Biostatistics (2018) **19**, 2, *pp.* 121–136
doi:10.1093/biostatistics/kxx027
Advance Access publication on June 15, 2017

High-dimensional multivariate mediation with application to neuroimaging data

OLIVER Y. CHÉN, CIPRIAN CRAINICEANU, ELIZABETH L. OGBURN, BRIAN S. CAFFO

Department of Biostatistics, Johns Hopkins University, USA

TOR D. WAGER

Department of Psychology and Neuroscience, University of Colorado Boulder, 345 UCB, Boulder, CO 80309-0345, USA

MARTIN A. LINDQUIST*

Department of Biostatistics, Johns Hopkins University, 615 N Wolfe St, Baltimore, MD 21205, USA
mlindqui@jhsph.edu

Summary

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 based on the proportion of the LSEM likelihood 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 single nucleotide polymorphisms (SNPs), or vectors of thousands of variables in large-scale epidemiological studies. We demonstrate the method using a functional magnetic resonance imaging 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 analyses 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, epidemiology, and

*To whom correspondence should be addressed.

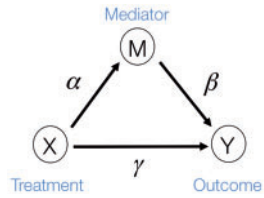

Fig. 1. The three-variable path diagram representing the standard mediation framework. The variables corresponding to X, Y, and M are all scalars, as are the path coefficients α, β, and γ.

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 *and others*, 1996; Ten Have *and others*, 2007; Albert, 2008; Jo, 2008; Sobel, 2008; VanderWeele and Vansteelandt, 2009; Imai *and others*, 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; Albert and Nelson, 2011; VanderWeele and Vansteelandt, 2014; Wang *and others*, 2013; Imai and Yamamoto, 2013; Daniel *and others*, 2015). 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 much larger than the number of observations. Such an extension is the focus of this work.

As a motivating example, consider functional magnetic resonance imaging (fMRI), an imaging modality where the signal of interest is the blood oxygenation level dependent (BOLD) response; a measure of the metabolic demands (i.e., oxygen consumption) of active neurons (Ogawa *and others*, 1990; Kwong *and others*, 1992; Lindquist, 2008). In fMRI experiments, a multivariate time series of 3D 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 *and others*, 2005; Bushnell *and others*, 2013). Recently, studies have focused on trial-by-trial modeling of the relationship between the intensity of noxious heat and self-reported pain (Wager *and others*, 2013; Atlas *and others*, 2014). In Woo *and others* (2015), for example, a series of thermal stimuli were applied at various temperatures (ranging from 44.3 to 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 *and others*, 2012). Standard mediation techniques are applicable to univariate mediators. An early approach to mediation in neuroimaging (Caffo *and others*, 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 *and others*, 2008; Wager *and others*, 2009b; Wager *and others*, 2009a). 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 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 we have a very high-dimensional vector of potential mediators $\mathbf{M} = (M^{(1)}, M^{(2)}, \ldots M^{(p)})^\intercal \in \mathbb{R}^p$. 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 (DMs) is developed. DM's linearly combine activity in different voxels into a smaller number of orthogonal components, with components ranked based upon 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 (Wold, 1985), 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 DMs 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 multivariate mediator setting. In Section 3, we introduce the DMs, and provide an estimation algorithm for estimating the DMs and their associated path coefficients when the mediator is high dimensional. In Section 4, we discuss a method for performing inference on the DMs. Finally, in Sections 5 and 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 $\mathbf{M} = (M^{(1)}, \cdots M^{(p)})^\intercal$ 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. We discuss this assumption in the context of fMRI in Section 7.

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

$$Y(x, \mathbf{M}(x)) - Y(x^*, \mathbf{M}(x^*)) = Y(x, \mathbf{M}(x)) - Y(x, \mathbf{M}(x^*)) + Y(x, \mathbf{M}(x^*)) - Y(x^*, \mathbf{M}(x^*)). \quad (2.1)$$

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:

$$\begin{aligned} Y(x, \mathbf{M}(x)) &\perp\!\!\!\perp X \\ Y(x, m) &\perp\!\!\!\perp \mathbf{M} | X \\ \mathbf{M}(x) &\perp\!\!\!\perp X \\ Y(x, m) &\perp\!\!\!\perp \mathbf{M}(x^*). \end{aligned} \tag{2.2}$$

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. Together, they are often referred to as sequential ignorability assumptions. 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 (2014) showed that under (2.2) the average direct and indirect effects are identified from the regression function for the observed data.

Suppose then (2.2) and the following model for the observed data hold:

$$\begin{aligned} E(M^{(j)} | X = x) &= a_0 + a_j x \qquad \text{for} \quad j = 1, \ldots, p \\ E(Y | X = x, \mathbf{M} = \mathbf{m}) &= b_0 + cx + b_1 m^{(1)} + b_2 m^{(2)} + \cdots + b_p m^{(p)}. \end{aligned} \tag{2.3}$$

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 article 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:

$$E(Y(x, \mathbf{m}) - Y(x^*, \mathbf{m})) = c(x - x^*) \tag{2.4}$$

$$E(Y(x, \mathbf{M}(x^*)) - Y(x^*, \mathbf{M}(x^*))) = c(x - x^*) \tag{2.5}$$

$$E(Y(x, \mathbf{M}(x)) - Y(x, \mathbf{M}(x^*))) = (x - x^*) \sum_{j=1}^{p} a_j b_j. \tag{2.6}$$

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 (2.2) hold, the right-hand sides of (2.5) and (2.6) identify causal mediation effects. When one or more of the assumptions in (2.2) fail to hold, or if the counterfactuals are not well defined, the right-hand sides of (2.5) and (2.6) may still be used in exploratory analysis to 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 (2.5) and (2.6), respectively; we are agnostic 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.

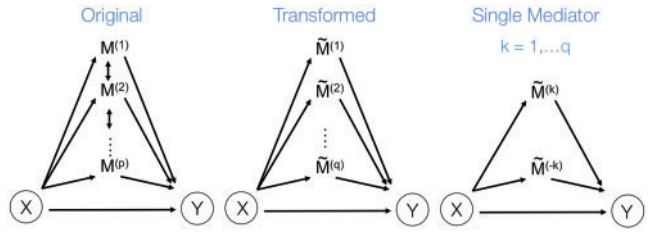

Fig. 2. (Left) The three-variable path diagram used to represent multivariate mediation. Here the p mediators are assumed to be correlated. (Center) A similar path diagram after an orthogonal transformation of the mediators, making the q transformed mediators uncorrelated with one another. (Right) Because the mediators are uncorrelated, a series of LSEMs, one for each transformed mediator $\tilde{M}^{(k)}$, can be used to estimate direct and indirect effects.

We also note that the model only considers the case where the entire vector $\mathbf{M}$ is subject to a single level of intervention (x). More generally, it would be possible to define potential mediators when some elements of $\mathbf{M}$ are subject to x and other elements of $\mathbf{M}$ are subject to x^*, which could support various decompositions of direct/indirect effects due to pathways acting through different $\mathbf{M}$. In the fMRI setting described herein we anticipate that the same elements of $\mathbf{M}$ (i.e., brain regions) are subject to both x and x^*. For example, warm and hot thermal stimuli give rise to activation in the same brain regions, but with different intensities.

Fitting the system (2.3) 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, allowing us to estimate the direct and indirect effects using a series of LSEMs, one for each transformed mediator; see Figure 2 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) potentially reducing the number of mediators needed to estimate the indirect effect. Finally, it is easier to assess potential treatment by mediator interactions using a lower dimensional summary of the mediator variables than it is doing this directly in the high-dimensional setting.

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, and an extension to the case when multiple trials are observed for each subject. We conclude with a discussion regarding estimation for the case when $p >> N$, where N represents the total number of observations.

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 $i = 1, \dots, n$, where n denotes the number of subjects. 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 $\mathbf{w}_k = (w_k^{(1)}, w_k^{(2)}, \dots w_k^{(p)})^\intercal \in \mathbb{R}^p$, for $k = 1, \dots, q$; and let $\tilde{\mathbf{M}} = \mathbf{MW} = (\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_2, \dots \tilde{\mathbf{M}}_n)^\intercal$ where $\tilde{\mathbf{M}}_i = \mathbf{M}_i^\intercal \mathbf{W} = (\tilde{M}_i^{(1)}, \dots, \tilde{M}_i^{(k)}, \dots \tilde{M}_i^{(q)})^\intercal$ with $\tilde{M}_i^{(k)} = \mathbf{M}_i^\intercal \mathbf{w}_k =$

$\sum_{j=1}^{p} M_i^{(j)} w_k^{(j)}$. We assume the relationship between the variables is given by the following LSEM:

$$\begin{aligned} \tilde{M}_i^{(k)} &= \alpha_0 + \alpha_k X_i + \epsilon_i \qquad \text{for} \quad k = 1, \ldots, q \\ Y_i &= \beta_0 + \gamma X_i + \beta_1 \tilde{M}_i^{(1)} + \beta_2 \tilde{M}_i^{(2)} + \ldots + \beta_q \tilde{M}_i^{(q)} + \xi_i, \end{aligned} \tag{3.1}$$

where ϵ_i and ξ_i are i.i.d. normal with mean 0 and variances σ_ϵ^2 and σ_ξ^2. The parameters of the LSEM can be estimated using linear regression. However, under the additional condition that the new transformed variables $\tilde{M}^{(k)}$ are orthogonal, we can estimate the parameters separately for each $\tilde{M}^{(k)}$. Thus, for each $k = 1, \ldots, q$ we can fit the following LSEM:

$$\begin{aligned} \tilde{M}_i^{(k)} &= \alpha_0 + \alpha_k X_i + \epsilon_i \\ Y_i &= \beta_0 + \gamma X_i + \beta_k \tilde{M}_i^{(k)} + \eta_i, \end{aligned} \tag{3.2}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $\eta_i \sim N(0, \sigma_\eta^2)$, for $i = 1, \ldots, n$. Using this notation, we can express the average direct and indirect effects defined in (2.5) and (2.6) as $\gamma(x - x^*)$ and $(x - x^*) \sum_{j=1}^{p} \alpha_j \beta_j$, respectively. Hence, these effects can be estimated either in the original or orthogonalized space. However, for large values of p the model expressed in (3.2) is computationally preferable to the one expressed in (2.3).

Let $\boldsymbol{\theta} := (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma) \in \mathbb{R}^5$ be the parameter vector for the LSEM in (3.2) for $k = 1$. We seek to simultaneously estimate $\boldsymbol{\theta}$ and find the first direction of mediation (DM) $\mathbf{w}_1$, defined as the coefficient vector of 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 voxel activations. Thus, similar to principal components analysis (PCA) (Andersen *and others*, 1999) or independent components analysis (McKeown *and others*, 1997) 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 (3.2). The *DMs* are defined as follows:

Step 1: The first DM is the vector $\mathbf{w}_1$, 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 second DM is the vector $\mathbf{w}_2$, 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.,

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

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.,

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

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 1 According to the model formulation the signs of the DMs are unidentifiable. A change in sign of $\tilde{\mathbf{M}}^{(k)}$ can be offset by a change in sign of both α_k and β_k.
Remark 2 As the transformed mediators are ranked based upon the proportion of the likelihood of the full LSEM explained, one could potentially limit the number of DMs computed to achieve dimension reduction. However, this could cause the estimate of the indirect effect to be biased.

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 $\boldsymbol{\theta}$ can be expressed as:

$$\log(\mathscr{L}(\Delta; \mathbf{w}_1, \boldsymbol{\theta})) \propto g_1(\Delta; \mathbf{w}_1, \boldsymbol{\theta}), \tag{3.3}$$

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:

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

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 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 λ as follows:

$$\hat{\mathbf{w}}_1|\boldsymbol{\theta}, \lambda = f_1(\Delta; \lambda, \boldsymbol{\theta}) \tag{3.4}$$

$$\hat{\lambda}|\boldsymbol{\theta} = \text{arg}_{\lambda \in \mathbb{R}^1} \left\{ f_2(\Delta; \lambda, \boldsymbol{\theta}) = 1 \right\} \tag{3.5}$$

$$\hat{\boldsymbol{\theta}}|\hat{\mathbf{w}}_1, \hat{\lambda} = \underset{\boldsymbol{\theta} \in \mathbb{R}^5}{\text{argmax}} L(\Delta; \hat{\mathbf{w}}_1, \boldsymbol{\theta}, \hat{\lambda}), \tag{3.6}$$

where $f_1(\Delta;\lambda,\boldsymbol{\theta}) = (\lambda\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta})$; $f_2(\Delta;\lambda,\boldsymbol{\theta}) = \left\|(\lambda\mathbf{I}+\boldsymbol{\psi}(\boldsymbol{\theta}))^{-1}\boldsymbol{\phi}(\boldsymbol{\theta})\right\|_2$, $\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{M}^\intercal\mathbf{M}\beta_1^2/\sigma_{\epsilon_1}^2 + \mathbf{M}^\intercal\mathbf{M}/\sigma_{\eta_1}^2$, 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 DM and path parameters as described in Algorithm 1. Further, in the Appendices A and B of supplementary material available at *Biostatistics* online we show that the estimated parameters are consistent and asymptotically normal (see Theorems 1 and 2).

Algorithm 1 First DM

Step 0: Initiate $\boldsymbol{\theta}_1$, denoted $\boldsymbol{\theta}_1^{(0)}$. **Step 1:** For each h, set:

$$\hat{\lambda}^{(h)}|\boldsymbol{\theta}_1^{(h)} = \arg_{\lambda\in\mathbb{R}^1}\left\{f_2(\Delta;\lambda,\boldsymbol{\theta}_1^{(h)}) = 1\right\} \tag{3.7}$$

$$\hat{\mathbf{w}}_1^{(h)}|\boldsymbol{\theta}_1^{(h)},\hat{\lambda}^{(h)} = f_1(\Delta;\hat{\lambda}^{(h)},\boldsymbol{\theta}_1^{(h)}) \tag{3.8}$$

$$\hat{\boldsymbol{\theta}}_1^{(h+1)}|\hat{\mathbf{w}}_1^{(h)},\hat{\lambda}^{(h)} = \arg\max_{\boldsymbol{\theta}_1\in\mathbb{R}^5}\left\{L(\Delta;\hat{\mathbf{w}}_1^{(h)},\boldsymbol{\theta}_1^{(h)},\hat{\lambda}^{(h)})\right\}. \tag{3.9}$$

Step 2: Repeat Step 1 until convergence; each time set $h = h+1$.

3.3. *Higher order DMs*

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

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

subject to

$$\left\{\boldsymbol{\theta}_k\in\mathbb{R}^{k+4},\mathbf{x}\in\bar{\mathbb{R}}^p : \mathbf{w}_k(\mathbf{x}) := \mathbf{x} - \sum_{k'=1}^{k-1}\frac{\langle\mathbf{x},\hat{\mathbf{w}}_{k'}\rangle}{\langle\hat{\mathbf{w}}_{k'},\hat{\mathbf{w}}_{k'}\rangle}\hat{\mathbf{w}}_{k'}\boldsymbol{\theta}_{k'}\right\},$$

where

$$g_k(\Delta,\hat{\mathbf{w}}_1,\ldots,\hat{\mathbf{w}}_{k-1};\mathbf{w}_k,\boldsymbol{\theta}_k) = -\Big\{\frac{1}{\sigma_\epsilon^2}\left\|\mathbf{y}-\beta_0-\mathbf{x}\gamma-\mathbf{M}\hat{\mathbf{w}}_1\beta_1-\ldots-\mathbf{M}\hat{\mathbf{w}}_{k-1}\beta_{k-1}-\mathbf{M}\mathbf{w}_k\beta_k\right\|_2 + \frac{1}{\sigma_\eta^2}\|\mathbf{M}\mathbf{w}_k-\alpha_0-\mathbf{x}\alpha_k\|_2\Big\}. \tag{3.10}$$

The performance of the projection approach is evaluated in Section 5.

3.4. *Multi-trial model*

In our setting, multiple observations (i.e., thermal stimulations and subsequent pain reports) are acquired for each subject. Here we describe a simple extension of the model described in (3.1). Throughout we assume that each subject has n_i observations, so that $N = \sum_{i=1}^{n} n_i$. Let X_{it}, $\tilde{M}_{it}$, and Y_{it} denote subject i's treatment, orthogonalized meditator, and outcome, respectively, on the t^{th} trial. While we ultimately want to express the relationship between these variables using a model that incorporates subject-specific random effects, here we use a simpler formulation that allows us to retain the estimation procedure outlined in the previous sections. We assume the relationship between the variables is given by the following LSEM:

$$\begin{aligned} \tilde{M}_{it}^{(k)} &= \alpha_0 + \alpha_k X_{it} + \epsilon_{it} \qquad \text{for} \quad k = 1, \ldots, q \\ Y_{it} &= \beta_0 + \gamma X_{it} + \beta_1 \tilde{M}_{it}^{(1)} + \beta_2 \tilde{M}_{it}^{(2)} + \ldots + \beta_q \tilde{M}_{it}^{(q)} + \xi_{it}, \end{aligned} \tag{3.11}$$

where $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ and $\xi_{it} \sim N(0, \sigma_\eta^2)$, for $t = 1, \ldots, n_i$ and $i = 1, \ldots, n$. Note that if each subject has a single observation, then $N = n$, and this model is equivalent to (3.1). As (3.11) generalizes (3.1), we use it in the continuation.

3.5. *High-dimensional DMs*

The estimation procedure described in Section 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 *and others*, 2010; Crainiceanu *and others*, 2011), which in contrast to singular value decomposition provides population-level information about $\mathbf{M}$. Throughout we assume that the data for each subject i is stored in an $n_i \times p$ matrix, $\overline{\mathbf{M}}_i$, whose t^{th} row contains voxel-wise activity for the measurements of the t^{th} trial for the i^{th} subject. All $\overline{\mathbf{M}}_i$ matrices are stacked vertically to form the $N \times p$ matrix $\mathbf{M}$, where $N = \sum_{i=1}^{n} n_i$.

The generalized PVD (GPVD) of $\overline{\mathbf{M}}_i$ (formally derived in Appendix C of the supplementary material available at *Biostatistics* online) is given by

$$\overline{\mathbf{M}}_i = \mathbf{U}_i^B \tilde{\mathbf{V}}_i \mathbf{D} + \mathbf{E}_i, \tag{3.12}$$

where $\mathbf{U}_i^B$ is an $n_i \times B$ 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, and $\mathbf{E}_i$ is an $n_i \times p$ matrix of residuals. Here the value of B is chosen based upon a criteria such as total variance explained, in a similar manner as in PCA. See supplementary material available at *Biostatistics* online for details on how to compute these components.

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

$$\breve{\mathbf{M}} = \left[(\mathbf{U}_1 \tilde{\mathbf{V}}_1)^\intercal, \cdots, (\mathbf{U}_n \tilde{\mathbf{V}}_n)^\intercal\right]^\intercal. \tag{3.13}$$

Let $\breve{\mathbf{w}} = \mathbf{D}\mathbf{w}$, where $\breve{\mathbf{w}}$ is $B \times 1$. Now $\mathbf{M}\mathbf{w} \approx \breve{\mathbf{M}}\breve{\mathbf{w}}$ and $\dim(\mathbf{w}) >> \dim(\breve{\mathbf{w}})$. Thus, we can use $\breve{\mathbf{M}}$ to estimate $\breve{\mathbf{w}}$, which is significantly less computationally intensive than using $\mathbf{M}$ to estimate $\mathbf{w}$. Since $\mathbf{D}$ can be obtained via GPVD, we can retrieve the original estimator of $\hat{\mathbf{w}}$ via the generalized inverse, i.e., $\hat{\mathbf{w}} = \mathbf{D}^- \breve{\mathbf{w}}^{est}$, where $\breve{\mathbf{w}}^{est}$ is the estimated $\breve{\mathbf{w}}$ and $^-$ indicates the generalized inverse.

4. INFERENCE

Here we discuss an approach for using a bootstrap procedure to perform inference. 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) Create a bootstrap sample $(x^*, y^*, \breve{\mathbf{M}}^*)$ consisting of n subjects, by randomly resampling subjects with replacement from the data set $(x, y, \breve{\mathbf{M}})$;
(2) Estimate the low-dimensional bootstrap DM $\hat{\breve{\mathbf{w}}}_k$ of length B for $k = 1, \ldots q$.
(3) Compute $\hat{\mathbf{w}}_k = \mathbf{D}^{-1}\hat{\breve{\mathbf{w}}}_k$, where $\hat{\mathbf{w}}_k$ is the high-dimensional bootstrap DM of length p;
(4) Repeat steps 1–3 R times. Stack values of $\hat{\mathbf{w}}_k$ vertically to form the $R \times p$ matrix $\hat{\mathbf{W}}_k^*$.

We seek to use the distribution to test whether specific elements of $\mathbf{w}_k$ differ from 0. Note the columns of $\hat{\mathbf{W}}_k^*$ are the bootstrap values of the k^{th} DM corresponding to voxel j, 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. To circumvent this we instead focus on obtaining a bootstrap distribution for $|\mathbf{w}_k|$.

We begin by performing the bootstrap procedure as outlined above. Next, we sort the bootstrap distributions for all voxels by their median, and choose voxels whose median values lies in either the second or third quartile. We randomly sample among these voxels (in our application we sample 10% of the voxels whose medians lie in these quartiles), and combine their distribution to create a pseudo-null distribution. Finally, we fit a half-normal distribution to the pseudo-null and use this distribution to estimate a p-value for each element of $\mathbf{w}_k$.

5. SIMULATION STUDY

We performed a series of simulation studies to investigate the efficacy of our approach. A description of these simulations and the results are in the Appendix D of supplementary material available at *Biostatistics* online.

6. AN FMRI STUDY OF THERMAL PAIN

The data comes from the fMRI study of thermal pain described in Section 1. Here 33 healthy, right-handed participants completed 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 data consists of 7 runs, consisting of between $58 - 75$ separate trials, in which participants experienced and rated the heat stimuli. 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 to 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 0–100. Each trial concluded with a 5–9 s resting period. For more information about the data acquisition and preprocessing, see Appendix E of supplementary material available at *Biostatistics* online.

A single trial analysis approach was used by constructing a GLM design matrix with separate regressors for each trial (Rissman *and others*, 2004). 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 (i) intercepts for each run; (ii) linear drift across time within each run; (iii) the six estimated head movement parameters (x, y, z, roll, pitch, and

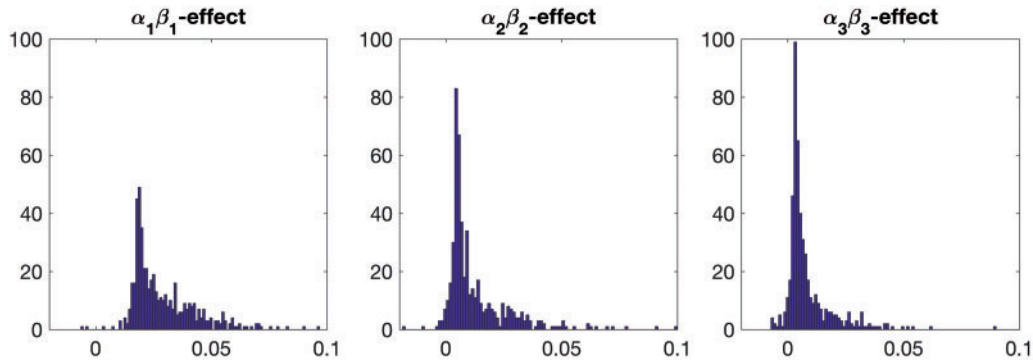

Fig. 3. Results of the fMRI study of thermal pain. Bootstrap distributions for $\alpha_k\beta_k$ corresponding to the first three DMs.

yaw), their mean-centered squares, derivatives, and squared derivative for each run; (iv) indicator vectors for outlier time points; (v) indicator vectors for the first two images in each run; and (vi) signal from white matter and ventricles. Using the results of the GLM analysis, whole-brain maps of activation were computed.

In summary, X_{it} and Y_{it} are the temperature level and pain rating, respectively, assigned on trial t to subject i, and $\mathbf{M}_{it} = (M_{it}^{(1)}, M_{it}^{(2)}, \ldots M_{it}^{(p)})^\intercal \in \mathbb{R}^p$ 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, $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, n_i\}$, where $n = 33$ and n_i takes subject-specific values between 58 and 75. The data was arranged in a matrix $\mathbf{M}$ of dimension $1149 \times 206\,777$, 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 1149.

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 3.5. We choose $\tilde{\mathbf{w}}_k$ 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 using GPVD, and the lower dimensional mediation matrix $\breve{\mathbf{M}}$ of dimension 1149×35, according to (3.13). The terms $(\mathbf{x}, \mathbf{y}, \breve{\mathbf{M}})$ were placed into the algorithm outlined in (3.7)–(3.9), using starting values $\boldsymbol{\theta}_1^{(0)} = \mathbf{0}$. Finally, values of $\hat{\mathbf{w}}_k$, of length 206 777, were computed using $\mathbf{D}^- \breve{\mathbf{w}}^{est}$.

We illustrate the results for the first three DMs. The parameter estimates are $\hat{\boldsymbol{\theta}}_1 = (1185.3, -39.4, -16.0, -0.00055, 0.45)$, $\hat{\boldsymbol{\theta}}_2 = (-522.1, 15.6, -16.3, -0.0005, 0.0001, 0.45)$, and $\hat{\boldsymbol{\theta}}_3 = (929.5, -32.2, -15.7, -0.0005, -0.0003, -4.4 \times 10^{-5}, 0.45)$. Figure 3 shows the bootstrap distribution for $\alpha_k\beta_k$ for $k = 1, 2, 3$. Clearly, the center of the distributions move toward 0 for increasing k, illustrating their decreasing contribution. However, for each value of k the contribution of $\alpha_k\beta_k$ remains significantly different from 0 at the 0.05 level.

Figure 4 shows the estimated weight maps $\hat{\mathbf{w}}_k$ for the first three DMs, thresholded using false discovery rate (FDR) correction with $q = 0.05$. These maps allow us to visualize latent brain components that are maximally effective as mediators. For each DM, we separate the maps according to whether the voxel-specific value is positive or negative. The maps are consistent with regions typically considered active in pain research, but also reveal some interesting structure that has not been uncovered by previous methods.

The first DM 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

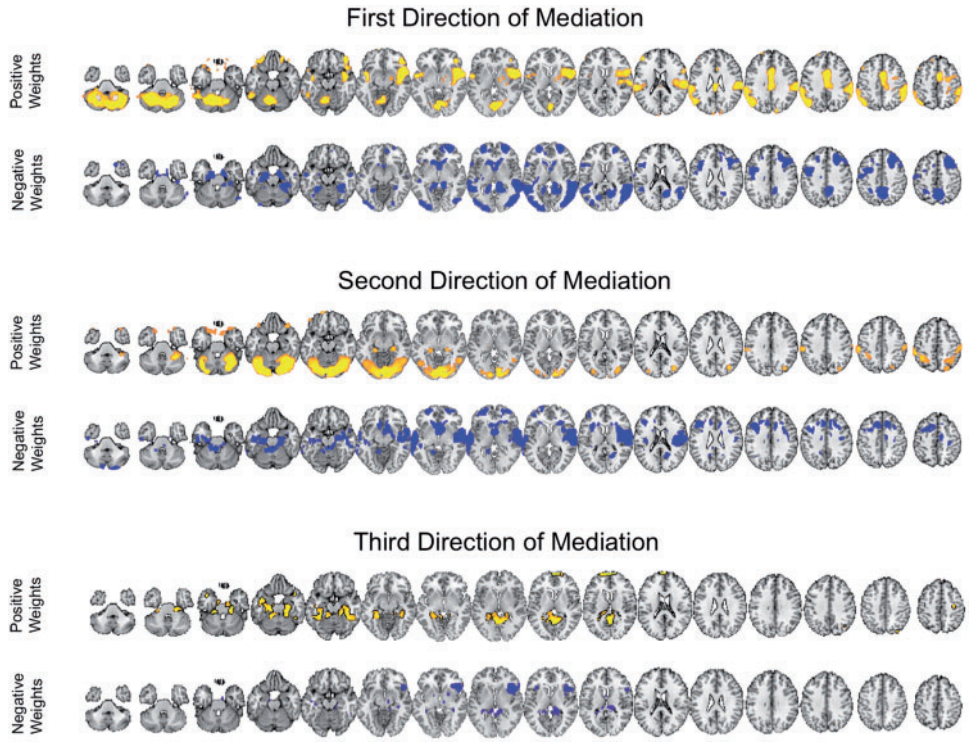

Fig. 4. Results of the fMRI study of thermal pain. Weight maps $\mathbf{w}_k$ corresponding to the first three DMs, mapped back onto the brain. Significant weights are separated into those with positive and negative values, respectively, for the each DM. All maps are thresholded using FDR correction with $q = 0.05$.

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 DM 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 DM involves parahippocampal cortex and anterior insula/Ventrolateral prefrontal cortex (VLPFC), both regions related to pain.

7. DISCUSSION

This article 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 strong untestable assumptions, namely sequential ignorability. For example, the assumption $Y(x, m) \perp\!\!\!\perp \mathbf{M}|X$ holds if the mediators are randomly assigned to the subjects. However, this is clearly not the case here, and instead, we must assume 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. In practice, this assumption could potentially be weakened by allowing for conditioning on potential confounders. However, no such covariates were available in this particular study. These caveats notwithstanding, we believe the proposed approach still has utility for performing exploratory mediation analysis and detecting regions that potentially mediate the relationship between treatment and outcome, allowing these regions to be explored further in more targeted studies. For further discussion of sequential ignorability assumptions in neuroimaging and the potential to intervene on the mediator without affecting the treatment, see Lindquist and Sobel (2011, 2012).

The nested counterfactuals defining mediation effects require an assumption that it is possible to intervene on the mediator without affecting the treatment. This assumption is meant to ensure the existence of independent mechanisms affecting treatment and mediator, to rule out, for example, cases where the treatment and the mediator must by necessity cooccur. Transcranial magnetic stimulation (TMS) manipulates the mediator $\mathbf{M}$ (i.e., brain activity) directly without affecting a treatment such as exposure to heat or pain. Using this technique, one can simulate temporary brain lesions while the subject performs certain tasks. TMS and other similar tools are not technologically advanced enough to allow us to intervene to set $\mathbf{M}$ to any user-specified value $\mathbf{m}$, but it is indeed hypothetically possible to activate or block brain activity on a voxel level. Intervening on a network of voxels is a thornier proposition, but it may be possible to activate a network through direct brain stimulation without affecting treatment.

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 (2.2). 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 $\mathbf{w}$ 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 $\mathbf{Mw} = 0$, making the estimate of β_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 type of generalized estimating equation (GEE). 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.

In this work, we have assumed that there are no covariates included in the proposed mediation model. However, in observational studies, one often needs to adjust for covariates. In future work, we intend to

extend the model and the method to allow for covariates. Finally, while in this work the likelihood is penalized using an L2 norm, in other situations one may require different penalties such as the L1 norm (or some combination of the two). The methods proposed can be extended in these directions, but again we leave this for future work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

NIH (R01EB016061, R01DA035484, and P41 EB015909) and NSF (0631637).

REFERENCES

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

ALBERT, J. M. AND NELSON, S. (2011). Generalized causal mediation analysis. *Biometrics* **67**, 1028–1038.

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**, 795–815.

ANGRIST, J. D., IMBENS, G. W. AND RUBIN, D. B. (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**, 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**, 1632–1648.

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

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**, 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**, 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**, 1140–1149.

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**, 775–790.

DANIEL, R. M., DE STAVOLA, B. L., COUSENS, S. N. AND VANSTEELANDT, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71**, 1–14.

HOLLAND, P. W. (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**, 309.

IMAI, K. AND YAMAMOTO, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis* **21**, 141–171.

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

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**, 5675–5679.

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

LINDQUIST, M. A. AND SOBEL, M. E. (2011). Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. *NeuroImage* **57**, 334–336.

LINDQUIST, M. A. AND SOBEL, M. E. (2012). Cloak and DAG: A response to the comments on our comment. *NeuroImage* **76**, 446–449.

LINDQUIST, M. A., SPICER, J., ASLLANI, I. AND WAGER, T. D. (2012). Estimating and testing variance components in a multi-level glm. *NeuroImage* **59**, 490–501.

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

MACKINNON, D. P. (2015). Mediation Analysis. *The Encyclopedia of Clinical Psychology*. 1–9.

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.

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**, 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**, 105.

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

PREACHER, K. J. AND HAYES, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods* **40**, 879–891.

RISSMAN, J., GAZZALEY, A. AND D'ESPOSITO, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage* **23**, 752–763.

ROBINS, J. M. 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. In: Shrout, P. (editor), *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. Oxford University Press, pp. 103–158.

RUBIN, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

SOBEL, M. E. (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**, 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**, 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. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods* **2**, 95–115.

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**, 1388–1397.

WAGER, T. D., DAVIDSON, M. L., HUGHES, B. L., LINDQUIST, M. A. AND OCHSNER, K.N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron* **59**, 1037–1050.

WAGER, T. D., VAN AST, V., DAVIDSON, M. L., LINDQUIST, M. A. AND OCHSNER, K. N. (2009*a*). Brain mediators of cardiovascular responses to social threat, Part II: Prefrontal subcortical pathways and relationship with anxiety. *NeuroImage* **47**, 836–851.

WAGER, T. D., WAUGH, C. E., LINDQUIST, M. A., NOLL, D. C., FREDRICKSON, B. L. AND TAYLOR, S. F. (2009*b*). 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.

WANG, W., NELSON, S. AND ALBERT, J. M. (2013). Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in medicine* **32**, 4211–4228.

WOLD, H. (2006). Partial Least Squares. *Encyclopedia of Statistical Sciences*. 9.

WOO, C. W., ROY, M., BUHLE, J. T. AND WAGER, T. D. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biology* **13**, e1002036.

[*Received August 25, 2016; revised March 18, 2017; accepted for publication May 7, 2017*]