High-dimensional Multivariate Mediation with Application to Neuroimaging Data

Oliver Y. Chén[1], Ciprian M. Crainiceanu[1], Elizabeth L. Ogburn[1], Brian S. Caffo[1], Tor D. Wager[2], Martin A. Lindquist[1]

[1] Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

[2] Department of Psychology and Neuroscience
University of Colorado Boulder

Abstract

Mediation analysis has become an important tool in the behavioral sciences for investigating the role of intermediate variables that lie in the path between a randomized treatment and an outcome variable. The influence of the intermediate variable on the outcome is often explored using structural equation models (SEMs), with model coefficients interpreted as possible effects. While there has been significant research on the topic in recent years, little work has been done on mediation analysis when the intermediate variable (mediator) is a high-dimensional vector. In this work we present a new method for exploratory mediation analysis in this setting called the directions of mediation (DMs). The first DM is defined as the linear combination of the elements of a high-dimensional vector of potential mediators that maximizes the likelihood of the SEM. The subsequent DMs are defined as linear combinations of the elements of the high-dimensional vector that are orthonormal to the previous DMs and maximize the likelihood of the SEM. We provide an estimation algorithm and establish the asymptotic properties of the obtained estimators. This method is well suited for cases when many potential mediators are measured. Examples of high-dimensional potential mediators are brain images composed of hundreds of thousands of voxels, genetic variation measured at millions of SNPs, or vectors of thousands of variables in large-scale epidemiological studies. We demonstrate the method using a functional magnetic resonance imaging (fMRI) study of thermal pain where we are interested in determining which brain locations mediate the relationship between the application of a thermal stimulus and self-reported pain.

Keywords directions of mediation, principal components analysis, fMRI, mediation analysis, structural equation models, high-dimensional data

1 Introduction

Mediation and path analysis have been pervasive in the social and behavioral sciences (e.g., Baron and Kenny (1986); MacKinnon (2008); Preacher and Hayes (2008)), and have found widespread use in many applications, including psychology, behavioral science, economics, decision-making, health psychology, epidemiology, and neuroscience. In the past couple of decades the topic has also begun to receive a great deal of attention in the statistical literature, particularly in the area of causal inference (e.g., Holland (1988); Robins and Greenland (1992); Angrist et al. (1996); Ten Have et al. (2007); Albert (2008); Jo (2008); Sobel (2008); VanderWeele and Vansteelandt (2009); Imai et al. (2010); Lindquist (2012); Pearl (2014)). When the effect of a treatment variable X on an outcome variable Y is at least partially directed through an intervening variable M, then M is said to be a mediator; see Figure 1 for an illustration of the corresponding path diagram. Mediation analysis allows one to parse the effects of the treatment on the outcome into separable direct and indirect effects. Here the direct effect is the influence of X on Y that is unmediated by M, the indirect effect is the influence mediated by M, and the total effect is the combination (the sum, when the effects are on the linear scale) of the direct and indirect effects. The influence of the intermediate variable on the outcome is often determined using linear structural equation models (SEMs), with model coefficients interpreted as effects. These models have important limitations, especially when the goal is to interpret the effects as causal (Ogburn (2012); VanderWeele (2015)). Under certain strong assumptions the coefficients of linear SEMs represent causal mediation effects, but when those assumptions are not met the effects estimated by these models may still suggest evidence of mediation that can be followed up with additional analyses.

One fundamental limitation of mediation analysis is that the mediating variable is assumed to be univariate. Multiple mediators can be tested either separately, or in some cases simul-

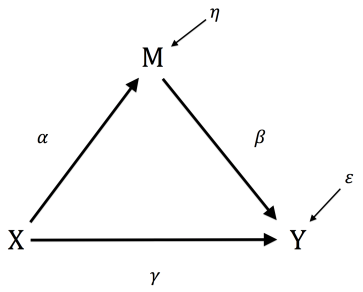


Figure 1: The three-variable path diagram representing the standard mediation framework. The variables corresponding to X, Y, and M are all scalars, as are the path coefficients α, β, and γ.

taneously (Preacher and Hayes, 2008). The latter approach is advantageous as it allows one to control for other potential mediating variables when assessing the indirect effect. This approach, however, becomes problematic if the different mediators are highly correlated; or if the total number of mediators is large. In recent years, many new applications measuring massive numbers of variables have appeared, including brain imaging, genetics, epidemiology, and public health studies. Applying mediation analysis to these applications requires a fundamental extension of that framework; such an extension is the focus of this work. We focus here on the definition and estimation of mediation effects in these settings; whether and when these effects may have causal interpretations will be the focus of future work.

As a motivating example, consider functional magnetic resonance imaging (fMRI), which is an imaging modality that allows researchers to measure changes in blood flow and oxygenation in the brain in response to neuronal activation (Ogawa et al. (1990); Kwong et al. (1992); Lindquist (2008)). In fMRI experiments, a multivariate time series of three dimensional brain volumes are obtained for each subject, where each volume consists of tens to hundreds of thousands of equally sized volume elements (voxels). A number of previous studies have used fMRI to investigate the relationship between painful heat and self-reported pain (Apkarian et al. (2005); Bushnell et al. (2013)). Recently, studies have focused on trial-by-trial modeling of the

relationship between the intensity of noxious heat and self-reported pain (Wager et al. (2013); Atlas et al. (2014)). In Woo et al. (2015), for example, a series of noxious thermal stimuli were applied at various temperatures (ranging from $44.3 - 49.3$ °C in 1 ° increments) to the left forearm of each of 33 subjects. In response, subjects gave subjective pain ratings at a specific time point following the offset of the stimulus. During the course of the experiment, brain activity in response to the thermal stimuli was measured across the entire brain using fMRI. One of the goals of the study was to search for brain regions whose activity level acts as potential mediators of the relationship between temperature and pain rating.

In this context, we are interested in whether the effect of temperature, X, on reported pain, Y, is mediated by the brain response, M. Here both X and Y are scalars, while M is the estimated brain activity measured over a large number of different voxels/regions. We assume that the values of M are either parameters or contrasts (linear combinations of parameters) obtained by fitting the general linear model (GLM), where for each subject, the relationship between the stimuli and the BOLD response is analyzed at the voxel level (Lindquist et al., 2012). Standard mediation techniques are applicable to univariate mediators, and the identification of univariate mediators has come to be known as Mediation Effect Parametric Mapping (Wager et al. (2008); Wager et al. (2009b); Wager et al. (2009a)) in the neuroimaging field. This approach, however, ignores the relationship between brain regions, and identifies a series of univariate mediators rather than an optimized, multivariate linear combination. A multivariate extension should focus on identifying latent brain components that are maximally effective as mediators.

Thus, we consider the same simple three-variable path diagram depicted in Figure 1, with the novel feature that the scalar potential mediator is replaced by a very high dimensional vector of potential mediators $\mathbf{M} = (M_1, M_2, \ldots M_p)$. In our motivating example, the vector of potential mediators consists of brain activity measured over a large number of possible regions. The goal is to find the linear combination of the entries of $\mathbf{M}$ (potential mediators) that provides

the strongest mediation signal as measured by the maximum likelihood criterion in the associated univariate SEM. To this end we propose a framework, which we denote the directions of mediation (DMs), which is philosophically similar to principal component analysis (PCA) but addresses a fundamentally different problem. The first direction of mediation, $\mathbf{w}_1$, is defined as the linear combination of the elements of $\mathbf{M}$ that maximizes the likelihood of the underlying three-variable path model. Like PCA, subsequent directions can thereafter be found that maximize the likelihood of the model, conditional on being orthogonal to the previous directions of mediation. In the brain imaging context, $\mathbf{w}_1$ provides the linear combination of the activation across brain regions. These values can be mapped back onto the brain and used to explore the relative contribution of different brain regions to the indirect effect of X on Y. The approach shares some similarities with partial least squares (PLS) (Wold (1982); Wold (1985); Krishnan et al. (2011)), which is a dimension reduction approach based on the correlation between a response variable (e.g. Y) and a set of explanatory variables (e.g. $\mathbf{M}$). In contrast, for DM the dimension reduction is based on the complete X-$\mathbf{M}$-Y relationship.

This article is organized as follows. In Section 2 we discuss the motivating thermal pain data set. In Sections 3.1 - 3.2 we review the standard mediation analysis framework and formulate a multivariate extension. In Sections 3.3 - 3.5 we provide an estimation algorithm and prove some asymptotic properties of the obtained estimates. In Section 3.6 we introduce a procedure for estimating the DM and its associated path coefficients when the mediator is high dimensional. In Section 4 we discuss a method for performing inference on the DM. Finally, in Sections 5 - 6 the efficacy of the approach is illustrated through simulations and an application to fMRI data.

2 Data Description

The data comes from the fMRI study of thermal pain described in the Introduction; see Woo et al. (2015) for an in-depth discussion. A total of 33 healthy, right-handed participants com-

pleted the study (age 27.9 ± 9.0 years, 22 females). All participants provided informed consent, and the Columbia University Institutional Review Board approved the study.

The experiment consisted of a total of nine runs. Seven runs were "passive", in which participants passively experienced and rated the heat stimuli, and two runs were "regulation", where the participants imagined the stimuli to be more or less painful than they actually were, in one run each (counterbalanced in order across participants). In this paper we consider only the seven passive runs, consisting of between $58 - 75$ separate trials (thermal stimulation repetitions). During each trial, thermal stimulations were delivered to the volar surface of the left inner forearm. Each stimulus lasted 12.5s, with 3s ramp-up and 2s ramp-down periods and 7.5s at the target temperature. Six levels of temperature, ranging from $44.3 - 49.3$ °C in increments of 1 °C, were administered to each participant. Each stimulus was followed by a $4.5 - 8.5$s long pre-rating period, after which participants rated the intensity of the pain on a scale of 0 to 100. Each trial concluded with a $5 - 9$s resting period.

Whole-brain fMRI data was acquired on a 3T Philips Achieva TX scanner at Columbia University. Structural images were acquired using high-resolution T1 spoiled gradient recall (SPGR) images with the intention of using them for anatomical localization and warping to a standard space. Functional EPI images were acquired with TR = $2,000$ms, TE = 20ms, field of view = 224mm, 64×64 matrix, $3 \times 3 \times 3$mm^3 voxels, 42 interleaved slices, parallel imaging, SENSE factor 1.5. For each subject, structural images were co-registered to the mean functional image using the iterative mutual information-based algorithm implemented in SPM8[1]. Subsequently, structural images were normalized to MNI space using SPM8's generative segment-and-normalize algorithm. Prior to preprocessing of functional images, the first four volumes were removed to allow for image intensity stabilization. Outliers were identified using the Mahalanobis distance for the matrix of slice-wise mean and the standard deviation values. The

[1] http://www.fil.ion.ucl.ac.uk/spm/

functional images were corrected for differences in slice-timing, and were motion corrected using SPM8. The functional images were warped to SPMs normative atlas using warping parameters estimated from coregistered, high resolution structural images, and smoothed with an 8mm FWHM Gaussian kernel. A high-pass filter of 180s was applied to the time series data.

A single trial analysis approach was used, by constructing a general linear model (GLM) design matrix with separate regressors for each trial (Rissman et al. (2004); Mumford et al. (2012)). Boxcar regressors, convolved with the canonical hemodynamic response function, were constructed to model periods for the thermal stimulation and rating periods for each trial. Other regressors that were not of direct interest included (a) intercepts for each run; (b) linear drift across time within each run; (c) the six estimated head movement parameters (x, y, z, roll, pitch, and yaw), their mean-centered squares, derivatives, and squared derivative for each run; (d) indicator vectors for outlier time points; (e) indicator vectors for the first two images in each run; (f) signal from white matter and ventricles. Using the results of the GLM analysis, whole-brain maps of activation were computed.

In summary, X_{ij} and Y_{ij} are the temperature level and pain rating, respectively, assigned on trial j to subject i, and $\mathbf{M}_{ij} = (M_{ij1}, M_{ij2}, \ldots M_{ijp})$ is the whole-brain activation measured over $p = 206,777$ voxels, defined as the regression parameter corresponding to the stimulus in the associated GLM. In addition, $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J_i\}$, where $I = 33$ and J_i takes subject-specific values between $58 - 75$.

3 Methods

In this section we review the standard approach to mediation analysis with linear SEMs, which is often used in behavioral sciences. Thereafter, we discuss the case when the mediator is multivariate and introduce the directions of mediation approach.

3.1 Mediation Analysis

Mediation analysis is often performed using the framework suggested by (Baron and Kenny, 1986), which is based on a linear structural equation model (LSEM). In this setting the variables X_i, Y_i, and M_i for $i = 1 \ldots n$, all take univariate scalar values. The LSEM corresponding to the path diagram in Figure 1, can be expressed as

$$M_i = \delta_1 + \alpha X_i + \epsilon_i \quad (1)$$

$$Y_i = \delta_2 + \gamma X_i + \beta M_i + \eta_i \quad (2)$$

for $i = 1 \ldots n$, where $\mathbb{E}(\epsilon | Z = z) = 0$ and $\mathbb{E}(\eta | Z = z, M = m) = 0$. Here the parameters of the LSEM can easily be estimated using a standard linear regression approach.

It can be shown that the total effect of X on Y, denoted τ, can be decomposed as follows:

$$\tau = \gamma + \alpha\beta.$$

Here the term γ represents the direct effect of X on Y, while $\alpha\beta$ represents the indirect (mediated) effect. To demonstrate mediation one can perform a hypothesis test to determine whether $\alpha\beta$ is significantly different from 0. This is typically performed using either the Sobel test (Sobel, 1982) or the bootstrap procedure (Shrout and Bolger, 2002).

3.2 Model Formulation

Now consider the case when X_i and Y_i are univariate variables, but $\mathbf{M}_i \in \mathbb{R}^p$ for all $i = 1, \ldots, n$. In the reminder of the paper, we use bold lower case symbols for column vectors, bold upper case symbols for matrices, with the remaining symbols, unless specified otherwise, scalars. We use n for the total number of observations and N for the total number of subjects, i.e., $n = \sum_{i=1}^{N} F_i$, where F_i is the number of observation for subject i. We denote the full dataset $\mathbf{D} = (\mathbf{X}, \mathbf{Y}, \mathbf{M})$, where $\mathbf{X} = (X_1, \ldots X_n)^\intercal \in \mathbb{R}^n$, $\mathbf{Y} = (Y_1, \ldots Y_n)^\intercal \in \mathbb{R}^n$, and

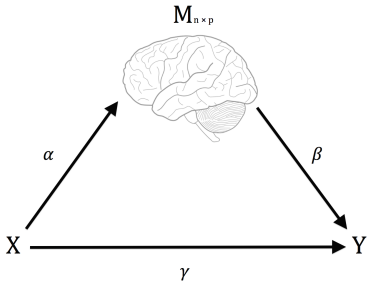


Figure 2: The three-variable path diagram used to represent the multivariate mediation framework. The variables corresponding to Z and Y are scalars, while the variable corresponding to M is a vector.

$\mathbf{M} = (\mathbf{M}_1, \dots \mathbf{M}_n)^\intercal \in \mathbb{R}^{n \times p}$. Let $\mathbf{w} \in \mathbb{R}^p$ be a vector that maps $\mathbf{M}$ onto $\mathbb{R}^n$, and $\boldsymbol{\theta} := (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma) \in \mathbb{R}^5$ the set of parameters of the LSEM described in Eqs. (1) - (2) fit using $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Mw}$ as input. This can be expressed as follows:

$$\mathbf{M}_i\mathbf{w} = \alpha_0 + \alpha_1 X_i + \eta_i. \tag{3}$$

$$Y_i = \beta_0 + \gamma X_i + \beta_1 \mathbf{M}_i\mathbf{w} + \epsilon_i \tag{4}$$

Throughout we assume that ϵ_i and η_i are both independent and identically distributed normal random variables with mean zero and variance σ^2_ϵ and σ^2_η, respectively.

Model estimation in this setting is complicated by the inclusion of the unknown parameter $\mathbf{w}$. Here we seek to find the value of $\mathbf{w}$ that maximizes the likelihood of the underlying LSEM described in Eqs. (3) - (4). The first direction of mediation is defined as the linear combination of the elements of $\mathbf{M}$ that maximizes the likelihood of the underlying LSEM. Like PCA, subsequent directions can be found that maximize the likelihood of the model, conditional on it being orthogonal to the previously found directions.

To illustrate, let $\mathscr{L}(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta})$ be the joint likelihood of the LSEM stated in Eqs. (3) - (4). The *Directions of Mediation* are formally defined as follows:

Step 1: The 1^{st} DM is the vector $\mathbf{w}_1$, with norm 1, which maximizes the conditional joint

likelihood $\mathscr{L}(\mathbf{D}, \boldsymbol{\theta}; \mathbf{w}_1)$, i.e.

$$\hat{\mathbf{w}}_1|\boldsymbol{\theta} = \underset{\left\{\mathbf{w}\in\mathbb{R}^p:\|\mathbf{w}\|=1\right\}}{\operatorname{argmax}} \left\{\mathscr{L}(\mathbf{D}, \boldsymbol{\theta}; \mathbf{w}_1)\right\}$$

Step 2: The 2^{nd} DM is the vector $\mathbf{w}_2$, with norm 1 and orthogonal to $\mathbf{w}_1$, which maximizes the conditional joint likelihood $\mathscr{L}(\mathbf{D}, \boldsymbol{\theta}, \mathbf{w}_1; \mathbf{w}_2)$, i.e.

$$\hat{\mathbf{w}}_2|\boldsymbol{\theta}, \mathbf{w}_1 = \underset{\left\{\begin{smallmatrix}\mathbf{w}\in\mathbb{R}^p:\|\mathbf{w}\|=1\\ \mathbf{w}_1\mathbf{w}^\intercal=0\end{smallmatrix}\right\}}{\operatorname{argmax}} \left\{\mathscr{L}(\mathbf{D}, \boldsymbol{\theta}, \mathbf{w}_1; \mathbf{w})\right\}$$

$$\vdots$$

Step k: The k^{th} DM is the vector $\mathbf{w}_k$, with norm 1 and orthogonal to $\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$, which maximizes the conditional joint likelihood $\mathscr{L}(\mathbf{D}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}; \mathbf{w}_k)$, i.e.

$$\hat{\mathbf{w}}_k|\boldsymbol{\theta}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1} = \underset{\left\{\begin{smallmatrix}\mathbf{w}\in\mathbb{R}^p:\|\mathbf{w}\|=1\\ \mathbf{w}_{k'}\mathbf{w}^\intercal=0\ \forall k'\in\{1,\ldots,k-1\}\end{smallmatrix}\right\}}{\operatorname{argmax}} \left\{\mathscr{L}(\mathbf{D}, \boldsymbol{\theta}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}; \mathbf{w})\right\}$$

Remark I: The norm constraint on $\mathbf{w}_1$ makes it separable from the slope parameter β_1.

Remark II: According to the model formulation the signs of the DMs are unidentifiable.

3.3 Estimation

In this Section we describe how to estimate the parameters of the DM model. Assuming joint normality, the joint $\log$ likelihood function for $\mathbf{w}_1$ and $\boldsymbol{\theta}$, $\mathscr{L}(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta})$, can be expressed as:

$$\begin{aligned}\mathscr{L}(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta}) &\propto g_1(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta}) \\ &:= -\Bigg\{\frac{(\mathbf{Y} - \beta_0 - \mathbf{X}\gamma_1 - \mathbf{M}\mathbf{w}_1\beta_1)^\intercal(\mathbf{Y} - \beta_0 - \mathbf{X}\gamma_1 - \mathbf{M}\mathbf{w}_1\beta_1)}{\sigma^2_{\epsilon_1}} \\ &+ \frac{(\mathbf{M}\mathbf{w}_1 - \alpha_0 - \mathbf{X}\alpha_1)^\intercal(\mathbf{M}\mathbf{w}_1 - \alpha_0 - \mathbf{X}\alpha_1)}{\sigma^2_{\eta_1}}\Bigg\}\end{aligned}$$

The goal is to find both the parameters of the LSEM and the first DM that jointly maximize $g_1(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta})$, under the constraint that the L_2 norm of $\mathbf{w}_1$ equals 1. Consider the Lagrangian

$$L(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta}, \lambda) = g_1(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta}) + \lambda(\mathbf{w}_1^\intercal \mathbf{w}_1 - 1).$$

The dual problem can be expressed:

$$(\hat{\mathbf{w}}_1, \hat{\boldsymbol{\theta}})|\lambda = \underset{\left\{\begin{smallmatrix}\mathbf{w}_1 \in \mathbb{R}^p \\ \boldsymbol{\theta} \in \mathbb{R}^5\end{smallmatrix}\right\}}{\operatorname{argmax}} L(\mathbf{D}; \mathbf{w}_1, \boldsymbol{\theta}, \lambda)$$

where λ is the Lagrange multiplier. To solve this problem we propose a method where λ is profiled out by one set of parameters of interest. We establish, under a regularity condition, the closed form solution for the path coefficients, the first DM, and λ.

Regularity Condition I: (N-0) The first partial derivatives of the objective function and the constraint function exist.

Under this condition, it can be shown that

$$\hat{\mathbf{w}}_1|\boldsymbol{\theta}, \lambda = (\lambda \mathbf{I} + \boldsymbol{\psi}(\boldsymbol{\theta}))^{-1} \boldsymbol{\phi}(\boldsymbol{\theta}) \tag{5}$$

$$\hat{\lambda}|\boldsymbol{\theta} = \arg_\lambda \left\{ [(\lambda \mathbf{I} + \boldsymbol{\psi}(\boldsymbol{\theta}))^{-1} \boldsymbol{\phi}(\boldsymbol{\theta})]^\intercal [(\lambda \mathbf{I} + \boldsymbol{\psi}(\boldsymbol{\theta})^{-1} \boldsymbol{\phi}(\boldsymbol{\theta})] = 1 \right\} \tag{6}$$

$$\hat{\boldsymbol{\theta}}|\hat{\mathbf{w}}_1, \hat{\lambda} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\mathbf{D}; \hat{\mathbf{w}}_1, \boldsymbol{\theta}, \hat{\lambda}) \tag{7}$$

where

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \frac{\mathbf{M}^\intercal \mathbf{M} \beta_1^2}{\sigma_{\epsilon_1}^2} + \frac{\mathbf{M}^\intercal \mathbf{M}}{\sigma_{\eta_1}^2}$$

and

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = \frac{\mathbf{M}^\intercal (\alpha_0 + \alpha_1 \mathbf{X})}{\sigma_{\eta_1}^2} + \frac{\mathbf{M}^\intercal (\mathbf{Y} - \beta_0 - \mathbf{X}\gamma_1)\beta_1}{\sigma_{\epsilon_1}^2}.$$

Using these results we outline an iterative procedure for jointly estimating the first direction of mediation and structural path parameters as follows:

1. Start with an initial value for $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}_1^{(0)}$.

2. For each k, set:

$$\hat{\lambda}^{(k)}|\boldsymbol{\theta}_1^{(k)} = \arg_{\lambda}\left\{[(\lambda\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}^{(k)}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta}^{(k)})]^{\intercal}[(\lambda\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}^{(k)}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta}^{(k)})] = 1\right\} \quad (8)$$

$$\hat{\mathbf{w}_1}^{(k)}|\boldsymbol{\theta}_1^{(k)}, \hat{\lambda}^{(k)} = (\hat{\lambda}^{(k)}\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}^{(k)}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta}^{(k)}) \quad (9)$$

$$\hat{\boldsymbol{\theta}}^{(k+1)}|\hat{\mathbf{w}_1}^{(k)} = \arg\max_{\boldsymbol{\theta}_1}\left\{g_1(\mathbf{X},\mathbf{Y},\boldsymbol{\theta}_1,\hat{\mathbf{w}_1}^{(k)})\right\} \quad (10)$$

3. Repeat step 2 until convergence, each time setting $k = k+1$.

3.4 Higher Order Directions of Mediation

We propose two approaches to obtain the higher order Directions of Mediation. The first uses additional penalty parameters, and the second uses subtraction and *Gram-Schmidt* projections. While the former approach is likely to achieve global maxima, the latter is computationally more efficient, and provides a good approximation of the higher order DMs. The performance of the projection approach is illustrated through extensive simulations in Section 5.

3.4.1 Penalty Approach

Estimates for the second direction of mediation, $\hat{\mathbf{w}}_2$, and the associated path coefficients, $\hat{\boldsymbol{\theta}}_2$, are obtained by computing:

$$(\hat{\mathbf{w}}_2, \hat{\boldsymbol{\theta}}_2)|\lambda_1,\lambda_2 = \underset{\mathbf{w}_2,\boldsymbol{\theta}_2}{\operatorname{argmax}}\left\{g_2(\mathbf{D},\hat{\mathbf{w}}_1;\mathbf{w}_2,\boldsymbol{\theta}_2) + \lambda_1\left(\mathbf{w}_2^{\intercal}\mathbf{w}_2 - 1\right) + \lambda_2\mathbf{w}_2^{\intercal}\hat{\mathbf{w}}_1\right\}$$

which guarantees that $\hat{\mathbf{w}}_2$, the estimate of $\mathbf{w}_2$ is of unit length and orthogonal to $\hat{\mathbf{w}}_1$. Here $\boldsymbol{\theta}_2 = \left(\alpha_{0,2}, \alpha_{1,2}, \beta_{0,1}, \beta_{1,1}^{(2)}, \beta_{1,2}^{(2)}, \gamma_2\right)$, where superscript (2) indicates that the parameter is specifically

associated with the second direction of mediation, and

$$g_2(\mathbf{D},\hat{\mathbf{w}}_1;\mathbf{w}_2,\boldsymbol{\theta}_2) := \left\{ \frac{(\mathbf{Y}-\beta_{0,2}-\mathbf{X}\gamma_2-\mathbf{M}\hat{\mathbf{w}}_1\beta_{1,1}^{(2)}-\mathbf{M}\mathbf{w}_2\beta_{1,2}^{(2)})^\intercal(\mathbf{Y}-\beta_{0,2}-\mathbf{X}\gamma_2-\mathbf{M}\hat{\mathbf{w}}_1\beta_{1,1}^{(2)}-\mathbf{M}\mathbf{w}_2\beta_{1,2}^{(2)})}{\sigma_{\epsilon_2}^2} + \frac{(\mathbf{M}\mathbf{w}_2-\alpha_{0,2}-\mathbf{X}\alpha_{1,2})^\intercal(\mathbf{M}\mathbf{w}_2-\alpha_{0,2}-\mathbf{X}\alpha_{1,2})}{\sigma_{\eta_2}^2} \right\}$$

Under regularity condition (N-0), it can be shown that

$$\hat{\lambda}_2|\boldsymbol{\theta}_2,\hat{\lambda}_1 = \arg_{\lambda_2}\left\{[(\lambda_1\mathbf{I}+\boldsymbol{\psi}_2(\boldsymbol{\theta}_2))^{-1}\boldsymbol{\phi}_2(\boldsymbol{\theta}_2,\lambda_2)]^\intercal\hat{\mathbf{w}}_1=0\right\} \tag{11}$$

$$\hat{\lambda}_1|\boldsymbol{\theta}_2,\hat{\lambda}_2 = \arg_{\lambda_1}\left\{[(\lambda_1\mathbf{I}+\boldsymbol{\psi}_2(\boldsymbol{\theta}_2))^{-1}\boldsymbol{\phi}_2(\boldsymbol{\theta}_2,\lambda_2)]^\intercal[(\lambda_1\mathbf{I}+\boldsymbol{\psi}_2(\boldsymbol{\theta}_2)^{-1}\boldsymbol{\phi}_2(\boldsymbol{\theta}_2,\lambda_2)]=1\right\} \tag{12}$$

$$\hat{\mathbf{w}}_2|\boldsymbol{\theta}_2,\lambda_1,\lambda_2 = (\lambda_1\mathbf{I}+\boldsymbol{\psi}_2(\boldsymbol{\theta}_2))^{-1}\boldsymbol{\phi}_2(\boldsymbol{\theta}_2,\lambda_2) \tag{13}$$

$$\hat{\boldsymbol{\theta}}_2|\hat{\mathbf{w}}_1,\hat{\lambda}_1,\hat{\lambda}_2 = \operatorname*{argmax}_{\boldsymbol{\theta}} L_2(\mathbf{D};\hat{\mathbf{w}}_1,\boldsymbol{\theta}_2,\hat{\lambda}_1,\hat{\lambda}_2) \tag{14}$$

where

$$\boldsymbol{\psi}(\boldsymbol{\theta}_2) = \frac{\mathbf{M}^\intercal\mathbf{M}\overset{(2)}{\beta_{1,2}^2}}{\sigma_{\epsilon_2}^2} + \frac{\mathbf{M}^\intercal\mathbf{M}}{\sigma_{\eta_2}^2}$$

and

$$\boldsymbol{\phi}(\boldsymbol{\theta}_2,\lambda_2) = \frac{\mathbf{M}^\intercal(\alpha_{0,2}+\alpha_{1,2}\mathbf{X})}{\sigma_{\eta_2}^2} + \frac{\mathbf{M}^\intercal(\mathbf{Y}-\beta_{0,2}-\mathbf{X}\gamma_2-\overset{(2)}{\beta_{1,1}}\mathbf{M}\hat{\mathbf{w}}_1)\overset{(2)}{\beta_{1,2}}}{\sigma_{\epsilon_2}^2} - \frac{1}{2}\lambda_2\hat{\mathbf{w}}_1.$$

Note that, unlike in Eqs. (5) - (7) where we only need to specify starting values for $\boldsymbol{\theta}$ and $\mathbf{w}_1$, here we need to specify a starting value for one of the λ's in (11) and (12). A convenient way is to solve for λ under the constraint that $\lambda_1 = \lambda_2$.

Estimates for the k^{th} direction of mediation, $\hat{\mathbf{w}}_k$, and the associated path coefficients, $\hat{\boldsymbol{\theta}}_k$, are similarly obtained by computing:

$$(\hat{\mathbf{w}}_k,\hat{\boldsymbol{\theta}}_k)|\lambda_1,\ldots,\lambda_k = \operatorname*{argmax}_{\mathbf{w}_k,\boldsymbol{\theta}_k}\left\{ g_k(\mathbf{D},\hat{\mathbf{w}}_1,\ldots,\hat{\mathbf{w}}_{k-1};\mathbf{w}_k,\boldsymbol{\theta}_k) + \lambda_1\left(\mathbf{w}_k^\intercal\mathbf{w}_k-1\right) + \sum_{j=2}^{k}\lambda_j\mathbf{w}_k^\intercal\hat{\mathbf{w}}_{j-1}\right\}.$$

which guarantees that $\hat{\mathbf{w}}_k$ is of unit length and orthogonal to all preceding $k-1$ DMs. Here $\boldsymbol{\theta}_k = \big(\alpha_{0,k}, \alpha_{1,k}, \beta_{0,k}, \beta_{1,1}^{(k)}, \ldots, \beta_{1,k}^{(k)}, \gamma_k\big)$, where super script (k) indicates that the parameter is specifically associated with the k^{th} direction of mediation, and

$$g_k(\mathbf{D}, \hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{k-1}; \mathbf{w}_k, \boldsymbol{\theta}_k) := \\ -\left\{ \frac{(\mathbf{Y} - \beta_{0,k} - \mathbf{X}\gamma_k - \sum_{i=1}^{k-1} \mathbf{M}\hat{\mathbf{w}}_i \beta_{1,i}^{(k)} - \mathbf{M}\mathbf{w}_k \beta_{1,k}^{(k)})^\intercal (\mathbf{Y} - \beta_{0,k} - \mathbf{X}\gamma_k - \sum_{i=1}^{k-1} \mathbf{M}\hat{\mathbf{w}}_i \beta_{1,i}^{(k)} - \mathbf{M}\mathbf{w}_k \beta_{1,k}^{(k)})}{\sigma_{\epsilon_k}^2} \right. \\ \left. + \frac{(\mathbf{M}\mathbf{w}_k - \alpha_{0,k} - \mathbf{X}\alpha_{1,k})^\intercal (\mathbf{M}\mathbf{w}_k - \alpha_{0,k} - \mathbf{X}\alpha_{1,k})}{\sigma_{\eta_k}^2} \right\}$$

The higher-order DMs and corresponding path coefficients can be estimated in an analogous manner as described above.

3.4.2 Projection Approach

While the penalty approach is likely to achieve global maximum, it is computationally difficult in high dimensions. As an alternative, we introduce a projection method, which is computationally more efficient, and provides a good approximation of the higher order DMs.

Estimates of the second direction of mediation, $\hat{\mathbf{w}}_2$, and the associated path coefficients, $\hat{\boldsymbol{\theta}}_2$, are obtained by computing:

$$(\hat{\mathbf{w}}_2, \hat{\boldsymbol{\theta}}_2)|\lambda = \underset{\left\{ \begin{smallmatrix} \mathbf{Z} \in \bar{\mathbb{R}}^p : \mathbf{w}_2(\mathbf{Z}) := \mathbf{Z} - \text{Proj}_{\hat{\mathbf{w}}_1}(\mathbf{Z}) \\ \boldsymbol{\theta}_2 \end{smallmatrix} \right\}}{\text{argmax}} \left\{ g_2(\mathbf{D}, \hat{\mathbf{w}}_1; \mathbf{w}_2(\mathbf{Z}), \boldsymbol{\theta}_2) - \lambda\big([\mathbf{w}_2(\mathbf{Z})]^\intercal [\mathbf{w}_2(\mathbf{Z})] - 1\big) \right\}$$

where $\text{Proj}_{\hat{\mathbf{w}}_1}(\mathbf{Z}) = \dfrac{\langle \mathbf{Z}, \hat{\mathbf{w}}_1 \rangle}{\langle \hat{\mathbf{w}}_1, \hat{\mathbf{w}}_1 \rangle} \hat{\mathbf{w}}_1$. Similarly, estimates for the k^{th} direction of mediation, $\hat{\mathbf{w}}_k$, and the associated path coefficients, $\hat{\boldsymbol{\theta}}_k$, are obtained by computing:

$$(\hat{\mathbf{w}}_k, \hat{\boldsymbol{\theta}}_k)|\lambda = \underset{\left\{ \begin{smallmatrix} \mathbf{Z} \in \bar{\mathbb{R}}^p : \mathbf{w}_k(\mathbf{Z}) := \mathbf{Z} - \sum_{i=1}^{k-1} \text{Proj}_{\hat{\mathbf{w}}_i}(\mathbf{Z}) \\ \boldsymbol{\theta}_k \end{smallmatrix} \right\}}{\text{argmax}} \left\{ g_k(\mathbf{D}, \hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{k-1}; \mathbf{w}_k, \boldsymbol{\theta}_k) - \lambda\big([\mathbf{w}_k(\mathbf{Z})]^\intercal [\mathbf{w}_k(\mathbf{Z})] - 1\big) \right\}$$

where $\text{Proj}_{\hat{\mathbf{w}}_i}(\mathbf{Z}) = \dfrac{\langle \mathbf{Z}, \hat{\mathbf{w}}_i \rangle}{\langle \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_i \rangle} \hat{\mathbf{w}}_i, \forall i \in \{1, \ldots, k-1\}$.

3.5 Asymptotic Properties

In this Section we provide asymptotic results related to the first DM. Define the parameter vector $\boldsymbol{\xi} = (\boldsymbol{\gamma}, \lambda) \in \boldsymbol{\Xi}$, where $\boldsymbol{\gamma} = (\boldsymbol{\theta}, \mathbf{w}) \in \boldsymbol{\Theta} \times \mathbf{w} = \boldsymbol{\Gamma}$ are our parameters of interest, and $\lambda \in \Lambda$ is a nuisance parameter. Due to regularity condition (N-0), we can express the dual function as:

$$G(\mathbf{D}; \boldsymbol{\theta}) := g_1(\mathbf{D}; \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})[\mathbf{w}(\boldsymbol{\theta})^\intercal \mathbf{w}(\boldsymbol{\theta}) - 1].$$

Define $D_0(\theta_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\dfrac{\partial \ell(D; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$, where $\ell(D; \boldsymbol{\theta}) = \dfrac{\partial G(D; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the first partial derivative of the Lagrangian, and $V(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\ell(D; \boldsymbol{\theta})\ell^\intercal(D; \boldsymbol{\theta})\right)$.

Theorem 1. *Under regularity conditions (N-1) - (N-12) [see Appendix], the structural path coefficient estimators are asymptotically consistent, and normally distributed, i.e.*

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$$

and

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Sigma(\boldsymbol{\theta}_0))$$

where $\Sigma(\boldsymbol{\theta}_0) = D_0^{-1}(\boldsymbol{\theta}_0) V(\boldsymbol{\theta}_0) [D_0^{-1}(\boldsymbol{\theta}_0)]^\intercal$.

Theorem 2. *Under regularity conditions (N-1) - (N-13) [see Appendix], the estimator of the first direction of mediation is asymptotically consistent, and normally distributed, i.e.*

$$\mathbf{w}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{w}(\boldsymbol{\theta}_0)$$

and

$$\sqrt{n}(\mathbf{w}(\hat{\boldsymbol{\theta}}) - \mathbf{w}(\boldsymbol{\theta}_0)) \to N\big(0, \Sigma^{\mathbf{w}}(\boldsymbol{\theta}_0)\big).$$

where $\Sigma^{\mathbf{w}}(\boldsymbol{\theta}_0) = [\nabla \mathbf{w}(\boldsymbol{\theta}_0)]^\intercal D_0^{-1}(\boldsymbol{\theta}_0) V(\boldsymbol{\theta}_0) [D_0^{-1}(\boldsymbol{\theta}_0)]^\intercal \nabla \mathbf{w}(\boldsymbol{\theta}_0)$.

Regularity conditions and proofs are included in the Appendix and supplementary materials, respectively. It is worth pointing out two aspects of *Theorem 2*. First, it is valid for $\dim(\mathbf{w}) \leq \dim(\boldsymbol{\theta})$ or $\dim(\mathbf{w}) \geq \dim(\boldsymbol{\theta})$. This lays the theoretical foundation for estimating the first DM when $\dim(\mathbf{w}) \gg \dim(\boldsymbol{\theta})$. Second, it involves an application of the multivariate delta method. The conditions required for the multivariate delta method are met, because under the stated regularity conditions, for every estimate of $\boldsymbol{\theta}$, we can find a unique function $\mathbf{f}(\cdot)$ of the estimate. Therefore, for a sufficiently large sample size n, there exists an estimate of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_n$, in the neighborhood of the true value. For that particular $\boldsymbol{\theta}_n$, we can find a unique mapping $\mathbf{f}_n(\cdot)$.

3.6 High-dimensional Directions of Mediation

The proposed method works well in the low-dimensional setting. In this section we propose two methods, based on using Singular Value Decomposition (SVD) and Population Value Decomposition (PVD), for estimating the directions of mediation when $\mathbf{M}$ is high dimensional. Throughout we assume that the data for each subject i is stored in an $F_i \times p$ matrix, $\mathbf{M}_i$, where the j^{th} row contains voxel-wise activity for the measurements of the j^{th} trail for the i^{th} subject. All $\mathbf{M}_i$ matrices are stacked vertically to form the $n \times p$ matrix $\mathbf{M}$, where $n = \sum_{i=1}^{N} F_i$.

3.6.1 Singular Value Decomposition Approach

Here we offer a step-by-step approach towards estimating the DM using SVD.

Step 1: Approximate the matrix $\mathbf{M}$ using

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\intercal} \approx \mathbf{U}_L\boldsymbol{\Sigma}_{L,R}\mathbf{V}_R^{\intercal}$$

where the $n \times L$-dimensional matrix $\mathbf{U}_L$ consists of the first L columns of the matrix $\mathbf{U}$, the $p \times R$-dimensional matrix $\mathbf{V}_R$ consists of the first R columns of the matrix $\mathbf{V}$, and $\boldsymbol{\Sigma}_{L,R}$ is obtained by retaining the first L rows and R columns of $\boldsymbol{\Sigma}$. The choice of L and R can be

based on various criteria, such as total variance explained or signal-to-noise ratio. Let us denote $\tilde{\mathbf{M}} = \mathbf{U}_L \mathbf{\Sigma}_{L,R}$ and $\tilde{\mathbf{w}} = \mathbf{V}_R^\intercal \mathbf{w}$. Note that $\mathbf{M}\mathbf{w} \approx \tilde{\mathbf{M}}\tilde{\mathbf{w}}$, and $\dim(\tilde{\mathbf{w}}) = R \ll p = \dim(\mathbf{w})$.

Step 2: Place $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{w}}$ into the DM framework:

$$\tilde{\mathbf{M}}_k \tilde{\mathbf{w}} = \alpha_0 + \alpha_1 X_k + \eta_k \tag{15}$$

$$Y_k = \beta_0 + \gamma X_k + \beta_1 \tilde{\mathbf{M}}_k \tilde{\mathbf{w}} + \epsilon_k \tag{16}$$

where ϵ_k and η_k are iid $\epsilon_k \sim N(0, \sigma_\epsilon^2), \eta_k \sim N(0, \sigma_\eta^2)$, respectively.

Step 3: Estimate (15) and (16) using the methods described in Section 3.3.

Since $\mathbf{V}_R$ can be obtained via SVD, we can retrieve the original estimator of the DM, $\hat{\mathbf{w}}$, using the generalized inverse, i.e., $\hat{\mathbf{w}} = \mathbf{V}_R^{-} \hat{\tilde{\mathbf{w}}}$, where $^-$ indicates the generalized inverse, or by minimizing the L_2 norm, $\hat{\mathbf{w}} = \min_{\mathbf{w}} \parallel \hat{\tilde{\mathbf{w}}} - \mathbf{V}_R \mathbf{w} \parallel$.

3.6.2 Generalized Population Value Decomposition Approach

While, the SVD approach is easy to implement, it only provides subject-specific information about $\mathbf{M}$. Population Value Decomposition (Crainiceanu et al., 2011) is a general method for conducting simultaneous dimensionality reduction of a large matrix, that also provides population-level information. While the PVD framework has a number of advantages, it assumes that the number of trails per subject is equal, which does not hold in many practical settings, including our motivating fMRI study. To address this issue, we introduce the Generalized Population Value Decomposition (GPVD), which allows the number of trials per subject to differ, while maintaining the dimension reduction benefits of the original.

Specifically, consider a subject-specific $F_i \times p$ matrix $\mathbf{M}_i$, where F_i denotes the number of trials per subject i, which may vary across subjects. The GPVD of $\mathbf{M}_i$ is given by

$$\mathbf{M}_i = \mathbf{U}_i \tilde{\mathbf{V}}_i \mathbf{D} + \mathbf{E}_i, \tag{17}$$

where $\mathbf{U}_i$ is an $F_i \times F_i$ unitary matrix, $\mathbf{V}_i$ is an $F_i \times B$ matrix of subject-specific coefficients,

$\mathbf{D}$ is a $B \times p$ population-specific matrix, $\mathbf{E}_i$ is an $F_i \times p$ matrix of residuals, and B is chosen based upon a criteria such as total variance explained. The difference between GPVD and SVD is that $\tilde{\mathbf{V}}_k$ is not necessarily diagonal; and the difference between GPVD and PVD is that $\mathbf{U}_k$ is subject-specific, and can have varying numbers of rows.

Here we introduce a step-by-step procedure for obtaining both the GPVD and DMs.

Step 1: Perform a subject-wise SVD.

1.1. For every subject i, compute the SVD: $\mathbf{M}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\intercal$, where $\mathbf{U}_i$, $\mathbf{\Sigma}_i$, and $\mathbf{V}_i$, are $F_i \times F_i$, $F_i \times F_i$, and $F_i \times p$ matrices, respectively.

1.2. Obtain the $F_i \times B$ matrix $\mathbf{U}_i^B$ by using the first B columns of $\mathbf{U}_i$, the $B \times B$ diagonal matrix $\mathbf{\Sigma}_i^B$ using the first B diagonal elements of $\mathbf{\Sigma}_i$, and the $B \times p$ matrix $\mathbf{V}_i^B$ using the first B columns of $\mathbf{V}_i$. Note in (17) we refer to $\mathbf{U}_i^B$ as $\mathbf{U}_i$.

1.3. Form a $p \times NB$ matrix $\mathbf{V} := [\mathbf{V}_1^B|, \ldots, |\mathbf{V}_N^B]$.

Step 2: Form the matrix $\mathbf{D}$.

When p is reasonably small, use SVD to compute $\mathbf{V}\mathbf{V}^\intercal = \breve{\mathbf{A}}\breve{B}^2\breve{\mathbf{A}}^\intercal$. The $p \times B$ matrix $\mathbf{D}$ is obtained using the first B columns of $\breve{\mathbf{A}}$. When p is large, performing the SVD is computationally impractical due to memory limitations. Here instead perform a block-wise SVD (Zipunnikov et al., 2011), and compute the matrix $\mathbf{D}$ as described above. Here it should be noted that $\mathbf{D}$ contains common features across subjects. At the population level $\mathbf{V} \approx \mathbf{D}(\mathbf{D}^\intercal\mathbf{V})$, and at the subject level $\mathbf{V}_i^B \approx \mathbf{D}(\mathbf{D}^\intercal\mathbf{V}_i^B)$.

Step 3: Compute $\tilde{\mathbf{V}}_i = \mathbf{\Sigma}_i^B(\mathbf{V}_i^B)^\intercal\mathbf{D}^\intercal$.

The GPVD in (17) can be summarized as follows:

$$\begin{aligned}\mathbf{M}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\intercal &\approx \mathbf{U}_i^B\mathbf{\Sigma}_i^B(V_i^B)^\intercal \\ &\approx \mathbf{U}_i^B\{\mathbf{\Sigma}_i^B(\mathbf{V}_i^B)^\intercal\mathbf{D}^T\}\mathbf{D} = \mathbf{U}_i^B\tilde{\mathbf{V}}_i^B\mathbf{D},\end{aligned} \tag{18}$$

where $\mathbf{U}_i^B$, $\mathbf{\Sigma}_i^B$, and $\mathbf{V}_i^B$ are obtained from Step 1, $\mathbf{D}$ from Step 2, and $\tilde{\mathbf{V}}_i^B$ from Step 3.

Remark: The first approximation in (18) is obtained by retaining the eigenvectors that explain most of the observed variability at the subject level. The second results from projecting the subject-specific right eigenvectors on the corresponding population-specific eigenvectors.

Using the GPVD framework, we can compute DMs in the high dimensional case as follows:

Step 1: Perform GPVD on the original $n \times p$ mediation matrix $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \vdots \\ \mathbf{M}_n \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1\tilde{\mathbf{V}}_1\mathbf{D} \\ \vdots \\ \mathbf{U}_n\tilde{\mathbf{V}}_n\mathbf{D} \end{bmatrix}$, where $N = \sum_{i=1}^{n} F_i$.

Step 2: Stack all $F_i \times B$ matrices $\mathbf{U}_i\tilde{\mathbf{V}}_i$ vertically to form an $n \times B$ matrix

$$\tilde{\mathbf{M}} = \begin{bmatrix} \mathbf{U}_1\tilde{V}_1 \\ \vdots \\ \mathbf{U}_n\tilde{V}_n \end{bmatrix}. \tag{19}$$

Let $\tilde{\mathbf{w}} = \mathbf{D}\mathbf{w}$, where $\tilde{\mathbf{w}}$ is $B \times 1$.

Step 3: Place $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{w}}$ into the LSEM equations:

$$\tilde{\mathbf{M}}_k\tilde{\mathbf{w}} = \alpha_0 + \alpha_1 X_k + \eta_k \tag{20}$$

$$Y_k = \beta_0 + \gamma X_k + \beta_1\tilde{\mathbf{M}}_k\tilde{\mathbf{w}} + \epsilon_k \tag{21}$$

Since $\mathbf{D}$ can be obtained via the GPVD, we can retrieve the original estimator of the high dimensional direction of mediation, $\hat{\mathbf{w}}$, via the generalized inverse, i.e.,

$$\hat{\mathbf{w}} = \mathbf{D}^{-}\hat{\tilde{\mathbf{w}}} \tag{22}$$

where $^{-}$ indicates the generalized inverse, or by computing $\hat{\mathbf{w}} = \min_{\mathbf{w}} \parallel \hat{\tilde{\mathbf{w}}} - \mathbf{D}\mathbf{w} \parallel$.

4 Inference

In low-dimensional settings, we can obtain variance estimates for the first DM and the path coefficients using *Theorems 1* and *2*. In high dimensional settings, variance estimation using the generalized inverse is under-estimated since the $\mathbf{D}$ obtained from (18) is random. Even if we were to adjust for this, the covariance estimation of $\mathbf{D}$ ($B \times p, B \ll p$) is computationally infeasible. Therefore, using the bootstrap to perform inference is a natural alternative.

Consider $\mathbf{M} = \tilde{\mathbf{M}}\mathbf{D}$, where $\mathbf{M}$ is $n \times p$, $\tilde{\mathbf{M}}$ is $n \times B$, $\mathbf{D}$ is $B \times p$, and $B < n \ll p$. The bootstrap procedure can be outlined as follows:

1. Bootstrap n rows from $\tilde{\mathbf{M}}$, stack them horizontally and form the $n \times B$ matrix $\tilde{\mathbf{M}}^{(j)}$;
2. Obtain $\hat{\tilde{\mathbf{w}}}^{(j)}$ from $\tilde{\mathbf{M}}^{(j)}$, where $\hat{\tilde{\mathbf{w}}}^{(j)}$ is the j^{th} bootstrap DM of length B;
3. Obtain $\hat{\mathbf{w}}^{(j)} = \mathbf{D}^{-1}\hat{\tilde{\mathbf{w}}}^{(j)}$, where $\hat{\mathbf{w}}^{(j)}$ is the high dimensional bootstrap DM of length p;
4. Repeat steps 1-3 J times. Stack all J values of $\hat{\mathbf{w}}^{(j)}$ vertically and form $\hat{\mathbf{w}}^* = \begin{bmatrix} \hat{\mathbf{w}}^{(1)} \\ \vdots \\ \hat{\mathbf{w}}^{(J)} \end{bmatrix}$, where $\hat{\mathbf{w}}^*$ is a $J \times p$ matrix.

Note $\hat{\mathbf{w}}^* = \begin{bmatrix} \hat{\mathbf{w}}_1 & \dots & \hat{\mathbf{w}}_p \end{bmatrix}$, where $\hat{\mathbf{w}}_k$ is the bootstrap values of the DM corresponding to voxel k, for $k \in \{1, \dots, p\}$, from which we can form a distribution. There will be two types of distributions: unimodal and bimodal. The occurrence of bimodal distributions is due to the fact that the signs of the DM are not identifiable. Hence, we obtain voxel-wise p-values for $k \in \{1, \dots, p\}$, by defining:

$$P_k = 2\mathbb{P}\big(t_{J-1} \geq | \, t_k \, | \, \big)$$

where $t_k = \min\left\{\frac{\hat{\mu}_{k,1}}{\hat{\sigma}_{k,1}}, \frac{\hat{\mu}_{k,2}}{\hat{\sigma}_{k,2}}\right\}$, $\hat{\mu}_{k,1}$ (resp. $\hat{\mu}_{k,2}$) and $\hat{\sigma}_{k,1}$ (resp. $\hat{\sigma}_{k,2}$) are the mean and standard deviation estimates of a mixed normal distribution. The *mixtools* package (Benaglia et al., 2009) in *R* includes EM-based procedures for estimating parameters from mixture distributions.

5 Simulation

5.1 Simulation Set-up

Here we describe a simulation study to investigate the efficacy of our approach. Assume that, for every subject $i \in \{1, \ldots, n\}$, the mediator vector $\mathbf{M}_i$ and the treatment X_i can be jointly simulated from an independent, identically distributed multivariate normal distribution with known mean and variance. In particular, let

$$\begin{pmatrix} \mathbf{M}_i^\intercal \\ X_i \end{pmatrix} \Bigg| \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim MVN\Big\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\Big\} \tag{23}$$

where $\boldsymbol{\mu} = \big((\boldsymbol{\mu}^{\mathbf{M}})^\intercal, \boldsymbol{\mu}^X\big)^\intercal$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^M & \boldsymbol{\Sigma}^{M,X} \\ \boldsymbol{\Sigma}^{X,M} & \Sigma^X \end{pmatrix}$. Here $\mathbf{M}_i = (m_1, \ldots, m_p)$ and $\boldsymbol{\Sigma}^{X,M}$ have dimensions $1 \times p$, $\boldsymbol{\mu}^{\mathbf{M}}$ and $\boldsymbol{\Sigma}^{M,X}$ have dimensions $p \times 1$, $\boldsymbol{\Sigma}^M$ has dimensions $p \times p$, and X_i, $\boldsymbol{\mu}^X$ and Σ^X are all scalar.

Conditioning on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we have

$$\big\{\mathbf{M}_i | X_i = x_i\big\} \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}), \tag{24}$$

where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}^M + \boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}(x_i - \mu^X)$, and $\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^M - \boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}\boldsymbol{\Sigma}^{X,M}$. From (3) :

$$\mathbb{E}(\mathbf{M}_i \mathbf{w}_1 | X_i = x_i) = \alpha_0 + \alpha_1 x_i.$$

Solving (3) and (24), we can write:

$$\begin{aligned} \alpha_0 &= \mathbf{w}_1[\boldsymbol{\mu}^M - \boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}\mu^X]; \\ \alpha_1 &= \mathbf{w}_1[\boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}]. \end{aligned} \tag{25}$$

Moreover,

$$\begin{aligned} \text{Var}(\mathbf{M}_i \mathbf{w}_1 | X_i = x_i) &= \boldsymbol{\sigma}_\eta \\ &= \mathbf{w}_1^\intercal \mathbf{Var}(\mathbf{M}_i | X_i = x_i)\mathbf{w}_1 \\ &= \mathbf{w}_1^\intercal \boldsymbol{\Sigma}^M - \boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}\boldsymbol{\Sigma}^{X,M}\mathbf{w}_1. \end{aligned}$$

Using these results we can outline the simulation process as follows:

1. Set the values for the mean $\begin{pmatrix} \boldsymbol{\mu}^M \\ \mu_x^M \end{pmatrix}$ and covariance $\begin{pmatrix} \boldsymbol{\Sigma}^M & \boldsymbol{\Sigma}^{M,X} \\ \boldsymbol{\Sigma}^{X,M} & \Sigma^X \end{pmatrix}$, and simulate n pairs of $(\mathbf{M}_i, X_i)$ according to (23) ;

2. Set the values for β_0, β_1, and γ_1, as well as $\mathbf{w}_1$. Compute α_0 and α_1 using (25) . Consider these to be the true path coefficients $\boldsymbol{\theta}_0$ and the first direction of mediation $\mathbf{w}_1$;

3. Simulate random error ϵ_i from a normal distribution with known mean and variance. Given $(\mathbf{M}_i, X_i)$, ϵ_i , and the path coefficients, generate Y_i, $i = 1, \ldots n$, according to (4) .

The generated data $\mathbf{D} = \{(X_i, \mathbf{M}_i, Y_i)\}_{i=1}^n$ from Steps 1 and 3 are used as input in the LSEM. The outputs of the algorithm are then compared with the true parameters.

We performed three sets of simulations.

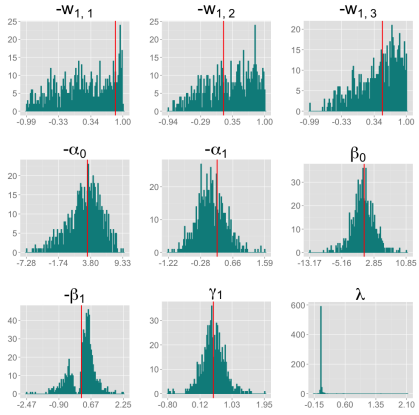
Simulation 1. Let $p = 3$, $\mathbf{w}_0 = (0.85, 0.17, 0.51)$, $\left((\boldsymbol{\mu}^{\mathbf{M}})^\intercal, \mu^X\right)^\intercal = (2, 3, 4, 5)$, $\boldsymbol{\Sigma}^{M,X} = (0.60, -0.90, 0.35)^\intercal$, and $\Sigma^X = 2.65$. Set the true path coefficients $(\beta_0, \beta_1, \gamma_1)$ equal to $(0.4, 0.2, 0.5)$. From (25) it follows that $(\alpha_0, \alpha_1) = (3.23, 0.20)$. Assuming $\epsilon_i \sim N(0, 1)$, we simulated $\{X_i, Y_i, \mathbf{M}_i\}_{i=1}^n$, with $n = 10, 100, 500$, and $1,000$. Each set of simulations was repeated $1,000$ times, and the parameter estimates were recorded.

Simulation 2. Let $p = 10$, $\mathbf{w}_0 = (0.42, 0.09, 0.25, 0.42, 0.17, 0.34, 0.51, 0.17, 0.17, 0.34)$, $\left((\boldsymbol{\mu}^{\mathbf{M}})^\intercal, \mu^X\right)^\intercal = (2, 3, 4, 5, 4, 6, 2, 5, 8, 1, 3)$, $\boldsymbol{\Sigma}^{M,X} = (-1.48, -0.51, -0.81, 0.98, -1.21, 0.53, -0.66, -0.73, -1.00, 0.29)^\intercal$, and $\Sigma^X = 5.10$. Set the true pathway coefficients $(\beta_0, \beta_1, \gamma_1)$ to $(0.4, 0.2, 0.5)$. From (25) it follows that $(\alpha_0, \alpha_1) = (11.08, -0.20)$. Assuming $\epsilon_i \sim N(0, 1)$, we simulated $\{X_i, Y_i, \mathbf{M}_i\}_{i=1}^n$, with $n = 100$, and $1,000$. Each set of simulations was repeated $1,000$ times, and the parameter estimates were recorded.

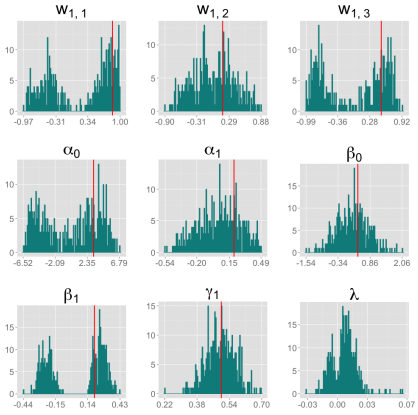
Simulation 3. Data are generated under the null hypothesis $\mathbf{w} = 0$, i.e. $\mathbf{Y}$ is generated assuming no mediation effect. The number of trials and voxels are chosen to match those in the experimental data. Let $\mathbf{X}$ be a vector of length $1,149$ taking values in the range $[36, 48.5]$. Let $(\beta_0, \gamma_1) = (-15, 0.5)$ and $\epsilon_i \sim N(0, 0.5)$. Generate $\mathbf{Y}_i$ according to (4) with $\mathbf{w} = 0$, and let

$\mathbf{M}_i[j] \sim N(m_i, s_i)$, where $m_i \sim N(2,5)$ and $s_i \sim N(20,5)$. Here $\mathbf{M}_i[j]$ represents the simulated value of the j^{th} voxel of trial i. Using the technique introduced in Section 4, we obtain p-values for the estimated DM from the bootstrap distribution for each voxel.

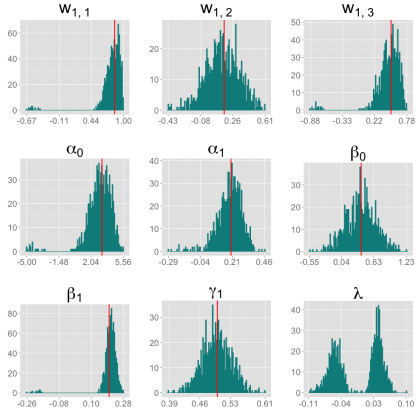
5.2 Simulation Results



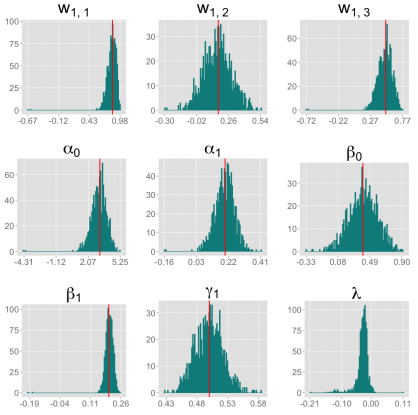
(a) 10 by 3



(b) 100 by 3



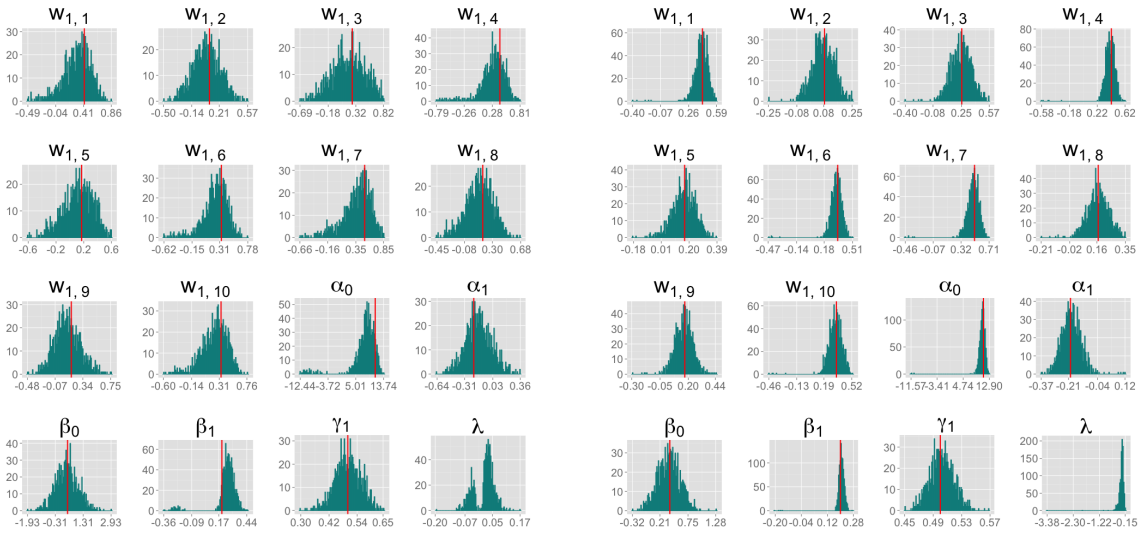
(c) 500 by 3



(d) 1000 by 3

Figure 3: Results for $p = 3$, when we increase sample size from 10 to 1,000 while keeping the ground truth values of $\mathbf{w}$ and $\boldsymbol{\theta} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_1)$ fixed. Red lines indicate truth.

Figures 3 and 4 show the results of Simulations 1 and 2. Figure 3 a-e display results for the case when $p = 3$, and the sample size is 10, 100, 500, and $1,000$, respectively. Figure 4 a-b display results for $p = 10$, and the sample size is 100 and $1,000$. As the sample size increases,



(a) 100 by 10 (b) 1000 by 10

Figure 4: Results for $p = 10$, when we increase sample size from 100 to 1,000 while keeping the ground truth values of $\mathbf{w}$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_1)$ fixed. Red lines indicate truth.

the estimates become more accurate, while the distribution becomes increasingly normal with a smaller standard deviation. The sign of the estimator is difficult to determine for smaller samples sizes, but becomes more consistent as the sample size increases.

		p	
		3	10
n	10	694	—
	100	387	923
	300	633	984
	500	897	1,000
	1,000	1,000	1,000

Table 1: The turn-out rate for different n and p combinations per 1,000 Simulations

Moreover, for fixed p, the turn-out rate (the number of estimating results an algorithm produces out of a fixed number of simulations) increases with n; see Table I. For fixed n, the turn-out rate improves with increasing p. The reason why some runs do not produce a result is that the function $\lambda(\boldsymbol{\theta})$ is not well behaved in small sample sizes, and the Newton-Raphson

optimization algorithm fails at one of the intermediary steps. When p is sufficiently large or high dimensional, the algorithm seems to improve. If $p \sim 3$, the algorithm runs better when we have sufficiently large sample size $n \sim 300$. Performance of the algorithm improves with more refined grid points, but this comes at the expense of computational efficiency.

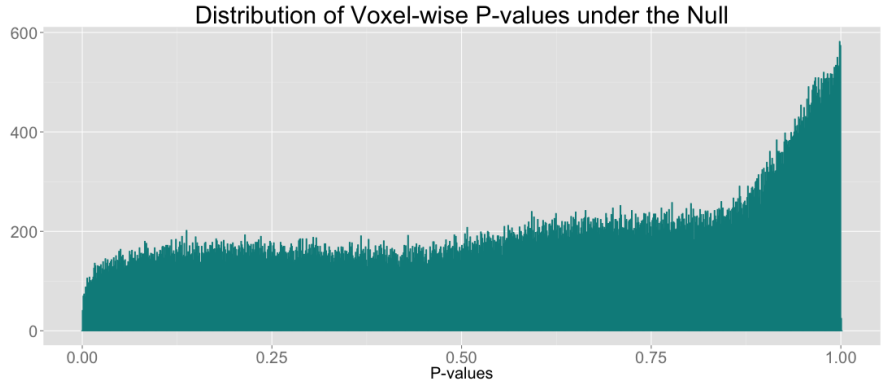


Figure 5: A histogram of voxel-wise p-values under the null that $\mathbf{w} = 0$, i.e. the response $\mathbf{Y}$ is generated only from the treatment effect $\mathbf{X}$ and measurement error assuming there is no mediation effect. Notice the distribution is approximately uniformly distributed. The heavy right tail is caused by the conservative method we used in estimating p-values. About 2.87% voxel-wise p-values are significant at the $\alpha = 0.05$ significance level.

Finally, the results of Simulation 3 are shown in Figure 5. The voxel-wise p-values are roughly uniformly distributed in the left tail, while the heavy right tail is likely caused by the conservative method used for estimating p-values. In addition, less than $100\alpha\%$ voxel-wise p-values are significant at the α significance level, suggesting that our approach provides adequate control of the false positive rate in the null setting.

6 Application to Data from an fMRI Study of Thermal Pain

Recall from Section 2 that the data structure is $\{X_{ij}, Y_{ij}, \mathbf{M}_{ij}\}$, where X_{ij} and Y_{ij} are the temperature and pain rating, respectively, assigned on trial j to subject i, and $\mathbf{M}_{ij} = (M_{ij1}, M_{ij2}, \ldots M_{ijp})$ is the whole-brain activation measured over p voxels. Note that $i \in \{1, \ldots, I\}$ and $j \in$

$\{1, \ldots, J_i\}$, where $I = 33$ and J_i takes subject-specific values between $58 - 75$.

The data was arranged in a matrix $\mathbf{M}$ of dimension $1,149 \times 206,777$, where each row consists of activation from a single trial on a single subject over $206,777$ voxels, and each column is voxel-specific. In particular, rows $1 - 48$ correspond to subject 1, rows $49 - 94$ correspond to subject 2, etc. The temperature level and reported pain are represented as the vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively, both of length $1,149$. The first DM corresponding to $(\mathbf{X}, \mathbf{Y}, \mathbf{M})$, is a vector of length $206,777$, whose estimation is computationally infeasible without first performing data reduction. Hence, we use the GPVD approach outlined in Section 3.6.2.

We choose $\tilde{\mathbf{w}}$ to have dimension 35, so that the number of rows of $\mathbf{D}$ is less than or equal to the minimum number of trials per subject, and this number ensures that 80% of the total variability of $\mathbf{M}$ is explained after dimension reduction. Furthermore, estimating a DM of length 35 is computationally feasible using our algorithm. The population-specific matrix $\mathbf{D}$ of dimension $35 \times 206,777$ was obtained according to (18), and the lower dimensional mediation matrix $\tilde{\mathbf{M}}$ of dimension $1,149 \times 35$, according to (19). The terms $(\mathbf{X}, \mathbf{Y}, \tilde{\mathbf{M}})$ were placed into the algorithm outlines in (8) - (10), using starting values $\boldsymbol{\theta}_1^{(0)} = 0.1 \times \mathbf{J}_5$, and $\mathbf{w}_1^{(0)} = 0.1 \times \mathbf{J}_{35}$. Finally, $\hat{\mathbf{w}}$, of length $206,777$, was computed using (22).

We compute the first two DMs and obtained estimates of $\hat{\boldsymbol{\theta}}_1 = (-3770, 96.31, -13.9, 0.00075, 0.40)$ and $\hat{\boldsymbol{\theta}}_2 = (-638.9 - 23.18 - 13.86, 0.00075, -1.19e - 07, 0.40)$. Figure 6 shows the weight maps for the first and second Directions of Mediation, thresholded using FDR correction with $q = 0.05$, separated according to whether the weight values were positive or negative.

The map is consistent with regions typically considered active in pain research, but also reveals some interesting structure that has not been uncovered by previous methods. The first direction of mediation shows positive weights on both targets of ascending nociceptive (pain-related) pathways, including the anterior cingulate, mid-insula, posterior insula, parietal operculum/S2, the approximate hand area of S1, and cerebellum. Negative weights were found in areas

often anti-correlated with pain, including parts of the lateral prefrontal cortex, parahippocampal cortex, and ventral caudate, and other regions including anterior frontal cortex, temporal cortex, and precuneus. These are associated with distinct classes of functions other than physical pain and are not thought to contain nociceptive neurons, but are still thought to play a role in mediating pain by processing elements of the context in which the pain occurs.

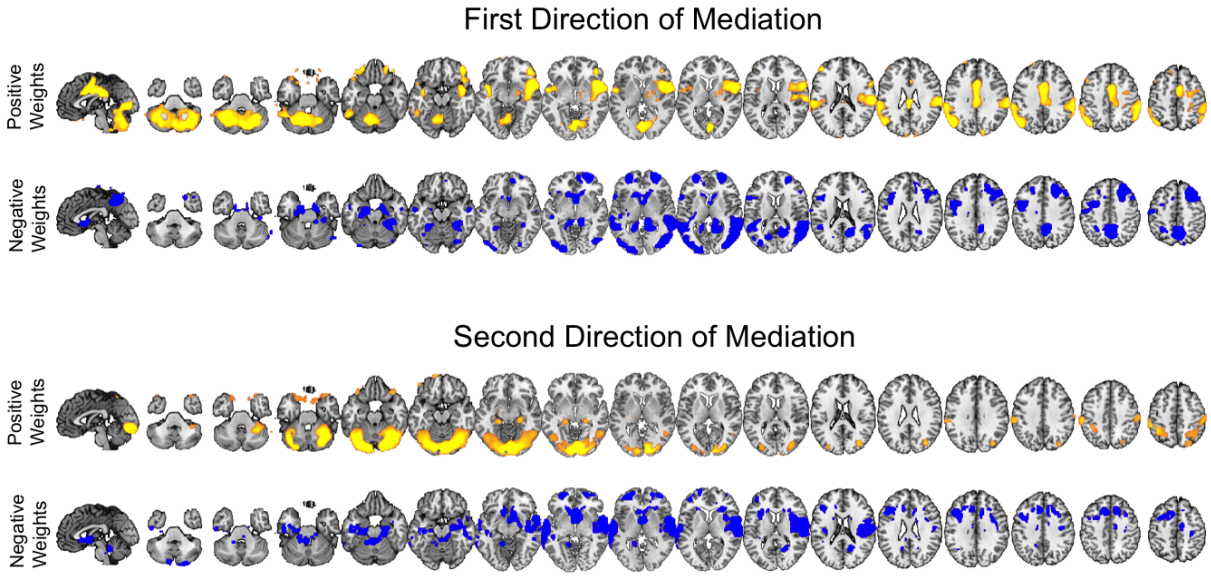


Figure 6: Weight maps for the first and second Directions of Mediation fit using data from the fMRI study of thermal pain. (A)-(B) Significant weights with positive and negative values, respectively, for the first DM. (C)-(D) Significant weights with positive and negative values, respectively, for the second DM. All maps are thresholded using FDR correction with $q = 0.05$.

The second direction of mediation is interesting because it also contains some nociceptive targets and other, non-nociceptive regions that partially overlap with and are partially distinct from the first direction. This component splits nociceptive regions, with positive weights on S1 and negative weights on the parietal operculum/S2 and amygdala, possibly revealing dynamics of variation among pain processing regions once the first direction of mediation is accounted for.

Positive weights are found on visual and superior cerebellar regions and parts of the hippocampus, and negative weights on the nucleus accumbens/ventral striatum and parts of dorsolateral and superior prefrontal cortex. The latter often correlate negatively with pain.

7 Discussion

This paper addresses the problem of mediation analysis in the high-dimensional setting. The first direction of mediation is the linear combination of the elements of a vector of potential mediators that maximizes the likelihood of the SEM. Subsequent directions can be found that maximizes the likelihood of the SEM conditional on being orthogonal to previous directions.

An interesting property of the DM framework is that the signs of the estimates are unidentifiable. To address this issue, there are two possible solutions. First, we can use Bayesian methods to apply a sign constraint based on prior knowledge. Second, if the magnitude of the voxel-wise mediation effect is of interest, we can consider a non-negativity constraint. This can be necessary because, under some circumstances, the coexistence of positive and negative elements of $\mathbf{w}$ might cancel out potential mediation effects. For example, assume $\mathbf{M} = (0.5, 0.4, 0.9)$ and $\mathbf{w} = (0.577, 0.577, -0.577)^\intercal$. Then $\mathbf{Mw} = 0$, making the estimate of β_1 unavailable. It, however, does not necessarily imply the non-existence of a mediation effect. In these circumstances, it is advantageous to impose a non-negativity constraint on $\mathbf{w}$ by choosing an injective mapping $\mathbf{w}_1 : \bar{\mathbb{R}}^p \mapsto [0, 1]$. For example, consider $\mathbf{w}_1(\mathbf{Z}) = \dfrac{\exp(\mathbf{Z})}{1 + \exp(\mathbf{Z})}$, where $\mathbf{Z} \in \bar{\mathbb{R}}^p$.

In many practical situations, the response $\mathbf{Y}$ and the mediator $\mathbf{M}$ are not necessarily normally distributed, but instead follow some distribution from the exponential family. It can be shown that we can estimate both the DMs and path coefficients under the exponential family setting using a GEE-like method. Essentially, conditioning on the DM, the direct and indirect causal pathway coefficient can be estimated using two sets of GEEs. The DM can then be estimated conditioning on the estimated pathway coefficients.

Finally, it may also be of interest to estimate an analogue to the DM and corresponding pathway coefficients that jointly maximize the indirect effect. This can be done in a similar fashion as we estimate the DM in the exponential family setting above. Future work will be to implement the aforementioned approaches.

8 Appendix

Here we provide the regularity conditions need for proving **Theorem 1** and **Theorem 2**. Detailed proofs can be found in the supplemental material. Let $\mathbf{D} = (\mathbf{X}, \mathbf{Y}, \mathbf{M})$ be a data triple, where $\mathbf{X} = (X_1, \ldots, X_n)^\intercal \in \mathbb{R}^n$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\intercal \in \mathbb{R}^n$, and $\mathbf{M} = (M_1, \ldots, M_n)^\intercal \in \mathbb{R}^{n\times p}$. Let $\mathbf{w} \in \mathbb{R}^p$ and $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma) \in \mathbb{R}^5$, be the parameters of interest. In particular, $\mathbf{w}$ maps $\mathbf{M}$ onto $\mathbb{R}^n$. Let $\lambda \in \mathbb{R}^1$ be a nuisance parameter. Consider the joint log-likelihood function $g(\cdot; \mathbf{w}, \boldsymbol{\theta})$ in (3.3) as the objective function.

Define the profiled Lagrangian $L(\mathbf{D}; \boldsymbol{\theta}) = g(\mathbf{D}; \boldsymbol{\theta}) + \lambda(\boldsymbol{\theta})\big(\mathbf{w}^\intercal(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta}) - 1\big)$ and $L(d; \boldsymbol{\theta}) = g(d; \boldsymbol{\theta}) + \dfrac{\lambda(\boldsymbol{\theta})\big(\mathbf{w}^\intercal(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta}) - 1\big)}{n}$, where $L(d; \boldsymbol{\theta}_0) \in \mathcal{P} = \{L(d; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is some properly defined space in $\mathbb{R}^5$. Define $\dot{L}(\mathbf{D}; \boldsymbol{\theta}) := \dfrac{\partial L}{\partial \boldsymbol{\theta}}(\mathbf{D}; \boldsymbol{\theta}) = \sum_{i=1}^n \ell(d_i, \boldsymbol{\theta})$, where $\ell(d, \boldsymbol{\theta}) = \dfrac{\partial g}{\partial \boldsymbol{\theta}}(d; \boldsymbol{\theta}) + \dfrac{\nabla^{\boldsymbol{\theta}}\lambda(\boldsymbol{\theta})[\mathbf{w}^\intercal(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta}) - 1] + 2\lambda(\boldsymbol{\theta})\nabla^{\boldsymbol{\theta}}\mathbf{w}(\boldsymbol{\theta})}{n}$. Further define $\hat{q}(\mathbf{D}; \boldsymbol{\theta}) = \dfrac{1}{n}\sum_{i=1}^n L(D_i; \boldsymbol{\theta})$, $q_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(L(d; \boldsymbol{\theta}))$, $\hat{\dot{q}}(\mathbf{D}; \boldsymbol{\theta}) = \dfrac{1}{n}\sum_{i=1}^n \ell(D_i; \boldsymbol{\theta})$, and $\dot{q}_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(\ell(d; \boldsymbol{\theta}))$.

Regularity Conditions II :

(N-1) $\boldsymbol{\theta}_0$ is in the interior of $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^5$;

(N-2) $q_0(\boldsymbol{\theta}) = 0$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;

(N-3) $L(d; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for all $d \in \mathcal{D}$. In particular, $g(\cdot; \boldsymbol{\theta})$, $\lambda(\boldsymbol{\theta})$, and $\mathbf{w}(\boldsymbol{\theta})$ are continuous;

(N-4) $\| L(d; \boldsymbol{\theta}) \| \leq d_0(d)$, $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbb{E}_{\boldsymbol{\theta}_0}[d_0(d)] < \infty$;

(N-5) $\mathbb{E}_{\boldsymbol{\theta}_0}\big(\dfrac{\partial g}{\partial \boldsymbol{\theta}}(d;\boldsymbol{\theta})\big) = 0$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;

(N-6) $\mathbb{E}_{\boldsymbol{\theta}_0}\left\{ \dfrac{\dfrac{\partial}{\partial \boldsymbol{\theta}}\{\lambda(\boldsymbol{\theta})[\mathbf{w}^\intercal(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta}) - 1]\}}{n} \right\} = o_p(n^{-1/2})$;

(N-7) $\ell(d, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for all $d \in \mathcal{D}$;

(N-8) $\parallel \ell(d, \boldsymbol{\theta}) \parallel \leq d_1(d)$, $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbb{E}_{\boldsymbol{\theta}_0}[d_1(d)] < \infty$;

(N-9) $\ell(d, \boldsymbol{\theta})$ is continuously differentiable in $\mathcal{N}_r(\boldsymbol{\theta}_0)$, where $\mathcal{N}_r(\boldsymbol{\theta}_0)$ is a r neighborhood of $\boldsymbol{\theta}_0$, $\mathcal{N}_r(\boldsymbol{\theta}_0) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < r\}$;

(N-10) $\parallel \dfrac{\partial \ell}{\partial \boldsymbol{\theta}}(d, \boldsymbol{\theta}) \parallel \leq d_2(d)$, $\forall \boldsymbol{\theta} \in \mathcal{N}_r(\boldsymbol{\theta}_0)$, and $\mathbb{E}_{\boldsymbol{\theta}_0}[d_2(d)] < \infty$;

(N-11) $D(\boldsymbol{\theta}_0)$ is non-singular, where $D_0(\boldsymbol{\theta}_0) := \mathbb{E}_{\boldsymbol{\theta}_0}[\dfrac{\partial \ell}{\partial \boldsymbol{\theta}}(d, \boldsymbol{\theta})]$;

(N-12) $B(\boldsymbol{\theta}_0) := \mathbb{E}_{\boldsymbol{\theta}_0}[\ell(d, \boldsymbol{\theta})\ell^\intercal(d, \boldsymbol{\theta})]$ exists;

Regularity Conditions III:

(N-13) $\lambda(\boldsymbol{\theta})$ and $\mathbf{w}(\boldsymbol{\theta})$ are continuously differentiable in $\mathcal{N}_r(\boldsymbol{\theta}_0)$.

Acknowledgement

This research was partially supported by NIH grants R01EB016061, R01DA035484 and P41 EB015909, as well as NSF grant 0631637. The authors would like to thank Tianchen Qian of Johns Hopkins Bloomberg School of Public Health (JHSPH) for his insightful comments on deriving the asymptotic property of the estimates, and Stephen Cristiano, Bin He, and Shen Xu of JHSPH for their valuable suggestions.

References and Notes

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in medicine*, 27(8):1282–1304.

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455.

Apkarian, A. V., Bushnell, M. C., Treede, R.-D., and Zubieta, J.-K. (2005). Human brain mechanisms of pain perception and regulation in health and disease. *European Journal of Pain*, 9(4):463–463.

Atlas, L. Y., Lindquist, M. A., Bolger, N., and Wager, T. D. (2014). Brain mediators of the effects of noxious heat on pain. *PAIN®*, 155(8):1632–1648.

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

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bushnell, M. C., Čeko, M., and Low, L. A. (2013). Cognitive and emotional control of pain and its disruption in chronic pain. *Nature Reviews Neuroscience*, 14(7):502–511.

Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M., and Punjabi, N. M. (2011). Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association*, 106(495).

Holland, P. (1988). Causal inference, path analysis and recursive structural equation models (with discussion). *Sociological Methodology*, 18:449–493.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, 13(4):314.

Krishnan, A., Williams, L. J., McIntosh, A. R., and Abdi, H. (2011). Partial least squares (pls) methods for neuroimaging: a tutorial and review. *Neuroimage*, 56(2):455–475.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., and Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23:439–464.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309.

Lindquist, M. A., Spicer, J., Asllani, I., and Wager, T. D. (2012). Estimating and testing variance components in a multi-level glm. *NeuroImage*, 59(1):490–501.

MacKinnon, D. P. (2008). Mediation analysis. *The Encyclopedia of Clinical Psychology*.

Mumford, J. A., Turner, B. O., Ashby, F. G., and Poldrack, R. A. (2012). Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–2643.

Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872.

Ogburn, E. L. (2012). Commentary on" mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables" by dylan small. *Journal of statistical research*, 46(2):105.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459.

Preacher, K. J. and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav res methods*, 40(3):879–891.

Rissman, J., Gazzaley, A., and D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, 23(2):752–763.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology*, 3:143–155.

Shrout, P. E. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological methods*, 7(4):422.

Sobel, M. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In Leinhardt, S., editor, *Sociological Methodology*, pages 290–312. Washington DC: American Sociological Association.

Sobel, M. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, 33:230–251.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, 63(3):926–934.

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

VanderWeele, T. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2:457–468.

Wager, T., Davidson, M., Hughes, B., Lindquist, M., and Ochsner, K. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, 59:1037–1050.

Wager, T., van Ast, V., Davidson, M., Lindquist, M., and Ochsner, K. (2009a). Brain mediators of cardiovascular responses to social threat, Part II: Prefrontal subcortical pathways and relationship with anxiety. *NeuroImage*, 47:836–851.

Wager, T., Waugh, C., Lindquist, M., Noll, D., Fredrickson, B., and Taylor, S. (2009b). Brain mediators of cardiovascular responses to social threat, Part I: Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. *NeuroImage*, 47:821–835.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fmri-based neurologic signature of physical pain. *New England Journal of Medicine*, 368(15):1388–1397.

Wold, H. (1982). Soft modelling: the basic design and some extensions. *Systems under indirect observation, Part II*, pages 36–37.

Wold, H. (1985). Partial least squares. *Encyclopedia of statistical sciences*.

Woo, C., Roy, M., Buhle, J., and Wager, T. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biology*, 13(1).

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20(4).