Statistics in Biosciences (2022) 14:432–451
https://doi.org/10.1007/s12561-021-09328-0

High-Dimensional Mediation Analysis with Applications to Causal Gene Identification

Qi Zhang[1]

Received: 23 April 2020 / Revised: 14 October 2021 / Accepted: 20 October 2021 /
Published online: 29 October 2021

Abstract

Mediation analysis has been a popular framework for elucidating the mediating mechanism of the exposure effect on the outcome in many disciplines including genetic studies. Previous literature in causal mediation primarily focused on the classical settings with univariate exposure and univariate mediator, with recent growing interests in high-dimensional mediator. In this paper, we study the mediation model with high-dimensional exposure, high-dimensional continuous mediator, and a continuous outcome. We introduce two procedures for mediator selection, MedFix and MedMix, and develop the corresponding causal effect tests. Our study is motivated by the causal gene identification problem in biomedical studies, where causal genes are defined as the genes that mediate the genetic effect. For this problem, the genetic variants are the high-dimensional exposure, the gene expressions the high-dimensional mediator, and the phenotype of interest the outcome. We evaluate the proposed methods using a mouse f2 dataset for diabetes study, and extensive real data-driven simulations. We show that the mixed model-based approach (MedMix) leads to higher accuracy in mediator selection with reasonable reproducibility across independent measurements of the response and is more robust against model misspecification. The R code and additional materials are available on Github (https://github.com/QiZhangStat/highMed).

Keywords Mediation analysis · Variable selection · Data integration · Integrative omics

✉ Qi Zhang
qi.zhang2@unh.edu

[1] Department of Mathematics and Statistics, University of New Hampshire, Durham, NH, USA

1 Introduction

Mediation analysis is a type of causal inference investigating the effect of certain exposures (**Z**) on an outcome (**Y**, Fig. 1a). It assumes a hypothesized causal chain in which all or a part of the exposure effect on the outcome may be attributed to its effect on some mediators (**M**) which in turns influence the outcome (indirect effect). These variables could be also influenced by some observed confounders (**X**).

Mediation analysis can be traced back to almost 100 years ago [52, 53], and has been considered in the causal inference context since 1990s [30, 32]. The classical mediation analysis focuses on the case of univariate exposure and univariate mediator. The statistical causal inference literature has made tremendous progress in the estimation, testing and understanding of the mediation effect in such settings [2, 27, 44, 46]. Mediation models with multiple mediators or multiple exposures have also been studied. For example, [45] proposed two analytic approaches for multiple mediators. [20] studied the joint analysis of the phenotype, the expression of a gene and the SNPs using mediation framework with the gene expression as the univariate mediator and the SNPs within the gene as the low-dimensional multivariate exposure. Recently, models with high-dimensional mediator have began to draw attention, and many researchers have developed various estimation procedures and significance tests for the effect of high-dimensional mediator[16, 19, 37, 57] in various applications, and proposed PCA-like concepts for summarizing the effects of high-dimensional mediators in brain imaging studies [5, 58]. A concurrent work [60] considered the setting with low-dimensional multivariate exposure and high-dimensional mediators, and provided consistent and asymptotic normal estimates. The focus of this paper is the sparse estimation and evaluation of high-dimensional mediator effects with high-dimensional exposure (Fig. 1b), which is different from all above.

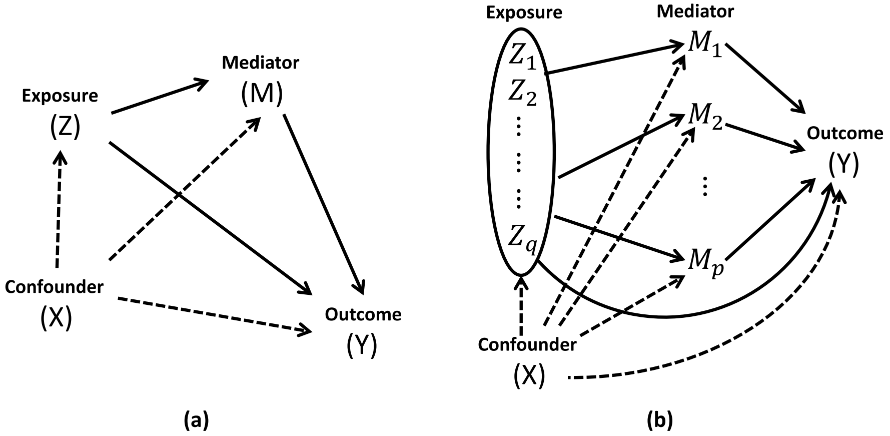

Fig. 1 Diagram of **a** Mediation analysis with univariate exposure and univariate mediator, and **b** Mediation analysis with high-dimensional exposure and high-dimensional mediators considered in this paper. The solid lines are links associated with the causal effects

One motivating example is the causal gene identification problem. How the genetic variants influence the phenotypes, and what molecular mechanism mediates such effects are the central questions in genomics. Along this direction, researchers have proposed numerous methods for identifying the *causal variant* at SNP level [22] or the *causal gene* at transcriptomic level [15, 19]. We focus on the latter problem in the natural genetic context. Specifically, we are interested in finding genes directly involved in the transcriptomic pathway from genetic variants to phenotype. Mediation models have been applied to this problem, most of which modeled one gene at a time [1, 61], or only considered univariate exposure [19, 37, 57].

In this paper, we study the mediation analysis with high-dimensional mediator and high-dimensional exposure. In particular, we are interested in mediator selection and evaluating the effect of the selected mediators. We propose two regularized regression-based approaches for this high-dimensional mediator selection problem, and develop relevant statistical tests for causal effects and measures of the mediation effect size. The first proposal is a "baseline" method that applies adaptive lasso [62] after modification under the conventional fixed effect regression framework for mediation analysis. The second approach is a novel method-based on high-dimensional linear mixed model. This proposal treats the direct effect as random effect, which reduces the model complexity and improves robustness against model misspecification.

We are interested in the effect of a multivariate exposure vector $\mathbf{Z} = (\mathbf{Z_1}, \dots, \mathbf{Z_q})$ on a univariate outcome $\mathbf{Y}$, and how it is mediated by the candidate mediators $\mathbf{M} = (\mathbf{M_1}, \dots, \mathbf{M_p})$ after being adjusted by length s confounders $\mathbf{X}$ (Fig. 1b). In the following, we formally introduce the model and define the causal effects for mediation models with high-dimensional exposure and high-dimensional candidate mediators using the counterfactual framework [34, 44].

Let $\mathbf{Y}(\mathbf{z}, \mathbf{m})$ denote the outcome when the exposure vector is set to $\mathbf{z}$ and the mediator vector is set to $\mathbf{m}$, and $\mathbf{M}(\mathbf{z})$ denote the value of the mediator vector if the exposure vector $\mathbf{Z}$ is set to $\mathbf{z}$. The total effect (TE) when the exposure is changed from $\mathbf{z}^\star$ to $\mathbf{z}$ is

$$\text{TE} = \mathbf{Y}(\mathbf{z}, \mathbf{M}(\mathbf{z})) - \mathbf{Y}(\mathbf{z}^\star, \mathbf{M}(\mathbf{z}^\star))$$

It can be decomposed into the natural indirect effect (NIE) and the natural direct effect (NDE)

$$\text{NIE} = \mathbf{Y}(\mathbf{z}, \mathbf{M}(\mathbf{z})) - \mathbf{Y}(\mathbf{z}, \mathbf{M}(\mathbf{z}^\star)), \quad \text{NDE} = \mathbf{Y}(\mathbf{z}, \mathbf{M}(\mathbf{z}^\star)) - \mathbf{Y}(\mathbf{z}^\star, \mathbf{M}(\mathbf{z}^\star))$$

To identify these causal effects, many causal assumptions for mediation model are necessary (Supplementary Table 1).

In this paper, we consider the following linear models for high-dimensional causal mediation.

$$E(\mathbf{Y}|\mathbf{X}, \mathbf{M}, \mathbf{Z}) = \mathbf{X}\alpha + \mathbf{M}\gamma + \mathbf{Z}\beta \quad \text{and} \quad E(\mathbf{M}|\mathbf{X}, \mathbf{Z}) = \mathbf{X}A + \mathbf{Z}B \tag{1}$$

where α, γ and β are vectors of length s, p and q, respectively; and A and B are matrices of size $s \times p$ and $q \times p$, respectively. The first equation in (1) is the *outcome model* and the rest are the *mediator models*. The motivating causal gene identification problem can be formulated as mediator selection, i.e., identifying $j \in \{1, \dots, p\}$ such that $\gamma_j \neq 0$ and $B_j \neq 0$ where B_j is the jth column of B. This is the primary goal of this paper.

Another goal of this paper is to evaluate the mediation effect. When the causal assumptions and model (1) hold, the total, natural indirect and natural direct effects are reduced to

$$\text{NDE} = (\mathbf{z} - \mathbf{z}^\star)^T \beta, \quad \text{NIE} = (\mathbf{z} - \mathbf{z}^\star)^T B\gamma, \quad \text{TE} = (\mathbf{z} - \mathbf{z}^\star)^T (B\gamma + \beta) \tag{2}$$

and the contribution of the individual mediators can also be defined as

$$\text{NIE}_j = (\mathbf{z} - \mathbf{z}^\star)^T B_j \gamma_j$$

The size of the overall mediation effect relative to the total effect is commonly defined as PM = NIE/TE in the literature [44]. For univariate exposure, PM is reduced to the ratio of the estimated coefficients of (1), as $(\mathbf{z} - \mathbf{z}^\star)$ in NIE and TE are canceled out. For multivariate exposure, if the only goal is testing whether TE, NDE or NIE exists, the test statistics do not explicitly depend on the specific realization of the exposure neither, as these tests are on the regression coefficients [20]. In our applications, we are interested in both testing the existence of the mediation effects and estimating its size. For the latter problem, the effect size measures in (2) and PM are all functions of $(\mathbf{z}, \mathbf{z}^\star)$. They could be potentially useful in genomic selection or personalized intervention when a specific exposure combination may be of interest. However, a population level overall measure of the mediation effect size is still needed to facilitate model assessment and interpretation in scientific research practice.

We propose to define such measure by treating the exposure as a random variable. In (2), let $\mathbf{z}^\star$ be a baseline exposure and kept fixed, and $\mathbf{z}$ be a randomly chosen exposure from the population under investigation. We use the variance of TE, NDE and NIE in (2) over the randomness of $\mathbf{z}$ as the population-level measures of effect sizes, and refer to them as the variance total effect (VTE), the variance direct effect (VDE) and the variance indirect effect (VIE). It immediately follows that

$$\text{VDE} = Var(\mathbf{Z}\beta),\ \text{VIE} = Var(\mathbf{Z}B\gamma),\ \text{VTE} = Var(\mathbf{Z}(B\gamma + \beta)) \tag{3}$$

Then we define the proportion of mediation effect simply as the **P**roportion of the **V**ariance **M**ediated (PVM), i.e.,

$$\text{PVM} = \text{VIE/VTE} \tag{4}$$

We remark that these definitions are specific to the population structure under investigation.

2 Regression Frameworks for High-Dimensional Mediation Analysis

With some abuse of notations, let Y denote the observed $n \times 1$ vector of the response, Z the $n \times q$ exposure matrix, M the $n \times p$ potential mediator matrix, and X the $n \times s$ confounder matrix including the intercept (Note that they represent data vector/matrices, which are different from the boldface $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{M}$ and $\mathbf{X}$ for random variables in the previous section). We assume that the columns of M and Z are centered and standardized, and potentially $q, p \gg n$, but presumably $s \ll n$. In genetics, Z could be the genotype matrix, M the gene expression matrix, and X includes the baseline covariates such as gender and age. Using the above notations, model (1) could be estimated using the following linear regression framework

$$\begin{aligned} Y =& X\alpha + M\gamma + Z\beta + \epsilon \\ M_j =& XA_j + ZB_j + \eta_j \quad j = 1, \dots, p \end{aligned} \tag{5}$$

where M_j is the jth column of matrix M, $\epsilon \sim N(0, \sigma^2 I_n)$, and $\eta_j \sim N(0, \sigma_j^2 I_n)$. We assume the regression coefficients are sparse with $s + sparsity(\gamma) + sparsity(\beta) = o(n^d)$, $s + sparsity(B_j) = o(n^d)$ for some $0 < d < 1$. We will introduce two regression-based methods for mediation analysis with high-dimensional exposure and mediators. Each procedure is consist of three steps: (a) parameter estimation by fitting the outcome model and mediator models separately, (b) causal effect testing, and (c) estimating PVM to measure the mediation effect of the selected mediators. We remark that the selected mediators need to have a nonzero estimate of the regression coefficients in the outcome model in step (a) and remain statistically significant in the related tests in step (b).

2.1 Baseline Proposal: Fixed Effect Model for Mediation (MedFix)

2.1.1 Parameter Estimation

One direct solution to (5) is applying a sparse regression technique. For the mediator models, various sparse regression procedures can be applied directly, such as adaptive lasso [62]. For the outcome model, the predictors are heterogeneous with M being continuous and Z being discrete, and it is unclear whether such heterogeneity will cause any excessive errors in the joint sparse selection. Thus we propose to introduce an additional tuning parameter to adjust such heterogeneity. In detail, we minimize the objective

$$\hat{\alpha}^{(fix)}, \hat{\gamma}^{(fix)}, \hat{\beta} = argmin_{(\alpha,\gamma,\beta)} \frac{1}{2} |Y - X\alpha - M\gamma - Z\beta|_2^2 + \rho(\gamma, \beta) \tag{6}$$

for the outcome model where

$$\rho_{\lambda,\theta}(\gamma, \beta) = \lambda(1-\theta) \sum_{j=1}^{p} w_j^{(\gamma)} |\gamma_j| + \lambda\theta \sum_{k=1}^{q} w_k^{(\beta)} |\beta_k|$$

Here $\lambda > 0$ is the overall penalty tuning parameter, $\theta \in (0, 1)$ adjusts the regularization levels on the two data types, and the weight vectors $w^{(\gamma)}$ and $w^{(\beta)}$ are normalized separately to be summed to p and q, respectively. It can be reformulated as adaptive lasso. The mediator models can be solved by directly applying adaptive lasso. The variable weight vectors in all cases are based on the initial lasso estimates of the regression coefficients (See Supplementary Notes for details). We refer to this solution to (5) using existing sparse regression procedure as **Med**iation analysis via **Fix**ed effect model (**MedFix**).

2.1.2 Causal Effect Testing

Next we will perform statistical tests for NIE, NDE, TE and the individual mediator effects NIE_j's. Let $\hat{\gamma}_j$ and s_j^2 be the estimated γ_j and the associated variance, and let $\hat{\beta}^{(nz)}$ and $\hat{B}_j^{(nz)}$ be the subvectors of the non-zero elements of the estimates of β and B_j by MedFix, respectively, and $\Sigma^{(\beta)}$ and $\Sigma^{(B_j)}$ their corresponding estimated variances. The tests for the causal effects can be reformulated as tests on the regression coefficients, for which we will take advantage of the oracle property of adaptive lasso [62]. We remark that a similar testing strategy has been used in the literature [57] where the exposure was univariate, while we focus on high-dimensional exposure in this paper.

Rejecting $H_{0,j} : NIE_j = 0$ requires rejecting both of $H_{0,\gamma_j} : \gamma_j = 0$ and $H_{0,B_j} : B_j = 0$. Let P_{γ_j} and P_{B_j} be their p-values, respectively. Thus we use their maximum $P_{med,j} = max(P_{\gamma_j}, P_{B_j})$ as the p-value for $H_{0,j}$. For MedFix, P_{γ_j} is given by a z-test as $P_{\gamma_j} = 2[1 - \Phi(|\hat{\gamma}_j|/s_j)]$ if $\hat{\gamma}_j \neq 0$, and $P_{\gamma_j} = 1$ otherwise. P_{B_j} is from the χ^2 likelihood ratio test comparing the selected mediator model for M_j and the corresponding reduced model with only X. A stepdown procedure [24] is applied to $P_{med,j}$ for $j \in \{1 \le j \le p; \hat{\gamma}_j \neq 0\}$ to control the false discovery proportion (FDP) in mediator selection.

Testing NDE is equivalently to testing $H_{0,\beta} : \beta = 0$. We calculate its p-value from a χ^2-test as $P_\beta = 1 - F_\beta((\hat{\beta}^{(nz)})^T(\Sigma^{(\beta)})^{-1}\hat{\beta}^{(nz)})$ where F_β is the cdf of a χ^2 distribution with degree of freedom= $|\hat{\beta}|_0$. We reject $H_{0,med} : NIE = 0$ if any of $H_{0,j}$ is rejected. The testing of TE can be done separately without involving mediators based on model $Y = X\alpha + Z\tilde{\beta} + \tilde{\epsilon}$ where $\tilde{\epsilon} \sim N(0, \tilde{\sigma}^2 I_n)$. It is equivalent to test $H_{0,\tilde{\beta}} : \tilde{\beta} = 0$ vs $H_{a,\tilde{\beta}} : \tilde{\beta} \neq 0$, for which many methods [36, 48, 59] are potentially applicable. We choose to use the bootstrap-based RPtest [36].

2.1.3 *PVM* Estimation

Let $\hat{B}$, $\hat{\beta}$ and $\hat{\gamma}$ be the estimates of the coefficients in (5). Since Z is already normalized, (3) can be estimated by $\widehat{VDE}_{fix} = n^{-1} \parallel Z\hat{\beta} \parallel^2$, $\widehat{VIE}_{fix} = n^{-1} \parallel Z(\hat{B}\hat{\gamma}) \parallel^2$, $\widehat{VTE}_{fix} = n^{-1} \parallel Z(\hat{B}\hat{\gamma} + \hat{\beta}) \parallel^2$ and $\widehat{PVM}_{fix} = \widehat{VIE}_{fix}/\widehat{VTE}_{fix}$.

2.2 MedMix: Mixed Effect Model for Mediation

In (5), γ models the association between the mediators and the outcome, and β and B_j's model the association between the exposure and the outcome/mediators, respectively. The sparse assumptions on β and B_j's are crucial for the initial estimation step of MedFix, but not necessary for hypothesis testing, as $H_{0,\beta}$ and H_{0,B_j} are hypotheses on the whole regression coefficient vectors instead of their individual elements. Thus, an alternative strategy is to bypass the sparse estimation of β and B_j completely, and model $Z\beta$ and ZB_j, the aggregate effects of Z in (5) as random effects instead. Consequently, it also reduces the number of candidate parameters in variable selection. The resultant linear mixed models are as the following.

$$\begin{aligned} Y =& X\alpha + M\gamma + u + \epsilon \\ M_j =& XA_j + v_j + \eta_j \quad j = 1, \dots, p \end{aligned} \tag{7}$$

where $\epsilon \sim N(0, \sigma^2 I_n)$, $\eta_j \sim N(0, \sigma_j^2 I_n)$ are noises, and u and v_j for $j = 1, \dots, p$ are all length n random genetic effect vectors such that $u \sim N(0, \tau K(Z))$ and $v_j \sim N(0, \tau_j K(Z))$. Each length n random effect vector is assumed to be independent with the noise vector in the same model. We assume the regression coefficient of the outcome model is sparse with $s + sparsity(\gamma) = o(n^d)$ for some $0 < d < 1$. $K(Z)$ is the $n \times n$ correlation matrix of the random effects u and v_j's. It depends on Z through a known function, and could be as simple as $q^{-1}ZZ^T$. The formulation of (7) is inspired by the quantitative genetics literature where modeling the genotype-trait association as random effects has led to huge success [18, 31]. In quantitative genetics, $K(Z)$ is the marker-based genetic relatedness matrix of the subjects [12, 47]. These random effects are length n vectors, and it is best to interpret the ith element of such a vector as the realized aggregate effect of all q exposures (genetic markers) on the outcome or the mediator for the ith subject.

2.2.1 Parameter Estimation

For model (7), the mediator selection for the outcome model is related to fixed effect selection for high-dimensional linear mixed model, which can be solved by minimizing the following penalized negative log-likelihood

$$Q_\lambda(\alpha, \gamma, \tau, \sigma^2) = \frac{1}{2}(Y - X\alpha - M\gamma)^T V^{-1}(Y - X\alpha - M\gamma) + \frac{1}{2}\log(|V|) + \rho_\lambda(\gamma) \tag{8}$$

where $V = \tau K(Z) + \sigma^2 I_n$, an $n \times n$ full rank matrix. For this potentially non-convex problem, we propose a novel variable selection algorithm which we will discuss in Section 2.3. Together with the mediator equations in (7) fitted using conventional linear mixed model technique (e.g., R package *rrBLUP*, [11]), we term this mediation analysis procedure based on (7) as **Med**iation analysis via **Mix**ed effect model (**MedMix**).

2.2.2 Causal Effect Testing

Causal effect testing for MedMix is similar to the workflow for MedFix, except the following differences. For testing the individual mediator effect, we replace $H_{0,B_j} : B_j = 0$ with $H_{0,\tau_j} : \tau_j = 0$, for which the p value P_{τ_j} is calculated using a SKAT-like score test [50]. Specifically, the score statistic for τ in the outcome model is $Q_\tau = (Y - X\hat{\alpha} - M\hat{\gamma})^T K(Z)(Y - X\hat{\alpha} - M\hat{\gamma})$, and the score statistic for τ_j is $Q_j = (M_j - X\hat{A}_j)^T K(Z)(M_j - X\hat{A}_j)$, where $K(Z)$ is the kernel matrix as used in our paper. Consequently, the p value for $H_{0,j}$ is $P_{med,j} = max(P_{\gamma_j}, P_{\tau_j})$, where P_{γ_j} is from the same asymptotic z test as for MedFix. For testing NDE, we propose to test $H_{0,\tau} : \tau = 0$ with its p-value P_τ given by a similar score test instead of testing $H_{0,\beta} : \beta = 0$. The testing of TE can be done separately without involving mediators based on model $Y = X\alpha + \tilde{u} + \tilde{\epsilon}$ where $\tilde{u} \sim N(0, \tilde{\tau}K(Z))$ and $\tilde{\epsilon} \sim N(0, \tilde{\sigma}^2 I_n)$. It is equivalent to test $H_{0,\tilde{\tau}} : \tilde{\tau} = 0$ vs $H_{a,\tilde{\tau}} : \tilde{\tau} > 0$, for which a simple score test with test statistics $Q_{\tilde{\tau}} = (Y - X\hat{\alpha})^T K(Z)(Y - X\hat{\alpha})$ will suffice.

2.2.3 *PVM* Estimation

Let $\hat{\gamma}$ be the estimated regression coefficients of the outcome model, and $\hat{v}_j = \hat{\tau}_j K(Z)(\hat{\tau}_j K(Z) + \hat{\sigma}_j^2 I_n)^{-1}(M_j - X\hat{A}_j), \hat{u} = \hat{\tau} K(Z)(\hat{\tau} K(Z) + \hat{\sigma}^2 I_n)^{-1}(Y - X\hat{\alpha} - M\hat{\gamma})$ the predicted random effects in (7). Note that in (7), v_j and u replace ZB_j and $Z\beta$ in (5). Define an $n \times p$ matrix $\Psi = (\hat{v}_1, \dots, \hat{v}_p)$, the MedMix-based estimates of the causal effects in (3) are $\widehat{VDE}_{mix} = n^{-1} \parallel \hat{u} \parallel^2$, $\widehat{VIE}_{mix} = n^{-1} \parallel \Psi\hat{\gamma} \parallel^2$, $\widehat{VTE}_{mix} = n^{-1} \parallel \Psi\hat{\gamma} + \hat{u} \parallel^2$ and $\widehat{PVM}_{mix} = \widehat{VIE}_{mix} / \widehat{VTE}_{mix}$.

2.3 Fixed Effect Selection Algorithm for MedMix

In this section, we discuss our proposed novel fixed effect selection algorithm for high-dimensional linear mixed model that was deployed by MedMix for the outcome model in (7).

Fixed effects selection has been studied in the literature [14, 17, 28, 29, 33, 35, 40, 54], most of which were on clustered data. An $L1$ penalized estimation procedure [35] was proposed for the fixed effect selection in high-dimensional linear mixed models. Focusing on the case of clustered data, they implemented their algorithm based on coordinate gradient descent for general random effect covariance structure, and distributed it as the R CRAN package *lmmlasso*. The model considered in this paper is different from the majority of the literature as there are no predefined clusters.

Fixed effect selection for linear mixed model naturally require the estimation of the variance components. Thus, it is also related to the joint estimation of the regression coefficient and the noise level for linear model. The aforementioned paper [35] can be regarded as such a procedure by maximizing the joint likelihood of all parameters. For *iid* noise, [38, 39] suggested that such one-stage maximum likelihood approach may cause bias in noise level estimation, and proposed the scaled lasso that iterates between estimating the regression coefficients and the noise level. Their procedure enjoys joint convexity, and the solution is almost always unique. Motivated by their success, we adopt a similar alternating optimization strategy [3] with a scaled adaptive lasso penalty

$$\rho_\lambda(\gamma) = \frac{\lambda}{\sigma}\sum_{j=1}^{p} w_j|\gamma_j| \tag{9}$$

where w_j's are the weights of the variables scaled to be summed to p.

Consider the objective function (8), with $V = \tau K(Z) + \sigma^2 I_n$ and the scaled adaptive lasso penalty (9). Suppose $K(Z)$ is full rank, and let its eigen decomposition be UDU^T. It follows that $V = UD(\tau, \sigma^2)U^T$ where $D(\tau, \sigma^2) = diag(\tau d_1 + \sigma^2, \dots, \tau d_n + \sigma^2)$ and d_i's are the known eigen values of $K(Z)$. Let $\tilde{Y} = U^T Y$, $\tilde{X} = U^T X$, $\tilde{M} = U^T M$, and $\tilde{y}_i$, $\tilde{x}_i$, and $\tilde{M}_i$ be their i'th row, respectively. (8) can be re-written as

$$Q_{\lambda,w}(\alpha, \gamma, \tau, \sigma^2) = \frac{1}{2}\sum_{i=1}^{n}\frac{(\tilde{y}_i - \tilde{x}_i\alpha - \tilde{M}_i\gamma)^2}{\tau d_i + \sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\log(\tau d_i + \sigma^2) + \frac{\lambda}{\sigma}\sum_{j=1}^{p} w_j|\gamma_j| \tag{10}$$

Reparametrize the above loss function by defining

$$\tilde{\alpha} = \sigma^{-1}\alpha, \quad \tilde{\gamma} = \sigma^{-1}\gamma, \rho = \sigma^{-1}, \quad and \quad \delta = \sigma^{-2}\tau,$$

then (10) becomes

$$\tilde{Q}_{\lambda,w}(\tilde{\alpha}, \tilde{\gamma}, \rho, \delta) = \frac{1}{2}\sum_{i=1}^{n}\frac{(\rho\tilde{y}_i - \tilde{x}_i\tilde{\alpha} - \tilde{M}_i\tilde{\gamma})^2}{\delta d_i + 1} - n\log(\rho) + \lambda\sum_{j=1}^{p} w_j|\tilde{\gamma}_j| + \frac{1}{2}\sum_{i=1}^{n}\log(\delta d_i + 1) \tag{11}$$

It is jointly convex when d_i's are all 0, corresponding to the case with iid noise as for the original scaled lasso. However, when the noise are correlated as in the linear mixed model under consideration, the joint convexity of this objective function depends on the parameters, the data and d_i's in a complicated way. Thus the direct optimization of (11) is not guaranteed to reach a local minimum, even though using a coordinate gradient descent assures the convergence to a stationary point [35].

In fact, (11) is the sum of a concave function (the last term) and a convex function (the sum of the others), for which a concave-convex procedure (CCCP [55]) can be applied. Consider an objective $Q(x) = Q_{vex}(x) + Q_{cav}(x)$ where $Q_{vex}(x)$ and $Q_{cav}(x)$ are a convex and a concave function, respectively. Let $\bigtriangledown Q_{cav}(x) = \frac{\partial Q}{\partial x}(x)$, and $x^{(t)}$ be the current estimate of x, then $Q^{(t)}(x) = Q_{vex}(x) + x \cdot \bigtriangledown Q_{cav}(x^{(t)})$ is a convex tight upper bound of $Q(x)$. CCCP minimizes this upper bound in each iteration, and eventually converges to a local minimum of the original objective. It has been used in the literature for implementing SCAD [13, 23].

For our algorithm, let $(\tilde{\alpha}^{(t)}, \tilde{\gamma}^{(t)}, \rho^{(t)}, \delta^{(t)})$ be a current estimate of $(\tilde{\alpha}, \tilde{\gamma}, \rho, \delta)$, a tight convex upper bound of (11) is

$$U^{(t)}_{\lambda,w}(\tilde{\alpha},\tilde{\gamma},\rho,\delta) = \frac{1}{2}\sum_{i=1}^{n}\frac{(\rho\tilde{y}_i - \tilde{x}_i\tilde{\alpha} - \tilde{M}_i\tilde{\gamma})^2}{\delta d_i + 1} - n\log(\rho) + \lambda\sum_{j=1}^{p} w_j|\tilde{\gamma}_j| + \delta\cdot\frac{1}{2}\sum_{i=1}^{n}\frac{d_i}{\delta^{(t)}d_i + 1} \tag{12}$$

In each iteration, we minimize this upper bound by alternatively updates the regression coefficients and the variance components related parameters. Our algorithm for minimizing (11) is summarized as the following.

1. *Initialization* Fixing λ, and let $(\rho^{(0)}, \delta^{(0)})$ be the initial estimates.
2. *Update* (α, γ): Given $(\rho^{(t)}, \delta^{(t)})$
   (a) *Estimate the variable weights* $\{w_j^{(t+1)}\}_{j=1}^{p}$ *for adaptive lasso using lasso:* See Supplementary Notes for details.
   (b) *Update the fixed effect estimate using adaptive lasso:*
   $$\tilde{\alpha}^{(t+1)}, \tilde{\gamma}^{(t+1)} = argmin_{(\tilde{\alpha},\tilde{\gamma})}\tilde{Q}_{\lambda,w^{(t+1)}}(\tilde{\alpha},\tilde{\gamma},\rho^{(t)},\delta^{(t)})$$
3. *Update* (ρ, δ): Given $(\tilde{\alpha}^{(t+1)}, \tilde{\gamma}^{(t+1)})$
   $$\rho^{(t+1)}, \delta^{(t+1)} = argmin_{(\rho,\delta)}U^{(t)}_{\lambda,w^{(t+1)}}(\tilde{\alpha}^{(t+1)},\tilde{\gamma}^{(t+1)},\rho,\delta)$$
4. *Stopping rule* Repeat steps 2-3 till the objective (11) converges.
5. *Output* Let $(\tilde{\alpha}^{(T)}, \tilde{\gamma}^{(T)}, \rho^{(T)}, \delta^{(T)})$ be the final estimates. The estimates of the original parameters are
   $$\hat{\sigma}^2 = \left(\rho^{(T)}\right)^{-2}, \quad \hat{\tau} = \delta^{(T)}\hat{\sigma}^2, \quad \hat{\alpha} = \hat{\sigma}\tilde{\alpha}^{(T)}, \quad and \quad \hat{\gamma} = \hat{\sigma}\tilde{\gamma}^{(T)}$$

3 Results

Each mediator model of MedFix has one penalty parameter for the adaptive lasso, and they are determined by minimizing BIC. The outcome model of MedFix has two tuning parameters (λ, θ). When $\theta = 0.5$, the overall penalty on the two data types in the outcome model of (5) are the same, and λ can be determined by minimizing *BIC*. We call this version of MedFix MedFix$_{0.5}$. Alternatively, (λ, θ) can be jointly selected by BIC, which we term as MedFix$_{\text{BIC}}$. The outcome model of MedMix has only one tuning parameter λ, and we determine it using *BIC* [35]. The mediator models of MedMix are low-dimensional linear mixed models that do not require penalty parameters, and they are fitted using R package *rrBLUP*. The random effect covariance matrix $K(Z)$ can be modeled in many ways. Since our application is in genetics, we deploy two estimates of the genetic relatedness matrix, the original linear proposal [47] and a shrinkage estimate [12]. MedMix with these two types of covariance matrix for the random effect are referred to as MedMix$_{\text{linear}}$ and MedMix$_{\text{shrink}}$, respectively. We will compare these four protocols in real data analysis and data-driven simulations. Recall that for any method, the selected mediators need to have a non-zero estimate of the regression coefficients in the outcome model in the estimation step and remain statistically significant in the related tests in the testing step. Throughout this paper, we select the mediators by controlling the false discovery proportion (FDP) with $P(\text{FDP} > \gamma) \leq \alpha$ where $\gamma = 0.2$ and $\alpha = 0.1$, unless specified otherwise.

3.1 Evaluation Based on Real Data

3.1.1 A Mouse f2 Cross Data

We analyze a mouse f2 cross data for diabetes study [42, 43, 50]. This dataset includes the genotype captured by SNP array with 2057 markers, microarray-based gene expression with about 40k probes from six tissues, and various clinical phenotypes. The phenotypes we use as outcome are the plasma insulin level at 10 weeks before sacrifice. We are interested in identifying potential causal genes for insulin level. Since insulin is secreted from islet, we use the islet gene expression as the candidate mediator, and genotype as the exposure. We use gender as a covariate in all mediator models and outcome models. There are 491 mice with all three data types. One unique feature of this dataset is that it includes the following three independent measures of plasma insulin at 10 week which we will refer to as **IA**, **IB**, and **IC**:

IA : “INSULIN (ng/mL) 10 wk” from lipomics measure, the primary insulin measurement.
IB : “insulin 10 wk” from lipomics, a reproduced measurement
IC : “Insulin (uIU/mL)” from RBM panel, a measurement using a different technology.

This enables us to compare methods based on the reproducibility of the analysis results across these phenotypes representing the independent measures of biologically identical signal, in the absence of “grand truth” in real data. We use all the genotype markers, and screen the microarray probes by keeping only those that share QTL and have reasonable correlations ($|corr| \geq 0.05$) with the outcome (See Supplementary Notes for details of data preprocessing). We conduct three separate analysis using IA, IB and IC as the phenotype, respectively. We primarily focus on the accuracy and reproducibility in mediator selection and PVM estimation in the main paper (Table 1), and the testing results and estimates of NIE, NDE and TE are in Supplementary Tables 2 and 3.

3.1.2 MedMix Tend to Select More Relevant Mediators

Since there is no well-established “grand truth” of causal genes for insulin, we use the known biological annotations as a proxy for evaluating the causal gene identification accuracy. For each gene, we record whether or not it is involved in a Gene Ontology (GO) term [6], KEGG pathway [21] or Medical Subject Headings (MeSH) term [25] for insulin, diabetes or islet/pancreas function. In Table 1, we present the numbers of the mediators selected for each insulin measure by each method, and

Table 1 Analysis of three independent measurements of insulin level before sacrifice

Model	# of selected (relevant) mediator				PVM			
	IA	IB	IC	Reproducible	IA	IB	IC	Std
MedFix$_{0.5}$	9 (3)	**4 (3)**	9 (1)	19 (3)	0.90	1.00	0.78	**0.11**
MedFix$_{\mathrm{BIC}}$	8 (3)	8 (3)	8 (1)	**18 (6)**	0.89	0.53	1.00	0.25
MedMix$_{\mathrm{linear}}$	8 (3)	11 (3)	14 (2)	26 (5)	0.53	0.65	0.82	0.15
MedMix$_{\mathrm{shrink}}$	**8 (4)**	12 (4)	**13 (3)**	24 (7)	0.53	0.66	0.76	**0.11**

For the columns of # of selected mediator genes, the numbers in parenthesis are the number of genes known to be relevant to pancreas function and/or diabetes determined by their affiliation to the relevant GO/KEGG/MeSH terms. In each column of IA, IB and IC, the method with the highest proportion of relevant genes is in boldface. The column Reproducible presents the number of mediators selected by at least two of the three phenotypes, and the total number of genes selected for any phenotype. The method with the highest such ratio is boldfaced. We present in the column Std the standard deviation of the estimated PVM's across three phenotypes for each method, and boldface the smallest ones

how many of them belong to a relevant term. We remark that this is not an enrichment analysis, and it is only based on their presence or absence in relevant terms. We find that MedMix$_{\mathrm{shrink}}$ has the highest proportion of selected mediators with such annotations for phenotypes IA and IC, MedFix$_{0.5}$ is the highest for IB.

3.1.3 MedMix Yield Reproducible Mediators and Estimates of Overall Mediation Effect Size

Since the three phenotypes measure the same biological signal, a good mediator selection procedure should output similar genes in these analyses. We refer to a mediator as "reproducible" if it is selected in at least two of the three analysis. We find that MedFix$_{\mathrm{BIC}}$ performs the best in terms of the ratio of the reproducible mediators and the mediator selected in any analysis, MedMix$_{\mathrm{shrink}}$ scores the close second (Table 1). We also calculate the standard deviation of the estimated PVM's across the three phenotypes for each method, and find that MedFix$_{\mathrm{BIC}}$ yields very large variation, MedFix$_{0.5}$ and MedMix$_{\mathrm{shrink}}$ perform the best, while *MedMix*$_{\mathrm{linear}}$ is in between towards the lower end (Table 1). Overall, we conclude that MedMix$_{\mathrm{shrink}}$ method generates reasonably reproducible results in the sense that it does not lead to very low reproducible ratio in gene selection, nor very high variation in PVM estimation.

3.1.4 MedMix and MedFix can Differentiate Causal Mediation Effect and Pleiotropy, but not Reverse Causation

It is known that insulin is secreted by islet. Thus the islet gene expressions are its most direct transcriptomic mediators. If the gene expressions from another tissue are used as the candidate mediator, the causal assumptions for mediation analysis would be violated. Any mediator genes selected in such analysis is most likely due to pleiotropy, spurious correlation or causal effect with reverse direction. It is difficult, if not impossible, for any quantitative model to automatically investigate the scientific appropriateness of the mediator candidates. Nevertheless, it is reasonable to expect a much smaller estimated PVM when irrelevant mediator candidates are used. We run MedFix and MedMix using gene expressions from each of the other five tissues (adipose, gastroc, hypothalamus, kidney and liver) as the candidate mediators, and calculate the corresponding PVM. If the output PVM from an irrelevant tissue is equal to or larger than the corresponding islet PVM, we label this case as a “failure” of separating the true and spurious causal mediation effects. We find that all methods output smaller PVMs in most cases when an irrelevant tissue is used as the potential mediators (Table 2). Out of the 15 analysis using an irrelevant tissue, MedMix$_{\text{linear}}$ only failed once, MedFix$_{0.5}$ and MedMix$_{\text{shrink}}$ fail twice, while MedFix$_{\text{BIC}}$ failed four times. Thus, MedMix methods perform better than MedFix methods. Most failures are for adipose and gastroc. It makes biological sense, because plasma insulin directly and indirectly (through glucose) regulates the expressions of a wide range of genes in adipose and gastroc [9, 10]. It reminds us that mediation analysis cannot distinguish the true causal mediation and the causal effect with reverse direction, which is a fundamental limitation of mediation models of all kinds.

3.2 Evaluation Based on Simulations

We simulate data using a pipeline based on the above f2 mouse data analysis with IA as the outcome. The main goal of our simulations is investigating the accuracy of mediator selection of MedFix and MedMix, and their robustness against model misspecification.

Table 2 Estimated PVM using gene expressions from various tissues

Outcome IA	Islet	Adipose	Gastroc	Hypothalamus	Kidney	Liver
MedFix$_{0.5}$	0.90	**1.00**	0.44	0.52	0.46	0.17
MedFix$_{\text{BIC}}$	0.89	0.62	**1.00**	0.62	0.24	0.22
MedMix$_{\text{linear}}$	0.53	**0.57**	0.30	0.18	0.26	0.07
MedMix$_{\text{shrink}}$	0.53	**0.85**	0.29	0.22	0.25	0.24
Outcome IB	Islet	Adipose	Gastroc	Hypothalamus	Kidney	Liver
MedFix$_{0.5}$	1.00	**1.00**	0.51	0.21	0.14	0.63
MedFix$_{\text{BIC}}$	0.53	**0.57**	**0.78**	0.07	0.22	**1.00**
MedMix$_{\text{linear}}$	0.65	0.11	0.29	0.09	0.16	0.11
MedMix$_{\text{shrink}}$	0.66	0.33	0.35	0.17	0.16	0.09
Outcome IC	Islet	Adipose	Gastroc	Hypothalamus	Kidney	Liver
MedFix$_{0.5}$	0.78	0.51	0.37	0.12	0.30	0.20
MedFix$_{\text{BIC}}$	1.00	0.22	0.08	0.11	0.27	0.27
MedMix$_{\text{linear}}$	0.82	0.45	0.29	0.04	0.13	0.13
MedMix$_{\text{shrink}}$	0.76	0.40	**1.00**	0.07	0.13	0.11

Only islet gene expressions are expected to directly regulate insulin. The cases where the reported PVM is larger than the PVM from the islet case are boldfaced

3.2.1 Real Data-Driven Simulation Model

We design a data-driven simulation model using the genotype, preprocessed islet gene expression and gender from the f2 mouse dataset as basis. We simulate the outcome Y and the mediators $\widetilde{M}$ using a hybrid of (5) and (7) controlled by a simulation parameter $\phi \in [0, 1]$. When $\phi = 0, 1$, (5) and (7) are the true models, respectively. When $\phi \in (0, 1)$, both MedMix and MedFix are misspecified. Therefore, this simulation model provides a fair framework for evaluating methods with different model setups, and a common platform for investigating their robustness against model misspecification. The other two simulation parameters that we investigate are (h, g), the strength of the mediation effect and the direct effect, respectively. In our simulations, we assume there are 15 true mediators, and 15 fake mediators with mediator-outcome link but without exposure-mediator link. The other ≈11 k candidate mediators have no mediator-outcome link, regardless whether it is controlled by the exposure (See Supplementary Notes for details of the data-driven simulation model, including the mathematical definitions of ϕ, h, g). We consider all 18 combinations of $g = 1, 2$, $h = 1, 2, 4$ and $\phi = 0, 0.5, 1$. For each scenario, we repeat the simulation for 100 replicates.

3.2.2 MedMix is More Accurate in Mediator Selection

We compare the mediator selection accuracy using the simulated data (Table 3), and find that in most of cases, MedMix models yield three to eight less false positives than MedFix methods. As a price, it may report up to one more false negatives. In

Table 3 The number of false positives and false negatives in mediator selection in simulations

(h, g, ϕ)	(1 , 1 , 0)	(1 , 2 , 0)	(2 , 1 , 0)	(2 , 2 , 0)	(4 , 1 , 0)	(4 , 2 , 0)
MedFix$_{0.5}$	4.47,0.02	4.32,0.04	4.78,0.02	4.87,0.00	6.19,0.01	5.14,0.01
MedFix$_{\text{BIC}}$	4.58,0.02	4.29,0.04	4.76,0.02	4.65,0.00	5.63,0.01	5.30,0.01
MedMix$_{\text{linear}}$	1.37,0.78	1.14,1.18	0.41,0.76	0.57,0.87	0.25,0.69	0.08,0.77
MedMix$_{\text{shrink}}$	1.13,0.80	0.98,1.29	0.36,0.76	0.41,0.89	0.41,0.70	0.07,0.80
(h, g, ϕ)	(1 , 1 , 0.5)	(1 , 2 , 0.5)	(2 , 1 , 0.5)	(2 , 2 , 0.5)	(4 , 1 , 0.5)	(4 , 2 , 0.5)
MedFix$_{0.5}$	7.06,0.23	7.42,1.66	7.92,0.08	7.66,0.16	7.12,0.08	8.56,0.06
MedFix$_{\text{BIC}}$	7.00,0.23	6.72,1.60	7.66,0.08	7.45,0.14	6.23,0.08	7.68,0.06
MedMix$_{\text{linear}}$	1.28,1.17	1.35,1.50	0.42,1.17	0.41,1.25	0.64,1.07	0.15,1.24
MedMix$_{\text{shrink}}$	1.31,1.15	1.03,1.50	0.39,1.19	0.32,1.22	0.63,1.07	0.10,1.26
(h, g, ϕ)	(1 , 1 , 1)	(1 , 2 , 1)	(2 , 1 , 1)	(2 , 2 , 1)	(4 , 1 , 1)	(4 , 2 , 1)
MedFix$_{0.5}$	8.23,1.20	6.24,5.73	7.90,0.18	8.05,1.00	8.05,0.14	7.28,0.22
MedFix$_{\text{BIC}}$	7.17,1.14	6.79,5.61	7.15,0.18	7.00,1.06	6.76,0.14	7.01,0.22
MedMix$_{\text{linear}}$	1.63,1.41	1.55,2.48	0.29,1.21	0.63,1.40	0.18,1.16	0.12,1.26
MedMix$_{\text{shrink}}$	1.35,1.44	1.26,2.55	0.25,1.20	0.57,1.39	0.26,1.17	0.15,1.27

There are 15 true mediators in all simulation settings terms of Hamming errors (the sum of false positives and false negatives), it is easy to conclude that MedMix is always superior.

3.2.3 MedMix is More Robust Against Model Misspecification

Another interesting observation in Table 3 is how each method performs when their model assumptions are violated. When MedFix model becomes misspecified as ϕ goes from zero to non-zero, an immediate sharp increase in errors is observed. In contrast, MedMix does not suffer under misspecified model as ϕ decreases from one to below one.

3.2.4 MedMix Provides Valid Control of False Discovery Proportion in Mediator Selection

We further compare MedMix and MedFix in terms of the false discovery proportion (FDP) control and the power for mediator selection. In our simulation study, the validity of FDP control is evaluated by the proportion of simulation replicates with $FDP > \gamma$, and the power is estimated by the proportion of true mediators being discovered after the statistical tests. Similar evaluation strategies have been used in the literature [8, 57]. We found that MedMix controls FDP at the nominal level, regardless whether its model is misspecified, while MedFix fails (Table 4). And this benefit outweighs its mild loss in power. It provides another reason for

Table 4 The empirical level of FDP control and power (in parentheses) at the nominal level $P(FDP > 0.2) < 0.1$ in mediator selection in simulations

(h, g, ϕ)	(1 , 1 , 0)	(1 , 2 , 0)	(2 , 1 , 0)	(2 , 2 , 0)	(4 , 1 , 0)	(4 , 2 , 0)
MedFix$_{0.5}$	0.54 (1.00)	0.59 (1.00)	0.59 (1.00)	0.61 (1.00)	0.71 (1.00)	0.61 (1.00)
MedFix$_{\text{BIC}}$	0.59 (1.00)	0.54 (1.00)	0.62 (1.00)	0.57 (1.00)	0.67 (1.00)	0.63 (1.00)
MedMix$_{\text{linear}}$	0.07 (0.95)	0.03 (0.92)	0.00 (0.95)	0.02 (0.94)	0.00 (0.95)	0.00 (0.95)
MedMix$_{\text{shrink}}$	0.04 (0.95)	0.05 (0.91)	0.00 (0.95)	0.00 (0.94)	0.02 (0.95)	0.00 (0.95)
(h, g, ϕ)	(1 , 1 , 0.5)	(1 , 2 , 0.5)	(2 , 1 , 0.5)	(2 , 2 , 0.5)	(4 , 1 , 0.5)	(4 , 2 , 0.5)
MedFix$_{0.5}$	0.82 (0.98)	0.90 ((0.89)	0.93 (0.99)	0.87 (0.99)	0.82 (0.99)	0.89 (1.00)
MedFix$_{\text{BIC}}$	0.83 (0.98)	0.79 (0.89)	0.83 (0.99)	0.87 (0.99)	0.74 (0.99)	0.86 (1.00)
MedMix$_{\text{linear}}$	0.07 (0.92)	0.07 (0.90)	0.00 (0.92)	0.01 (0.92)	0.02 (0.93)	0.00 (0.92)
MedMix$_{\text{shrink}}$	0.07 (0.92)	0.03 (0.90)	0.00 (0.92)	0.00 (0.92)	0.03 (0.93)	0.00 (0.92)
(h, g, ϕ)	(1 , 1 , 1)	(1 , 2 , 1)	(2 , 1 , 1)	(2 , 2 , 1)	(4 , 1 , 1)	(4 , 2 , 1)
MedFix$_{0.5}$	0.92 (0.92)	0.94 (0.62)	0.87 (0.99)	0.89 (0.93)	0.92 (0.99)	0.84 (0.99)
MedFix$_{\text{BIC}}$	0.83 (0.92)	0.91 (0.63)	0.78 (0.99)	0.87 (0.93)	0.77 (0.99)	0.74 (0.99)
MedMix$_{\text{linear}}$	0.09 (0.91)	0.14 (0.83)	0.01 (0.92)	0.02 (0.91)	0.00 (0.92)	0.01 (0.92)
MedMix$_{\text{shrink}}$	0.04 (0.90)	0.10 (0.83)	0.01 (0.92)	0.01 (0.91)	0.01 (0.92)	0.00 (0.92)

There are 15 true mediators in all simulation settings our advocacy of the use of mixed model framework for high-dimensional mediator selection.

3.2.5 MedMix is More Accurate in PVM Estimation

In Supplementary Table 4, we present $\sqrt{MSE}$ in PVM estimation. We find that MedMix provides smaller MSE in estimating PVM in most of the cases, and are comparable to MedFix in the rest.

3.2.6 The Fix Effect Selection Algorithm in MedMix is on Par or Better than Adaptive Lasso in Variable Selection

To investigate the consistency of the fix effect selection algorithm in MedMix, we evaluate its performance in terms of the L2 loss in estimating γ. We remark that this estimated γ is not further truncated based on the subsequent testing step in MedMix, because the primary goal of this analysis is the consistency of the fixed effect selection algorithm in Sect. 2.3 itself. In Supplementary Table 5, we find that MedMix outperforms MedFix in most of the settings, except when $\phi = 0$, the two perform similarly. In these cases, MedFix is the true model, and MedMix is misspecified. Since MedFix is an application of adaptive lasso, we have shown that the proposed fix effect selection algorithm in MedMix, as a variable selection algorithm, is on par of adaptive lasso when the simulation model is in favor of adaptive lasso, and superior to it in the other settings.

3.2.7 MedMix is Robust Against the Perturbations of Initial Estimates

MedMix requires initial estimates of $(\rho^{(0)}, \delta^{(0)})$. This is equivalent to providing initial estimates of the variance components $(\tau^{(0)}, (\sigma^{(0)})^2)$, where $\rho^{(0)} = 1/\sigma^{(0)}$ and $\delta^{(0)} = \tau^{(0)}(\sigma^{(0)})^{-2}$. The default of MedMix is $(\tau^{(0)}, (\sigma^{(0)})^2) = (0.5 \cdot var(Y), 0.5 \cdot var(Y))$. To investigate the sensitivity of MedMix on the initial estimates, we repeat the analysis of Sects. 3.2.5 and 3.2.6 with perturbed initial values. The initial estimates used are $(\tau^{(0)}, (\sigma^{(0)})^2) = (c \cdot var(Y), d \cdot var(Y))$ where $(c, d) \in \{(0.1, 0.1), (0.5, 0.5), (0.9, 0.9), (0.1, 0.9), (0.9, 0.1)\}$. The errors of MedMix with these initial estimates are normalized by (divided by) the performance at the default setting ($(c, d) = (0.5, 0.5)$), and presented in Supplementary Tables 6-7. In most cases, we find that the variations in performance are acceptable (ranging between 0.8 to 1.2).

4 Discussion

In this paper, we study the mediation analysis with high-dimensional candidate mediator and high-dimensional exposure, and its application to causal gene detection. We develop two procedures for mediator selection and their associated causal effect tests and mediation effect measure. MedFix is a "baseline" method that applies adaptive lasso with an additional tuning parameter for the outcome model balancing the penalties on the exposure and the mediator variables. MedMix is based on high-dimensional linear mixed models, for which we also develop a new fixed effect selection algorithm for the outcome model. In the absence of rigorous theoretical analysis, the proposed methods perform well empirically. In our data-driven simulation studies, we show that MedMix lead to more accurate mediator selection, and are more robust against model misspecification. We apply these methods to causal gene identification from an f2 mouse dataset for diabetes study, and find that MedMix is more likely to select higher proportions of relevant genes, and yields reasonably reproducible results in mediator selection and PVM estimation.

In the motivating genetic studies, the exposures are discrete, and the mediators and the outcome are continuous with gaussian noise. There is no technical barriers to apply the proposed methods to data with continuous exposure. On the other hand, significant extensions are needed to apply the proposed framework to the cases with discrete or heavy tail mediators or outcomes, or allowing exposure-mediator interaction. We focus on mediator selection and the evaluation of the joint mediation effect of the selected mediators in this paper, and do not discuss the roles of the individual mediators. When the mediators may affect each other causally, quantifying their effects requires stronger assumptions [7], and different models under these assumptions may be needed for detailed studies on the mediation mechanism. Yet the proposed method, along with the existing methods with similar causal assumptions [16, 57] remain useful in quantifying the marginal contribution of each mediator to the overall mediation effect, even though their power may become lower in detecting the signals related to antagonistic epistatic interactions [16]. One potential extension of the proposed work along this direction is to combine the linear mixed model considered here with DAG-based mediation (e.g., [4]). We leave out the important issue of post-selection inference of this paper. This is because existing works in post-selection inference [26, 41, 49, 51, 56] cannot be directly applied to our mediation problem, or perform poorly in our exploratory simulation studies (results not shown), and a more comprehensive investigation on this topic is beyond the scope of this paper, even for a simpler mediation model setup. Another future work we will pursue is to extend the current framework to the joint analysis of multiple outcomes. In the context of causal gene identification, such high-dimensional mediation framework for multiple phenotype will dissect the pleiotropy of the phenotyeps, and lead to refined phenotypical causal networks with genetic basis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12561-021-09328-0.

Acknowledgements We thank Dr. Alan Attie for easy access to the data and insightful discussions on the biology of diabetes, and the Holland Computing Center (HCC) at University of Nebraska Lincoln for computation resources and technical supports. We also thank the editor, associate editor and the two anonymous reviewers for their valuable suggestions and comments.

Funding This work was supported by NSF ABI (Award# DBI-1564621), NSF EPSCoR (RII) Track II (Award# OIA-1736192), NU Collaborative System Science Seed Grant to QZ when he was faculty in Department of Statistics at University of Nebraska Lincoln.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Barfield R, Shen J, Just AC, Vokonas PS, Schwartz J, Baccarelli AA, VanderWeele TJ, Lin X (2017) Testing for the indirect effect under the null for genome-wide mediation analyses. Genet Epidemiol 41(8):824–833
2. Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51(6):1173
3. Bezdek JC, Hathaway RJ (2003) Convergence of alternating optimization. Neural Parallel Sci Comput 11(4):351–368
4. Chakrabortty A, Nandy P, Li H (2018) Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. arXiv preprint arXiv:1809.10652
5. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA (2017) High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics 19(2):121–136
6. Consortium GO (2004) The gene ontology (go) database and informatics resource. Nucleic Acids Res 32(1):D258–D261
7. Daniel R, De Stavola B, Cousens S, Vansteelandt S (2015) Causal mediation analysis with multiple mediators. Biometrics 71(1):1–14
8. Dezeure R, Bühlmann P, Meier L, Meinshausen N (2015) High-dimensional inference: Confidence intervals, p-values and r-software hdi. Stat Sci 533–558
9. Ducluzeau PH, Perretti N, Laville M, Andreelli F, Vega N, Riou JP, Vidal H (2001) Regulation by insulin of gene expression in human skeletal muscle and adipose tissue: evidence for specific defects in type 2 diabetes. Diabetes 50(5):1134–1142
10. Elbein SC, Kern PA, Rasouli N, Yao-Borengasser A, Sharma NK, Das SK (2011) Global gene expression profiles of subcutaneous adipose and muscle from glucose-tolerant, insulin-sensitive, and insulin-resistant individuals matched for bmi. Diabetes 60(3):1019–1029
11. Endelman JB (2011) Ridge regression and other kernels for genomic selection with r package rrblup. Plant Genome 4(3):250–255
12. Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. Genes Genomes Genet 2(11):1405–1413
13. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360
14. Fan Y, Li R (2012) Variable selection in linear mixed effects models. Ann Stat 40(4):2043
15. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ et al (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47(9):1091
16. Gao Y, Yang H, Fang R, Zhang Y, Goode EL, Cui Y (2019) Testing mediation effects in high-dimensional epigenetic studies. Front Genet 10:1195
17. Ghosh A, Thoresen M (2018) Non-concave penalization in linear mixed-effect models and regularized selection of fixed effects. AStA Adv Stat Anal 102(2):179–210
18. Hayes BJ, Bowman PJ, Chamberlain A, Goddard M (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92(2):433–443
19. Huang YT, Pan WC (2016) Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics 72(2):402–413
20. Huang YT, VanderWeele TJ, Lin X (2014) Joint analysis of snp and gene expression data in genetic association studies of complex diseases. Ann Appl Stat 8(1):352
21. Kanehisa M, Goto S (2000) Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
22. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet 10(10):e1004722

23. Kim Y, Choi H, Oh HS (2008) Smoothly clipped absolute deviation on high dimensions. J Am Stat Assoc 103(484):1665–1673
24. Lehmann E, Romano JP (2005) Generalizations of the familywise error rate. Ann Stat 33(3):1138–1154
25. Lipscomb CE (2000) Medical subject headings (mesh). Bull Med Libr Assoc 88(3):265
26. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014) A significance test for the lasso. Ann Stat 42(2):413
27. MacKinnon D (2012) Introduction to statistical mediation analysis. Routledge, London
28. Müller S, Scealy JL, Welsh AH et al (2013) Model selection in linear mixed models. Stat Sci 28(2):135–167
29. Pan J, Shang J (2018) Adaptive lasso for linear mixed model selection via profile log-likelihood. Commun Stat 47(8):1882–1900
30. Pearl J (2001) Proceedings of the seventeenth conference on uncertainty in artificial intelligence. Direct and indirect effects
31. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44(2):217
32. Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 143–155
33. Rohart F, San Cristobal M, Laurent B (2014) Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. Comput Stat Data Anal 80:209–222
34. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688
35. Schelldorfer J, Bühlmann P, De Geer SV (2011) Estimation for high-dimensional linear mixed-effects models using 1-penalization. Scand J Stat 38(2):197–214
36. Shah RD, Bühlmann P (2018) Goodness-of-fit tests for high dimensional linear models. J R Stat Soc Ser B 80(1):113–135
37. Sohn MB, Li H et al (2019) Compositional mediation analysis for microbiome studies. Ann Appl Stat 13(1):661–681
38. Sun T, Zhang CH (2010) Comments on: 1-penalization for mixture regression models. TEST 19(2):270–275
39. Sun T, Zhang CH (2012) Scaled sparse linear regression. Biometrika 99(4):879–898
40. Tan Z, Roche K, Zhou X, Mukherjee S (2018) Scalable algorithms for learning high-dimensional linear mixed models. arXiv preprint arXiv:1803.04431
41. Taylor J, Tibshirani R (2018) Post-selection inference for-penalized likelihood models. Can J Stat 46(1):41–61
42. Tian J, Keller MP, Oler AT, Rabaglia ME, Schueler KL, Stapleton DS, Broman AT, Zhao W, Kendziorski C, Yandell BS et al (2015) Identification of the bile transporter slco1a6 as a candidate gene that broadly affects gene expression in mouse pancreatic islets. Genetics 201(3):1253–1262
43. Tu Z, Keller MP, Zhang C, Rabaglia ME, Greenawalt DM, Yang X, Wang IM, Dai H, Bruss MD, Lum PY et al (2012) Integrative analysis of a cross-loci regulation network identifies app as a gene regulating insulin secretion from pancreatic islets. PLoS Genet 8(12):e1003107
44. VanderWeele T (2015) Explanation in causal inference: methods for mediation and interaction. Oxford University Press, Oxford
45. VanderWeele T, Vansteelandt S (2014) Mediation analysis with multiple mediators. Epidemiol Methods 2(1):95–115
46. VanderWeele TJ (2011) Controlled direct and mediated effects: definition, identification and bounds. Scand J Stat 38(3):551–563
47. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423
48. Verzelen N, Villers F (2010) Goodness-of-fit tests for high-dimensional gaussian linear models. Ann Stat 38(2):704–752
49. Wang H, Zhong PS, Cui Y (2018) Empirical likelihood ratio tests for COE cients in high dimensional heteroscedastic linear models
50. Wang P, Dawson JA, Keller MP, Yandell BS, Thornberry NA, Zhang BB, Wang IM, Schadt EE, Attie AD, Kendziorski C (2011) A model selection approach for expression quantitative trait loci (eqtl) mapping. Genetics 187(2):611–621
51. Wasserman L, Roeder K (2009) High dimensional variable selection. Ann Stat 37(5A):2178

52. Wright S (1918) On the nature of size factors. Genetics 3(4):367
53. Wright S (1934) The method of path coefficients. Ann Math Stat 5(3):161–215
54. Xu P, Wang T, Zhu H, Zhu L (2015) Double penalized h-likelihood for selection of fixed and random effects in mixed effects models. Stat Biosci 7(1):108–128
55. Yuille AL, Rangarajan A (2003) The concave-convex procedure. Neural Comput 15(4):915–936
56. Zhang CH, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B 76(1):217–242
57. 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
58. Zhao Y, Lindquist MA, Caffo BS (2018) Sparse principal component based high-dimensional mediation analysis. arXiv preprint arXiv:1806.06118
59. Zhong PS, Chen SX (2011) Tests for high-dimensional regression coefficients with factorial designs. J Am Stat Assoc 106(493):260–274
60. Zhou RR, Wang L, Zhao SD (2020) Estimation and inference for the indirect effect in high-dimensional linear mediation models. Biometrika 107(3):573–589
61. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM et al (2016) Integration of summary data from GWAS and EQTL studies predicts complex trait gene targets. Nat Genet 48(5):481
62. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429