High Dimensional Mediation Analysis with Applications to Causal Gene Identification

Qi Zhang*
Department of Statistics, University of Nebraska Lincoln, NE 68583, USA
Email: qi.zhang@unl.edu

December 14, 2018

Abstract

Mediation analysis has been a popular framework for elucidating the mediating mechanism of the exposure effect on the outcome. 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 and high dimensional mediator, and introduce two procedures for mediator selection, MedFix and MedMix. MedFix is our new application of adaptive lasso with one additional tuning parameter. MedMix is a novel mediation model based on high dimensional linear mixed model, for which we also develop a new variable selection algorithm. Our study is motivated by the causal gene identification problem, 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 in extensive real data driven simulations, and apply them to causal gene identification from a mouse f2 dataset for diabetes study. We show that the mixed model based approach leads to higher accuracy in mediator selection and mediation effect size estimation, and is more reproducible across independent measurements of the response and more robust against model misspecification.

1 Introduction

Mediation analysis is a type of causal inference investigating the effect of certain exposures (Z) on an outcome (Y, Figure 1). 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). Each of the outcome and the mediators could be also influenced by some observed confounders (X). But these confounders are assumed to have no effect on the exposure assignments. Mediation analysis was originated in social science (Baron and Kenny, 1986; MacKinnon, 2012), and has been widely used in many applications.

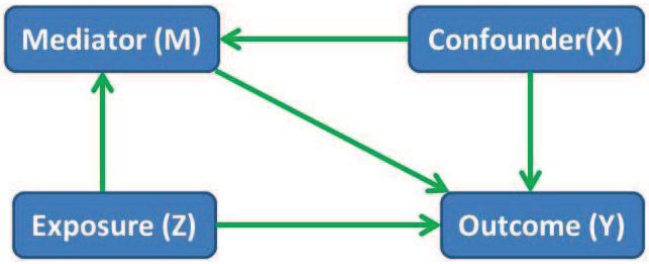

Figure 1: Diagram of the general mediation analysis

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 (Baron and Kenny, 1986; MacKinnon, 2012; VanderWeele, 2011, 2015). Mediation models with multiple mediators or multiple exposures have also been studied. For example, VanderWeele and Vansteelandt (2014) proposed two analytic approaches for multiple mediators. Huang *et al.* (2014) 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. Huang and Pan (2016) proposed a significance test for the effect of high dimensional mediator. Chén *et al.* (2017) proposed direction of mediation, a PCA-like concept for estimating the mediation effects, and applied it to brain imaging studies. In the context of microbiome studies, Sohn and Li (2017) proposed a sparse model for the estimation and testing for the mediation effects. Zhao and Luo (2016) proposed a two-layer SEM framwork for mediation analysis with correlated additive errors, and applied it to fMRI studies. The focus of this paper, however, is the scenario with high dimensional mediator and high dimensional exposure, which is different from all above.

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 (Kichaev *et al.*, 2014) or the *causal gene* at transcriptomic level (Huang and Pan, 2016; Gamazon *et al.*, 2015). 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 and Mendelian randomization methods have been applied to this problem, most of which modeled one gene at a time (Zhu *et al.*, 2016; Gusev *et al.*, 2016; Barfield *et al.*, 2017), or only considered univariate exposure (Huang and Pan, 2016; Sohn and Li, 2017).

In this paper, we study the mediation analysis with high dimensional mediator and high dimensional exposure. In detail, let the potential mediator matrix M be of the size $n \times p$, and the exposure matrix Z be $n \times q$ with potentially $p \gg n$ and $q \gg n$ in the mediation diagram (Figure 1). We are interested in identifying a small subset of the columns of M that mediate the effect of Z on Y. We propose two regularized solutions for this high dimensional mediator selection problem. The first one is a new application of adaptive lasso (Zou, 2006) 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.

The rest of the paper is organized as follows. In Section 2, we start with the model setup for the mediation analysis with high dimensional mediator and high dimensional exposure, proceed to our two proposed procedures for this problem, including a new variable selection algorithm for high dimensional linear mixed model, and finish at our proposed measure of the mediation effect size. In Section 3, we compare the two frameworks in our extensive data driven simulation studies. In Section 4, we apply the proposed methods to causal gene identification from an f2 mouse dataset for diabetes study. We conclude in Section 5 with discussions on our future work.

2 Model Setup and Methods

Let Y be an $n \times 1$ 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. We assume that the columns of M and Z are centered and standardized, and potentially $q, p \gg n$, but presumably $s < n$. In genomic studies, 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, the high dimensional mediation model could be written as the following

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

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 in the outcome model, i.e., identifying the non-zero elements of the sparse vector γ. This is the primary goal of this paper.

2.1 The Fixed Effect Model Based Procedure

One naive solution is applying a sparse regression technique to the outcome model of (1) for the joint sparse estimation of (γ, β), and minimize the following objective function

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

While any sparse regression procedure can be applied, we utilize adaptive lasso for our implementation. Specifically, we use the penalty

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

for the outcome model. In 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 bias in the joint sparse selection. Thus we introduce another tuning parameter $\theta \in (0, 1)$ in the above penalty term to adjust the regularization levels on the two data types. The adaptive lasso weight vectors $w^{(\gamma)}$ and $w^{(\beta)}$ are normalized separately to be summed to p and q, respectively. The mediator models can be solved by directly applying adaptive lasso. We refer to this naive solution to (1) using existing sparse regression procedure as **Med**iation analysis via **Fix**ed effect model (**MedFix**).

2.2 MedMix: the Mixed Effect Model Formulation

In (1), γ models the association between the mediators and the outcome, and the sparse selection of γ is the main focus of this paper. β and B_j's model the association between genotype and the phenotype/transcriptomics, respectively. The sparse assumptions on β and B_j's are crucial for MedFix, but not scientifically necessary for causal gene identification. In genomics, most of these genetic effects are very weak, but may not be exactly zeros. It is unclear whether forcing them to be zero in sparse selection will cause unpredictable biases in mediator selection

Since we are only interested in the variable selection for γ, an alternative strategy is to model the effects of Z in (1) as random effects to reduce the dimension of the parameter space, which lead to the following linear mixed models

$$\begin{aligned} Y &= X\alpha + M\gamma + u + \epsilon \\ M_j &= XA_j + v_j + \eta_j \quad j = 1, \ldots, p \end{aligned} \tag{3}$$

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 $j = 1, \ldots, p$ are all length n random genetic effect vectors such that

$$u \sim N(0, \tau K(Z)) \quad and \quad v_j \sim N(0, \tau_j K(Z)).$$

The formulation of (3) is inspired by the quantitative genetics literature where modeling the genotype-trait association as random effects in linear mixed models has led to huge success (Hayes *et al.*, 2009; Riedelsheimer *et al.*, 2012). In these applications, $K(Z)$, the $n \times n$ covariance matrix for the random effects is the marker-based genetic relatedness matrix of the subjects (VanRaden, 2008; Endelman and Jannink, 2012).

Now the mediator selection for the outcome model is translated into a fixed effect selection problem for a high dimensional linear mixed model. We will solve it by minimizing the following penalized negative log-likelihood

$$Q_\lambda(\alpha, \gamma, \tau, \sigma^2) = \frac{1}{2}(Y - X\alpha - M\gamma)^T V^{-1}(Y - X\alpha - M\gamma) + \frac{1}{2}\log(|V|) + \rho_\lambda(\gamma) \tag{4}$$

where $V = \tau K(Z) + \sigma^2 I_n$. We remark that this optimization problem is potentially non-convex, even if the penalty term is convex. However, it is the sum of a convex and a concave function after reparametrization, for which an alternative optimization algorithm in the spirit of the scaled lasso (Sun and Zhang, 2012) combined with a concave-convex procedure (CCCP, Yuille and Rangarajan 2003) could be applied.

The mediator equations in (3) are conventional linear mixed models, and we solve them using R package *rrBLUP* (Endelman, 2011). Together with the minimizer of (4) to be detailed in the next section, we term this approach based on (3) as **Med**iation analysis via **Mix**ed effect model (**MedMix**).

2.3 Implementation Details

MedFix has two tuning parameters (λ, θ). When $\theta = 0.5$, the penalty on the two data types in the outcome model of (1) are the same, and MedFix is equivalent to the direct application of adaptive lasso, for which λ is 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 $MedFix_{BIC}$.

MedMix has only one tuning parameter λ, and we determine it using BIC. However, 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 proposal (VanRaden, 2008) and a shrinkage estimate (Endelman and Jannink, 2012). MedMix with these two types of covariance matrix for the random effect are referred to as $MedMix_{linear}$ and $MedMix_{shrink}$, respectively.

We will compare these four protocols in simulations and real data analysis.

2.4 Estimating the size of mediation effect

In this section, we study the estimation of the mediation effect size for high dimensional mediator and high dimensional exposure.

Using the counter-factual notations, the total effect (TE), the natural direct effect (NDE) and the natural indirect effect (NIE) when changing the exposure (the genotype vector) from $z^\star$ to z are derived from (1) as the following,

$$TE = (z - z^\star)(B\gamma + \beta) \quad NDE = (z - z^\star)^T \beta \quad and \quad NIE = (z - z^\star)^T B\gamma \tag{5}$$

where z and $z^\star$ are two length q exposure vectors from the population under investigation, and B is an $q \times p$ matrix whose jth column is B_j.

The proportion of the mediation effect is often defined as $PM = NIE/TE$ in the literature (VanderWeele, 2015). For univariate exposure, PM is not a function of the exposure under model (1) as $(z - z^\star)$ in NIE and TE are canceled out when the ratio is calculated. For multivariate exposure, if the 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 (Huang *et al.*, 2014). In our applications, our genotype and gene expression data are for the whole genome. Hence it could be assumed that TE exists, and consequently at least one of NDE and NIE exists. The subsequent question to be answered is how much of TE is mediated, which is an estimation problem. For this problem, the effect size measures in (5) and PM are all functions of $(z, z^\star)$. Such measures could potentially be useful in genomic selection or personalized intervention when a specific exposure combination may be of interest. However, a population-wise 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 (5), let $z^\star$ be an arbitrarily chosen baseline exposure and kept fixed, and z be a random chosen exposure from the population under investigation. We use the variance of TE, NDE and NIE in (5) over the randomness of z as the measure 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

$$VTE = (B\gamma+\beta)^T Var(z)(B\gamma+\beta) \quad VDE = \beta^T Var(z)\beta \quad and \quad VIE = (B\gamma)^T Var(z)(B\gamma) \tag{6}$$

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

$$PVM = VIE/VTE \tag{7}$$

We remark that VTE is not necessarily the sum of VIE and VDE due to the potential correlatedness of the estimated direct and the indirect effects in real data application.

In the context of (1), $Var(z)$ could be estimated using $Z^T Z$ if Z is normalized. Then (6) is reduced to

$$VTE = \| Z(B\gamma + \beta) \|^2 \quad VDE = \| Z\beta \|^2 \quad and \quad VIE = \| Z(B\gamma) \|^2 \tag{8}$$

For the mixed model (3), define the $n \times p$ matrix $\Psi = (v_1, \ldots, v_p)$ as the exposure effects on mediators. The linear mixed model version of the estimates of the causal effects in (6) are

$$VTE = \| \Psi\gamma + u \|^2 \quad VDE = \| u \|^2 \quad and \quad VIE = \| \Psi\gamma \|^2 \tag{9}$$

For either MedFix or MedMix, the effect sizes are estimated by plugging in the parameter estimates from the corresponding models.

When there is no confounders, VTEs defined in (8)-(9) are simply the L2 norm of the predicted outcomes, and VDEs and VIEs are the contributions of the direct effect and the mediators, respectively. Thus the proposed measure of the mediation effect size could be seen as the proportion of the exposure effect explained by the mediator component of the outcome model. This is also related to the statistical literature in deriving appropriate R^2 measures for particular models (Xu, 2003; Gelman and Pardoe, 2006; Jaeger *et al.*, 2017), and the heritability estimation in the quantitative genetics literature (Schweiger *et al.*, 2016). Comparing with these works, our goal is different, and our measures of the effect sizes only depend on the contribution of each component to the predicted values. Our proposed measures are not tailored to any specific underlying model or estimation procedure, and provide a fair standard for mediation effect size estimation and model comparison.

3 Simulations

We perform simulation studies driven by the analysis of a mouse f2 cross dataset (See Section 4 for details of the real dataset). The goals of our simulations are investigating the accuracy of mediator selection and PVM estimation of MedFix and MedMix, and their robustness against model mis-specification.

3.1 Real Data Driven Simulation Model

We simulate the outcome and the realized true mediators using the following hybrid of models (1) and (3).

$$\begin{aligned} Y &= X\alpha + M\gamma + [\phi u + (1-\phi)Z\beta] + \epsilon \\ M_j &= \phi(XA_j^{(mix)} + v_j) + (1-\phi)(XA_j^{(fix)} + ZB_j) + \eta_j \quad for \quad j = 1, \ldots, p \quad s.t., \gamma_j \neq 0 \end{aligned} \tag{10}$$

When $\phi = 0, 1$, (1) and (3) are the true models, respectively. When $\phi \in (0, 1)$, both of them are misspecified. Thus our simulation setting provides a common platform for investigating their model robustness.

In the above simulation model, we use the 491×2057 normalized genotype matrix from the real data as Z, the intercept and sex from the real data as the columns of X, and fixed $\alpha = (1, 0.2)^T$.

For γ, the sparse regression coefficient for the mediators, we set $\gamma = h\gamma_0$ where h a simulation parameter representing the strength of the mediation effect, and γ_0 is a sparse vector with 22 non-zero elements whose coordinates represent the selected islet genes from the real data analysis by any method under comparison using the primary insulin measure (IA) as the outcome. So the true mediators are the same across all simulations. The values of the 22 nonzero elements of γ_0 are $(1, -1, 2, -2, 1, -1, 2, -2, 1, -1, 2, -2, 1, -1, 2, -2, 1, -1, 2, -2, 1, -1)$.

For β, the sparse direct effect regression coefficient in (1), we assume $\beta = g\beta_0$ where g a simulation parameter representing the strength of the direct effect, and β_0 is a sparse vector with 8 non-zero elements whose coordinates are equally spaced across the 2057 markers, and their values are $(1, -1, 2, -2, 1, -1, 2, -2)$.

We then simulate the direct effect satisfying (3) from $u \sim N\left(0, q^{-1} \parallel \beta \parallel^2 ZZ^T\right)$.

The 491×11712 candidate mediator matrix M in (10) has the same entries as in the normalized islet gene expression matrix in the real data, except for the columns corresponding to the 22 true mediator genes. For the true mediators, we simulate M_j based on (10) as the following. We use the predicted gene expression values by MedMix and MedFix from the real data analysis as $XA_j^{(mix)} + v_j$ and $XA_j^{(fix)} + ZB_j$, respectively. We then simulate $\eta_j \sim N(0, \sigma_j^2 I_n)$ where σ_j^2 is the mean squared error in predicting the expression of jth gene in the real data using the predicted values $\tilde{M}_j^{(\phi)} = \phi(XA_j^{(mix)} + v_j) + (1-\phi)(XA_j^{(fix)} + ZB_j)$. This choice of the noise variance is for matching the variation of the simulated gene expression with that in the real data.

Finally, we simulate $\epsilon \sim N(0, I_n)$, and calculate Y as in (10).

There are three simulation parameters (h, g, ϕ), representing the strengths of the mediation effect, direct effect and the mixing weight of (3) in (10). We consider all combinations of $g = 1, 2$, $h = 2, 4$ and $\phi = 0, 0.5, 1$. For each scenario, we repeat the simulation for 20 replicates. The true mediators are the same across all simulations. For each simulation, we define the true PVMs as the PVM calculated using Equation (7) and without the noise term. But they still vary across simulation replicates due to the randomness in simulating u.

3.2 Simulation Results

We first compare the mediator selection performance in Table 1. As expected, all methods perform better when h increases, as it represents the strength of the mediation effect. Since the direct effect is treated as part of the correlated "noise" in the mediator selection process, all methods suffer when g increases.

When $\phi = 0$, the simulation model is the same as the model underlying MedFix (1), and MedMix is mis-specified. However, MedFix and MedMix have comparable mediator selection performance, except probably for $(2, 2, 0)$ where MedFix shows moderate benefit. In contrast, when $\phi = 1$ and the MedMix model (3) is the true model, MedMix clearly lead to less false positives and less false negatives in mediator selection in most settings. When $\phi = 0.5$, neither (1) nor (3) is the true model, and both of MedFix and MedMix are misspecified. In most of these cases, MedMix performs better than MedFix, except for $(4, 1, 0.5)$ where the two methods are comparable. Overall, these results suggest that mediator selection by MedMix is more robust than MedFix against model misspecification, and yields more accurate results than MedFix in a wider range of settings.

Additionally, MedFix_{BIC} makes less errors than $\text{MedFix}_{0.5}$ in all cases, which illustrates the advantage of including the additional tuning parameter θ in MedFix comparing with the direct application of adaptive lasso. When comparing MedMix_{linear} and MedMix_{shrink}, MedMix_{shrink} performs slightly better than or comparable to MedMix_{linear} in most scenarios except for $(2, 1, 1)$, which hints the potential benefit of using the shrinkage estimate of the genetic relatedness matrix. However, it remains unclear whether such choice would lead to universal improvement for MedMix.

We then compare their performance in PVM estimation measured by the absolute error (absolute value of the difference between the true PVM and its estimate, Figure 2), and observe a similar phenomenon. MedMix lead to overall smaller errors than MedFix when $\phi = 0.5, 1$.

4 Analysis of a mouse f2 cross data

We analyze a mouse f2 cross data for diabetes study (Wang *et al.*, 2011; Tu *et al.*, 2012; Tian *et al.*, 2015). This dataset include the genotype, gene expressions from six tissues, and various clinical phenotypes. The phenotypes we focus on as the response variable are the plasma insulin level at 10 weeks right before sacrifice. We are interested in identifying potential causal genes for

Table 1: The number of of false positives and false negatives in mediator selection in simulations. There are 22 true mediators in all simulation settings.

(h, g, ϕ)	(2 , 1 , 0)	(2 , 2 , 0)	(4 , 1 , 0)	(4 , 2 , 0)
$\text{MedFix}_{0.5}$	1.8 , 0	1.3 , 0	0.6 , 0	1.6 , 0
MedFix_{BIC}	0.9 , 0	1.5 , 0	0.2 , 0	0.4 , 0
MedMix_{linear}	0.8 , 0	3 , 0	0.5 , 0	0.4 , 0
MedMix_{shrink}	1 , 0	2 , 0	0.2 , 0	0.8 , 0
(h, g, ϕ)	(2 , 1 , 0.5)	(2 , 2 , 0.5)	(4 , 1 , 0.5)	(4 , 2 , 0.5)
$\text{MedFix}_{0.5}$	3.5 , 0	6 , 0.4	1.2 , 0	2.5 , 0
MedFix_{BIC}	2.8 , 0	4.5 , 0.4	0.8 , 0	1.9 , 0
MedMix_{linear}	1.8 , 0	3.1 , 0	1.1 , 0	0.8 , 0
MedMix_{shrink}	2 , 0	2 , 0	1.2 , 0	0.8 , 0
(h, g, ϕ)	(2 , 1 , 1)	(2 , 2 , 1)	(4 , 1 , 1)	(4 , 2 , 1)
$\text{MedFix}_{0.5}$	5.2 , 0.3	8 , 3	1.8 , 0	5.8 , 0
MedFix_{BIC}	3.8 , 0.4	6.3 , 3.7	2.1 , 0	4.2 , 0
MedMix_{linear}	2.5 , 0	5.2 , 0	1.5 , 0	2.8 , 0
MedMix_{shrink}	3.7 , 0	4.5 , 0	1.4 , 0	1.7 , 0

insulin level. Since insulin is secreted from islet, we use the islet gene expression as the candidate mediator, and genotype as the exposure. 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 reprodcibility of the analysis results across these phenotypes representing the independent measurements of biologically identical signal, in the absence of "grand truth" in real data applications.

4.1 Data preprocessing

Mediation effect implies the existence of the association between the response and the exposure, the mediator and the exposure, and the mediator and the response. We use these assumptions to filter genes that enter our analysis.

We only keep genes that satisfies the following three criteria. (1) Islet expression shares QTL with the phenotype under investigation (lod score$\geq$ 3). (2) Islet expression correlates with the phenotype ($|corr| \geq 0.05$). (3) Islet expression has larger correlation with the phenotype than the expression in the other five tissues. The original microarray data has about 40k probes, and about 11k genes survive the filtering process for each insulin measure.

4.2 Real Data Analysis Results

We first compare the mediator selection results. Since there is no well-established "grand truth" of causal genes for insulin, we use the known biological annotations as a proxy. For each gene, we record whether or not it is involved in a Gene Ontology (GO) term (Consortium, 2004), KEGG pathway (Kanehisa and Goto, 2000) or Medical Subject Headings (MeSH) term (Lipscomb, 2000) for insulin, diabetes or islet/pancreas function. In Table 2, we present the numbers of the mediators selected for each insulin measure by each method, and how many of them belong to a relevant biological annotation term. We remark that this is not an enrichment analysis, and

Table 2: Analysis of three independent measurements of insulin level before sacrifice. 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 affilication to the relevant GO/KEGG/MeSH terms. In each column, the method with the highest proportion of relevant genes is in boldface. The gene selection consistency between two datasets for each method is measured by Jaccard index. For each pair of datasets, the method with the best consistency is in boldface. For PVM, the row with the smallest variability among the three phenotypes is in boldface. We find that $MedMix$ has higher mediator selection accuracy and better reproducibility in mediator selection and PVM estimation.

	# of selected mediator			Mediator consistency			PVM		
Model	IA	IB	IC	(IA,IB)	(IA,IC)	(IB,IC)	IA	IB	IC
MedFix$_{0.5}$	22(9)	**20(11)**	32(8)	0.40	0.20	0.16	1.00	0.82	0.76
MedFix$_{BIC}$	19(8)	15(8)	32(8)	0.42	0.21	0.12	0.65	0.54	0.62
MedMix$_{linear}$	**16(7)**	27(12)	**21(10)**	0.39	**0.28**	0.20	0.71	0.85	0.84
MedMix$_{shrink}$	**16(7)**	18(8)	26(11)	**0.48**	0.24	**0.22**	**0.71**	**0.77**	**0.80**

it is only based on the membership of individual genes to relevant terms. Our results show that $MedMix_{linear}$ has the highest proportion of selected mediators with such annotations for phenotypes IA and IC, $MedFix_{0.5}$ is the highest for IB, and $MedMix_{shrink}$ ties with $MedMix_{linear}$ for IA.

Since the three phenotypes measure the same biological signal, a good mediator selection procedure should output similar genes in these analyses. For each pair of these phenotypes, we calculate the Jaccard index of the selected mediators, and find that MedMix always lead to better variable selection consistency across phenotypes (Table 2). We also find that the Jaccard indices are higher between IA and IB than with IC, potentially due to the aforementioned different technology used in IC measurement. Overall, the mixed model based methods select higher proportions of relevant genes and are more reproducible across independent measures of insulin.

Another aspect of cross phenotype reproducibility lays in PVM estimation. For each method, we calculate the variance of the PVM's for the three phenotypes, and find that $MedMix_{shrink}$ lead to the smallest variation in PVM estimates (Table 2).

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 detect 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 models using the expressions of the preprocessed genes from each of the other five tissues (adipose, gastroc, hypothalamus, kidney and liver) as the candidate mediator, and calculate PVM. If the output PVM from an irrelevant tissue is larger than 50% of the corresponding PVM from islet, we label it as a "failure". We find that for hypothalamus, kidney and liver, only MedFix fails in many cases (Table 3). It suggests that the mixed model based high dimensional mediation models are more likely to distinguish between the relevant tissue and the irrelevant ones. On the other hand, all methods fail for adipose and gastroc. This observation makes biological sense, because plasma insulin directly and indirectly (through glucose) regulates the expressions of a wide range of genes in adpose and gastroc (Ducluzeau *et al.*, 2001; Elbein *et al.*, 2011). It reminds us that mediation cannot determine the direction of the causality, which is a fundamental limitation of mediation analysis framework of all kinds.

Table 3: Multi-tissue comparison of the mediation proportion using the same set of genes as the mediator candidates but the expression from different tissues. Only islet gene expressions are expected to directly regulate insulin. The values in boldface are the cases where the mediation model fails in the sense that the reported PVM is larger than 50% of the PVM from the correct model using islet expressions. We find that MedMix models fail less often than MedFix models.

			IA			
	Islet	Adipose	Gastroc	Hypothalamus	Kidney	Liver
$\text{MedFix}_{0.5}$	1.00	**0.75**	**0.67**	0.29	0.41	0.35
MedFix_{BIC}	0.65	**0.88**	**1.00**	0.26	0.24	0.14
MedMix_{linear}	0.71	**0.68**	**0.38**	0.22	0.32	0.26
MedMix_{shrink}	0.71	**0.68**	**0.38**	0.33	0.35	0.29
			IB			
	Islet	Adipose	Gastroc	Hypothalamus	Kidney	Liver
$\text{MedFix}_{0.5}$	0.82	**0.61**	**0.57**	0.32	0.37	**0.44**
MedFix_{BIC}	0.54	**0.42**	0.26	**0.49**	0.16	**0.53**
MedMix_{linear}	0.85	**0.68**	**0.43**	0.17	0.40	0.41
MedMix_{shrink}	0.77	**0.66**	**0.45**	0.17	0.33	0.29
			IC			
	Islet	Adipose	Gastroc	Hypothalamus	Kidney	Liver
$\text{MedFix}_{0.5}$	0.76	**0.48**	**0.44**	0.24	**0.41**	**0.40**
MedFix_{BIC}	0.62	**0.61**	**0.73**	0.16	**0.36**	**0.34**
MedMix_{linear}	0.84	**0.56**	**0.81**	0.12	0.28	0.31
MedMix_{shrink}	0.80	**0.57**	**0.80**	0.16	0.27	0.31

5 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 a measure of mediation effect size. MedFix is an application of 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 develop a new fixed effect selection algorithm for the outcome model. In our data driven simulation studies, we show that MedMix lead to more accurate mediator selection and PVM estimation in most settings, and is 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 more reproducible results in causal gene identification and PVM estimation.

We focus on the estimation of the joint mediation effect of the selected mediators in this paper, and do not discuss the effects of the individual mediators. Due to the potential mediator-mediator interaction, quantifying such individual effects requires strong assumptions, even when there are only a few mediators, and their order in the causal chain is known (Daniel *et al.*, 2015). Thus we leave it for future research. We do not consider the exposure-exposure interaction, exposure-mediator interaction or nonlinear causal effects neither. Generalizations along this direction could be made potentially by extending (1) to a high dimensional additive model or by modeling $K(Z)$ in (3) with an appropriate kernel function (Morota and Gianola, 2014). Another future work we will pursue is to extend the current framework to the joint analysis of multiple responses. In the context of causal gene identification, such multiple phenotype high dimensional mediation framework will dissect the pleiotropy of the phenotyeps, and lead to refined phenotypical causal networks with genetic basis.

Acknowledgements

We thank Dr. Alan Attie for providing easy access to the data and insightful discussions on the biology of diabetes, and the Holland Computing Center (HCC) at UNL for computation resources and technical supports. This work has been supported by NSF ABI (Award# DBI-1564621), NSF EPSCoR (RII) Track II (Award# OIA-1736192) and NU Collaborative System Science Seed Grant to QZ.

The source code could be made available upon request

References

Barfield, R., Shen, J., Just, A. C., Vokonas, P. S., Schwartz, J., Baccarelli, A. A., VanderWeele, T. J., and Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic epidemiology*, **41**(8), 824–833.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, **51**(6), 1173.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2017). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, **19**(2), 121–136.

Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, **32**(suppl_1), D258–D261.

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

Ducluzeau, P.-H., Perretti, N., Laville, M., Andreelli, F., Vega, N., Riou, J.-P., and 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.

Elbein, S. C., Kern, P. A., Rasouli, N., Yao-Borengasser, A., Sharma, N. K., and Das, S. K. (2011). Global gene expression profiles of subcutaneous adipose and muscle from glucose-tolerant, insulin-sensitive, and insulin-resistant individuals matched for bmi. *Diabetes*, page DB_101270.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *The Plant Genome*, **4**(3), 250–255.

Endelman, J. B. and Jannink, J.-L. (2012). Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics*, **2**(11), 1405–1413.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., *et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, **47**(9), 1091.

Gelman, A. and Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, **48**(2), 241–251.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., *et al.* (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, **48**(3), 245.

Hayes, B. J., Bowman, P. J., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, **92**(2), 433–443.

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.

Huang, Y.-T., VanderWeele, T. J., and Lin, X. (2014). Joint analysis of snp and gene expression data in genetic association studies of complex diseases. *The annals of applied statistics*, **8**(1), 352.

Jaeger, B. C., Edwards, L. J., Das, K., and Sen, P. K. (2017). An r 2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, **44**(6), 1086–1105.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, **10**(10), e1004722.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, **88**(3), 265.

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

Morota, G. and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in genetics*, **5**, 363.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A. E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature genetics*, **44**(2), 217.

Schweiger, R., Kaufman, S., Laaksonen, R., Kleber, M. E., März, W., Eskin, E., Rosset, S., and Halperin, E. (2016). Fast and accurate construction of confidence intervals for heritability. *The American Journal of Human Genetics*, **98**(6), 1181–1192.

Sohn, M. B. and Li, H. (2017). Compositional mediation analysis for microbiome studies. *bioRxiv*, page 149419.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, **99**(4), 879–898.

Tian, J., Keller, M. P., Oler, A. T., Rabaglia, M. E., Schueler, K. L., Stapleton, D. S., Broman, A. T., Zhao, W., Kendziorski, C., Yandell, B. S., *et al.* (2015). Identification of the bile transporter slco1a6 as a candidate gene that broadly affects gene expression in mouse pancreatic islets. *Genetics*, pages genetics–115.

Tu, Z., Keller, M. P., Zhang, C., Rabaglia, M. E., Greenawalt, D. M., Yang, X., Wang, I.-M., Dai, H., Bruss, M. D., Lum, P. Y., *et al.* (2012). Integrative analysis of a cross-loci regulation network identifies app as a gene regulating insulin secretion from pancreatic islets. *PLoS genetics*, **8**(12), e1003107.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

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

VanderWeele, T. J. (2011). Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian Journal of Statistics*, **38**(3), 551–563.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, **91**(11), 4414–4423.

Wang, P., Dawson, J. A., Keller, M. P., Yandell, B. S., Thornberry, N. A., Zhang, B. B., Wang, I.-M., Schadt, E. E., Attie, A. D., and Kendziorski, C. (2011). A model selection approach for expression quantitative trait loci (eqtl) mapping. *Genetics*, **187**(2), 611–621.

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in medicine*, **22**(22), 3527–3541.

Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, **15**(4), 915–936.

Zhao, Y. and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749*.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., *et al.* (2016). Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, **48**(5), 481.

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

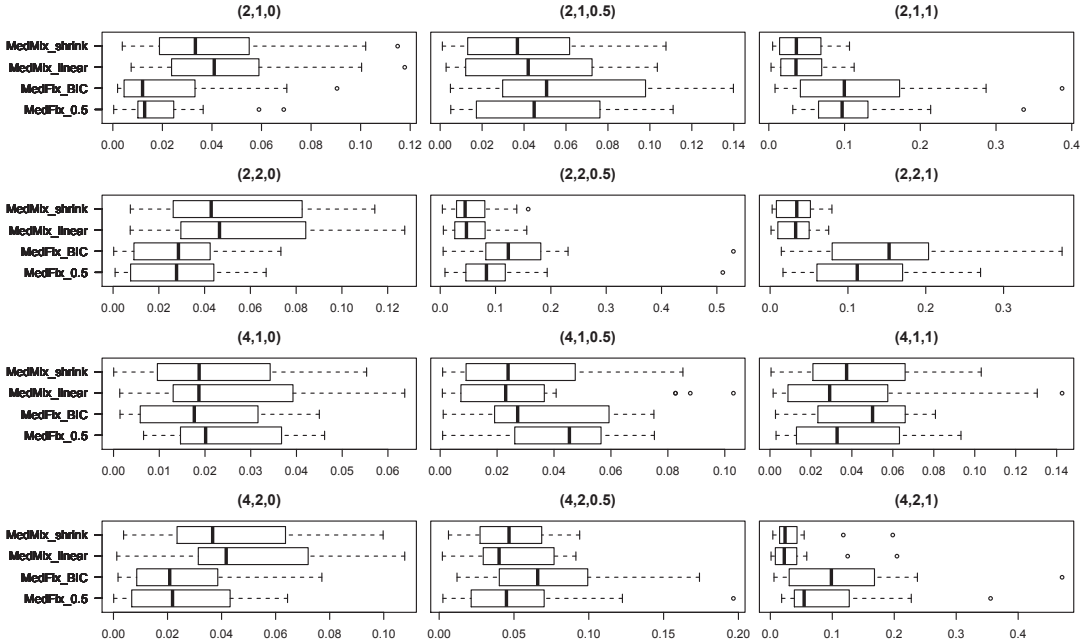

Figure 2: Boxplots of the absolute error in PVM estimation (the absolute value of the difference of the estimate and the true PVM) for all 12 combinations of (h, g, ϕ).