# 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 is an important tool in the behavioral sciences for investigating the role of intermediate variables that lie in the path between a treatment and an outcome variable. The influence of the intermediate variable on the outcome is often explored using a linear structural equation model (LSEM), with model coefficients interpreted as possible effects. While there has been significant research on the topic, little work has been done when the intermediate variable (mediator) is a high-dimensional vector. In this work we introduce a novel method for identifying potential mediators in this setting called the directions of mediation (DMs). DMs linearly combine potential mediators into a smaller number of orthogonal components, with components ranked by the proportion of the LSEM likelihood (assuming normally distributed errors) each accounts for. 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 $X$ on an outcome $Y$ is at least partially directed through an intervening variable $M$, then $M$ is said to be a mediator. The three-variable path diagram shown in Figure 1 illustrates this relationship. The influence of the intermediate variable on the outcome is frequently ascertained using linear structural equation models (LSEMs), with the model coefficients interpreted as causal effects; see below for discussion of the assumptions under which this interpretation is warranted. Typically, interest centers on parsing the effects of the treatment on the outcome into separable direct and indirect effects, representing the influence of $X$ on $Y$ unmediated and mediated by $M$, respectively.

To date most research in mediation analysis has been devoted to the case of a single mediator, with some attention given to the case of multiple mediators (e.g., Preacher and Hayes (2008); VanderWeele and Vansteelandt (2013)). However, high dimensional mediation has received scarce attention. Recent years have seen a tremendous increase of new applications measuring massive numbers of variables, including brain imaging, genetics, epidemiology, and public health studies. It has therefore become increasingly important to develop methods to deal with mediation in the high-dimensional setting, i.e., when the number of mediators is

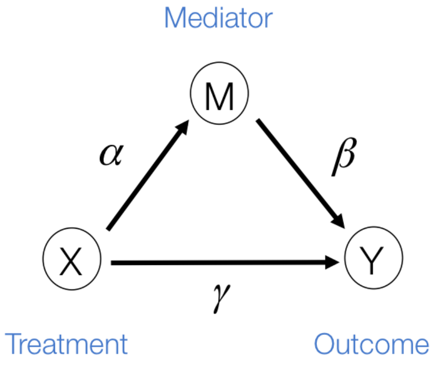


Figure 1: The three-variable path diagram representing the standard mediation framework. The variables corresponding to $X$, $Y$, and $M$ are all scalars, as are the path coefficients $\alpha$, $\beta$, and $\gamma$.

much larger than the number of observations. Such an extension is the focus of this work. It is important to emphasize that even though we focus on high dimensional mediators in the context of LSEMs, the principles extend to any other model-based approach to mediation.

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 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 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 act 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, $\mathbf{M}$. Here both $X$ and $Y$ are scalars, while $\mathbf{M}$ is the estimated brain activity measured over a large number of different voxels/regions. We assume that the values of $\mathbf{M}$ 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. An early approach to mediation in neuroimaging (Caffo et al., 2008) took the route of re-expressing the multivariate images into targeted, simpler, composite summaries on which mediation analysis was performed. In contrast, the identification of univariate mediators on a voxel-wise basis 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 voxels, 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 may be maximally effective as mediators, i.e. those that are simultaneously most predictive of the outcome and predicted by the treatment.

Thus, in this work 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 $\mathbf{M} = (M^{(1)}, M^{(2)}, \dots M^{(p)})^\intercal \in \mathbb{R}^p$. While an LSEM can be used to estimate mediation effects (defined precisely below), in this setting there are too many mediators to allow reasonable interpretation (unless the model coefficients are highly structured) and there are many more mediators than subjects, precluding estimation using standard procedures. To overcome these problems, a new model, called the directions of mediation (DM) is developed. DM's linearly combine activity in different voxels into a smaller number of orthogonal components, with components ranked by the proportion of the LSEM likelihood (assuming normally distributed errors) each accounts for. Ideally, the components form a small number of uncorrelated mediators that represent interpretable networks of voxels. 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 define direct and indirect effects for the multiple mediator setting. In Section 3 we introduce the directions of mediation, and provide an estimation algorithm 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 the fMRI study of thermal pain.

## 2 A Mutivariate Causal Mediation Model

Let $X$ denote an exposure/treatment for a given subject (e.g., thermal pain), and $Y$ an outcome (e.g., reported pain). Suppose there are multiple mediators $\mathbf{M} = (M^{(1)}, \cdots M^{(p)})$ in the path between treatment and outcome; in the fMRI context, the mediators are $p$ dependent activations over the $p$ voxels. Here we assume for simplicity that each subject is scanned under one condition.

Using potential outcomes notation (Rubin (1974)), let $\mathbf{M}(x)$ denote the value of the mediators if treatment $X$ is set to $x$. Similarly, let $Y(x, \mathbf{m})$ denote the outcome if $X$ is set to $x$ and $\mathbf{M}$

is set to $\mathbf{m}$. The controlled unit direct effect of $x$ vs. $x^*$ is defined as $Y(x, \mathbf{m}) - Y(x^*, \mathbf{m})$, the natural unit direct effect as $Y(x, \mathbf{M}(x^*)) - Y(x^*, \mathbf{M}(x^*))$, and the natural unit indirect effect as $Y(x, \mathbf{M}(x)) - Y(x, \mathbf{M}(x^*))$. Note that for these nested counterfactuals to be well-defined it must be hypothetically possible to intervene on the mediator without affecting the treatment.

The total unit effect is the sum of the natural unit direct and unit indirect effects, i.e.

$$Y(x, \mathbf{M}(x)) - Y(x^*, \mathbf{M}(x^*)) = Y(x, \mathbf{M}(x)) - Y(x, \mathbf{M}(x^*)) + Y(x, \mathbf{M}(x^*)) - Y(x^*, \mathbf{M}(x^*)) \tag{1}$$

Note that the direct effect could also be defined as $Y(x, \mathbf{M}(x)) - Y(x^*, \mathbf{M}(x))$. In general, this would lead to a different decomposition of the total effect; however, as we consider linear models below, this is not of further concern. Suppose the following four assumptions hold for the set of mediators:

$$\begin{aligned} Y(x, \mathbf{M}(x)) &\perp\!\!\!\perp X \\ Y(x, m) &\perp\!\!\!\perp \mathbf{M}|X \\ \mathbf{M}(x) &\perp\!\!\!\perp X \\ Y(x, m) &\perp\!\!\!\perp \mathbf{M}(x^*). \end{aligned} \tag{2}$$

In words, these assumptions imply there is no confounding for the relationship between: (i) treatment $X$ and outcome $Y$; (ii) mediators $\mathbf{M}$ and outcome $Y$; (iii) treatment $X$ and mediators $\mathbf{M}$; and (iv) no confounding for the relationship between mediator and outcome that is affected by the treatment. See Robins and Richardson (2010) and Pearl (2014) for detailed discussion of these assumptions, and for a critical evaluation of these assumptions in the high-dimensional setting see Huang and Pan (2015). VanderWeele and Vansteelandt (2013) showed that under (2) the average direct and indirect effects are identified from the regression function for the observed data. Suppose then (2) and the following model for the observed data hold:

$$\begin{aligned} E(M^{(j)}|X=x) &= \alpha_0 + \alpha_j x \qquad \text{for} \quad j=1,\ldots,p \\ E(Y|X=x,\mathbf{M}=\mathbf{m}) &= \beta_0 + \gamma x + \beta_1 M^{(1)} + \beta_2 M^{(2)} + \cdots + \beta_p M^{(p)}. \end{aligned} \tag{3}$$

Note that this model encodes the assumptions of linear relations among treatment, mediators, and outcome and, importantly, the absence of any treatment-mediator interaction in the outcome regression. When the treatment interacts with one or more of the mediators, the LSEM framework considered in this paper is not appropriate for mediation analysis (Ogburn, 2012).

The average controlled direct effect, average natural direct effect and average indirect effect are expressed as follows:

$$E(Y(x,\mathbf{m}) - Y(x^*,\mathbf{m})) = \gamma(x-x^*) \tag{4}$$

$$E(Y(x,\mathbf{M}(x^*)) - Y(x^*,\mathbf{M}(x^*))) = \gamma(x-x^*) \tag{5}$$

$$E(Y(x,\mathbf{M}(x)) - Y(x,\mathbf{M}(x^*))) = (x-x^*)\sum_{j=1}^{p}\alpha_j\beta_j. \tag{6}$$

Note the average controlled direct effect and natural direct effect are equivalent whenever there is no treatment-mediator interaction, as is assumed throughout.

When the counterfactuals are well-defined and the assumptions in (2) hold, the right hand sides of (5) and (6) identify causal mediation effects. When one or more of the assumptions in (2) fail to hold, or if the counterfactuals are not well-defined, the right hand sides of (5) and (6) may still be used in exploratory analysis to help identify potential mediators. For example, they could identify linear combinations of voxels that correspond to specific brain functions, suggesting mediation through correlates of those brain functions. Throughout, for simplicity, we use "direct effect" and "indirect effect" to refer to the right hand sides of (5) and (6), respectively; we are agnostic throughout as to whether these expressions can be interpreted causally or should be taken as exploratory. Similarly, we use "mediator" agnostically to refer to variables that temporally follow treatment and precede outcome and potentially may lie on a causal pathway between them.

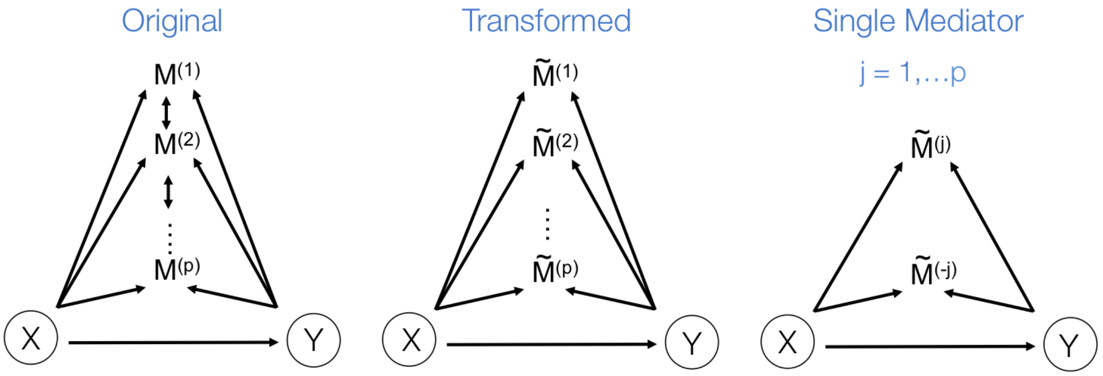


Figure 2: (Left) The three-variable path diagram used to represent multivariate mediation. Here the $p$ mediators are assumed to be correlated. (Center) A similar path diagram after an orthogonal transformation of the mediators. Now the $p$ mediators are independent of one another, allowing for the use of a series of LSEMs (Right), one for each transformed mediator, to estimate direct and indirect effects.

Fitting the system (3) is straightforward if the number of mediators is small. However, the estimates become unstable as $p$ increases, and in fMRI the number of mediators will greatly exceed the sample size. Therefore we seek an orthogonal transformation of the mediators. This both simplifies and stabilizes the parameter estimates in the model (3), allowing us to estimate the direct and indirect effects using a series of LSEMs, one for each transformed mediator; see Fig. 3 for an illustration. The novelty of our approach lies in choosing the transformation so that the transformed mediators are ranked by the proportion of the likelihood of the full LSEM that they account for. This has the benefit of potentially: (i) providing more interpretable mediators (i.e. linear combinations of voxels rather then individual voxels); and (ii) reducing the number of mediators needed to estimate the indirect effect.

# 3 Directions of Mediation

In this section we introduce a transformation of the space of mediators, determined by finding linear combinations of the original mediators that (i) are orthogonal; and (ii) are chosen to maximize the likelihood of the underlying three-variable SEM. We first formulate the model before introducing an estimation algorithm. We conclude with a discussion regarding estimation for the case when $p >> n$.

## 3.1 Model Formulation

Let $X_i$ and $Y_i$ denote univariate variables, and $\mathbf{M}_i = (M_i^{(1)}, M_i^{(2)}, \dots M_i^{(p)})^\intercal \in \mathbb{R}^p$, for $i = 1, \dots, n$. We denote the full dataset $\Delta = (\mathbf{x}, \mathbf{y}, \mathbf{M})$, where $\mathbf{x} = (X_1, \dots X_n)^\intercal \in \mathbb{R}^n$, $\mathbf{y} = (Y_1, \dots Y_n)^\intercal \in \mathbb{R}^n$, and $\mathbf{M} = (\mathbf{M}_1, \dots \mathbf{M}_n)^\intercal \in \mathbb{R}^{n \times p}$. Now let $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots \mathbf{w}_q) \in \mathbb{R}^{p \times q}$ be a linear transformation matrix, where $\mathbf{w}_d = (w_d^{(1)}, w_d^{(2)}, \dots w_d^{(p)})^\intercal \in \mathbb{R}^p$, for $d = 1, \dots, q$; and let $\tilde{\mathbf{M}} = \mathbf{M}\mathbf{W} = (\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_2, \dots \tilde{\mathbf{M}}_n)^\intercal$ where $\tilde{\mathbf{M}}_i = \mathbf{M}_i^\intercal \mathbf{W} = (\tilde{M}_i^{(1)}, \dots, \tilde{M}_i^{(d)}, \dots \tilde{M}_i^{(q)})^\intercal$ with $\tilde{M}_i^{(d)} = \mathbf{M}_i^\intercal \mathbf{w}_d = \sum_{k=1}^p M_i^{(k)} w_d^{(k)}$. We assume the relationship between the variables is given by the following LSEM:

$$\begin{aligned} \tilde{M}_i^{(j)} &= \alpha_0 + \alpha_j X_i + \epsilon_i \qquad \text{for} \quad j = 1, \dots, q \\ Y_i &= \beta_0 + \gamma X_i + \beta_1 \tilde{M}_i^{(1)} + \beta_2 \tilde{M}_i^{(2)} + \dots + + \beta_p \tilde{M}_i^{(q)} + \xi_i \end{aligned} \tag{7}$$

where $\epsilon_i$ and $\xi_i$ are i.i.d. bivariate normal with mean $0$ and variances $\sigma_\epsilon^2$ and $\sigma_\xi^2$. The parameters of the LSEM can be estimated using linear regression. However, under the additional condition that the new transformed variables $\tilde{M}^{(j)}$ are orthogonal, we can estimate the parameters separately for each $\tilde{M}^{(j)}$. Thus, for each $j = 1, \dots, q$ we can fit the following LSEM:

$$\begin{aligned} \tilde{M}_i^{(j)} &= \alpha_0 + \alpha_j X_i + \epsilon_i \\ Y_i &= \beta_0 + \gamma X_i + \beta_j \tilde{M}_i^{(j)} + \eta_i \end{aligned} \tag{8}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $\eta_i \sim N(0, \sigma_\eta^2)$, for $i = 1, \dots, n$.

Let $\boldsymbol{\theta} := (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma) \in \mathbb{R}^5$ be the parameter vector for the LSEM in (8) for $j = 1$. We seek to simultaneously estimate $\boldsymbol{\theta}$ and find the first direction of mediation (DM) $\mathbf{w}_1$, defined as the linear combination of the elements of $\mathbf{M}$ that maximizes the likelihood of the underlying LSEM. In our motivating example, $\mathbf{w}_1$ is a linear combination of the voxel activations. Thus, similar to principal components analysis (PCA) (Andersen et al. (1999)) or independent components analysis (ICA) (McKeown et al. (1997); Calhoun et al. (2001)) when applied to fMRI data, the weights can be mapped back onto the brain, with the resulting maps interpreted as coherent networks that together act as mediators of the relationship between treatment and outcome. Also like PCA, subsequent directions can be found that maximize the likelihood of the model, conditional on these being orthogonal to the previous directions.

To formalize, let $\mathscr{L}(\Delta; \mathbf{w}_1, \boldsymbol{\theta})$ be the joint likelihood of the SEM stated in (3). The *Directions of Mediation* are defined as follows:

*Step 1*: The $1^{st}$ DM is the vector $\mathbf{w}_1 \in \mathbb{R}^p$, with norm 1, that maximizes the conditional joint likelihood $\mathscr{L}(\Delta, \boldsymbol{\theta}; \mathbf{w}_1)$, i.e.

$$\hat{\mathbf{w}}_1 | \boldsymbol{\theta} = \text{argmax} \left\{ \mathscr{L}(\Delta, \boldsymbol{\theta}; \mathbf{w}_1) \right\},$$

subject to

$$\left\{ \mathbf{w}_1 \in \mathbb{R}^p : \|\mathbf{w}_1\|_2 = 1 \right\}.$$

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

$$\hat{\mathbf{w}}_2 | \boldsymbol{\theta}, \mathbf{w}_1 = \text{argmax} \left\{ \mathscr{L}(\Delta, \boldsymbol{\theta}, \mathbf{w}_1; \mathbf{w}) \right\}$$

subject to

$$\left\{ \mathbf{w}_2 \in \mathbb{R}^p : \|\mathbf{w}_2\|_2 = 1, \mathbf{w}_1 \mathbf{w}_2^\intercal = 0 \right\}.$$

$$\vdots$$

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

$$\hat{\mathbf{w}}_k | \boldsymbol{\theta}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1} = \operatorname{argmax} \left\{ \mathscr{L}(\Delta, \boldsymbol{\theta}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}; \mathbf{w}) \right\}$$

subject to

$$\left\{ \mathbf{w}_k \in \mathbb{R}^p : \|\mathbf{w}_k\|_2 = 1, \mathbf{w}_{k'} \mathbf{w}_k^\intercal = 0, \forall k' \in \{1, \ldots, k-1\} \right\}.$$

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

## 3.2 Estimation

Here we describe how to estimate the parameters associated with the first DM. Assuming joint normality, the joint $\log$ likelihood function for $\mathbf{w}_1$ and $\boldsymbol{\theta}$, $\mathscr{L}(\Delta; \mathbf{w}_1, \boldsymbol{\theta})$, can be expressed as:

$$\mathscr{L}(\Delta; \mathbf{w}_1, \boldsymbol{\theta}) \propto g_1(\Delta; \mathbf{w}_1, \boldsymbol{\theta}), \tag{9}$$

where $g_1(\Delta; \mathbf{w}_1, \boldsymbol{\theta}) \equiv -\left\{ \frac{1}{\sigma_\epsilon^2} \|\mathbf{y} - \beta_0 - \mathbf{x}\gamma_1 - \mathbf{M}\mathbf{w}_1\beta_1\|_2 + \frac{1}{\sigma_\eta^2} \|\mathbf{M}\mathbf{w}_1 - \alpha_0 - \mathbf{x}\alpha_1\|_2 \right\}$. The goal is to find both the parameters of the LSEM and the first DM that jointly maximize $g_1(\Delta; \mathbf{w}_1, \boldsymbol{\theta})$, under the constraint that the $L_2$ norm of $\mathbf{w}_1$ equals 1. Consider the Lagrangian

$$L(\Delta; \mathbf{w}_1, \boldsymbol{\theta}, \lambda) = g_1(\Delta; \mathbf{w}_1, \boldsymbol{\theta}) + \lambda(\|\mathbf{w}_1\|_2 - 1).$$

The dual problem can be expressed:

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

where $\lambda$ is the Lagrange multiplier. To solve this problem we propose a method where $\lambda$ is profiled out by one set of parameters of interest. We establish, under the assumption that the first partial derivatives of the objective function and the constraint function exist, the closed form solution for the path coefficients, the first DM, and $\lambda$ as follows:

$$\hat{\mathbf{w}}_1|\boldsymbol{\theta},\lambda \;=\; f_1(\Delta;\lambda,\boldsymbol{\theta}) \tag{10}$$

$$\hat{\lambda}|\boldsymbol{\theta} \;=\; \arg_{\lambda\in\mathbb{R}^1}\left\{f_2(\Delta;\lambda,\boldsymbol{\theta})=1\right\} \tag{11}$$

$$\hat{\boldsymbol{\theta}}|\hat{\mathbf{w}}_1,\hat{\lambda} \;=\; \underset{\boldsymbol{\theta}\in\mathbb{R}^5}{\operatorname{argmax}}\, L(\Delta;\hat{\mathbf{w}}_1,\boldsymbol{\theta},\hat{\lambda}) \tag{12}$$

where $f_1(\Delta;\lambda,\boldsymbol{\theta}) = (\lambda\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta})$; $f_2(\Delta;\lambda,\boldsymbol{\theta}) = \|(\lambda\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta})\|_2$, $\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{M}^\intercal\mathbf{M}\beta_1^2/\sigma_{\epsilon_1}^2 + \mathbf{M}^\intercal\mathbf{M}/\sigma_{\eta_1}^2$, and $\boldsymbol{\phi}(\boldsymbol{\theta}) = \mathbf{M}^\intercal(\alpha_0+\alpha_1\mathbf{x})/\sigma_{\eta_1}^2 + \mathbf{M}^\intercal(\mathbf{y}-\beta_0-\mathbf{x}\gamma_1)\beta_1/\sigma_{\epsilon_1}^2$. Using these results we outline an iterative procedure for jointly estimating the first direction of mediation and path parameters as described in Algorithm 1. Further, in the Supplemental Material we show that the estimated parameters are consistent and asymptotically normal (see Theorems 1 and 2).

**Algorithm 1** First DM

```
Step 0: Initiate $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}_1^{(0)}$.
Step 1: For each $k$, set:
    $\hat{\lambda}^{(k)}|\boldsymbol{\theta}_1^{(k)} = \arg_{\lambda\in\mathbb{R}^1}\left\{f_2(\Delta;\lambda,\boldsymbol{\theta}_1^{(k)})=1\right\}$   (13)
    $\hat{\mathbf{w}}_1^{(k)}|\boldsymbol{\theta}_1^{(k)},\hat{\lambda}^{(k)} = f_1(\Delta;\hat{\lambda}^{(k)},\boldsymbol{\theta}_1^{(k)})$   (14)
    $\hat{\boldsymbol{\theta}}_1^{(k+1)}|\hat{\mathbf{w}}_1^{(k)},\hat{\lambda}^{(k)} = \arg\max_{\boldsymbol{\theta}_1\in\mathbb{R}^5}\left\{L(\Delta;\hat{\mathbf{w}}_1^{(k)},\boldsymbol{\theta}_1^{(k)},\hat{\lambda}^{(k)})\right\}.$   (15)
Step 2: Repeat Step 1 until convergence; each time set $k=k+1$.
```

## 3.3 Higher Order Directions of Mediation

To estimate higher order DMs we investigated two alternative approaches. The first uses additional penalty parameters (one for each additional constraint), and the second 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 higher order DMs; thus we focus on this approach here. Using this approach, estimates of the $k^{\text{th}}$ direction of mediation, $\hat{\mathbf{w}}_k$, and the associated path coefficients, $\hat{\boldsymbol{\theta}}_k$, are obtained by computing:

$$(\hat{\mathbf{w}}_k, \hat{\boldsymbol{\theta}}_k)|\lambda = \operatorname{argmax} \Big\{ g_k(\Delta, \hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{k-1}; \mathbf{w}_k, \boldsymbol{\theta}_k) - \lambda\big( \|\mathbf{w}_k(\mathbf{x})\|_2 - 1\big) \Big\},$$

subject to

$$\Big\{ \boldsymbol{\theta}_k \in \mathbb{R}^{k+4}, \mathbf{x} \in \bar{\mathbb{R}}^p : \mathbf{w}_k(\mathbf{x}) := \mathbf{x} - \sum_{i=1}^{k-1} \text{Proj}_{\hat{\mathbf{w}}_i}(\mathbf{x}) \boldsymbol{\theta}_k \Big\}$$

where $\text{Proj}_{\hat{\mathbf{w}}_i}(\mathbf{x}) = \dfrac{\langle \mathbf{x}, \hat{\mathbf{w}}_i \rangle}{\langle \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_i \rangle} \hat{\mathbf{w}}_i$, $\forall i \in \{1, \ldots, k-1\}$. The performance of the projection approach is evaluated through extensive simulations in Section 5.

## 3.4 High-dimensional Directions of Mediation

The estimation procedure described in 3.2 works well in the low-dimensional setting, but becomes cumbersome as $p$ increases. Therefore it is critical to augment it with a matrix decomposition technique. Here we use a generalized version of Population Value Decomposition (PVD) (Caffo et al., 2010; Crainiceanu et al., 2011), which in contrast to Singular Value Decomposition (SVD) provides population-level information about $\mathbf{M}$. We begin by introducing the generalized version of PVD and thereafter illustrate its use in estimating the DMs. Throughout we assume that the data for each subject $i$ is stored in an $T_i \times p$ matrix, $\mathbf{M}_i$, whose $j^{\text{th}}$ row contains voxel-wise activity for the measurements of the $j^{\text{th}}$ trail for the $i^{\text{th}}$ subject. All $\mathbf{M}_i$ matrices are stacked vertically to form the $n \times p$ matrix $\mathbf{M}$, where $n = \sum_{i=1}^{N} T_i$.

### 3.4.1 Generalized PVD

The PVD framework assumes that the number of trials per subject is equal, which is not the case in many practical settings. To address this issue, we introduce Generalized Population Value Decomposition (GPVD), which allows the number of trials per subject to differ, while maintaining the dimension reduction benefits of the original. The GPVD of $\mathbf{M}_i$ is given by

$$\mathbf{M}_i = \mathbf{U}_i^B \tilde{\mathbf{V}}_i \mathbf{D} + \mathbf{E}_i, \tag{16}$$

where $\mathbf{U}_i^B$ is an $T_i \times B$ matrix, $\tilde{\mathbf{V}}_i$ is an $B \times B$ matrix of subject-specific coefficients, $\mathbf{D}$ is a $B \times p$ population-specific matrix, $\mathbf{E}_i$ is an $T_i \times p$ matrix of residuals. Here $B$ is chosen based upon a criteria such as total variance explained.

Below we introduce a step-by-step procedure for obtaining the GPVD.

**Step 1:** For each subject $i$, use SVD to compute: $\mathbf{M}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\intercal \approx \mathbf{U}_i^B \mathbf{\Sigma}_i^B (\mathbf{V}_i^B)^\intercal$ where $\mathbf{U}_i^B$ consists of the first $B$ columns of $\mathbf{U}_i$, $\mathbf{\Sigma}_i^B$ consists of the first $B$ diagonal elements of $\mathbf{\Sigma}_i$, and $\mathbf{V}_i^B$ consists of first $B$ columns of $\mathbf{V}_i$.

**Step 2:** Form the $p \times NB$ matrix $\mathbf{V} := [\mathbf{V}_1^B, \dots, \mathbf{V}_N^B]$. When $p$ is reasonably small, use SVD to compute the eigenvectors of $\mathbf{V}$. The $p \times B$ matrix $\mathbf{D}$ is obtained using the first $B$ eigenvectors. When $p$ is large, performing 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 before. Here $\mathbf{D}$ 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:** The GPVD in (16) can be summarized as follows:

$$\begin{aligned} \mathbf{M}_i &= \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\intercal \approx \mathbf{U}_i^B \mathbf{\Sigma}_i^B (\mathbf{V}_i^B)^\intercal \\ &\approx \mathbf{U}_i^B \underbrace{\{\mathbf{\Sigma}_i^B (\mathbf{V}_i^B)^\intercal \mathbf{D}^T\}}_{\tilde{V}_i} \mathbf{D} = \mathbf{U}_i^B \tilde{\mathbf{V}}_i \mathbf{D}, \end{aligned} \tag{17}$$

where $\mathbf{U}_i^B$, $\mathbf{\Sigma}_i^B$, and $\mathbf{V}_i^B$ are obtained from Step 1, and $\mathbf{D}$ from Step 2. The first approximation in (17) 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.

### 3.4.2 Estimation using GPVD

To estimate the DMs, perform GPVD on $\mathbf{M} = [\mathbf{M}_1^{\top}, \cdots, \mathbf{M}_n^{\top}]^{\top} = \left[(\mathbf{U}_1\tilde{\mathbf{V}}_1\mathbf{D})^{\top}, \cdots, (\mathbf{U}_n\tilde{\mathbf{V}}_n\mathbf{D})^{\top}\right]^{\top}$. Next, stack all $T_i \times B$ matrices $\mathbf{U}_i\tilde{\mathbf{V}}_i$ vertically to form an $n \times B$ matrix

$$\breve{\mathbf{M}} = \left[(\mathbf{U}_1\tilde{V}_1)^{\top}, \cdots, (\mathbf{U}_n\tilde{V}_n)^{\top}\right]^{\top} \tag{18}$$

Let $\breve{\mathbf{w}} = \mathbf{D}\mathbf{w}$, where $\breve{\mathbf{w}}$ is $B \times 1$. Finally, place $\breve{\mathbf{M}}$ and $\breve{\mathbf{w}}$ into (7). Since $\mathbf{D}$ can be obtained via GPVD, we can retrieve the original estimator of the high dimensional direction of mediation, $\hat{\mathbf{w}}$, via the generalized inverse, i.e.,

$$\hat{\mathbf{w}} = \mathbf{D}^{-}\breve{\mathbf{w}}^{est} \tag{19}$$

where $\breve{\mathbf{w}}^{est}$ is the estimated $\breve{\mathbf{w}}$ and $^{-}$ indicates the generalized inverse.

# 4 Inference

In low-dimensional settings, we can obtain variance estimates for the first DM and the path coefficients using *Theorems 1* and *2* from the Supplemental material. In high dimensional settings, variance estimation using the generalized inverse is under-estimated since the $\mathbf{D}$ obtained from (17) 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 $\mathbf{M} = \breve{\mathbf{M}}\mathbf{D}$, where $\mathbf{M}$ is $n \times p$, $\breve{\mathbf{M}}$ 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 $\breve{\mathbf{M}}$, stack them horizontally and form the $n \times B$ matrix $\breve{\mathbf{M}}_{(j)}$;

2. Obtain $\hat{\breve{\mathbf{w}}}_{(j)}$ from $\breve{\mathbf{M}}_{\ (j)}$, where $\hat{\breve{\mathbf{w}}}_{(j)}$ is the $j^{th}$ bootstrap DM of length $B$;

3. Obtain $\hat{\mathbf{w}}_{(j)} = \mathbf{D}^{-1}\hat{\breve{\mathbf{w}}}_{(j)}$, 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 $\hat{\mathbf{W}}^* = (\hat{\mathbf{w}}_{(1)}, \ldots, \hat{\mathbf{w}}_{(J)})^\intercal$, where $\hat{\mathbf{W}}^*$ is a $J \times p$ matrix.

Note the columns of $\hat{\mathbf{W}}^*$ are the bootstrap values of the DM corresponding to voxel $k$, 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 $k \in \{1, \ldots, p\}$, by defining:

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

where $t_k = \min\left\{\dfrac{\hat{\mu}_{k,1}}{\hat{\sigma}_{k,1}}, \dfrac{\hat{\mu}_{k,2}}{\hat{\sigma}_{k,2}}\right\}$, $\hat{\mu}_{k,1}$ (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 Study

## 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

$$\begin{pmatrix} \mathbf{M}_i \\ X_i \end{pmatrix} \Bigg| \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_{p+1}\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\big) \tag{20}$$

where $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^M \\ \mu^X \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^M & \boldsymbol{\Sigma}^{M,X} \\ \boldsymbol{\Sigma}^{X,M} & \Sigma^X \end{pmatrix}$. Conditioning on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we have

$$\big\{\mathbf{M}_i | X_i = x_i\big\} \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}), \tag{21}$$

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 (7) :

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

Solving (7) and (21), we can write:

$$\begin{aligned}\alpha_0 &= \mathbf{w}_1[\boldsymbol{\mu}^M - \boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}\mu^X];\\ \alpha_1 &= \mathbf{w}_1[\boldsymbol{\Sigma}^{M,X}[\Sigma^X]^{-1}]. \end{aligned} \tag{22}$$

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 $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and simulate $n$ pairs of $(\mathbf{M}_i, X_i)$ according to (20) ;

2. Set the values for $\beta_0, \beta_1$, and $\gamma_1$, as well as $\mathbf{w}_1$. Compute $\alpha_0$ and $\alpha_1$ using (22) . Consider these to be the true path coefficients $\boldsymbol{\theta}_1$ and the first direction of mediation $\mathbf{w}_1$;

3. Simulate random error $\epsilon_i$ from a normal distribution with known mean and variance. Given $(\mathbf{M}_i, X_i)$, $\epsilon_i$ , and the path coefficients, generate $Y_i$, $i = 1, \ldots n$, according to (7).

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

Below we outline the four simulation studies that were performed.

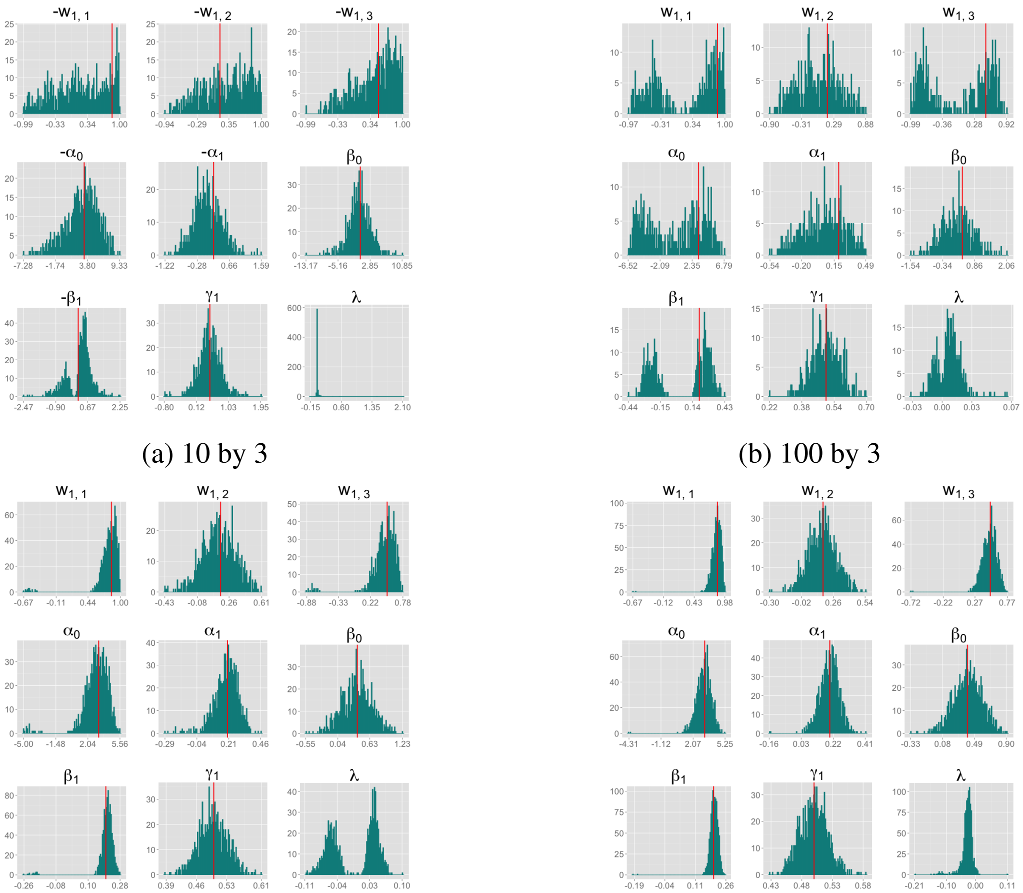
**Simulation 1.** Let $p = 3$, $\mathbf{w}_0 = (0.85, 0.17, 0.51)$, $\boldsymbol{\mu} = (2, 3, 4, 5)$, $\boldsymbol{\Sigma}^{M,X} = (0.60, -0.90, 0.35)^\intercal$, and $\Sigma^X = 2.65$. Set the true path coefficients $(\beta_0, \beta_1, \gamma_1)$ equal to $(0.4, 0.2, 0.5)$. From (22) it follows that $(\alpha_0, \alpha_1) = (3.23, 0.20)$. 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 $p = 10$, $\mathbf{w}_0 = (0.42, 0.09, 0.25, 0.42, 0.17, 0.34, 0.51, 0.17, 0.17, 0.34)$, $\boldsymbol{\mu} = (2, 3, 4, 5, 4, 6, 2, 5, 8, 1, 3)$, $\boldsymbol{\Sigma}^{M,X} = (-1.48, -0.51, -0.81, 0.98, -1.21, 0.53, -0.66, -0.73, -1.00, 0.29)^\intercal$, and $\Sigma^X = 5.10$. Set the true pathway coefficients $(\beta_0, \beta_1, \gamma_1)$ to $(0.4, 0.2, 0.5)$. From (22) 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. Consider $\mathbf{X}$, a vector of length $1,149$, that ranges between $[44.3, 49.3]$ (both values chosen to mimic the fMRI data studied in the next section). Consider $(\beta_0, \gamma_1) = (-15, 0.5)$ and $\epsilon_i \sim N(0, 0.5)$. Generate $\mathbf{Y}_i$ according to (7) with $\mathbf{w} = 0$, and let $M_i^{(j)} \sim N(m_i, s_i)$, where $m_i \sim N(2, 5)$ and $s_i \sim N(20, 5)$. Here $M_i^{(j)}$ 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. Fixing $\mathbf{X}$, we independently generate $(\mathbf{W}, \mathbf{Y})$ 100 times, each time obtaining voxel-specific p-values.

**Simulation 4.** Let $p = 10,000$ and $n = 1,000$. First simulate $\mathbf{X}$ from a truncated normal distribution $N^+(46.8, 2)$, truncated to take values in the range between $44.3$ and $49.3$. Next construct $\mathbf{M}$ under the assumption there are $1,000$ active and $9,000$ non-active voxels. This is achieved by simulating a vector of length $1,000$, corresponding to the active voxels, from a $N(1.5, 0.5)$ distribution, truncated to takes values in the range between $1$ and $2$. These values were placed between two vectors of zeros each of length $4,500$, corresponding to non-active voxels, giving a vector of voxel-wise activity of length $10,000$. Noise from a $N(0, 0.1)$ distribution was added to each voxel. This procedure was repeated for each of the $n$ subjects. Entries of $\mathbf{w}$ were set to weigh the voxels according to a Gaussian function, constrained to have norm $1$, centered at the middle voxel and designed to overlap in support with the $500$ centermost voxels. Finally, $\mathbf{Y}$ is simulated according to (8), where $(\beta_0, \gamma, \beta_1) = (-0.5, 0.12, 0.5)$ and $\eta_i \sim N(0, 0.5)$.

## 5.2 Simulation Results



(a) 10 by 3

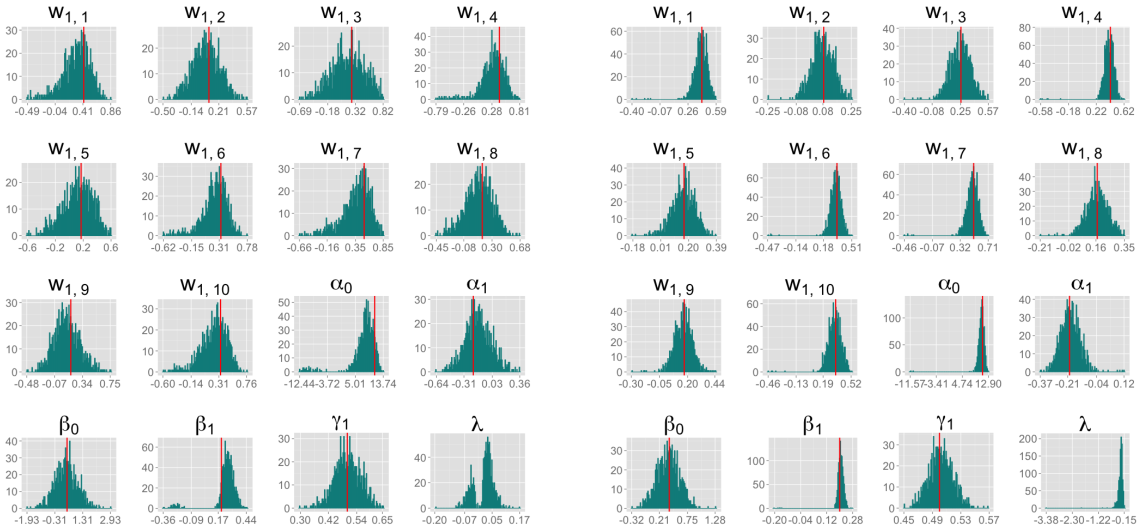
(b) 100 by 3

(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-d 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, 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.



(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.

| | | 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 (e.g., $n \sim 300$). Performance of the algorithm improves with more refined grid points, but this comes at the expense of computational efficiency.

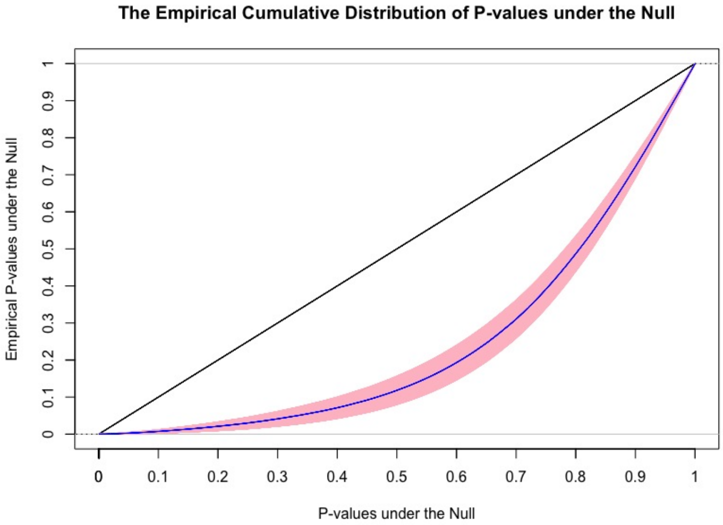


Figure 5: The empirical p-value plotted against the theoretical p-value. The straight line indicates exact correspondence between the two, and $95\%$ confidence bands are shown in pink.

The results of Simulation 3 are shown in Figure 5. Here the empirical p-values under the null, represented by the portion of voxels that fall below a certain threshold, are plotted against the theoretical p-values. $95\%$ confidence bounds are shown in pink. Clearly, the approach provides adequate control of the false positive rate in the null setting, albeit with somewhat over-conservative results. Finally, Fig. 6 shows bootstrap confidence bands for the estimated first direction of mediation from 100 bootstrap repetitions. Recall that the mediator is designed to to have $1,000$ active voxels. Clearly, the estimated first direction of mediation is consistent with the simulated signal.

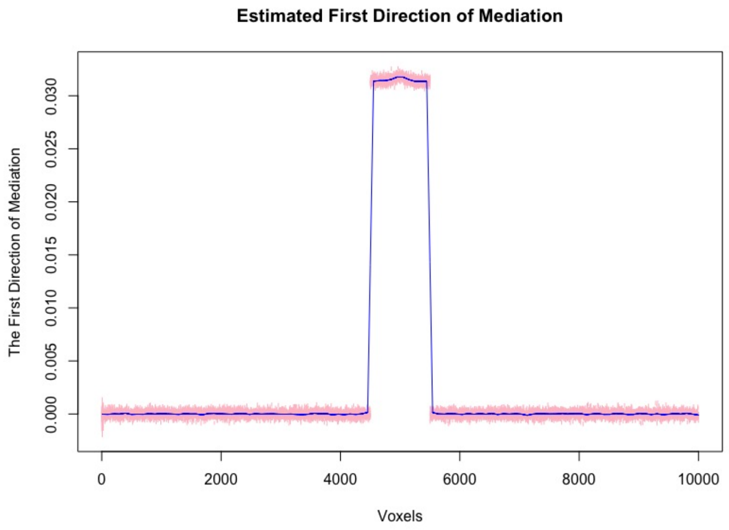


Figure 6: Bootstrap confidence bands of the estimated first direction of mediation computed from $100$ bootstrap repetitions. The simulated mediator is designed to show strong activations in the center $1,000$ voxels. The estimated first direction of mediation is consistent with the simulated signals (blue line).

# 6 An fMRI Study of Thermal Pain

## 6.1 Data Description

The data comes from the fMRI study of thermal pain described in the Introduction. A total of $33$ healthy, right-handed participants completed the study (age $27.9 \pm 9.0$ 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 $12.5$s, with $3$s ramp-up and $2$s ramp-down periods and $7.5$s 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 = $2000$ms, TE = $20$ms, field of view = $224$mm, $64 \times 64$ 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 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 $180$s 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.

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

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 $\mathbf{M}_{ij} = (M_{ij}^{(1)}, M_{ij}^{(2)}, \ldots M_{ij}^{(p)})^\intercal \in \mathbb{R}^p$ 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 $58 - 75$. 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. The temperature level and reported pain are represented as the vectors $\mathbf{x}$ and $\mathbf{y}$, respectively, both of length $1,149$.

## 6.2 Results

Each DM corresponding to $\Delta = (\mathbf{x}, \mathbf{y}, \mathbf{M})$, 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 3.4. We choose $\tilde{\mathbf{w}}$ to have dimension $B = 35$, to ensure that the number of rows of $\mathbf{D}$ is less than or equal to the minimum number of trials per subject. This value ensures that $80\%$ of the total variability of $\mathbf{M}$ is explained after dimension reduction. The population-specific matrix $\mathbf{D}$ of dimension $35 \times 206,777$ was obtained according to (17), and the lower dimensional mediation matrix $\tilde{\mathbf{M}}$ of dimension $1,149 \times 35$, according to (18). The terms $(\mathbf{x}, \mathbf{y}, \tilde{\mathbf{M}})$ were placed into the algorithm outlined in (13) - (15), 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 (19).

We compute the first three DMs and obtained estimates of $\hat{\boldsymbol{\theta}}_1 = (-3769.30, 96.32, -13.86, 0.00075, 0.40)$, $\hat{\boldsymbol{\theta}}_2 = (-695.85, -24.11, -13.86, 0.00075, -1.06 \times 10^{-7}, 0.40)$, and $\hat{\boldsymbol{\theta}}_3 = (1.35, -0.03, -13.86, 0.00075, -3.585 \times 10^{-7}, -5.5 \times 10^{-9}, 0.40)$. Figure 7 shows the weight maps for the first three Directions of Mediation, thresholded using FDR correction with $q = 0.05$, separated according to whether the weight values were positive or negative.

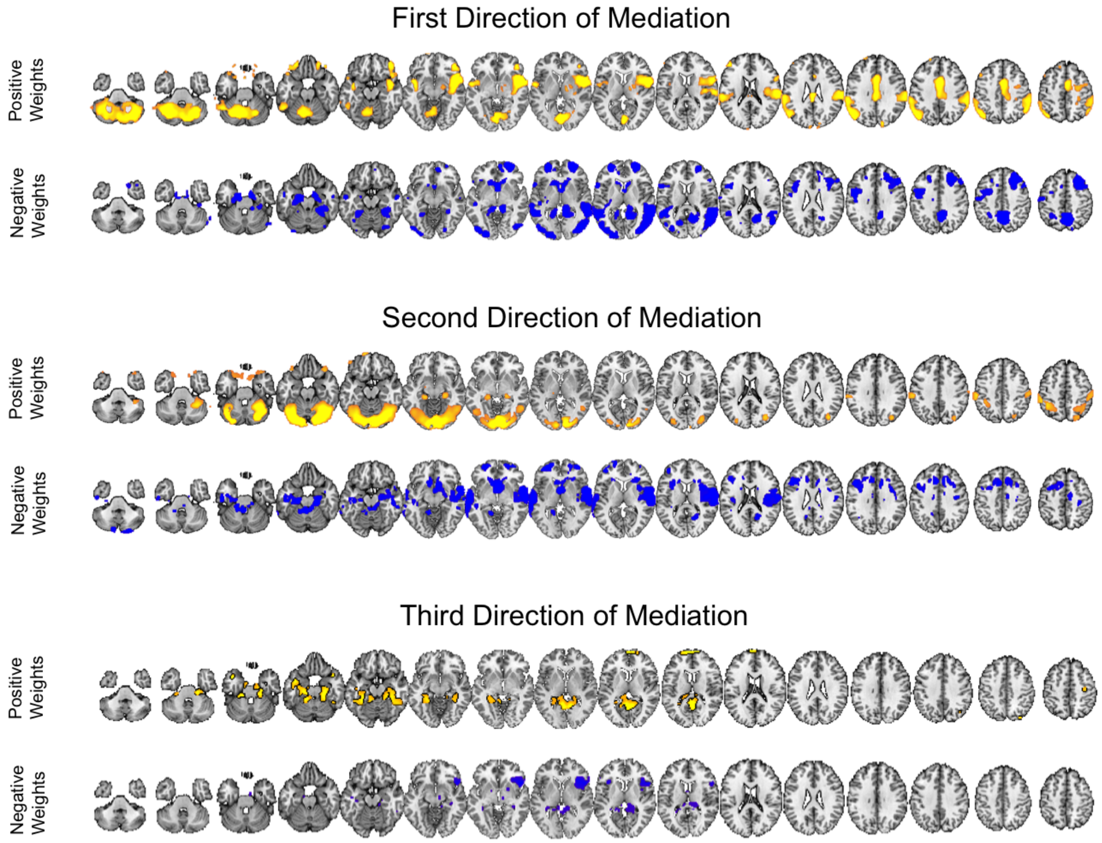


Figure 7: Weight maps for the first three Directions of Mediation fit using data from the fMRI study of thermal pain. Significant weights are separated into those with positive and negative values, respectively, for the each DM. All maps are thresholded using FDR correction with $q = 0.05$.

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.

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.

Finally, the third direction of mediation involves parahippocampal cortex and anterior insula/VLPFC, both regions related to pain.

# 7 Discussion

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

The causal interpretation for the parameters of the DM approach rests on a strong untestable assumption, namely sequential ignorability. For example, the assumption $Y(x, m) \perp\!\!\!\perp \mathbf{M}|X$ would be valid if the mediators were randomly assigned to the subjects. However, this is not the case here, and instead, we must assume that they behave as if they were. This assumption is unverifiable in practice and ultimately depends on context. In the neuroimaging setting, its validity may differ across brain regions, making causal claims more difficult to access. That said, we believe the proposed approach still has utility for performing exploratory mediation analysis and detecting sets of regions that potentially mediate the relationship between treatment and outcome, allowing these regions to be explored further in more targeted studies.

It should further be noted that when deriving the direct and indirect effect in section 2 we assumed each subject was scanned under one condition. However, in most fMRI experiments subjects are scanned under multiple conditions, as in our motivating pain data set. Extension of the casual model to this case will allow for single subject studies of mediation in which unit direct effects on the mediators and unit total effects on outcomes are observed. In some instances, the observability of these unit effects can be used to estimate both single subject and population averaged models under weaker and/or alternative conditions than those in 2. We leave this extension for future work. In addition, in our motivating example the mediator is brain activation measured with error. Thus, an extension would be to modify the model to deal with systematic errors of measurement in the mediating variable (Sobel and Lindquist (2014)).

One 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. For example, through re-parameterization, as by setting $w = \exp(v)$. This can be necessary because, under some circumstances, the coexistence of positive and negative elements of w could 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 $\mathbf{M}\mathbf{w} = 0$, making the estimate of $\beta_1$ unavailable. It, however, does not necessarily imply the non-existence of a mediation effect.

In many settings, 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 this setting using a GEE-like method. Essentially, conditioning on the DM, the path coefficient can be estimated using two sets of GEEs. The DM can then be estimated conditioning on the estimated coefficients.

# Acknowledgement

This research was partially supported by NIH grants R01EB016061, R01DA035484 and P41 EB015909, and 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, Haoyu Zhang, 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.

Andersen, A. H., Gash, D. M., and Avison, M. J. (1999). Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. *Magnetic Resonance Imaging*, 17(6):795–815.

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.

Caffo, B., Chen, S., Stewart, W., Bolla, K., Yousem, D., Davatzikos, C., and Schwartz, B. S. (2008). Are brain volumes based on magnetic resonance imaging mediators of the associations of cumulative lead dose with cognitive function? *American journal of epidemiology*, 167(4):429–437.

Caffo, B. S., Crainiceanu, C. M., Verduzco, G., Joel, S., Mostofsky, S. H., Bassett, S. S., and Pekar, J. J. (2010). Two-stage decompositions for the analysis of functional connectivity for fmri with application to alzheimer's disease risk. *NeuroImage*, 51(3):1140–1149.

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001). A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151.

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.

Huang, Y.-T. and Pan, W.-C. (2015). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*.

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*.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. (1997). Analysis of fMRI data by blind separation into independent spatial components. Technical report, DTIC Document.

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.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158.

Rubin, D. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66:688–701.

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

Sobel, M. E. and Lindquist, M. A. (2014). Causal inference for fmri time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *Journal of the American Statistical Association*, 109(507):967–976.

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

VanderWeele, T. and Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

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).