational statistics & data analysis*, 142:106835.

Zhao, Y. and Luo, X. (2022). Pathway lasso: pathway estimation and selection with high-dimensional mediators. *Statistics and its interface*, 15(1):39–50.

Zhou, R. R., Wang, L., and Zhao, S. D. (2020). Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3):573–589.

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.

Appendix

We provide the proofs of Lemma 1, Propositions 1–3, and Theorem 1 in Sections A–E, respectively. The supplementary simulation results of Pathway LASSO is presented in Appendix F.

A Proof of Lemma 1

To prove Lemma 1, we first state conclusions that will be needed in the following derivation. In particular, under the model (1), we have

$$\mathbb{E}[M_j \mid T = t, \boldsymbol{X} = \boldsymbol{x}] = \alpha_j^* t + \boldsymbol{\zeta}_{Mj}^{*\top}\boldsymbol{x}, \quad j = 1, \ldots, p, \tag{15}$$

$$\mathbb{E}[Y \mid T = t, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{X} = \boldsymbol{x}] = \eta^* t + \boldsymbol{\beta}^{*\top}\boldsymbol{m} + \boldsymbol{\zeta}_Y^{*\top}\boldsymbol{x}, \tag{16}$$

where $\boldsymbol{\zeta}_{Mj}^*$ represents the jth column of $\boldsymbol{\zeta}_M^*$. For $j = 1, \ldots, p$, we have

$$
\begin{aligned}
\mathbb{E}[M_j(t) \mid \boldsymbol{X} = \boldsymbol{x}] &= \mathbb{E}[M_j(t) \mid T = t, \boldsymbol{X} = \boldsymbol{x}] \qquad \text{(by Condition 2 (i))} \\
&= \mathbb{E}[M_j \mid T = t, \boldsymbol{X} = \boldsymbol{x}] \qquad \text{(by Condition 1)} \\
&= \alpha_j^* t + \boldsymbol{\zeta}_{Mj}^{*\top}\boldsymbol{x}. \qquad \text{(by (15))}
\end{aligned}
\tag{17}
$$

To derive $\delta_j(t'; , t)$, note that $\mathbb{E}\{Y(t, M_j(t'), \boldsymbol{M}_{-j}(t))\} = \mathbb{E}[\mathbb{E}[Y(t, M_j(t'), \boldsymbol{M}_{-j}(t)) \mid \boldsymbol{X}]]$. We first examine

$$
\begin{aligned}
&\mathbb{E}[Y(t, M_j(t'), \boldsymbol{M}_{-j}(t)) \mid \boldsymbol{X} = \boldsymbol{x}] \\
&= \int_{\mathbb{R}^p} \mathbb{E}[Y(t, m, \boldsymbol{w}) \mid M_j(t') = m, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{M}_{-j}(t) = \boldsymbol{w}] \, \mathrm{d}F_{t',t|\boldsymbol{x}}(m, \boldsymbol{w}),
\end{aligned}
\tag{18}
$$

where $F_{t',t|\boldsymbol{x}}$ represents the distribution of $(M_j(t'), \boldsymbol{M}_{-j}(t))$ conditional on $\boldsymbol{X} = \boldsymbol{x}$. Since Condition 2 (i) implies $Y(t, m, \boldsymbol{w}) \perp\!\!\!\perp T \mid \{M_j(t') = m', \boldsymbol{M}_{-j}(t'') = \boldsymbol{w}'', \boldsymbol{X} = \boldsymbol{x}\}$,

$$(18) = \int_{\mathbb{R}^p} \mathbb{E}[Y(t, m, \boldsymbol{w}) \mid M_j(t') = m, \boldsymbol{M}_{-j}(t) = \boldsymbol{w}, T = t, \boldsymbol{X} = \boldsymbol{x}] \, \mathrm{d}F_{t',t|\boldsymbol{x}}(m, \boldsymbol{w})$$

$$= \int_{\mathbb{R}^p} \mathbb{E}[Y(t, m, \boldsymbol{w}) \mid M_j(t) = m, \boldsymbol{M}_{-j}(t) = \boldsymbol{w}, T = t, \boldsymbol{X} = \boldsymbol{x}] \, \mathrm{d}F_{t',t|\boldsymbol{x}}(m, \boldsymbol{w}), \tag{19}$$

where the last equation follows by Condition 2 (iii). By Condition 1,

$$(19) = \int_{\mathbb{R}^p} \mathbb{E}[Y \mid M_j = m, \boldsymbol{M}_{-j} = \boldsymbol{w}, T = t, \boldsymbol{X} = \boldsymbol{x}] \, \mathrm{d}F_{t',t|\boldsymbol{x}}(m, \boldsymbol{w}) \tag{20}$$

$$= \int_{\mathbb{R}^p} \left(\eta^* t + \beta_j^* m + \boldsymbol{\beta}_{-j}^{*\top}\boldsymbol{w} + \boldsymbol{\zeta}_Y^{*\top}\boldsymbol{x}\right) \mathrm{d}F_{t',t|\boldsymbol{x}}(m, \boldsymbol{w}) \qquad \text{(by (16))}$$

$$= \eta^* t + \alpha_j^* \beta_j^* t' + \boldsymbol{\beta}_{-j}^{*\top}\boldsymbol{\alpha}_{-j}^* t + (\boldsymbol{\beta}^{*\top}\boldsymbol{\zeta}_M^{*\top} + \boldsymbol{\zeta}_Y^{*\top})\boldsymbol{x}, \qquad \text{(by (17))} \tag{21}$$

where $\boldsymbol{\alpha}^*_{-j}$ and $\boldsymbol{\beta}^*_{-j}$ represent the vectors $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ excluding their j-th elements, respectively. Combining (18)–(21), we obtain $\mathbb{E}\left[Y(t, M_j(t'), \boldsymbol{M}_{-j}(t))\right] = \eta^* t + \alpha_j^* \beta_j^* t' + \boldsymbol{\beta}^{*\top}_{-j} \boldsymbol{\alpha}^*_{-j} t + (\boldsymbol{\beta}^{*\top} \boldsymbol{\zeta}^{*\top}_M + \boldsymbol{\zeta}^{*\top}_Y) \mathbb{E}(\boldsymbol{X})$. Similar conclusion can be derived for $\mathbb{E}\left[Y(t, M_j(t), \boldsymbol{M}_{-j}(t))\right]$ by letting $t' = t$. In summary, $\delta_j(t'; t) = \mathbb{E}\left[Y(t, M_j(t'), \boldsymbol{M}_{-j}(t))\right] - \mathbb{E}\left[Y(t, M_j(t), \boldsymbol{M}_{-j}(t))\right] = \alpha_j^* \beta_j^* (t' - t)$ is obtained.

B   Proof of Proposition 1

When applying the loss functions defined in (8), we can rewrite the optimizations (3) and (4) as

$$\left(\hat{\boldsymbol{\theta}}_{AP}, \hat{\boldsymbol{\theta}}_{AU}\right) = \underset{\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU}}{\operatorname{argmin}} \|\mathbf{R}_A - \mathbf{D}_{AP}\boldsymbol{\theta}_{AP} - \mathbf{D}_{AU}\boldsymbol{\theta}_{AU}\|_{\mathrm{F}}^2 + \bar{\mathcal{P}}_A(\boldsymbol{\theta}_{AP}) \tag{22}$$

for $A \in \{M, Y\}$. Since $\mathbf{P}_{AU}^{\perp}$ is a projection matrix onto the column space orthogonal to that of $\mathbf{D}_{AU}$, by properties of projection matrices (Seber and Lee, 2003), we can decompose $\|\mathbf{R}_A - \mathbf{D}_{AP}\boldsymbol{\theta}_{AP} - \mathbf{D}_{AU}\boldsymbol{\theta}_{AU}\|_{\mathrm{F}}^2 = \ell_{A,1}(\boldsymbol{\theta}_{AP}) + \ell_{A,2}(\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU})$, where $\mathbf{P}_{AU} = \mathrm{I}_{n \times n} - \mathbf{P}_{AU}^{\perp}$,

$$\ell_{A,1}(\boldsymbol{\theta}_{AP}) = \|\mathbf{P}_{AU}^{\perp}(\mathbf{R}_A - \mathbf{D}_{AP}\boldsymbol{\theta}_{AP})\|_{\mathrm{F}}^2,$$
$$\ell_{A,2}(\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU}) = \|\mathbf{P}_{AU}(\mathbf{R}_A - \mathbf{D}_{AP}\boldsymbol{\theta}_{AP}) - \mathbf{D}_{AU}\boldsymbol{\theta}_{AU}\|_{\mathrm{F}}^2,$$

and $\ell_{A,1}(\cdot)$ does not depend on $\boldsymbol{\theta}_{AU}$. Plugging the above decomposition into (22), we obtain

$$\left(\hat{\boldsymbol{\theta}}_{AP}, \hat{\boldsymbol{\theta}}_{AU}\right) = \underset{\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU}}{\operatorname{argmin}} \ell_{A,1}(\boldsymbol{\theta}_{AP}) + \ell_{A,2}(\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU}) + \bar{\mathcal{P}}_A(\boldsymbol{\theta}_{AP}), \tag{23}$$

which is unique by the convexity of the penalized loss function.

By equivalently optimizing over $\hat{\boldsymbol{\theta}}_{AU}$ and $\hat{\boldsymbol{\theta}}_{AP}$ sequentially, we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{AP} &= \underset{\boldsymbol{\theta}_{AP}}{\operatorname{argmin}} \ \ell_{A,1}(\boldsymbol{\theta}_{AP}) + \bar{\mathcal{P}}_A(\boldsymbol{\theta}_{AP}) + \underset{\boldsymbol{\theta}_{AU}}{\min} \ell_{A,2}(\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU}) \\
&= \underset{\boldsymbol{\theta}_{AP}}{\operatorname{argmin}} \ \ell_{A,1}(\boldsymbol{\theta}_{AP}) + \bar{\mathcal{P}}_A(\boldsymbol{\theta}_{AP}),
\end{aligned} \tag{24}$$

where the second equation follows by the property of ordinary least squares regression that $\min_{\boldsymbol{\theta}_{AU}} \ell_{A,2}(\boldsymbol{\theta}_{AP}, \boldsymbol{\theta}_{AU}) = 0$ for each given $\boldsymbol{\theta}_{AP}$. Therefore, the first equation in (9) is proved. Similarly, by sequential optimization, we know

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{AU} &= \underset{\boldsymbol{\theta}_{AU}}{\operatorname{argmin}} \ \ell_{A,1}(\hat{\boldsymbol{\theta}}_{AP}) + \bar{\mathcal{P}}_A(\hat{\boldsymbol{\theta}}_{AP}) + \ell_{A,2}(\hat{\boldsymbol{\theta}}_{AP}, \boldsymbol{\theta}_{AU}) \\
&= \underset{\boldsymbol{\theta}_{AU}}{\operatorname{argmin}} \ \ell_{A,2}(\hat{\boldsymbol{\theta}}_{AP}, \boldsymbol{\theta}_{AU}) = \mathbf{D}_{AU}^{\dagger}(\mathbf{R}_A - \mathbf{D}_{AP}\hat{\boldsymbol{\theta}}_{AP}),
\end{aligned}$$

which follows by the solution of ordinary least squares regression.

Remark 4. *When $\mathcal{P}_M(\cdot) = 0$, $\hat{\boldsymbol{\alpha}}_n$ is given by*

$$\hat{\alpha}_{nj} = \mathcal{T}\left(\frac{\mathbf{M}_{nj}^{\top} \mathbf{P}_{\bar{\boldsymbol{X}}_n}^{\perp} \mathbf{T}_n}{\|\mathbf{P}_{\bar{\boldsymbol{X}}_n}^{\perp} \mathbf{T}_n\|_2^2}; \ \frac{\lambda_{n\alpha}}{2\hat{w}_{nj,\alpha}\|\mathbf{P}_{\bar{\boldsymbol{X}}_n}^{\perp} \mathbf{T}_n\|_2^2}\right), j = 1, \ldots, p, \tag{25}$$

where $\mathbf{M}_{nj}$ is the jth column of $\mathbf{M}_n$ and $\mathbf{P}_{\bar{\boldsymbol{X}}_n}^{\perp} = \mathrm{I}_{n \times n} - \boldsymbol{X}_n(\boldsymbol{X}_n^{\top}\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^{\top}$.

Proof. By Proposition 1, optimization (3) can be written as

$$\hat{\boldsymbol{\alpha}}_n = \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{P}_{\boldsymbol{X}_n}^{\perp}(\mathbf{M}_n - \mathbf{T}_n\boldsymbol{\alpha}^\top)\|_{\mathrm{F}}^2 + \lambda_{n\alpha} \sum_{j=1}^{p} \frac{|\alpha_j|}{\hat{w}_{nj,\alpha}}$$

$$= \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \sum_{j}^{p} \left\{ \|\mathbf{P}_{\boldsymbol{X}_n}^{\perp}\mathbf{M}_{nj} - \alpha_j\mathbf{P}_{\boldsymbol{X}_n}^{\perp}\mathbf{T}_n\|_2^2 + \lambda_{n\alpha} \frac{|\alpha_j|}{\hat{w}_{nj,\alpha}} \right\}.$$

Hence, each element $\hat{\alpha}_{nj}$ in $\hat{\boldsymbol{\alpha}}_n$ is the solution to the following optimization problem:

$$\hat{\alpha}_{nj} = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}}\{\|\mathbf{P}_{\boldsymbol{X}_n}^{\perp}\mathbf{M}_{nj} - \alpha_j\mathbf{P}_{\boldsymbol{X}_n}^{\perp}\mathbf{T}_n\|_2^2 + \lambda_{n\alpha}|\alpha_j|/\hat{w}_{nj,\alpha}\},$$

which has a closed form given by (25). □

C Proof of Proposition 2

First, to show $\hat{\alpha}_{nj}^0$ constructed in (7) satisfies Condition 3(ii), note that $|\hat{\alpha}_{nj}^0| \leq l_{nj,\alpha} = l_0 \cdot \hat{se}(\hat{\alpha}_{nj,o})$. Thus, $1/\hat{\alpha}_{nj}^0 = O_p(1/\hat{se}(\hat{\alpha}_{nj,o})) = O_p(\sqrt{n})$, which follows from the third term in (26) in Lemma 2 below.

Lemma 2. *Under the conditions of Proposition 2, OLS estimators $\hat{\alpha}_{nj,o}$ and $\hat{\beta}_{nj,o}$ satisfy*

$$\sqrt{n}(\hat{\alpha}_{nj,o} - \alpha_j^*) = O_p(1), \qquad \sqrt{n}\hat{se}(\hat{\alpha}_{nj,o}) = O_p(1), \qquad \frac{1}{\sqrt{n}\hat{se}(\hat{\alpha}_{nj,o})} = O_p(1), \qquad (26)$$

$$\sqrt{n}(\hat{\beta}_{nj,o} - \beta_j^*) = O_p(1), \qquad \sqrt{n}\hat{se}(\hat{\beta}_{nj,o}) = O_p(1), \qquad \frac{1}{\sqrt{n}\hat{se}(\hat{\beta}_{nj,o})} = O_p(1). \qquad (27)$$

Proof. See Section C.1. □

Second, we show that $\hat{\alpha}_{nj}^0$ satisfies Condition 3(i). For $\hat{\alpha}_{nj}^0$ in (7), we can equivalently rewrite it as $\hat{\alpha}_{nj}^0 = \hat{\alpha}_{nj,o}\mathbb{1}_{\{|\hat{\alpha}_{nj,o}|\geq l_{nj,\alpha}\}} + l_{nj,\alpha}\mathbb{1}_{\{|\hat{\alpha}_{nj,o}|<l_{nj,\alpha}\}}$, where $\mathbb{1}$ represents the indicator function. Therefore,

$$\hat{\alpha}_{nj}^0 - \alpha_j^*$$
$$= (\hat{\alpha}_{nj,o} - \alpha_j^*)\mathbb{1}_{\{|\hat{\alpha}_{nj,o}|\geq l_{nj,\alpha}\}} + \{l_{nj,\alpha} - \hat{\alpha}_{nj,o} + (\hat{\alpha}_{nj,o} - \alpha_j^*)\}\mathbb{1}_{\{|\hat{\alpha}_{nj,o}|<l_{nj,\alpha}\}}$$
$$= \hat{\alpha}_{nj,o} - \alpha_j^* + (l_{nj,\alpha} - \hat{\alpha}_{nj,o})\mathbb{1}_{\{|\hat{\alpha}_{nj,o}|<l_{nj,\alpha}\}}. \qquad (28)$$

Thus,

$$|\sqrt{n}(\hat{\alpha}_{nj}^0 - \alpha_j^*)| \leq \sqrt{n}|\hat{\alpha}_{nj,o} - \alpha_j^*| + 2\sqrt{n}l_{nj,\alpha} = \sqrt{n}|\hat{\alpha}_{nj,o} - \alpha_j^*| + 2l_0\sqrt{n}\hat{se}(\hat{\alpha}_{nj,o}). \qquad (29)$$

By the first and second terms in (26) in Lemma 2, we can conclude that (29) $= O_p(1)$. We can similarly prove that $\hat{\beta}_{nj}^0$ satisfies Condition 3.

C.1 Proof of Lemma 2

Proof of (26). We first show $\sqrt{n}(\hat{\alpha}_{nj,o} - \alpha_j^*) = O_p(1)$. We use $\mathbf{E}_{nj}$ to denote the jth column of $\mathbf{E}_n$. Note that $\hat{\alpha}_{nj,o}$ is the OLS for the coefficient of $\mathbf{T}_n$ under the linear regression $\mathbf{M}_{nj} \sim (\mathbf{T}_n, \mathbf{X}_n)$. By Frisch–Waugh–Lovell Theorem, we know

$$\hat{\alpha}_{nj,o} = (\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n)^{-1} \mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{M}_{nj} = \alpha_j^* + (\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n)^{-1} \mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{E}_{nj}, \tag{30}$$

where the second equation follows by the model $\mathbf{M}_{nj} = \mathbf{T}_n \alpha_j^* + \mathbf{X}_n \boldsymbol{\zeta}_{M,j} + \mathbf{E}_{nj}$ in (2). By the independence between $\mathbf{E}_n$ and $\mathbf{D}_M$, and Condition 4, we have that

$$\frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n}{\sqrt{n}} \to_d \mathcal{N}\left(\mathbf{0}_p, \sigma_T^2 \mathbf{\Sigma}\right), \tag{31}$$

where $\to_d$ denotes convergence in distribution. Then we have

$$\sqrt{n}(\hat{\alpha}_{nj,o} - \alpha_j^*) = \left(\frac{\|\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n\|_2^2}{n}\right)^{-1} \frac{\mathbf{E}_{nj}^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n}{\sqrt{n}} = O_p(1).$$

We next prove the second and third terms in (26). Let $\hat{\Sigma}_{jj}$ denote the estimator of Σ_{jj} under OLS. By the property of OLS (Seber and Lee, 2003) and Condition 4, we have $\hat{\Sigma}_{jj} \to_p \Sigma_{jj}$ as $n \to \infty$ and

$$\hat{se}^2(\hat{\alpha}_{nj,o}) = (\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n)^{-1} \hat{\Sigma}_{jj}. \tag{32}$$

Then by Slutsky's theorem and Condition 4,

$$n\hat{se}^2(\hat{\alpha}_{nj,o}) \to_p \Sigma_{jj}/\sigma_T^2 \quad \text{and} \quad \frac{1}{n\hat{se}^2(\hat{\alpha}_{nj,o})} \to_p \frac{\sigma_T^2}{\Sigma_{jj}},$$

where $\Sigma_{jj}/\sigma_T^2 > 0$ is fixed. Therefore, the second and third terms in (26) can be obtained.

Proof of (27). We first show $\sqrt{n}(\hat{\beta}_{nj,o} - \beta_j^*) = O_p(1)$. Note $\hat{\boldsymbol{\beta}}_{n,o}$ is the regression coefficient vector of $\mathbf{M}_n$ under the linear regression $\mathbf{Y}_n \sim \mathbf{M}_n + \mathbf{D}_M$. By Frisch–Waugh–Lovell Theorem, we know

$$\hat{\boldsymbol{\beta}}_{n,o} = (\mathbf{M}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{M}_n)^{-1} \mathbf{M}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{Y}_n = \boldsymbol{\beta}^* + (\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n)^{-1} \mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \boldsymbol{\epsilon}_n, \tag{33}$$

where the second equation follows by the model $\mathbf{Y}_n = \mathbf{M}_n \boldsymbol{\beta}^* + \mathbf{D}_M(\eta^*, \boldsymbol{\zeta}_Y^{*\top})^\top + \boldsymbol{\epsilon}_n$ by (2). To finish the proof, it suffices to show

$$\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n/n \to_p \mathbf{\Sigma} \qquad \text{and} \qquad \mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \boldsymbol{\epsilon}_n/\sqrt{n} \to_d \mathcal{N}(\mathbf{0}_p, \sigma^2 \mathbf{\Sigma}), \tag{34}$$

where the latter is $O_p(1)$. By the law of large numbers and the central limit theorem, we know $\mathbf{E}_n^\top \mathbf{E}_n/n \to_p \mathbf{\Sigma}$ and $\mathbf{E}_n^\top \boldsymbol{\epsilon}_n/\sqrt{n} \to_d \mathcal{N}(\mathbf{0}_p, \sigma^2 \mathbf{\Sigma})$. Thus, by $\mathbf{P}_{\mathbf{D}_M}^\perp = \mathbf{I}_{n \times n} - \mathbf{P}_{\mathbf{D}_M}$ and Slutsky's theorem, it remains to show

$$\frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M} \mathbf{E}_n}{n} = \frac{\mathbf{E}_n^\top \mathbf{X}_n (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{E}_n}{n} + \frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n (\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n)^{-1} \mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{E}_n}{n}$$

$$\frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M} \boldsymbol{\epsilon}_n}{\sqrt{n}} = \frac{\mathbf{E}_n^\top \mathbf{X}_n (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \boldsymbol{\epsilon}_n}{\sqrt{n}} + \frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n (\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n)^{-1} \mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \boldsymbol{\epsilon}_n}{\sqrt{n}}$$

are $o_p(1)$, where the equations hold by $\mathbf{P}_{\mathbf{D}_M} = \mathbf{P}_{(\mathbf{T}_n, \mathbf{X}_n)} = \mathbf{P}_{\mathbf{X}_n} + \mathbf{P}_{\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n}$. In particular, the conclusion follows because by Condition 4 and Markov's inequality, we have $\mathbf{X}_n^\top \mathbf{X}_n / n \to \mathbf{\Sigma}_X$, $\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n / n \to \sigma_T^2$, $\mathbf{X}_n^\top \mathbf{E}_n = O_p(\sqrt{n})$, $\mathbf{X}_n^\top \boldsymbol{\epsilon}_n / n = O_p(\sqrt{n})$, $\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{E}_n = O_p(\sqrt{n})$, and $\mathbf{T}_n^\top \mathbf{P}_{\mathbf{X}_n}^\perp \boldsymbol{\epsilon}_n = O_p(\sqrt{n})$.

We next prove the second and third terms in (27). Let $\hat{\sigma}^2$ denote the estimator of σ^2 under OLS. By the properties of OLS and Condition 4, we have $\hat{\sigma}^2 \to_p \sigma^2$ as $n \to \infty$ and

$$\hat{se}^2(\hat{\beta}_{nj,o}) = \left((\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n)^{-1} \right)_{jj} \hat{\sigma}^2. \tag{35}$$

Then by (34) and (35),

$$n\hat{se}^2(\hat{\beta}_{nj,o}) \to_p \sigma^2 (\mathbf{\Sigma}^{-1})_{jj} \quad \text{and} \quad \frac{1}{n\hat{se}^2(\hat{\beta}_{nj,o})} \to_p \frac{1}{\sigma^2 (\mathbf{\Sigma}^{-1})_{jj}},$$

where $(\mathbf{\Sigma}^{-1})_{jj}$, the jth diagonal element of $\mathbf{\Sigma}^{-1}$, is positive and fixed by Condition 4. Therefore, the second and third terms in (27) can be obtained.

D Proof of Theorem 1

This section is organized as follows. Section D.1 defines notation and lemmas to be used in the proof of Theorem 1. Section D.2 provides the main proof of Theorem 1. Sections D.3 and D.4 prove Lemmas 3 and 4 given in Section D.1.

D.1 Notation and Preliminary Lemmas

Notation. For an index set, we use a superscript c to denote its complement, and we use the index set itself as a subscript to a vector or matrix to represent its corresponding subvector or submatrix. For example, $\mathcal{A}^{*c}$ represents the complement of $\mathcal{A}^* \subseteq \{1, \ldots, p\}$, $\boldsymbol{\alpha}_{\mathcal{A}^*}^*$ represents the subvector of $\boldsymbol{\alpha}^*$ consisting of the elements with indices in $\mathcal{A}^*$, and $\mathbf{\Sigma}_{\mathcal{A}^*}$ represents the submatrix of $\mathbf{\Sigma}$ consisting of the elements with both row and column indices in $\mathcal{A}^*$. Additionally, we define

$$\mathcal{A}_\alpha^* = \left\{ j : \alpha_j^* \neq 0, \ \ j = 1, \ldots, p \right\} \quad \text{and} \quad \mathcal{A}_\beta^* = \left\{ j : \beta_j^* \neq 0, \ \ j = 1, \ldots, p \right\}$$

as the index sets of mediators with nonzero exposure-to-mediator effects and with nonzero mediator-to-outcome effects.

Lemma 3. *Under the conditions in Theorem 1, as the sample size $n \to \infty$,*

$$\frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}} \to_p \begin{cases} 0, & \text{for } j \in \mathcal{A}_\alpha^* \\ \infty, & \text{for } j \in \mathcal{A}_\alpha^{*c} \end{cases} \quad \text{and} \quad \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \to_p \begin{cases} 0, & \text{for } j \in \mathcal{A}_\beta^* \\ \infty, & \text{for } j \in \mathcal{A}_\beta^{*c} \end{cases}. \tag{36}$$

Proof. See Section D.3. □

Lemma 4. *Under the conditions in Theorem 1, the JAP estimators of $\boldsymbol{\alpha}^*$ satisfy:*

$$\sqrt{n}\left(\hat{\boldsymbol{\alpha}}_{n,\mathcal{A}_\alpha^*} - \boldsymbol{\alpha}_{\mathcal{A}^*}^*\right) \to_d \mathcal{N}\left(\mathbf{0}, \sigma_T^{-2}\boldsymbol{\Sigma}_{\mathcal{A}_\alpha^*}\right), \quad and \quad \sqrt{n}\hat{\boldsymbol{\alpha}}_{n,\mathcal{A}_\alpha^{*c}} \to_p \mathbf{0}, \tag{37}$$

and the JAP estimators of $\boldsymbol{\beta}^*$ satisfy:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_\beta^*} - \boldsymbol{\beta}_{\mathcal{A}_\beta^*}^*\right) \to_d \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_{\mathcal{A}_\beta^*}^{-1}\right) \quad and \quad \sqrt{n}\hat{\boldsymbol{\beta}}_{n,\mathcal{A}_\beta^{*c}} \to_p \mathbf{0}. \tag{38}$$

Proof. See Section D.4. $\qquad\square$

D.2   Proof of Theorem 1

First, to prove $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_{\mathrm{F}}^2 + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\|_{\mathrm{F}}^2 \to_p 0$, note that Lemma 4 implies $\hat{\alpha}_{nj} - \alpha_j^* \to_p 0$ and $\hat{\beta}_{nj} - \beta_j^* \to_p 0$ for $j = 1, \dots, p$. Then the conclusion follows from Slutsky's theorem.

We next prove (12). As $\{\hat{\mathcal{A}}_n = \mathcal{A}^*\}^c = \{\hat{\mathcal{A}}_n \supseteq \mathcal{A}^*\}^c \cup \{\hat{\mathcal{A}}_n \subseteq \mathcal{A}^*\}^c$, by Boole's inequality, to prove (12), it suffices to show

$$\lim_{n\to\infty} \Pr(\{\hat{\mathcal{A}}_n \supseteq \mathcal{A}^*\}^c) = 0, \tag{39}$$

$$\lim_{n\to\infty} \Pr(\{\hat{\mathcal{A}}_n \subseteq \mathcal{A}^*\}^c) = 0. \tag{40}$$

(i) Proof of (39).   Note

$$\Pr(\{\hat{\mathcal{A}}_n \supseteq \mathcal{A}^*\}^c) = \Pr(\cup_{j\in\mathcal{A}^*}\{j \notin \hat{\mathcal{A}}_n\}) \leqslant \sum_{j\in\mathcal{A}^*} \Pr(j \notin \hat{\mathcal{A}}_n)$$

$$\leqslant \sum_{j\in\mathcal{A}^*} \{\Pr(\hat{\alpha}_{nj} = 0) + \Pr(\hat{\beta}_{nj} = 0)\},$$

where the last inequality follows as $j \notin \hat{\mathcal{A}}_n$ is equivalent to $\{\hat{\alpha}_{nj} = 0\} \cup \{\hat{\beta}_{nj} = 0\}$ based on our construction. Note

$$\Pr(\hat{\beta}_{nj} = 0) = \Pr\left(\beta_j^* = -(\hat{\beta}_{nj} - \beta_j^*)\right) \leq \Pr\left(\sqrt{n}\,|\beta_j^*| = \sqrt{n}\left|\hat{\beta}_{nj} - \beta_j^*\right|\right). \tag{41}$$

By the first part of (38), (41) $\to 0$ as $n \to \infty$ for any $j \in \mathcal{A}^*$. Similarly, we can use (37) to show that $\lim_{n\to\infty} \Pr(\hat{\alpha}_{nj} = 0) = 0$ for all $j \in \mathcal{A}^*$. In summary, $\lim_{n\to\infty} \Pr(\{\hat{\mathcal{A}}_n \supseteq \mathcal{A}^*\}^c) = 0$, and (39) is proved.

(ii) Proof of (40).   Since $\{\hat{\mathcal{A}}_n \subseteq \mathcal{A}^*\}^c = \{\text{There exists } j \notin \mathcal{A}^* \text{ such that } j \in \hat{\mathcal{A}}_n\} = \cup_{j\notin\mathcal{A}^*}\{j \in \hat{\mathcal{A}}_n\}$, and $\{j \in \hat{\mathcal{A}}_n\} = \{\hat{\alpha}_{nj} \neq 0\} \cap \{\hat{\beta}_{nj} \neq 0\}$ by our construction, we have

$$\Pr(\{\hat{\mathcal{A}}_n \subseteq \mathcal{A}^*\}^c) \leqslant \sum_{j\notin\mathcal{A}^*} \Pr(\{\hat{\alpha}_{nj} \neq 0\} \cap \{\hat{\beta}_{nj} \neq 0\})$$

$$= \sum_{j\in(\mathcal{A}_\alpha^*\cap\mathcal{A}_\beta^*)^c} \Pr(\{\hat{\alpha}_{nj} \neq 0\} \cap \{\hat{\beta}_{nj} \neq 0\}) \quad (\text{by } \mathcal{A}^* = \mathcal{A}_\alpha^*\cap\mathcal{A}_\beta^*)$$

$$\leqslant \sum_{j\in\mathcal{A}_\alpha^{*c}} \Pr(\hat{\alpha}_{nj} \neq 0) + \sum_{j\in\mathcal{A}_\beta^{*c}} \Pr(\hat{\beta}_{nj} \neq 0). \tag{42}$$

31

First, $\hat{\alpha}_{nj} \neq 0$, by Remark 4, is equivalent to

$$\frac{|\mathbf{M}_{nj}^\top \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n|}{\|\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n\|_2^2} > \frac{\lambda_{n\alpha}}{2\hat{w}_{nj,\alpha} \|\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n\|_2^2}.$$

Multiplying both sides of the above inequality by $\|\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n\|_2^2/\sqrt{n}$ and substituting the model of $\mathbf{M}_{nj}$ in (2), we obtain

$$\left| \frac{\|\mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n\|_2^2}{\sqrt{n}} \alpha_j^* + \frac{\mathbf{E}_{nj} \mathbf{P}_{\mathbf{X}_n}^\perp \mathbf{T}_n}{\sqrt{n}} \right| > \frac{1}{2} \frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}}. \tag{43}$$

For $j \in \mathcal{A}_\alpha^{*c}$, $\alpha_j^* = 0$, thus the left hand side of (43) is $O_p(1)$ by (31), while the right hand side of (43) $\to_p \infty$ as $n \to \infty$ by (36).

Second, by the Karush–Kuhn–Tucker (KKT) condition and (9), we know the estimate $\hat{\boldsymbol{\beta}}_n$ satisfies

$$-2\mathbf{D}_M^\top \mathbf{P}_{\mathbf{D}_M}^\perp (\mathbf{Y}_n - \mathbf{M}_n \hat{\boldsymbol{\beta}}_n) + \lambda_{n\beta} \sum_{j=1}^p \frac{1}{\hat{w}_{nj,\beta}} \frac{\partial |\beta|}{\partial \beta}\bigg|_{\beta = \hat{\beta}_{nj}} = 0.$$

For $\hat{\beta}_{nj} \neq 0$, it reduces to

$$2\mathbf{M}_{nj}^\top \mathbf{P}_{\mathbf{D}_M}^\perp (\mathbf{Y}_n - \mathbf{M}_n \hat{\boldsymbol{\beta}}_n) = \frac{\lambda_{n\beta}}{\hat{w}_{nj,\beta}} \operatorname{sign}(\hat{\beta}_{nj}).$$

Dividing both sides of the equation by $2\sqrt{n}$ and substituting $\mathbf{Y}_n$ and $\mathbf{M}_n$ using (2), we obtain

$$\frac{\mathbf{E}_{nj}^\top \mathbf{P}_{\mathbf{D}_M}^\perp \boldsymbol{\epsilon}_n}{\sqrt{n}} + \frac{\mathbf{E}_{nj}^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n}{n} \sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_n) = \frac{1}{2} \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \operatorname{sign}(\hat{\beta}_{nj}). \tag{44}$$

For any $j \in \mathcal{A}_\beta^{*c}$, the left hand side of (44) is $O_p(1)$ by (34) and the first part of (38), while the absolute value of the right hand side of (44) $\to_p \infty$ as $n \to \infty$ by (36). Therefore, for any $j \in \mathcal{A}_\beta^{*c}$,

$$\Pr\left(j \in \hat{\mathcal{A}}_{n,\beta} \right) \leq \Pr\left(\left| \frac{\mathbf{E}_{nj}^\top \mathbf{P}_{\mathbf{D}_M}^\perp \boldsymbol{\epsilon}_n}{\sqrt{n}} + \frac{\mathbf{E}_{nj}^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n}{n} \sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_n) \right| = \frac{1}{2} \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \right)$$

$$\to 0 \quad \text{as} \quad n \to \infty.$$

In summary, we obtain (42) $\to 0$ as $n \to \infty$, which finishes the proof.

D.3   Proof of Lemma 3

First, to prove $\frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}} \to_p \infty$ for $j \in \mathcal{A}_\alpha^{*c}$, since $\lambda_{n\alpha} \gg n^{1/2-\eta_\alpha}$, it suffices to show $n^{\eta_\alpha}\hat{w}_{nj,\alpha} = O_p(1)$ for $j \in \mathcal{A}_\alpha^{*c}$. Note that for $j \in \mathcal{A}_\alpha^{*c}$, $\sqrt{n}\hat{\alpha}_{nj}^0 = \sqrt{n}(\hat{\alpha}_{nj}^0 - \alpha_j^*) = O_p(1)$ by Condition 3(i), then

$$n^{\eta_\alpha}\hat{w}_{nj,\alpha} = n^{\frac{2\eta_\alpha - \gamma_\alpha}{2}} |\sqrt{n}\hat{\alpha}_{nj}^0|^{\gamma_\alpha} |\hat{\beta}_{nj}^0|^{\gamma_\alpha} + |\sqrt{n}\hat{\alpha}_{nj}^0|^{2\eta_\alpha} = O_p(1),$$

which follows by $\gamma_\alpha > 2\eta_\alpha$.

Second, we show $\frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}} \to_p 0$ for $j \in \mathcal{A}_\alpha^*$. Following Condition 3(i) and Slutsky's theorem, as $n \to \infty$,

$$\hat{w}_{nj,\alpha} = |\hat{\alpha}_{nj}^0 \hat{\beta}_{nj}^0|^{\gamma_\alpha} + |\hat{\alpha}_{nj}^0|^{2\eta_\alpha} \to_p |\alpha_j^* \beta_j^*|^{\gamma_\alpha} + |\alpha_j^*|^{2\eta_\alpha} > 0 \quad \text{for } j \in \mathcal{A}_\alpha^*.$$

Since $\lambda_{n\alpha} \ll n^{1/2}$, $\frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}} \to_p 0$ for $j \in \mathcal{A}_\alpha^*$.

Following similar arguments, we can prove the conclusions for $\frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}}$.

D.4 Proof of Lemma 4

(i) Proof of (38). To prove (38), we use the formula for penalized coefficients in Proposition 1. In particular, under the loss function in (11), we have $\boldsymbol{\theta}_{YP} = \boldsymbol{\beta}$, $\bar{\mathcal{P}}_Y(\boldsymbol{\beta}) = \lambda_{n\beta} \sum_{j=1}^p |\beta_j|/\hat{w}_{nj,\beta}$, and

$$\ell_{Y,1}(\boldsymbol{\beta}) = \|\mathbf{P}_{\mathbf{D}_M}^\perp (\mathbf{Y}_n - \mathbf{M}_n \boldsymbol{\beta})\|_{\mathrm{F}}^2$$
$$= \|\mathbf{P}_{\mathbf{D}_M}^\perp \{\mathbf{M}_n(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \boldsymbol{\epsilon}_n\}\|_{\mathrm{F}}^2 = \|\mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \mathbf{P}_{\mathbf{D}_M}^\perp \boldsymbol{\epsilon}_n\|_{\mathrm{F}}^2,$$

where the second and third equations follow by the model of $\mathbf{Y}_n$ and $\mathbf{M}_n$ in (2).

Define $\hat{\mathbf{u}}_n^\beta = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$. By (24), we have $\hat{\mathbf{u}}_n^\beta = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} V_n^\beta(\mathbf{u})$, where we define

$$V_n^\beta(\mathbf{u}) \equiv \left\{ \ell_{Y,1}(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}) + \bar{\mathcal{P}}_Y(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}) \right\} - \{\ell_{Y,1}(\boldsymbol{\beta}^*) + \bar{\mathcal{P}}_Y(\boldsymbol{\beta}^*)\}$$
$$= \mathbf{u}^\top \left(\frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \mathbf{E}_n}{n} \right) \mathbf{u} - 2\mathbf{u}^\top \frac{\mathbf{E}_n^\top \mathbf{P}_{\mathbf{D}_M}^\perp \boldsymbol{\epsilon}_n}{\sqrt{n}}$$
$$+ \sum_{j=1}^p \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \sqrt{n} \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right), \tag{45}$$

which follows by plugging the formulae of $\ell_{Y,1}(\cdot)$ and $\bar{\mathcal{P}}_Y(\cdot)$. To obtain the asymptotics of $\hat{\mathbf{u}}_n^\beta$, we derive the limit of $V_n^\beta(\cdot)$.

First, in $V_n^\beta(\mathbf{u})$, $\frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \sqrt{n} \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right) = 0$ if $u_j = 0$. When $u_j \neq 0$, its convergence can be discussed in two cases. For any $j \in \mathcal{A}_\beta^*$ and fixed $u_j \neq 0$, by (36),

$$\left| \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \sqrt{n} \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right) \right| \leq \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} |u_j| \to_p 0 \quad \text{as} \quad n \to \infty. \tag{46}$$

For any $j \in \mathcal{A}_\beta^{*c}$ and fixed $u_j \neq 0$,

$$\frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} \sqrt{n} \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right) = \frac{\lambda_{n\beta}}{\sqrt{n}\hat{w}_{nj,\beta}} |u_j| \to_p \infty \quad \text{as} \quad n \to \infty. \tag{47}$$

By (34) and Slutsky's theorem, we have that for any fixed $\mathbf{u}$, as $n \to \infty$,

$$V_n^\beta(\mathbf{u}) \to_d V^\beta(\boldsymbol{u}) \equiv \begin{cases} \mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u} - 2\mathbf{u}^\top \boldsymbol{W} & \text{if } u_j = 0 \text{ for all } j \in \mathcal{A}_\beta^{*c}, \\ \infty & \text{otherwise,} \end{cases} \tag{48}$$
$$= \begin{cases} \mathbf{u}_{\mathcal{A}_\beta^*}^\top \boldsymbol{\Sigma}_{\mathcal{A}_\beta^*} \mathbf{u}_{\mathcal{A}_\beta^*} - 2\mathbf{u}_{\mathcal{A}_\beta^*}^\top \boldsymbol{W}_{\mathcal{A}_\beta^*} & \text{if } u_j = 0 \text{ for all } j \in \mathcal{A}_\beta^{*c}, \\ \infty & \text{otherwise,} \end{cases}$$

33

where $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, and $\boldsymbol{W}_{\mathcal{A}_\beta^*}$ denotes its subvector satisfying $\boldsymbol{W}_{\mathcal{A}_\beta^*} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_{\mathcal{A}_\beta^*})$. Since $\boldsymbol{\Sigma}_{\mathcal{A}_\beta^*}$ is positive definite, $V^\beta(\boldsymbol{u})$ has a unique minimizer. Following the epi-convergence results in Geyer (1994), we have

$$\hat{\mathbf{u}}_n^\beta = \underset{\mathbf{u}}{\operatorname{argmin}}\, V_n^\beta(\mathbf{u}) \to_d \underset{\mathbf{u}}{\operatorname{argmin}}\, V^\beta(\boldsymbol{u}) \quad \Rightarrow \quad \begin{cases} \hat{\mathbf{u}}_{n,\mathcal{A}_\beta^*}^\beta \to_d \boldsymbol{\Sigma}_{\mathcal{A}_\beta^*}^{-1}\boldsymbol{W}_{\mathcal{A}_\beta^*}, \\ \hat{\mathbf{u}}_{n,\mathcal{A}_\beta^{*c}}^\beta \to_d \mathbf{0}. \end{cases} \tag{49}$$

Therefore, (38) is proved.

(ii) Proof of (37). The proof is similar to the proof of (38). To prove (37), we use the simplified formula for penalized coefficients in Proposition 1. Specifically, under the loss function in (8) with $\mathcal{P}_M(\cdot) = 0$, we have $\boldsymbol{\theta}_{MP} = \boldsymbol{\alpha}$, $\bar{\mathcal{P}}_M(\boldsymbol{\alpha}) = \lambda_{n\alpha}\sum_{j=1}^p |\alpha_j|/\hat{w}_{nj,\alpha}$, and

$$\ell_{M,1}(\boldsymbol{\alpha}) = \|\mathbf{P}_{\mathbf{X}_n}^\perp(\mathbf{M}_n - \mathbf{T}_n\boldsymbol{\alpha}^\top)\|_F^2 = \|\mathbf{P}_{\mathbf{X}_n}^\perp\mathbf{E}_n + \mathbf{P}_{\mathbf{X}_n}^\perp\mathbf{T}_n(\boldsymbol{\alpha}^{*\top} - \boldsymbol{\alpha}^\top)\|_F^2,$$

where the second equation follows by the model of $\mathbf{M}_n$ in (2).

Define $\hat{\mathbf{u}}_n^\alpha = \sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*)$. By (24), we have $\hat{\mathbf{u}}_n^\alpha = \operatorname{argmin}_{\mathbf{u}\in\mathbb{R}^p} V_n^\alpha(\mathbf{u})$, where we define

$$V_n^\alpha(\mathbf{u}) \equiv \left(\ell_{M,1}(\boldsymbol{\alpha}^* + \frac{\mathbf{u}}{\sqrt{n}}) + \bar{\mathcal{P}}_M(\boldsymbol{\alpha}^* + \frac{\mathbf{u}}{\sqrt{n}})\right) - (\ell_{M,1}(\boldsymbol{\alpha}^*) + \bar{\mathcal{P}}_M(\boldsymbol{\alpha}^*))$$

$$= \frac{\|\mathbf{P}_{\mathbf{X}_n}^\perp\mathbf{T}_n\|_2^2}{n}\mathbf{u}^\top\mathbf{u} - 2\mathbf{u}^\top\frac{\mathbf{E}_n^\top\mathbf{P}_{\mathbf{X}_n}^\perp\mathbf{T}_n}{\sqrt{n}} + \sum_{j=1}^p \frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}}\sqrt{n}\left(\left|\alpha_j^* + \frac{u_j}{\sqrt{n}}\right| - \left|\alpha_j^*\right|\right), \tag{50}$$

which follows by plugging the formulae of $\ell_{M,1}(\cdot)$ and $\bar{\mathcal{P}}_M(\cdot)$. We derive the limit of $V_n^\alpha(\cdot)$ in (50) to obtain the asymptotics in (37).

For the third term in (50), following analogous reasoning to the discussion above, we can show that for any fixed $u_j \neq 0$, as $n \to \infty$,

$$\frac{\lambda_{n\alpha}}{\sqrt{n}\hat{w}_{nj,\alpha}}\sqrt{n}\left(\left|\alpha_j^* + \frac{u_j}{\sqrt{n}}\right| - \left|\alpha_j^*\right|\right) \to_p \begin{cases} 0 & \text{if } j \in \mathcal{A}^*, \\ \infty & \text{if } j \in \mathcal{A}^{*c}. \end{cases}$$

Combining with Condition 4 and (31), by Slutsky's theorem, we have that for any fixed $\mathbf{u}$, as $n \to \infty$,

$$V_n^\alpha(\mathbf{u}) \to_d V^\alpha(\mathbf{u}) \equiv \begin{cases} \sigma_T^2\mathbf{u}^\top\mathbf{u} - 2\mathbf{u}^\top\boldsymbol{W}' & \text{if } u_j = 0 \text{ for all } j \in \mathcal{A}_\alpha^{*c}, \\ \infty & \text{otherwise,} \end{cases} \tag{51}$$

$$= \begin{cases} \sigma_T^2\mathbf{u}_{\mathcal{A}_\alpha^*}^\top\mathbf{u}_{\mathcal{A}_\alpha^*} - 2\mathbf{u}_{\mathcal{A}_\alpha^*}^\top\boldsymbol{W}'_{\mathcal{A}_\alpha^*} & \text{if } u_j = 0 \text{ for all } j \in \mathcal{A}_\alpha^{*c}, \\ \infty & \text{otherwise,} \end{cases}$$

where $\boldsymbol{W}' \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\boldsymbol{\Sigma})$, and $\boldsymbol{W}'_{\mathcal{A}_\alpha^*}$ denotes its subvector satisfying $\boldsymbol{W}'_{\mathcal{A}_\alpha^*} \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\boldsymbol{\Sigma}_{\mathcal{A}_\alpha^*})$. Since $\boldsymbol{\Sigma}_{\mathcal{A}_\alpha^*}$ is positive definite, $V^\alpha(\boldsymbol{u})$ has a unique minimizer. Following the epi-convergence results in Geyer (1994), we have

$$\hat{\mathbf{u}}_n^\alpha = \underset{\mathbf{u}}{\operatorname{argmin}}\, V_n^\alpha(\mathbf{u}) \to_d \underset{\mathbf{u}}{\operatorname{argmin}}\, V^\alpha(\mathbf{u}) \quad \Rightarrow \quad \begin{cases} \hat{\mathbf{u}}_{n,\mathcal{A}_\alpha^*}^\alpha \to_d \sigma_T^{-2}\boldsymbol{W}'_{\mathcal{A}_\alpha^*}, \\ \hat{\mathbf{u}}_{n,\mathcal{A}_\alpha^{*c}}^\alpha \to_d \mathbf{0}. \end{cases} \tag{52}$$
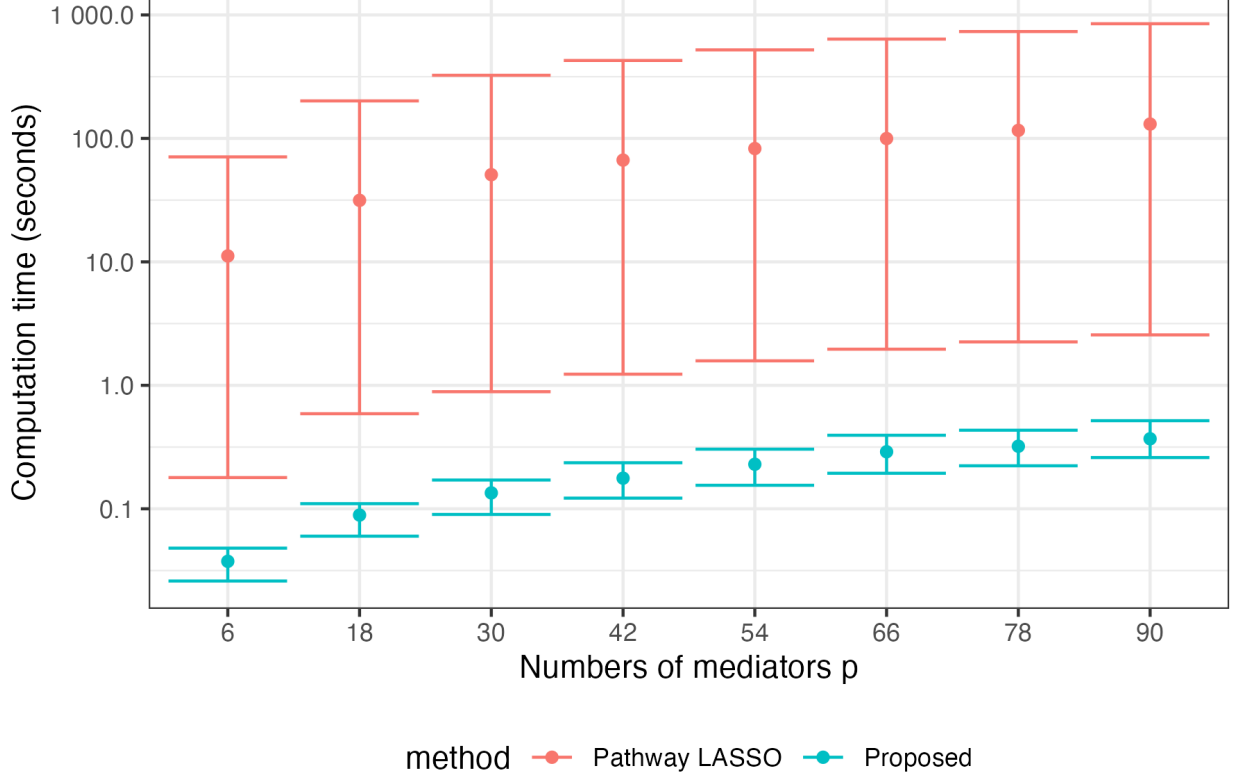
Therefore, (37) is proved.

Figure 8: Empirical computation time for fitting Pathway LASSO and JAP once across 100 repetitions. The dots represent the average computation time, and the error bars indicate the 0.05 and 0.95 quantiles.

E   Proof of Proposition prop:penaltyscale

We have that

$$\frac{\hat{w}_{nj,\alpha}}{\hat{w}_{\mathrm{AL},nj,\alpha}} = 1 + |\hat{\alpha}_{nj}^0|^{\gamma_\alpha - 2\eta_\alpha}|\hat{\beta}_{nj}^0|^{2\gamma_\alpha} \quad \text{and} \quad \frac{\hat{w}_{nj,\beta}}{\hat{w}_{\mathrm{AL},nj,\beta}} = 1 + |\hat{\alpha}_{nj}^0|^{2\gamma_\beta}|\hat{\beta}_{nj}^0|^{\gamma_\beta - 2\eta_\beta}. \tag{53}$$

When $0 < 2\eta_\alpha < \gamma_\alpha$ and $0 < 2\eta_\beta < \gamma_\beta$, (13) follows from Condition 1 and Slutsky's Theorem.

F   Pathway LASSO

In this section, we present supplementary simulation studies on Pathway LASSO.

Setup. We adopt a data generation mechanism that is a special case of that in Section 6 with a focus on Case (I) of $\boldsymbol{E}$ with $\rho = 0$. We set $n = 2000$ and $\eta^* = 1$. For the pairwise coefficients $\{(\alpha_j^*, \beta_j^*) : j = 1, \ldots, p\}$, we still consider the six groups of patterns as defined in (14) and Table 1 with $\delta = 2^{-1.5}$. Under each setting, we simulate 100 independent datasets.

Computation time. The computation time of Pathway LASSO grows rapidly in p, making it prohibitive for the $p = 150$ setting considered in Section 6. To illustrate, we next

consider a smaller range of $p \in \{6, 18, 30, 42, 54, 66, 78, 90\}$. Figure 8 compares the empirical computation time between Pathway LASSO and JAP when they are fitted once. Notably, Pathway LASSO is almost always 300 times more time-consuming than JAP. For our simulations in Section 6, including Pathway LASSO is even more prohibitive as a larger $p = 150$ is used, and cross-validation is applied for optimal penalty parameter λ. As an example, when $p = 90$, selecting the optimal λ from 40 candidates using 5-fold cross-validation takes Pathway LASSO over 14 hours, whereas JAP completes the same task in under 3 minutes.

Selection accuracy. Due to the computational burden discussed above, we evaluate the selection accuracy of Pathway LASSO for $p \in \{6, \ldots, 90\}$, which are smaller than $p = 150$ in Section 6, and their default range of hyperparameters is considered. Specifically, their penalty hyperparameter λ is chosen from values ranging from 0.0014 to 100000, formed by combining two uniformly logarithmically spaced sequences, featuring a denser grid for smaller magnitudes and a sparser grid for larger magnitudes. When fitting JAP for comparison, we let $\lambda_{n\alpha}$ and $\lambda_{n\beta}$ in (3) and (4) take the same range of forty values as λ in Pathway LASSO, while fixing $\gamma_\alpha = \eta_\alpha = \eta_\alpha = \eta_\beta = 1$ in (5). Across all p values, the empirical accuracy of selecting $\mathcal{A}^*$ of JAP under the tuned hyperparameters is 1, and that of Pathway LASSO under all values of λ are 0. In future studies, it could be of interest to improve the accuracy of Pathway LASSO by exploring a larger range of candidate hyperparameters.