HHS Public Access

Author manuscript

Stat Interface. Author manuscript; available in PMC 2023 January 01.

Published in final edited form as:

Stat Interface. 2022 ; 15(1): 39–50. doi:10.4310/21-sii673.

Pathway Lasso: Pathway Estimation and Selection with High-Dimensional Mediators

Yi Zhao,

Department of Biostatistics and Health Data Science, Indiana University School of Medicine, 410 West 10th Street, Indianapolis, IN USA

Xi Luo*

Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, 1200 Pressler Street, Houston, TX USA

Abstract

In many scientific studies, it becomes increasingly important to delineate the pathways through a large number of mediators, such as genetic and brain mediators. Structural equation modeling (SEM) is a popular technique to estimate the pathway effects, commonly expressed as the product of coefficients. However, it becomes unstable and computationally challenging to fit such models with high-dimensional mediators. This paper proposes a sparse mediation model using a regularized SEM approach, where sparsity means that a small number of mediators have a nonzero mediation effect between a treatment and an outcome. To address the model selection challenge, we innovate by introducing a new penalty called *Pathway Lasso*. This penalty function is a convex relaxation of the non-convex product function for the mediation effects, and it enables a computationally tractable optimization criterion to estimate and select pathway effects simultaneously. We develop a fast ADMM-type algorithm to compute the model parameters, and we show that the iterative updates can be expressed in closed form. We also prove the asymptotic consistency of our Pathway Lasso estimator for the mediation effect. On both simulated data and an fMRI data set, the proposed approach yields higher pathway selection accuracy and lower estimation bias than competing methods.

Keywords

Convex optimization; Mediation analysis; Structural equation modeling; Path analysis

1. INTRODUCTION

Mediation analysis is widely applied in social, economic, and biological sciences to assess the effect of a treatment or exposure on an outcome of interest passing through intermediate variables (mediators). Recently, it becomes increasingly popular to study the decomposition of the total treatment effect on the outcome through multiple mediation pathways. This

* Corresponding author. rossi.stat@gmail.com.

paper studies the problem of pathway selection and effect estimation under the setting of a large number of pathways.

Classical mediation analysis usually involves one mediator [for example, 4, 16, 31]. Recent extensions studied the multiple mediator setting [17, 12, 19, 32], though most of the methods are designed for relatively low-dimensional data. For the setting with more than two mediators, structural equation models (SEMs) are commonly employed where all the mediators are entered as predictors in regression-type models [24, 40]. This paper focuses on the estimation and computational problems under the SEM framework, and extends to the setting of high-dimensional mediators where the number of mediators is close to or larger than the sample size.

To improve estimation stability, reducing high-dimensional mediators into linear combinations has been studied. Huang and Pan [15] employed principal component analysis (PCA) to reduce the multivariate mediation model into multiple independent single-mediator models. Chén et al. [11] employed a matrix decomposition method, where the linear projection is optimized by maximizing the joint likelihood of the SEMs. Zhao et al. [41] recently extended the proposal in Huang and Pan [15] with sparse PCA to improve the interpretability of the mediator PCs. As linear combinations of the original mediators are used in these methods, they provided limited interpretability for each mediation pathway.

Besides dimension reduction, regularization is another general approach for high dimensional problems. In a related problem, Shojaie and Michailidis [27] employed Lasso [30] to penalize each connection in directed acyclic graphs. In another related problem on latent factor structural equation models, various regularization penalty choices were considered, including Lasso [18] and Bayesian Lasso [14]. Zhang et al. [38] utilized the MCP [13] regularization criterion to estimate and test the mediator to outcome path effect under the high-dimensional mediator setting in epigenetic studies. These approaches, however, do not address the problem of regularizing the effects of mediation pathways, which are commonly represented as products of two parameters [31]. The product of two parameters is a non-convex function that is not considered by existing convex regularization methods including various Lasso-type penalties. In this study, we will introduce a new convex penalty, named *Pathway Lasso*, to directly regularize the pathway effects.

We motivate our method by studying brain pathways using task-related functional magnetic resonance imaging (fMRI). In the early attempts of modeling neurological images as mediators between experimental stimuli and psychological outcomes, univariate mediation analysis is a widely employed approach, where univariate summaries were extracted from the multivariate images fitted separately in univariate mediator models under the assumption that the summaries are independent [8, 33, 3]. In recent studies, Chén et al. [11] and Zhao et al. [41] proposed orthogonalization approaches to transform the high-dimensional mediator candidates into independent directions/components as mediators with limited interpretability. In this study, we propose a regularized approach to perform mediation analysis with high-dimensional mediators. This paper circumvents the following limitations. First, it allows modeling correlated mediators directly, which is the setting where analyzing multiple brain regions as mediators. Second, it allows direct and more straightforward

interpretation of each mediator pathway whereas linear combinations (e.g. via principal component analysis) can be less interpretable.

The contributions of this paper are the following: (1) this is among the first attempts to model high-dimensional mediation pathways *jointly*; (2) we propose a general multiple mediator model under the SEM framework, which relaxes the ordering assumption of the mediators; (3) we introduce a novel convex penalty, *Pathway Lasso* penalty, for the non-convex function of the product, and this penalty enables simultaneous pathway selection and pathway effect estimation; (4) we propose an alternating direction method of multipliers (ADMM) type algorithm and study the solutions in closed form; (5) theoretical analysis shows that the proposed *Pathway Lasso* estimators consistently estimate the total mediation effect; (6) we demonstrate the robustness and advantages of the proposed methods through simulation studies and a publicly available fMRI data set.

This paper is organized as follows. In Section 2, we present the model with multiple dependent mediators. We introduce an ℓ_1'-regularization on the mediation pathways to select the mechanisms in Section 3. To estimate the parameters, an ADMM combined with augmented Lagrangian algorithm is developed. In Section 4, we compare the performance of our approaches with the marginal SEM approach through simulation studies. The proposed methods are applied to an open-source fMRI data set in Section 5. Section 6 summarizes this paper with discussions. The supplementary materials collect the technical proofs and additional results.

2. MODEL

2.1 A marginal model with multiple mediators

In this section, we first introduce a marginal model with multiple mediators which does not require identifying the temporal ordering of the mediators. In many scientific studies, the ordering of the mediators is usually unknown. In fMRI experiments, the ordering of brain mediators is generally hard to determine due to the low temporal resolution of the technique.

Let $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^n$, $\mathbf{M}_j = (M_{ij}, \ldots, M_{nj})^\top \in \mathbb{R}^n$ (for $j = 1, \ldots, K$) and $\mathbf{R} = (R_1, \ldots, R_n)^\top \in \mathbb{R}^n$ denote the observational data of treatment assignment (Z), K mediators (M_j's) and the outcome (Y) of n subjects, respectively. Under the linear SEM (LSEM) framework, we propose the following

$$\begin{aligned} &\mathbf{M}_1 = \mathbf{Z}A_1 + \mathbf{E}_{11}, \ \cdots, \ \mathbf{M}_K = \mathbf{Z}A_K + \mathbf{E}_{1K}, \\ &\mathbf{R} = \mathbf{Z}C + \mathbf{M}_1 B_1 + \cdots + \mathbf{M}_K B_k + \mathbf{E}_2, \end{aligned} \tag{1}$$

where $A_1, \ldots, A_K$, $B_1, \ldots, B_K$ and C are model coefficients; $\mathbf{E}_{11}, \ldots, \mathbf{E}_{1K}$ are model errors in the mediator models $\left(\mathbf{E}_{1j} = (E_{11j}, \ldots, E_{1nj})^\top \in \mathbb{R}^n \text{ for } j = 1, \ldots, K\right)$, which are assumed to be normally distributed with mean zero and covariance matrix $\mathbf{\Sigma}_1$, and $\mathbf{E}_{1j}$'s are independent of $\mathbf{Z}$; and $\mathbf{E}_2 = (E_{21}, \ldots, E_{2n})^\top \in \mathbb{R}^n$ is the model error in the outcome model normally distributed with mean zero and variance σ_2^2, and $\mathbf{E}_2$ is independent of $\mathbf{Z}, \mathbf{M}_1, \ldots, \mathbf{M}_K$. $\mathbf{E}_2$ is assumed to be independent of $\{\mathbf{E}_{11}, \ldots, \mathbf{E}_{1K}\}$. Here, for simplicity, the data are assumed

to be centered, and thus, the intercept terms are dropped. Under model (1), we allow the mediators to be dependent of each other, as long as their dependence structure is captured by the error correlation matrix.

In this study, we are interested in estimating and identifying nonzero path effects. The product, A_jB_j, is interpreted as the path effect of mediator M_j (as shown in Figure 1(b)), for $j = 1, \ldots, K$, and $\sum_{j=1}^{K} A_jB_j$ is the overall path effect of the treatment on the outcome through the mediators. C denotes the treatment effect not through the mediators. These interpretations build on a strong and potentially unrealistic assumption that all mediators are sequentially ignorable [17]. Formal interpretations of these individual coefficients are beyond the scope of this paper, as it involves studying theoretical assumptions for a large number of potential outcomes under various combinations of treatment and mediator assignments, which grows exponentially with the number of mediators. A special case of the model when $K = 2$ coincides with the most common two-mediator model in practice [17]. When $K < n$ and the mediators are conditionally independent given Z ($\mathbf{\Sigma}_1$ is a diagonal matrix), model (1) is equivalent to the marginal mediation analysis [31]. However, the proposed method scales well with general K, even for $K > n$. In the next section, we link this model to a special scenario where the sequential ordering of the mediators is known.

2.2 A special model with sequential mediators

In this section, we consider a scenario that the sequential ordering or dependence of the mediators is fully specified. That is, $M_1 \rightarrow M_2 \rightarrow \cdots \rightarrow M_K$. Though it is unlikely that the ordering is known in most experimental data with many mediators, such as the fMRI data application, we use it to demonstrate that the model parameters still provide meaningful insight. Suppose the underlying mechanism is represented using a diagram as in Figure 1(a). In the figure, the mediators are related in such a way that "earlier" M's (with smaller subscript) may affect "later" ones. The LSEM representation of this diagram is

$$\begin{aligned} &\mathbf{M}_1 = \mathbf{Z}a_1 + \epsilon_{11},\ \mathbf{M}_2 = \mathbf{Z}a_2 + \mathbf{M}_1 d_{12} + \epsilon_{12},\ \cdots, \\ &\mathbf{M}_K = \mathbf{Z}a_K + \mathbf{M}_1 d_{1K} + \cdots + \mathbf{M}_{K-1} d_{K-1,K} + \epsilon_{1K}, \\ &\mathbf{R} = \mathbf{Z}c + \mathbf{M}_1 b_1 + \mathbf{M}_2 b_2 + \cdots + \mathbf{M}_K b_K + \epsilon_2, \end{aligned} \tag{2}$$

where $a_1, \ldots, a_K, b_1, \ldots, b_K, c, d_{12}, \ldots, d_{K-1,K}$ are the model coefficients; and $\boldsymbol{\epsilon}_{11}, \ldots, \boldsymbol{\epsilon}_{1K}$ and $\boldsymbol{\epsilon}_2$ are the model errors, which are assumed to be mutually independent.

We show that the model coefficients and errors in the proposed marginal model (1) have the following relationship with the ones in model (2):

$$\mathbf{A} = \mathbf{a}(\mathbf{I}_K - \Delta)^{-1}, \mathbf{B} = \mathbf{b}, C = c,$$

$$\mathbf{E}_1 = \epsilon_1(\mathbf{I}_K - \Delta)^{-1}, \mathbf{E}_2 = \epsilon_2,$$

where $\mathbf{A} = (A_1, \ldots, A_K) \in \mathbb{R}^{1\times K}$, $\mathbf{a} = (a_1, \ldots, a_K) \in \mathbb{R}^{1\times K}$, $\mathbf{B} = (B_1, \ldots, B_K)^\top \in \mathbb{R}^K$, $\mathbf{b} = (b_1, \ldots, b_K)^\top \in \mathbb{R}^K$; $\mathbf{E}_1 = (\mathbf{E}_{11}, \ldots, \mathbf{E}_{1K}) \in \mathbb{R}^{n\times K}$, $\epsilon_1 = (\epsilon_{11}, \ldots, \epsilon_{1K}) \in \mathbb{R}^{n\times K}$;

$$\Delta = \begin{pmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1K} \\ & 0 & d_{23} & \cdots & d_{2K} \\ & & \ddots & & \vdots \\ & & & \ddots & d_{K-1,K} \\ & 0 & & & 0 \end{pmatrix} \in \mathbb{R}^{K\times K}$$

is the weighted adjacency matrix of the mediators in model (2), which is an upper-triangular matrix; and $\mathbf{I}_K$ is the K-dimensional identity matrix. Matrix $(\mathbf{I}_K - \mathbf{\Delta})^{-1}$ is called the influence matrix [27], where the (l, j) element represents the influence of mediator M_l on mediator M_j (for $l < j$) with a self-influence of one when $l = j$. In model (2), a_j denotes the direct effect of Z on M_j. Multiplying the a_j's with the influence matrix, A_j accounts the total effect of Z on M_j, and thus A_jB_j is the path effects of Z on R through M_j, for $j = 1, \ldots, K$. Under model (2), $\boldsymbol{\epsilon}_{11}, \ldots, \boldsymbol{\epsilon}_{1K}$ and $\boldsymbol{\epsilon}_2$ are mutually independent, and thus $\mathbf{E}_{1j}$ is independent of $\mathbf{E}_2$ for each $j = 1, \ldots, K$. Assume $\mathrm{Var}(\epsilon_{1j}) = \xi_{1j}^2\mathbf{I}_n$, then

$$\mathrm{Cov}[\mathrm{vec}(\epsilon_1)] = \mathrm{diag}\left\{\xi_{11}^2, \ldots, \xi_{1K}^2\right\} \otimes \mathbf{I}_n \triangleq \Xi \otimes \mathbf{I}_n,$$
$$\mathrm{Cov}[\mathrm{vec}(\mathbf{E}_1)] = \left(\mathbf{I}_K - \Delta^\top\right)^{-1} \Xi \left(\mathbf{I}_K - \Delta\right)^{-1} \otimes \mathbf{I}_n \triangleq \Sigma_1 \otimes \mathbf{I}_n,$$

where vec(·) is the vectorization operator of a matrix and ⊗ is the Kronecker product operator. Thus, $\mathbf{\Sigma}_1 = (\mathbf{I}_K - \mathbf{\Delta}^>)^{-1}\mathbf{\Xi}(\mathbf{I}_K - \mathbf{\Delta})^{-1}$. Under the Gaussian error assumption, model (1) therefore accounts for the dependence between the mediators through the error correlations. A special case is that all the mediators are independent [17], where $\mathbf{\Delta}$ is a zero matrix. Under this case, the derivation above shows that the errors in model (1) are also mutually independent.

3. METHOD

In this section, we introduced a regularized estimator of the path effects. Let $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_K) \in \mathbb{R}^{n\times K}$, model (1) can be written in matrix form as

$$\begin{aligned} \mathbf{M} &= \mathbf{Z}A + \mathbf{E}_1, \\ \mathbf{R} &= \mathbf{Z}C + \mathbf{M}B + \mathbf{E}_2\,. \end{aligned} \tag{3}$$

We consider continuous outcome and mediators with normally distributed errors as

$$\mathrm{vec}(\mathbf{E}_1) \sim \mathcal{N}_{n\times K}(0, \Sigma_1 \otimes \mathbf{I}_n), \mathbf{E}_2 \sim \mathcal{N}_n\left(0, \sigma_2^2\mathbf{I}_n\right),$$

and $\mathbf{E}_1$ and $\mathbf{E}_2$ are independent.

3.1 A Pathway Lasso method

Using the likelihood formulation, we first define a convex loss function of $(\mathbf{A}, \mathbf{B}, C)$ as

$$\ell(\mathbf{A},\mathbf{B},C) = \mathrm{tr}\left[\boldsymbol{\Omega}_1(\mathbf{M}-\mathbf{Z}\mathrm{A})^\top(\mathbf{M}-\mathbf{Z}\mathrm{A})\right] + w_2(\mathbf{R}-\mathbf{Z}C-\mathbf{M}\mathrm{B})^\top(\mathbf{R}-\mathbf{Z}C-\mathbf{M}\mathrm{B}), \tag{4}$$

where $\Omega_1 = \Sigma_1^{-1} \succ 0$ (positive-definite) is the inverse covariance matrix of the mediator errors, and $w_2 = \sigma_2^{-2} > 0$ is the inverse variance of the outcome error. In this paper, we are not interested in estimating $\boldsymbol{\Omega}_1$ or w_2 for the purpose of point estimation. Replacing them with unit variance will not affect the consistency of the least-squares type estimators, as long as all the variables are standardized to the unit scale [35]. Though the statistical inference (e.g. efficiency) of the estimates will be affected, this paper focuses on the problems of point estimation and model selection. Thus, we do not treat them as parameters to be estimated simultaneously with (**A**,**B**,*C*), but rather we replace them with any reasonable covariance estimates before running the optimization algorithm to be introduced below. In practice, one can replace $\boldsymbol{\Omega}_1$ and w_2 with an identity matrix and one, respectively, by standardizing the data to unit scale. This choice corresponds to a special case of banded covariance matrix estimation [6]. Other choices can be used depending on different structural assumptions (see a review on covariance matrix estimation by Cai et al. [9] and a finite sample study of the Lasso error variance by Reid et al. [25]). In the simulation study in Section 4, the robustness of this simplification is examined.

To estimate and select the pathway effects A_jB_j, for $j = 1, \ldots, K$, we propose to minimize the following penalized criterion

$$\begin{aligned} & f(\mathbf{A},\mathbf{B},C) \\ &= \frac{1}{2}\ell + \lambda\left\{\sum_{j=1}^{K}\left(|A_jB_j| + \phi\left(A_j^2+B_j^2\right)\right) + |C|\right\} \\ &+ \omega\left\{\sum_{j=1}^{K}\left(|A_j| + |B_j|\right)\right\} \end{aligned} \tag{5}$$

$$= \frac{1}{2}\ell + \lambda P_1(\mathbf{A},\mathbf{B},C) + \omega P_2(\mathbf{A},\mathbf{B}), \tag{6}$$

where $\lambda, \phi, \omega \geq 0$ are the tuning parameters. The first penalty term P_1 aims to stabilize and shrink the estimates for the pathway effects A_jB_j and C, and the second term P_2 aims to provide additional shrinkage to the individual A_j and B_j. In particular, P_1 aims to provide a convex penalty for the parameter of interest, A_jB_j. The combination of P_1 and P_2 is similar in spirit to the elastic net [42]. It should be noted that the method will also work if the tuning parameters vary with j. We here use the same parameters for simplicity as all the variables in the data are standardized to unit scale across j.

When $\phi = 0$, P_1 intuitively shrinks the pathway effect $|A_jB_j|$ and $|C|$ towards zero via a Lasso-type regularization. However, P_1 is not convex under this setting. The following theorem shows that that P_1 is convex if and only if $\phi \geq 1/2$.

Theorem 3.1.—For $a, b \in \mathbb{R}$, if and only if when $\phi \geq 1/2$,

$$v(a,b) = \left|ab\right| + \phi\left(a^2 + b^2\right) \tag{7}$$

is a convex function. When $\phi > 1/2$, it is strictly convex.

Figure 2 shows the 3D plot of three different penalty functions and the contour plot of penalty function (7) with different choices of ϕ. From the figures, we can see that $|ab|$ is not a convex function while $|ab| + (a^2 + b^2)/2$ is, and it is very different from the ℓ_1 penalty $|a| + |b|$. The contour plot indicates that ϕ determines the convexity of the penalty function. The penalty P_1 is non-differentiable at the points where $ab = 0$. The contour of P_1 approaches the ℓ_2 (or ridge) penalty when $\phi \to \infty$, and it approaches the ℓ_1 penalty when $\phi = 1/2$. In Section 4, we will examine the choice of ϕ through simulation studies.

The proposed Pathway Lasso penalty P_1 also differs from the Group Lasso penalty $\sqrt{A_j^2 + B_j^2}$ [37]. In the mediation model, the study interest is in shrinking A_jB_j towards zero for each j. There are four possible scenarios for A_j and B_j: (i) $A_j = B_j = 0$; (ii) $A_j = 0$ and $B_j \neq 06$; (iii) $A_j \neq 0$ and $B_j = 0$; (iv) $A_j \neq 0$ and $B_j \neq 0$. The indirect effect A_jB_i for the jth pathway is zero under scenarios (i)–(iii) and is nonzero under scenario (iv). The pathway penalty estimates allow all the four possible scenarios, and the penalty P_1 encourages scenarios (i)–(iii) because under such it is not differentiable. On the contrary, the Group Lasso only allows two scenarios (i) and (iv). Ignoring the other two scenarios in the Group Lasso may lead to model misspecification and biased estimation. For example, suppose the true model of A_j and B_j falls under scenario (ii), and thus M_j is a potential mediator-outcome confounding factor with no mediation effect. The Group Lasso estimate would either yield an incorrect nonzero mediation effect or introduce model bias by dropping the confounder M_j for the outcome model.

To illustrate the difference in shrinkage effects between the Pathway Lasso penalty P_1 and other popular choices, we plot the solution $(\hat{a}, \hat{b})$ to the following toy optimization problem

$$\min_{a,b} |a^* - a|^2 + |b^* - b|^2 + \lambda \text{pen}(a,b), \tag{8}$$

where pen(·,·) is a penalty function that is set to the Pathway Lasso, Lasso, or Group Lasso. Figure 3 compares the estimated product $\hat{a}\hat{b}$ as λ varies. To compare across different penalties, we fix the magnitude of the estimates $(|\hat{a}| + |\hat{b}|)$ on the x-axis. This figure shows that the Lasso shrinks ab aggressively towards zero because $\hat{a}\hat{b}$ goes to zero faster under the Lasso than under the Pathway Lasso as the magnitude $|\hat{a}| + \left|\hat{b}\right|$ decreases and the resulting product estimate is nonzero only when $|\hat{a}| + \left|\hat{b}\right|$ is relatively large. On the other hand, the Group Lasso provides gradual shrinkage to the product but can yield a zero product value only when $|\hat{a}| = \left|\hat{b}\right| = 0$. The Pathway Lasso shrinks ab less aggressively than the Lasso, and can yield a zero product even when $|\hat{a}| = \left|\hat{b}\right| \neq 0$.

3.2 An alternating direction method of multipliers

To estimate the model parameters, we observe that the objective function f consists of two parts, *i*) the differentiable loss function $\ell 2$, and *ii*) the non-differentiable penalty function. We propose to employ the alternating direction method of multipliers (ADMM), which is well suited to large-scale statistical problems [7]. The ADMM form of minimizing (5) is

$$\begin{aligned} &\text{minimize } u(\mathbf{A},\mathbf{B},C)+v(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma),\\ &\text{subject to } \mathbf{A}=\alpha,\ \mathbf{B}=\boldsymbol{\beta},\ C=\gamma, \end{aligned} \tag{9}$$

where

$$\begin{aligned} u(\mathbf{A},\mathbf{B},C)=&\frac{1}{2}\mathrm{tr}\left\{\Omega_1(\mathbf{M}-\mathbf{Z}A)^{\top}(\mathbf{M}-\mathbf{Z}A)\right\}\\ &+\frac{1}{2}w_2(\mathbf{R}-\mathbf{Z}C-\mathbf{M}B)^{\top}(\mathbf{R}-\mathbf{Z}C-\mathbf{M}B) \end{aligned} \tag{10}$$

is the differentiable loss function, and

$$\begin{aligned} v(\alpha,\beta,\gamma)=&\lambda\left\{\sum_{j=1}^{K}|\alpha_j\beta_j|+\sum_{j=1}^{K}\phi_j\left(\alpha_j^2+\beta_j^2\right)+|\gamma|\right\}\\ &+\omega\left\{\sum_{j=1}^{K}|\alpha_j|+\sum_{j=1}^{K}|\beta_j|\right\} \end{aligned} \tag{11}$$

is the non-differentiable regularization function, in which $\alpha=(\alpha_1,\ldots,\alpha_K)\in\mathbb{R}^{1\times K}$, $\beta=(\beta_1,\ldots,\beta_K)^{\top}\in\mathbb{R}^K$ and $\gamma\in\mathbb{R}$.

Following the ADMM approach, we introduce the augmented Lagrange function, $\mathscr{L}$, to enforce the constraints in the following optimization

$$\begin{aligned} &\mathscr{L}(\mathbf{A},\mathbf{B},C,\alpha,\beta,\gamma,\rho,\nu_r)\\ &=u(\mathbf{A},\mathbf{B},C)+v(\alpha,\beta,\gamma)\\ &+\sum_{r=1}^{3}\left\{\nu_r h_r(\mathbf{A},\mathbf{B},C,\alpha,\beta,\gamma)+\rho h_r^2(\mathbf{A},\mathbf{B},C,\alpha,\beta,\gamma)\right\}, \end{aligned} \tag{12}$$

where $h_1(\mathbf{A},\mathbf{B},C,\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)=\mathbf{A}-\boldsymbol{\alpha}$, $h_2(\mathbf{A},\mathbf{B},C,\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)=\mathbf{B}-\boldsymbol{\beta}$, and $h_3(\mathbf{A},\mathbf{B},C,\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)=C-\gamma$. We summarize the algorithm in Algorithm 1, where the parameters are updated iteratively. The solution for the updates of $\mathbf{A}$, $\mathbf{B}$, and C are provided in explicit forms in Section A.2 of the supplementary materials. The subproblem for updating $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is decomposed into K optimization problems for each coordinate. Each optimization problem is of the same form and has explicit solutions presented in Lemma A1 in the supplementary materials. In the lemma, ϕ_1 and ϕ_2 are required to be greater than λ to ensure the convexity of the objective function as demonstrated in Theorem 3.1. The solutions show that the shrinkage effect towards $ab=0$, including cases where only one of the (a, b) parameters is zero (conditions (5) and (6) in Lemma A1).

Algorithm 1

An algorithm of solving problem (9) using augmented Lagrangian method.

Given the results from the sth step, for the $(s+1)$th step,

$$\mathbf{A}^{(s+1)} = \arg\min_{\mathbf{A}} \mathscr{L}\left(\mathbf{A}, \mathbf{B}^{(s)}, C^{(s)}, \alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}, \rho, \nu_r^{(s)}\right);$$

$$\mathbf{B}^{(s+1)} = \arg\min_{\mathbf{B}} \mathscr{L}\left(\mathbf{A}^{(s+1)}, \mathbf{B}, C^{(s)}, \alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}, \rho, \nu_r^{(s)}\right);$$

$$C^{(s+1)} = \arg\min_{C} \mathscr{L}\left(\mathrm{A}^{(s+1)}, \mathrm{B}^{(s+1)}, C, \alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}, \rho, \nu_r^{(s)}\right);$$

$$\begin{pmatrix}\alpha^{(s+1)}\\ \beta^{(s+1)}\end{pmatrix} = \arg\min_{\alpha,\beta} \mathscr{L}\left(\mathrm{A}^{(s+1)}, \mathrm{B}^{(s+1)}, C^{(s+1)}, \alpha, \beta, \gamma^{(s)}, \rho, \nu_r^{(s)}\right);$$

$$\gamma^{(s+1)} = \arg\min_{\gamma} \mathscr{L}\left(\mathbf{A}^{(s+1)}, \mathbf{B}^{(s+1)}, C^{(s+1)}, \alpha^{(s+1)}, \beta^{(s+1)}, \gamma, \rho, \nu_r^{(s)}\right);$$

$$\nu_r^{(s+1)} = \nu_r^{(s)} + 2\rho h_r\left(\mathbf{A}^{(s+1)}, \mathbf{B}^{(s+1)}, C^{(s+1)}, \alpha^{(s+1)}, \beta^{(s+1)}, \gamma^{(s+1)}\right), \text{ for } r = 1; 2; 3.$$

The Lasso penalty (P_2) has been studied extensively in the literature. We here thus focus on analyzing the behavior of the penalty P_1. Table A2 in the supplementary materials lists the solutions with the penalty P_1 alone, which corresponds to setting $\omega = 0$ in Table A1. As shown in Section A.2 of the supplementary materials, $\phi_1 = \phi_2 = 2\lambda\phi + 2\rho$ in Algorithm 1, where $\phi \geq 1/2$ by Theorem 3.1 and ρ is the Lagrangian multiplier in the algorithm. In ADMM algorithms, ρ can be either fixed or increasing. The following proposition shows that the solutions by Algorithm 1 converge to zero when $\lambda \to \infty$, as long as $\rho/\lambda \to 0$. We use fixed $\rho = 1$ for simplicity.

Proposition 3.1.— *When $\phi_i = \kappa_i\lambda + \theta_i$, where $\kappa_i \geq 1$ is a constant, $\theta_i > 0$ and $\theta_i/\lambda \to 0$ as $\lambda \to \infty$ ($i = 1, 2$), the following problem is minimized at $a = b = 0$ when $\lambda \to \infty$:*

$$\min_{a,b \in \mathbb{R}} imize\ \lambda|ab| + \frac{1}{2}\phi_1 a^2 + \frac{1}{2}\phi_2 b^2 - \mu_1 a - \mu_2 b\,. \tag{13}$$

3.3 Consistency of the Pathway Lasso estimator

In this section, we prove that the proposed P_1 and P_2 penalties are prediction consistent under regularity conditions. The models in (2) include equations with the mediator M_j's serving as the response as well as equations serving as the explanatory variables. To simplify the presentation and focus on the path effect of the mediators, we assume that the direct effect, C, is known, and define the prediction loss based on the indirect pathways. We use ($\mathbf{A}^*$,$\mathbf{B}^*$) to denote the true coefficients. Let $W = R - ZC$. We first introduce the following definition.

Definition 3.1.— *For a treatment assignment Z, we define the intermediate prediction of W as*

$$\widehat{W} = Z\widehat{\mathbf{A}}\widehat{\mathbf{B}} = Z\left(\sum_{j=1}^{K} \widehat{A}_j \widehat{B}_j\right), \tag{14}$$

where $\widehat{\mathbf{A}}$ *and* $\widehat{\mathbf{B}}$ *are estimates of* $\mathbf{A}$ *and* $\mathbf{B}$, *respectively*.

In this definition, the outcome is predicted by the effect through the mediators based on the model estimate $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$. The estimated mean squared intermediate prediction error (MSIPE) is defined as

$$\widehat{\text{MSIPE}}[W(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})] = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{W}_i - W_i^*\right)^2, \tag{15}$$

where $W^* = Z\mathbf{A}^*\mathbf{B}^*$ is the intermediate outcome under the true parameters. It is well known that the Lagrange formulation (5) is equivalent to

$$\underset{\mathbf{A},\mathbf{B},C}{\text{minimize}} \ \ell(\mathbf{A}, \mathbf{B}, C),$$

$$\text{subject to} \ \ P_1(\mathbf{A}, \mathbf{B}, C) \le \theta_1,$$
$$P_2(\mathbf{A}, \mathbf{B}) \le \theta_2,$$

where θ_1 and θ_2 are tuning parameters that have correspondence to the Lagrange multipliers λ and ω. We derive the theory using the above formulation with the tuning parameters θ_1 and θ_2, because they are easier to interpret and analyze with minimal assumptions [10]. Under the following assumptions, we prove prediction consistency using either the P_1 penalty alone or a linear combination of P_1 and P_2, hereafter referred to as P_1 and $P_1 + P_2$, respectively.

Assumption (1)—The treatment vector $\mathbf{Z}$ is generated by a probability distribution with finite variance and the entries of $\mathbf{Z}$ are bounded so that $|Z_i| \le G$, for $i = 1, \ldots, n$, almost surely.

Assumption (2)—The tuning parameters are chosen such that

$$P_1(\mathbf{A}^*, \mathbf{B}^*) = \sum_{j=1}^{K}\left\{\left|A_j^* B_j^*\right| + \phi\left(A_j^{*2} + B_j^{*2}\right)\right\} \le \theta_1;$$

$$P_2(\mathbf{A}^*, \mathbf{B}^*) = \sum_{j=1}^{K}\left(\left|A_j^*\right| + \left|B_j^*\right|\right) \le \theta_2.$$

Assumption (3)—The model is correctly specified that

$$\mathbf{M} = Z\mathbf{A}^* + \mathbf{E}_1, \tag{16}$$

$$W = \mathbf{M}\mathbf{B}^* + E_2, \tag{17}$$

where $\mathbf{M} = (M_1, \ldots, M_K)$, $\mathbf{E}_1 = (E_{11}, \ldots, E_{1K})$; and $\mathbf{E}_1$ and E_2 are independent and normally distributed with mean zero and covariance matrix $\mathbf{\Sigma}_1$ and variance σ_2^2, respectively, with $\max\{\sigma_{11}, \ldots, \sigma_{1K}, \sigma_2\} \leq \sigma < \infty$, where $\text{diag}(\Sigma_1) = \left\{\sigma_{11}^2, \ldots, \sigma_{1K}^2\right\}$.

Theorem 3.2.—Under Assumptions (1)–(3), the estimated MSIPE is bounded, such that

i. *under the penalty formulation* P_1,

$$\mathbb{E}\left\{\widehat{\text{MSIPE}}(W(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}))\right\} \leq 2\theta_1(1 + s_2\kappa)G\sigma\sqrt{\frac{2\log(2K)}{n}}, \tag{18}$$

ii. *under the penalty formulation* $P_1 + P_2$,

$$\mathbb{E}\left\{\widehat{\text{MSIPE}}(W(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}))\right\} \leq 2\theta_1(1 + \theta_2)G\sigma\sqrt{\frac{2\log(2K)}{n}}, \tag{19}$$

where $s_2 = |\mathcal{S}_2|$ *is the cardinality of set* $\mathcal{S}_2 = \left\{j : B_j^* \neq 0\right\}$ *which is the support of* $\mathbf{B}^*$, *and the true* B_j^* *is bounded by* κ *such that* $\left|B_j^*\right| \leq \kappa$ *(for* $\forall j \in \mathcal{S}_2$*).*

The rates above are the same as the prediction loss bounds for standard Lasso regression without coherence-type assumptions on the design matrix [26, 10]. This prediction consistency result implies the consistency of estimating the mediation effect as stated in the following corollary.

Corollary 3.1.—*Assume* $\mathbb{E}Z^2 = q > 0$, *then the estimate of the total mediation effect is consistent in the* ℓ_2*-norm,*

i. *under the penalty formulation* P_1,

$$\left\|\widehat{\mathbf{A}}\widehat{B} - \mathbf{A}^*\mathbf{B}^*\right\|_2^2 \leq \frac{1}{q}\left\{8\theta_1^2G^2\sqrt{\frac{2\log(2)}{n}} + 2\theta_1(1 + s_2\kappa)G\sigma\sqrt{\frac{2\log(2K)}{n}}\right\}, \tag{20}$$

ii. *under the penalty formulation* $P_1 + P_2$,

$$\left\|\widehat{\mathbf{A}}\widehat{\mathbf{B}} - \mathbf{A}^*\mathbf{B}^*\right\|_2^2 \leq \frac{1}{q}\left\{8\theta_1^2G^2\sqrt{\frac{2\log(2)}{n}} + 2\theta_1(1 + \theta_2)G\sigma\sqrt{\frac{2\log(2K)}{n}}\right\}, \tag{21}$$

where $\widehat{\mathbf{A}}\widehat{\mathbf{B}} = \sum_{j=1}^{K} \hat{A}_j \hat{B}_j$ is the estimate of the total mediation effect and $\mathbf{A}^*\mathbf{B}^* = \sum_{j=1}^{K} A_j^* B_j^*$ is the true total mediation effect.

Corollary 3.1 shows that the Pathway Lasso estimator (under either P_1 or P_1+P_2) of the mediation effect is consistent. Note the upper bound for P_1 depends explicitly on $s_2\kappa$, instead of the corresponding term θ_2 for $P_1 + P_2$. Though the difference is minimal theoretically, this may explain the difference in numerical performance as we show in Section 4.

4. SIMULATION STUDY

In this section, we compare the proposed Pathway Lasso (PathLasso) method with a marginal SEM approach. In the marginal SEM approach, the Baron-Kenny (BK) [4] mediation analysis is applied to each mediator separately. When the mediators are orthogonal or independent [31], the parameters are equivalent to the proposed marginal model with multiple mediators (1). The BK estimators are biased under the setting with dependent mediators. The pathway effects or the product estimators are tested by the delta method [28] and significant pathways are selected by controlling the false discovery rate [5]. In the simulation study, we generate $n = 50$ samples and vary the number of mediators with $K = 20,50,200$. For the proposed method, we consider four approaches with *a*) $\lambda = 0$ (P_2 penalty only); *b*) $\omega = 0$ (P_1 penalty only); *c*) $\omega = 0.1\lambda$; *d*) $\omega = \lambda$.

In the simulation, Z is firstly generated following a Bernoulli distribution with a probability of 0.5 to be one. The M_j's and R are then generated following model (1). In the models, for $j = 1, \ldots, K$, parameters $\{A_j, B_j\}$ are generated as presented in Figure 4(a) to include four types of mediators, (1) mediator with nonzero mediation effect ($A_jB_j \neq 0$, blue nodes in Figure 4(b)), (2) mediator with $A_j \neq 0$ but $B_j = 0$ (yellow nodes in Figure 4(b)), (3) mediator with $A_j = 0$ but $B_j \neq 0$ (green nodes in Figure 4(b)), and (4) mediator with $A_j = B_j = 0$ ($j = 12, \ldots, K$, not shown in the figure). C is set to be the maximum of A_jB_j. For all model errors, the standard deviation is set to be 200. In the M_j models, the covariance matrix $\mathbf{\Sigma}_1$ is set to be a sparse matrix with sparsity level $(1-1/K)$ (i.e. $1/K$ of the off-diagonal entries are randomly chosen to be nonzero) and the off-diagonal entries, ρ_M, are chosen from $\{0,\pm 0.4\}$. In the estimation, we set $\mathbf{W}_1$ to be an identity matrix and w_2 to be one after standardizing the data. With the existence of nonzero off-diagonal elements, the goal is to examine the robustness of this choice. The tuning parameter, λ, is set to be a sequence of values between 10^{-5} and 10^4; and the tuning parameter, ϕ, which controls the convexity of the penalty function, is set to vary from $\{0.5, 1, 2, 5, 10\}$. To compare the performance of various methods without setting the tuning parameter or p-value thresholds, we first employ the following metrics: (1) receiver operating characteristic (ROC) curves, and (2) the mean squared error (MSE) of the total mediation effect **AB** estimates. For a fair comparison, we compare the MSE under the same ℓ_1 norm of the estimated pathway effects for all methods.

In Section B.1 of the supplementary material, we show that the proposed method is not sensitive to the choice of ϕ in selecting mediation pathways or estimating the pathway effects. We fix $\phi = 2$ for the following simulations as it yields slightly better performance than the rest. Figure 5 compares the performance of all the considered methods with/without

error correlations between the mediators. The ROC curves of the BK method are almost the same as the diagonal line, even when the number of mediators is less than the number of observations, indicating that this multiple testing for marginal mediation effect approach loses the power of identifying the significant mechanisms regardless of the dependencies between the mediators. From the figure, we notice that the P_2 penalty, which adds ℓ_1 penalty on each A_j and B_j, improves the performance in identifying the pathways, while the P_1 penalty significantly decreases the mean squared error in estimating the mediation effects, as expected from Corollary 3.1. All the PathLasso methods yield a higher area under the ROC curve than the BK method. This suggests that for this high dimensional mediator problem, the regularization approach attains more reliable and higher statistical power in selecting mediators that have significant mediation effects. From the discussion in Section 2.2, the proposed marginal model reparameterizes the dependency between the mediators into the correlations among model errors. From the figure, the PathLasso methods perform similarly under different values of ρ_M. This demonstrates that setting $\mathbf{\Omega}_1$ and w_2 to identity is robust under varying dependence between the mediators.

For tuning parameter selection, we propose to employ the variable selection stability criterion introduced in Sun et al. [29]. The performance is presented in Table B1 in the supplementary materials. It is also observed that the P_1 penalty yields a lower mean squared error in estimating the mediation effects and adding P_2 helps improve the selection accuracy.

5. AN FMRI STUDY

We apply the proposed method to a task-fMRI data set obtained from the OpenfMRI database (accession number is ds000002). The task is a probabilistic classification learning (PCL) task using “weather prediction” [2, 1]. The goal is to investigate the mechanisms of PCL in human and to examine how the memory systems interact during the task. The experiment was designed to effectively distinguish neural responses to stimuli, delay, and negative and positive feedback components. In this study, we focus on identifying the brain mechanism that is associated with response delay, which is measured by the time to react. We use the data from a right-handed English-speaking participant aged between 21 to 26 in a healthy condition. The experiment was repeated for two scans (runs) to examine the test–retest reliability of fMRI. In this study, we also take the advantage of this two-scan design to evaluate the reliability of the proposed approach. Each scan consists of $n = 80$ trials with fifty PCL trials and thirty baseline trials. In each PCL trial, a stimulus was presented at randomized locations. The participant would respond by pressing either the left button for a “sun” prediction or the right button for a “rain” prediction. Baseline trials were included to control for visual stimulation, button press, and computer response to the button press. We consider the reaction time as the outcome (R) and aim to further identify brain pathways that have an intermediate effect on the reaction time when comparing PCL ($Z = 1$) and baseline ($Z = 0$) trials. We consider the single-trial activation [20] from $K = 128$ brain regions of interest (ROIs) as the mediator candidates (M_j’s). These brain regions are grouped into eleven functional modules [23]. More details of fMRI data processing are presented in Section C of the supplementary materials. As the BK method cannot efficiently identify significant mediation pathways, we only present the results from the PahtLasso (with $\phi = 2$) methods.

A reliability study demonstrates that the PathLasso under P_1 penalty obtains more stable results in both effect estimation and pathway selection (see Section C.2 of the supplementary materials). Using the tuning parameter selected by selection stability [29], six ROIs are selected by the PathLasso ($\omega = 0$) in both runs (Table 1 and Figure 6). The total effect of PCL trials on the reaction time is estimated as 0.405 in run 1 and 0.571 in run 2 compared to baseline trials. Compared to the button-pressing task, it takes a longer time during the PCL task as the brain is expected to take more time to process the stimuli and to make a prediction. From Table 1, all six regions have a positive mediation effect on the reaction time. The estimate of AB, A, and B are consistent between the two runs in terms of the direction (either both are positive or both are negative). Compared to the baseline trial, the two executive control regions, one in the prefrontal cortex and one in the medial temporal lobe, are less activated during the PCL trials. Existing studies have shown that the medial frontal and parietal cortex are deactivated when the task involves visual stimuli. The MTL, which is one of the major memory systems, was also identified to be deactivated during the classification learning task [22, 2]. The negative estimate of B suggests that the deactivation of these two regions increases the reaction time. The rest four ROIs are more activated during the PCL trials, and this activation further increases the time to respond. The classification learning task is a nondeclarative memory procedure. The opposite activation patterns in the striatum and the parietal regions support the competing role of two memory systems during learning [22]. It has been discovered that the activation in the visual cortex, which is involved in processing sensory feedback related to the motor response, is positively correlated with reaction time [36]. Applying the proposed method, we identify six potential mediation pathways that reveal the brain mechanisms of longer reaction time in the PCL task.

6. DISCUSSION

In this study, we propose a general marginal model for multiple dependent mediators under the SEM framework. A novel convex penalty is introduced for shrinkage estimation and pathway selection. We develop an ADMM algorithm to estimate the parameters and provide an explicit solution to the iterative updates. The simulation studies indicate that the Pathway Lasso method is robust and performs better than the marginal mediation approach in identifying significant pathway mechanisms. The numeric merits are further illustrated using a task-fMRI data set, on which the Pathway Lasso shows higher replicability in both effect estimation and pathway selection.

The new Pathway Lasso penalty is introduced to achieve the simultaneous pathway selection and pathway effect estimation purpose. The current study shows that the estimator is consistent in the ℓ_2-norm. For a Lasso procedure, the consistency of variable selection usually requires the incoherence assumption in the design matrix [21, 34, 39]. This incoherence assumption may not hold in a mediation problem, as the mediators can be highly dependent. We will leave the study of sparsistency of the method to future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and two anonymous referees for their valuable comments. R code is available on GitHub at https://github.com/zhaoyi1026/PathwayLasso. Luo was partially supported by National Institutes of Health grants R01EB022911 and R01MH110449 and National Science Foundation grant DMS 1557467.

REFERENCES

[1]. Aron AR, Gluck MA, and Poldrack RA (2006). Long-term test–retest reliability of functional MRI in a classification learning task. Neuroimage, 29(3):1000–1006. [PubMed: 16139527]

[2]. Aron AR, Shohamy D, Clark J, Myers C, Gluck MA, and Poldrack RA (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. Journal of Neurophysiology, 92(2):1144–1152. [PubMed: 15014103]

[3]. Atlas LY, Lindquist MA, Bolger N, and Wager TD (2014). Brain mediators of the effects of noxious heat on pain. PAIN®, 155(8):1632–1648. [PubMed: 24845572]

[4]. Baron RM and Kenny DA (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6):1173. [PubMed: 3806354]

[5]. Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1):289–300.

[6]. Bickel PJ and Levina E (2004). Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli, 10(6):989–1010.

[7]. Boyd S, Parikh N, Chu E, Peleato B, and Eckstein J (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning, 3(1):1–122.

[8]. Caffo B, Chen S, Stewart W, Bolla K, Yousem D, Davatzikos C, and Schwartz BS (2007). Are brain volumes based on magnetic resonance imaging mediators of the associations of cumulative lead dose with cognitive function? American Journal of Epidemiology, 167(4):429–437. [PubMed: 18079133]

[9]. Cai TT, Ren Z, and Zhou HH (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electronic Journal of Statistics, 10(1):1–59.

[10]. Chatterjee S (2013). Assumptionless consistency of the lasso. arXiv preprint arXiv:1303.5817.

[11]. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, and Lindquist MA (2017). High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics, 19(2):121–136.

[12]. Daniel R, De Stavola B, Cousens S, and Vansteelandt S (2015). Causal mediation analysis with multiple mediators. Biometrics, 71(1):1–14. [PubMed: 25351114]

[13]. Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360.

[14]. Guo R, Zhu H, Chow S-M, and Ibrahim JG (2012). Bayesian lasso for semiparametric structural equation models. Biometrics, 68(2):567–577. [PubMed: 22376150]

[15]. 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. [PubMed: 26414245]

[16]. Imai K, Keele L, and Yamamoto T (2010). Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science, 25(1):51–71.

[17]. 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.

[18]. Jacobucci R, Grimm KJ, and McArdle JJ (2016). Regularized structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal, 23(4):555–566. [PubMed: 27398019]

[19]. Lin S-H and VanderWeele TJ (2017). Interventional approach for path-specific effects. Journal of Causal Inference, 5(1).

[20]. Lindquist MA (2008). The statistical analysis of fMRI data. Statistical Science, 23(4):439–464.

[21]. Meinshausen N and Bühlmann P (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, pages 1436–1462.

[22]. Poldrack RA, Clark J, Pare-Blagoev E, Shohamy D, Moyano JC, Myers C, and Gluck MA (2001). Interactive memory systems in the human brain. Nature, 414(6863):546–550. [PubMed: 11734855]

[23]. Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, and Schlaggar BL (2011). Functional network organization of the human brain. Neuron, 72(4):665–678. [PubMed: 22099467]

[24]. Preacher KJ and Hayes AF (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40(3):879–891. [PubMed: 18697684]

[25]. Reid S, Tibshirani R, and Friedman J (2016). A study of error variance estimation in lasso regression. Statistica Sinica, pages 35–67.

[26]. Rigollet P, Tsybakov A, et al. (2011). Exponential screening and optimal rates of sparse estimation. The Annals of Statistics, 39(2):731–771.

[27]. Shojaie A and Michailidis G (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 97(3):519–538. [PubMed: 22434937]

[28]. Sobel ME (1982). Asymptotic confidence intervals for indirect effects in structural equation models. Sociological methodology, 13(1982):290–312.

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

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

[31]. VanderWeele TJ (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press.

[32]. Vansteelandt S and Daniel RM (2017). Interventional effects for mediation analysis with multiple mediators. Epidemiology (Cambridge, Mass.), 28(2):258.

[33]. Wager TD, Davidson ML, Hughes BL, Lindquist MA, and Ochsner KN (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron, 59(6):1037–1050. [PubMed: 18817740]

[34]. Wainwright M (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory, 55(5):2183–2202.

[35]. White H (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48(4):817–838.

[36]. Yarkoni T, Barch DM, Gray JR, Conturo TE, and Braver TS (2009). Bold correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fmri analysis. PLoS one, 4(1):e4257. [PubMed: 19165335]

[37]. Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67.

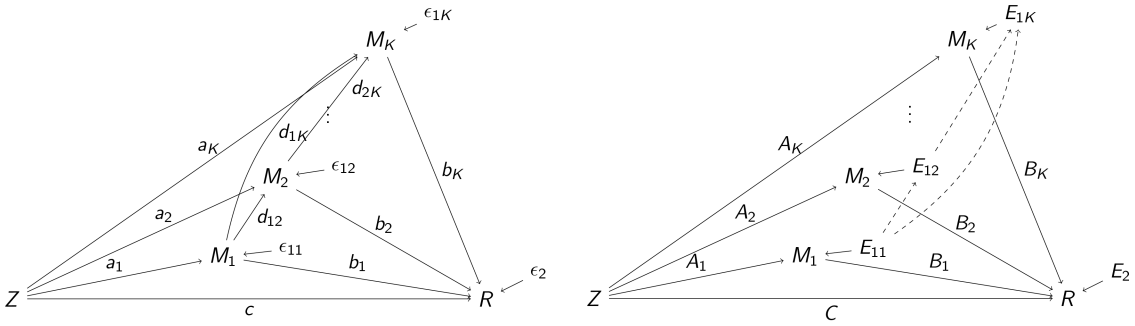
[38]. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, Zhang W, Schwartz J, Just A, and Colicino E (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics, page btw351.

[39]. Zhao P and Yu B (2006). On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563.

[40]. Zhao SD, Cai TT, and Li H (2014). More powerful genetic association testing via a new statistical framework for integrative genomics. Biometrics, 70(4):881–890. [PubMed: 24975802]

[41]. Zhao Y, Lindquist MA, and Caffo BS (2020). Sparse principal component based high-dimensional mediation analysis. Computational Statistics & Data Analysis, 142:106835. [PubMed: 32863492]

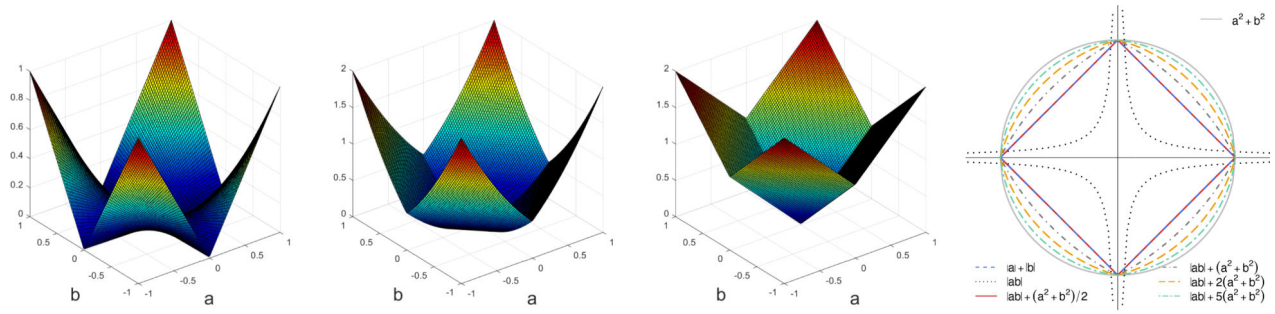
[42]. Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320.



(a) The sequential model. (b) The proposed marginal model.

Figure 1.

Diagram of (a) the sequential model and (b) the proposed marginal model with multiple dependent mechanisms. Z is the treatment, R is the outcome, and M_j's are the mediators.



(a) $|ab|$. (b) $|ab| + (a^2 + b^2)/2$. (c) $|a| + |b|$. (d) Contour plot of (7).

Figure 2.

3D plot of different penalty functions and the contour plot of function (7) under different choices of ϕ.

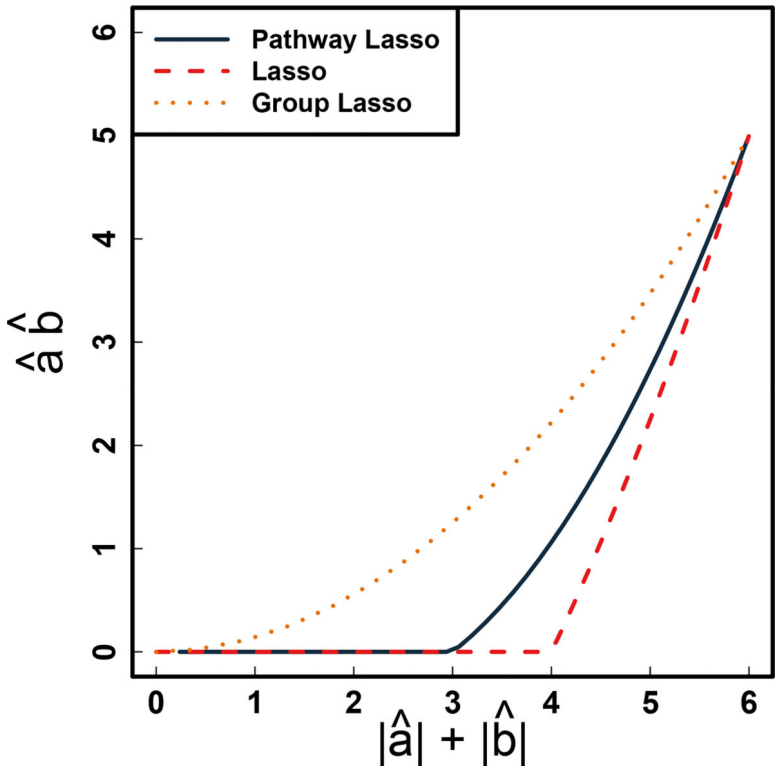
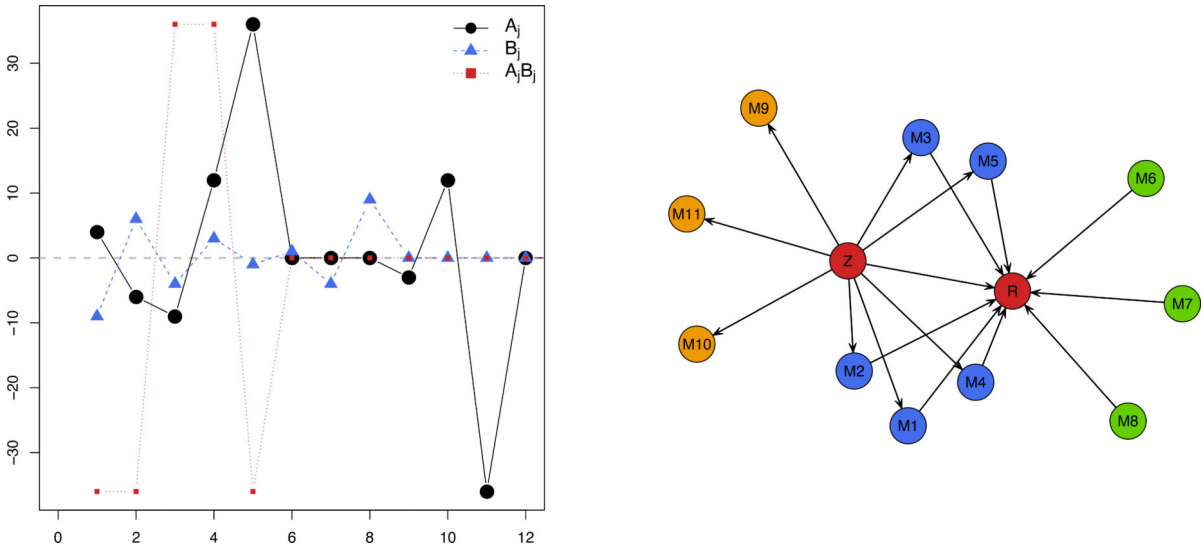


Figure 3.

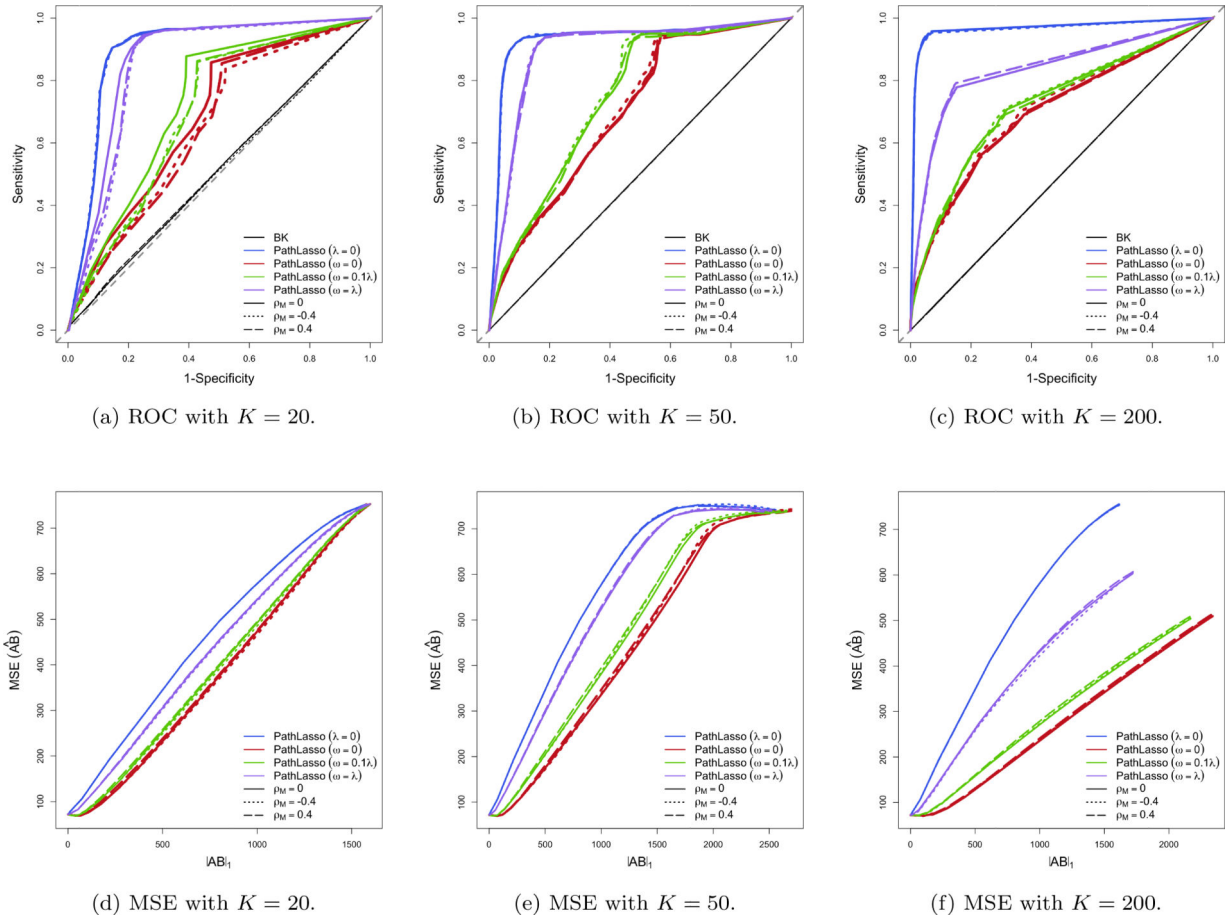
Comparing the shrinkage effect on the product, ab, under different penalty choices in (8) *as* λ *varies. The true a** = 5 *and b** = 1.



(a) Parameter setting in the simulation study.

(b) True causal mediation pathways.

Figure 4.
Parameter setting and the true underlying causal mediation pathways in the simulation study.



(a) ROC with $K = 20$. (b) ROC with $K = 50$. (c) ROC with $K = 200$.

(d) MSE with $K = 20$. (e) MSE with $K = 50$. (f) MSE with $K = 200$.

Figure 5.

Performance comparison under each ρ_M in ROC curves ((a)-(c)) and mean squared errors (MSE) in estimating the total mediation effect ((d)-(f)) with various numbers of mediators (K).

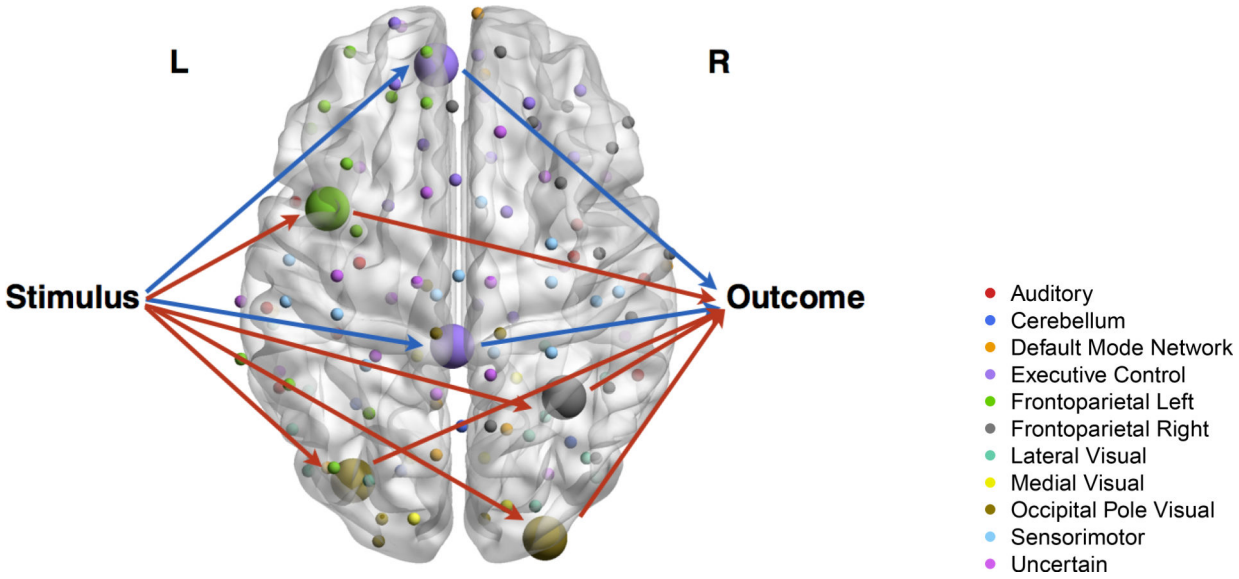


Figure 6.

Six (out of 128) brain regions identified with significant brain pathway effects on the reaction time under the probabilistic classification learning task in both runs.

Table 1.

Estimate of mediation effect in the probabilistic classification learning task. The study was repeated for two runs. In the table, (x,y,z) is the coordinate of the brain region in the MNI space. (EC: executive control, OPV: occipital pole visual, FPL: frontoparietal left, FPR: frontoparietal right.)

x	*y*	*z*	Module	*AB*		*A*		*B*	
				run1	run2	run1	run2	run1	run2
−2	−37	44	EC	0.0017	0.0021	−0.1271	−0.0761	−0.0134	−0.0279
−7	51	−1	EC	0.0036	0.0019	−0.0790	−0.0540	−0.0458	−0.0361
27	−97	−13	OPV	0.0012	0.0012	0.0725	0.0562	0.0160	0.0216
−33	−79	−13	OPV	0.0037	0.0012	0.1450	0.0470	0.0255	0.0249
−41	6	33	FPL	0.0020	0.0014	0.0440	0.0526	0.0465	0.0274
33	−53	44	FPR	0.0028	0.0014	0.0922	0.0395	0.0304	0.0346