High-dimensional Multivariate Mediation with Application to Neuroimaging Data

Supplementary Materials

Appendix A. Theoretical Properties

Let $\mathbf{D} = (\mathbf{X}, \mathbf{Y}, \mathbf{M})$ be a data triple, where $\mathbf{X} = (X_1, \ldots, X_n)^\intercal \in \mathbb{R}^n$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\intercal \in \mathbb{R}^n$, and $\mathbf{M} = (M_1, \ldots, M_n)^\intercal \in \mathbb{R}^{n\times p}$. Let $\mathbf{w} \in \mathbb{R}^p$ and $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma) \in \mathbb{R}^5$, be the parameters of interest. In particular, $\mathbf{w}$ maps $\mathbf{M}$ onto $\mathbb{R}^n$. Let $\lambda \in \mathbb{R}^1$ be a nuisance parameter. Consider the joint $\log$-likelihood function $g(\cdot; \mathbf{w}, \boldsymbol{\theta})$ in equation (3.9) as the objective function.

Define the profiled Lagrangian $L(\mathbf{D}; \boldsymbol{\theta}) = g(\mathbf{D}; \boldsymbol{\theta}) + \lambda(\boldsymbol{\theta})\big(\mathbf{w}^\intercal(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta}) - 1\big)$ and $L(d; \boldsymbol{\theta}) = g(d; \boldsymbol{\theta}) + 1/n\lambda(\boldsymbol{\theta})\big(\mathbf{w}^\intercal(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta}) - 1\big)$, where $L(d; \boldsymbol{\theta}_0) \in \mathcal{P} = \{L(d; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is some properly defined space in $\mathbb{R}^5$. Define $\dot{L}(\mathbf{D}; \boldsymbol{\theta}) := \dfrac{\partial L}{\partial \boldsymbol{\theta}}(\mathbf{D}; \boldsymbol{\theta}) = \sum_{i=1}^n \ell(d_i, \boldsymbol{\theta})$, where $\ell(d, \boldsymbol{\theta}) = \dfrac{\partial g}{\partial \boldsymbol{\theta}}(d; \boldsymbol{\theta}) + 1/n\big[\nabla^{\boldsymbol{\theta}}\lambda(\boldsymbol{\theta})(\|\mathbf{w}(\boldsymbol{\theta})\|_2 - 1) + 2\lambda(\boldsymbol{\theta})\nabla^{\boldsymbol{\theta}}\mathbf{w}(\boldsymbol{\theta})\big]$. Further define $\hat{q}(\mathbf{D}; \boldsymbol{\theta}) = 1/n\sum_{i=1}^n L(D_i; \boldsymbol{\theta})$, $q_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(L(d; \boldsymbol{\theta}))$, $\hat{\dot{q}}(\mathbf{D}; \boldsymbol{\theta}) = 1/n\sum_{i=1}^n \ell(D_i; \boldsymbol{\theta})$, and $\dot{q}_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(\ell(d; \boldsymbol{\theta}))$.

A.1. Regularity Conditions

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

Regularity Conditions II :

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

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

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

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

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

(N-6) $\mathbb{E}_{\boldsymbol{\theta}_0}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \{1/n\lambda(\boldsymbol{\theta})(\|\mathbf{w}(\boldsymbol{\theta})\|_2 - 1)\} \right\} = o_p(n^{-1/2})$;

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

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

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

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

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

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

Regularity Conditions III:

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

A.2. Asymptotic Properties

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

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

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

Theorem 1. *Under regularity conditions (N-1) - (N-12), the structural path coefficient estimator, $\hat{\boldsymbol{\theta}}^{PMLE}$, is asymptotically consistent, and normally distributed, i.e.*

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

and

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

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

Theorem 2. *Under regularity conditions (N-1) - (N-13), the estimator of the first direction of mediation, $\mathbf{w}^{DM1}$, is asymptotically consistent, and normally distributed, i.e.*

$$\mathbf{w}^{DM1} \xrightarrow{p} \mathbf{w}_0$$

and

$$\sqrt{n}(\mathbf{w}^{DM1} - \mathbf{w}_0) \to N\big(0, \Sigma^{\mathbf{w}}(\boldsymbol{\theta}_0)\big).$$

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

It is worth pointing out two aspects of *Theorem 2*. First, it is valid for $\dim(\mathbf{w}) \leq \dim(\boldsymbol{\theta})$ or $\dim(\mathbf{w}) > \dim(\boldsymbol{\theta})$. See also a discussion regarding the identifiability of $\mathbf{w}$ in the discussion section in the paper. Second, the conditions required for the multivariate delta method are met. Under regularity conditions, for every estimate of $\boldsymbol{\theta}$, we can find a unique function of the estimate. For a sufficiently large n, there exists an estimate $\hat{\boldsymbol{\theta}}_n$ in the neighborhood of the true value. For that particular $\hat{\boldsymbol{\theta}}_n$, we can find a unique mapping f_n.

Appendix B. Proofs of the Theoretical Properties

Lemma B.1. (Consistency Theorem) Suppose that $Q(\boldsymbol{\theta}; \mathbf{Z}_n)$ is continuous in $\boldsymbol{\theta}$ and there exists a function $Q_0(\boldsymbol{\theta})$ such that: $Q_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$; $\boldsymbol{\Theta}$ is compact; $Q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$; and $Q(\boldsymbol{\theta}; \mathbf{Z}_n)$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$. Then $\hat{\boldsymbol{\theta}}(\mathbf{Z}_n)$ defined as the value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ which for each $\mathbf{Z}_n = \mathbf{z}_n$ maximizes the objective function $Q(\boldsymbol{\theta}; \mathbf{Z}_n)$ satisfies $\hat{\boldsymbol{\theta}}(\mathbf{Z}_n) \xrightarrow{p} \theta_0$.

Proof of Lemma B.1. See Theorem 2.1 in (Newey and McFadden, 1994). ■

Lemma B.2. Consider a compact space $\boldsymbol{\Theta}$. Let $L(z, \mathbf{w}, \boldsymbol{\theta}, \lambda) = f^{\text{obj}}(z, \mathbf{w}, \boldsymbol{\theta}) + f^{\text{pen}}(z, \mathbf{w}, \lambda)$, where $f^{\text{obj}}(z, \mathbf{w}, \boldsymbol{\theta})$ is an objective function and $f^{\text{pen}}(z, \mathbf{w}, \lambda) = \dfrac{\lambda\{f^{\text{cons}}(\mathbf{w}) - c\}}{n}$ is a penalization function, for some constant c. If both the objective function and the penalization function can be profiled by $\boldsymbol{\theta}$, defined as $f^{\text{obj}}(z, \boldsymbol{\theta})$ and $f^{\text{pen}}(\boldsymbol{\theta}) := \dfrac{\lambda(\boldsymbol{\theta})\{f^{\text{cons}}(\boldsymbol{\theta}) - c\}}{n}$; the objective function is a $\log$ likelihood function; both $f^{\text{obj}}(z, \boldsymbol{\theta})$ and $f^{\text{pen}}(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$; and there exists a function $d_0(z)$ such that $\mid L(z, \boldsymbol{\theta}) \mid := \mid f^{\text{obj}}(z, \boldsymbol{\theta}) + f^{\text{pen}}(\boldsymbol{\theta}) \mid \leq d_0(z)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $z \in \mathcal{Z}$, and $\mathbb{E}_{\boldsymbol{\theta}_0}[d_0(x)] < \infty$, then

i. $q_0(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}_0}[L(z, \boldsymbol{\theta})]$ is continuous in $\boldsymbol{\theta}$;

ii. $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mid q(\boldsymbol{\theta}; \mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid \xrightarrow{p} 0$, where $q(\boldsymbol{\theta}; \mathbf{Z}_n) := \dfrac{1}{n} L(\mathbf{Z}_n, \boldsymbol{\theta})$.

Note: the above Lemma can be stated in a more general case where there are multiple sets of parameters and several constraint functions.

Proof of Lemma B.2. Consider the regularity conditions stated in the Appendix.

$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, choose a sequence $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}$, such that $\boldsymbol{\theta}_k \to \boldsymbol{\theta}$. By (N-3), we have $L(x; \boldsymbol{\theta}_k) \to L(x; \boldsymbol{\theta})$. By (N-4) and the dominated convergence theorem (DCT), $q_0(\boldsymbol{\theta}_k) := \mathbb{E}_{\boldsymbol{\theta}_0}(L(Z, \boldsymbol{\theta}_k)) \to \mathbb{E}_{\boldsymbol{\theta}_0}(L(Z, \boldsymbol{\theta})) = q_0(\boldsymbol{\theta})$. Hence, $q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. $L(z, \boldsymbol{\theta})$ is uniformly continuous since $f^{\text{obj}}(z, \boldsymbol{\theta})$ and $f^{\text{prof}}(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$. Hence,

$$\Delta(z,\delta) = \sup_{\{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2):\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|<\delta\}} \mid L(z,\boldsymbol{\theta}_1) - L(z,\boldsymbol{\theta}_2) \mid \to 0$$

as $\delta \to 0$. By (N-4), $\Delta(z,\delta) \leq 2d_0(z), \forall\delta$. By DCT, $\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z,\delta)] \to 0$ as $\delta \to 0$.

Define $B(\boldsymbol{\theta}_j,\delta) = \{\tilde{\boldsymbol{\theta}} : \| \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \| < \delta\}$. Since $\boldsymbol{\Theta}$ is compact, for every fixed δ, $\exists$ a subcover $\{B(\boldsymbol{\theta}_j,\delta), j = 1,\ldots,J\}$ such that $\bigcup_{j=1}^{J<\infty} B(\boldsymbol{\theta}_j,\delta) \supset \boldsymbol{\Theta}$. Then, we have:

$$\begin{aligned} \mid q(\boldsymbol{\theta};\mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid \quad &\leq \quad \mid q(\boldsymbol{\theta};\mathbf{Z}_n) - q(\boldsymbol{\theta}_j;\mathbf{Z}_n) \mid & (1)\\ &+ \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid & (2)\\ &+ \mid q_0(\boldsymbol{\theta}_j) - q_0(\boldsymbol{\theta}) \mid . & (3) \end{aligned}$$

Choose $\boldsymbol{\theta}_j$ such that $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_j;\delta)$. Since $\| \boldsymbol{\theta} - \boldsymbol{\theta}_j \| < \delta$, then:

(1) $=\mid \frac{1}{n}\sum_{i=1}^n\{L(Z_i,\boldsymbol{\theta}) - L(Z_i,\boldsymbol{\theta}_j)\} \mid \leq \frac{1}{n}\sum_{i=1}^n \mid L(Z_i,\boldsymbol{\theta}) - L(Z_i,\boldsymbol{\theta}_j) \mid \leq \frac{1}{n}\sum_{i=1}^n \Delta(Z_i,\delta)$.

Next, (2) $< \max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid$; and choose δ to be small, then (3) $\leq \sup_{\{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2):\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|<\delta\}} \mid q_0(\boldsymbol{\theta}_1) - q_0(\boldsymbol{\theta}_2) \mid \leq \epsilon^*(\delta)$, where $\epsilon(\delta) \to 0$ as $\delta \to 0$.

Combining (1) - (3), we have:

$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mid q(\boldsymbol{\theta};\mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid \leq \frac{1}{n}\sum_{i=1}^n \Delta(Z_i,\delta) + \max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid + \epsilon^*(\delta)$. Choose $\delta_1 \in \{\delta : \epsilon^*(\delta) \leq \frac{\epsilon}{3}\}$. Then for any $\delta < \delta_1$, we have:

$$\begin{aligned} &P_{\boldsymbol{\theta}_0}[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mid q(\boldsymbol{\theta};\mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid > \epsilon] \\ &\quad \leq P_{\boldsymbol{\theta}_0}[\frac{1}{n}\sum_{i=1}^n \Delta(Z_i,\delta) + \max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid > \frac{2\epsilon}{3}] \\ &\quad \leq P_{\boldsymbol{\theta}_0}[\frac{1}{n}\sum_{i=1}^n \Delta(Z_i,\delta) > \frac{\epsilon}{3}] + & \text{(4a)} \\ &\qquad\quad P_{\boldsymbol{\theta}_0}[\max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid > \frac{\epsilon}{3}] & \text{(4b)} \end{aligned}$$

Note that (4a) $= P_{\boldsymbol{\theta}_0}[\frac{1}{n}\sum_{i=1}^n\{\Delta(Z_i,\delta) - \mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)]\} + \mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)] > \frac{\epsilon}{3}]$, where $\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)] \to 0$ as $\delta \to 0$. Choose $\delta_2 \in \{\delta : \mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)] < \frac{\epsilon}{6}\}$. Take $\delta < \min(\delta_1,\delta_2)$. Then:

$$P_{\boldsymbol{\theta}_0}[\frac{1}{n}\sum_{i=1}^{n}\{\Delta(Z_i,\delta)-\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)]\}>\frac{\epsilon}{6}]:=(4)'$$

By the Weak Law of Large Numbers (WLLN), $\exists\ N_1(\epsilon,\xi)$ such that $\forall\ n>N_1(\epsilon,\xi)$, $(4)<(4)'<\frac{\xi}{2}$.

Consider the finite subcover $\{B(\boldsymbol{\theta}_j,\delta), j=1,\dots,J\}$ for δ considered above. Note that:

$$(5)=P_{\boldsymbol{\theta}_0}[\bigcup_{j=1}^{J}\{\mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n)-q_0(\boldsymbol{\theta}_j)\mid>\frac{\epsilon}{3}\}]\leq\sum_{j=1}^{J}P_{\boldsymbol{\theta}_0}[\mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n)-q_0(\boldsymbol{\theta}_j)\mid>\frac{\epsilon}{3}].$$

By the WLLN, $\forall\boldsymbol{\theta}_j$ and $\forall\epsilon,\xi>0$, $\exists\ N_{2j}(\epsilon,\xi)$ such that $\forall n>N_{2j}(\epsilon,\xi)$: $P_{\boldsymbol{\theta}_0}[\mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n)-q_0(\boldsymbol{\theta}_j)\mid>\frac{\epsilon}{3}]\leq\frac{\xi}{2J}$. Let $N_2(\epsilon,\xi)=\max_{j\in\{1,\dots,J\}}\{N_{2j}\}$. Then, $\forall n>N_2(\epsilon,\xi)$, we have: $\sum_{j=1}^{J}P_{\boldsymbol{\theta}_0}[\mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n)-q_0(\boldsymbol{\theta}_j)\mid>\frac{\epsilon}{3}]<\frac{\xi}{2}$. Hence, (4b) $<\frac{\xi}{2}$.

(4a) and (4b) show that $\exists$ an $N(\epsilon,\xi)=\max\big(N_1(\epsilon,\xi),N_2(\epsilon,\xi)\big)$ such that $\forall n>N(\epsilon,\xi)$, $P_{\boldsymbol{\theta}_0}[\sup_{\boldsymbol{\theta}\in\Theta}\mid q(\boldsymbol{\theta};\mathbf{Z}_n)-q_0(\boldsymbol{\theta})\mid]<\xi$. ■

B.3. Proof of Theorem 1. Define $Q(\boldsymbol{\theta};\mathbf{Z}_n):=\frac{1}{n}L(\boldsymbol{\theta};\mathbf{Z}_n)$ and $Q_0(\boldsymbol{\theta}):=\mathbb{E}_{\boldsymbol{\theta}_0}(L(\boldsymbol{\theta}_0;\mathbf{Z}_n))$, $L(\boldsymbol{\theta};z):=f^{\text{obj}}\big(\boldsymbol{\theta};z\big)+\frac{\lambda(\boldsymbol{\theta})(f^{\text{cons}}-c)}{n}$, and $\hat{\boldsymbol{\theta}}(\mathbf{Z}_n):=\text{argmax}_{\boldsymbol{\theta}}\{L(\mathbf{Z}_n;\boldsymbol{\theta})\}$, henceforth $\hat{\boldsymbol{\theta}}$.

I. Consistency

To show

$$\hat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0, \tag{5}$$

by Lemma B.1, it suffices to show: (a) $Q(\boldsymbol{\theta};\mathbf{Z}_n)$ is continuous in $\boldsymbol{\theta}$; (b) $Q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$; and (c) $\sup_{\boldsymbol{\theta}\in H}\mid Q(\boldsymbol{\theta};\mathbf{Z}_n)-Q_0(\boldsymbol{\theta})\mid\xrightarrow{p}0$. Note that (a) is implied by (N-3); (N-3), (N-4), and Lemma B.2 give (d) $q_0(\boldsymbol{\theta}):=\mathbb{E}_{\boldsymbol{\theta}_0}(L(z,\boldsymbol{\theta}))$ is continuous in $\boldsymbol{\theta}$; and (e) $\sup_{\boldsymbol{\theta}\in H}\parallel\hat{q}(\boldsymbol{\theta};\mathbf{Z}_n)-q_0(\boldsymbol{\theta})\parallel\xrightarrow{p}0$, where $\hat{q}(\boldsymbol{\theta};\mathbf{Z}_n)=\frac{1}{n}\sum_{i=1}^{n}L(Z_i,\boldsymbol{\theta})$. Then (d) implies (b); (e) implies (c). ■

II. Asymptotic Normality

Let $\ell(z;\boldsymbol{\theta}):=\frac{\partial f^{\text{obj}}(z;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+\frac{\nabla^{\boldsymbol{\theta}}\lambda(\boldsymbol{\theta})(f^{\text{cons}}(\boldsymbol{\theta})-c)+\lambda(\boldsymbol{\theta})\nabla^{\boldsymbol{\theta}}f^{\text{cons}}(\boldsymbol{\theta})}{n}$, $\hat{q}(\mathbf{Z}_n;\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^{n}L(Z_i;\boldsymbol{\theta})$,

$q_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(L(z;\boldsymbol{\theta}))$, $\hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta}) = \dfrac{1}{n}\sum_{i=1}^n \ell(Z_i;\boldsymbol{\theta})$, and $\dot{q}_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(\ell(z;\boldsymbol{\theta}))$.

$\hat{\boldsymbol{\theta}}$ satisfies: $\dfrac{\partial L(\mathbf{Z}_n,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}/_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$. Hence, $\hat{\dot{q}}(\mathbf{Z}_n;\hat{\boldsymbol{\theta}}) = 0$. (N-2) and (N-6) give,

$$\dot{q}_0(\boldsymbol{\theta}) = o_p(n^{-1/2}). \tag{6}$$

Expanding $\hat{\dot{q}}(\mathbf{Z}_n;\hat{\boldsymbol{\theta}})$, we have:

$$0 = \hat{\dot{q}}(\mathbf{Z}_n;\hat{\boldsymbol{\theta}}) = \hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta}_0) + D_n^*(\mathbf{Z}_n)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{7}$$

where $D_n^*(\mathbf{Z}_n) = \dfrac{\partial\hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}/_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$, for some $\boldsymbol{\theta}^*$ in the open interval bounded by $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$.

(N-9), (N-10) and Lemma B.2 give: $\sup \parallel \hat{D}(\mathbf{Z}_n;\boldsymbol{\theta}) - D_0(\boldsymbol{\theta}) \parallel \xrightarrow{p} 0$, where $D_0(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}\big(\nabla^{\boldsymbol{\theta}}\ell(z;\boldsymbol{\theta})\big)$ and $D_0(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\big(\nabla^{\boldsymbol{\theta}}\ell(z;\boldsymbol{\theta}_0)\big)$.

Since $\boldsymbol{\theta}^* \in \mathcal{N}_r(\boldsymbol{\theta}_0)$ w.p.1., then $D_n^*(\mathbf{Z}_n) \xrightarrow{p} D_0(\boldsymbol{\theta}_0)$. From (7), we have $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\{D_n^*(\mathbf{Z}_n)\}^{-1}\{\sqrt{n}\hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta}_0)\}$; From (6), we have $\sqrt{n}\hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta}_0) = \sqrt{n}\left(\dfrac{\sum_{i=1}^n \ell(Z_i;\boldsymbol{\theta}_0)}{n} - \dot{q}_0(\boldsymbol{\theta})\right) + \sqrt{n}\dot{q}_0(\boldsymbol{\theta}) \to N\big(0, V(\boldsymbol{\theta}_0)\big)$, where $V(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\ell(z;\boldsymbol{\theta})\ell^\intercal(z;\boldsymbol{\theta})\right)$. It follows:

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

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

B.4 Proof of Theorem 2.

I. Consistency. (N-0), and Eqs. (5)-(7) in Section 3.3 show that, for every fixed estimate $\hat{\boldsymbol{\theta}}$, there is an unique $\mathbf{w} : \mathbb{R}^{\dim(\hat{\boldsymbol{\theta}})} \longmapsto \mathbb{R}^{\dim(\mathbf{w})}$, such that $\mathbf{w} = \mathbf{w}(\hat{\boldsymbol{\theta}})$. Hence, under (N-13), (5) gives: $\mathbf{w}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{w}(\boldsymbol{\theta}_0)$.

II. Asymptotic Normality. Under (N-12) and (N-13), and by the Multivariate Delta Method,

$$\sqrt{n}(\mathbf{w}(\hat{\boldsymbol{\theta}}) - \mathbf{w}(\boldsymbol{\theta}_0)) \to N\left(0, [\nabla\mathbf{w}(\boldsymbol{\theta})]^\intercal\Sigma(\boldsymbol{\theta}_0)\nabla\mathbf{w}(\boldsymbol{\theta})\right). \qquad ■$$

Appendix C. Generalized Population Value Decomposition

The Population Value Decomposition (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 $\overline{\mathbf{M}}_i$ is given by

$$\overline{\mathbf{M}}_i = \mathbf{U}_i^B \tilde{\mathbf{V}}_i \mathbf{D} + \mathbf{E}_i, \tag{8}$$

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.

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

Step 1: For each subject i, use SVD to compute: $\overline{\mathbf{M}}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\intercal \approx \mathbf{U}_i^B \mathbf{\Sigma}_i^B (\mathbf{V}_i^B)^\intercal$ 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 *and others*, 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 (8) can be summarized as follows:

$$\begin{aligned} \overline{\mathbf{M}}_i &= \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\intercal \approx \mathbf{U}_i^B \mathbf{\Sigma}_i^B (\mathbf{V}_i^B)^\intercal \\ &\approx \mathbf{U}_i^B \underbrace{\{\mathbf{\Sigma}_i^B (\mathbf{V}_i^B)^\intercal \mathbf{D}^T\}}_{\tilde{V}_i} \mathbf{D} = \mathbf{U}_i^B \tilde{\mathbf{V}}_i \mathbf{D}, \end{aligned} \tag{9}$$

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 (9) 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.

Appendix D. Simulation Study

D.1 Simulation Set-up

Here we describe a series of simulation studies to investigate the efficacy of our approach. Assume that, for every observation, the mediator $\mathbf{M}_{it}$ and treatment X_{it} can be jointly simulated from an independent, identically distributed multivariate normal distribution with known mean and variance.

In particular, let

$$\begin{pmatrix}\mathbf{M}_{it}\\ X_{it}\end{pmatrix}\Bigg| \boldsymbol{\mu}, \mathbf{\Sigma} \sim N_{p+1}\big(\boldsymbol{\mu}, \mathbf{\Sigma}\big) \tag{10}$$

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

$$\big\{\mathbf{M}_{it}|X_{it} = x_{it}\big\} \sim N(\bar{\boldsymbol{\mu}}, \bar{\mathbf{\Sigma}}), \tag{11}$$

where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}^M + \mathbf{\Sigma}^{M,X}[\Sigma^X]^{-1}(x_{it} - \mu^X)$, and $\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}^M - \mathbf{\Sigma}^{M,X}[\Sigma^X]^{-1}\mathbf{\Sigma}^{X,M}$. From (3.7):

$$\mathbb{E}(\mathbf{M}_{it}^\intercal \mathbf{w}_1 | X_{it} = x_{it}) = \alpha_0 + \alpha_1 x_{it}.$$

Solving (3.7) and (11), we can write:

$$\begin{aligned}\alpha_0 &= \mathbf{w}_1[\boldsymbol{\mu}^M - \mathbf{\Sigma}^{M,X}[\Sigma^X]^{-1}\mu^X];\\ \alpha_1 &= \mathbf{w}_1[\mathbf{\Sigma}^{M,X}[\Sigma^X]^{-1}].\end{aligned} \tag{12}$$

Moreover,

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

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 $(\mathbf{M}_{it}, X_{it})$ according to (10) ;

2. Set the values for β_0, β_1, and γ_1, as well as $\mathbf{w}_1$. Compute α_0 and α_1 using (12) . Consider these to be the true path coefficients $\boldsymbol{\theta}_1$ and the first direction of mediation $\mathbf{w}_1$;

3. Simulate random error η_{it} from a normal distribution with known mean and variance. Given $(\mathbf{M}_{it}, X_{it})$, η_{it}, and the path coefficients, generate Y_{it}, for $t = 1, \ldots n_i$ and $i = 1, \ldots n$, according to (3.17).

The data generated from Steps 1 and 3 are used as input in the LSEM.

We perform four different simulations. Simulations 1 and 2 evaluate the estimation accuracy of the algorithm described in Section 3.2, focusing primarily on the parameters associated with the first DM. Simulations 3 and 4 evaluate the performance of the GPVD and bootstrap in the high-dimensional setting. Here we attempt to mimic aspects of the fMRI data example analyzed in Section 6. Below we outline the four simulations in turn.

Simulation 1. Let $p = 3$, $\mathbf{w}_1 = (0.85, 0.17, 0.51)$, $\boldsymbol{\mu} = (2, 3, 4, 5)$, $\boldsymbol{\Sigma}^{M,X} = (0.60, -0.90, 0.35)^\intercal$, and $\Sigma^X = 2.65$. Set the true path coefficients $(\beta_0, \beta_1, \gamma)$ equal to $(0.4, 0.2, 0.5)$. From (12) it follows that $(\alpha_0, \alpha_1) = (3.23, 0.20)$. Assuming $\eta_{it} \sim N(0, 1)$, we simulated X_{it}, Y_{it}, and $\mathbf{M}_{it}$, with $n_i = 1\ \forall i$ and $n = 10, 100, 500$, and $1,000$. This procedure is repeated $1,000$ times, and each time the parameter estimates associated with the first DM are recorded.

Simulation 2. Let $p = 10$, $\mathbf{w}_1 = (0.42, 0.09, 0.25, 0.42, 0.17, 0.34, 0.51, 0.17, 0.17, 0.34)$, $\boldsymbol{\mu} = (2, 3, 4, 5, 4, 6, 2, 5, 8, 1, 3)$, $\boldsymbol{\Sigma}^{M,X} = (-1.48, -0.51, -0.81, 0.98, -1.21, 0.53, -0.66, -0.73, -1.00, 0.29)^{\intercal}$, and $\Sigma^X = 5.10$. Set the true pathway coefficients $(\beta_0, \beta_1, \gamma)$ to $(0.4, 0.2, 0.5)$. From (12) it follows that $(\alpha_0, \alpha_1) = (11.08, -0.20)$. Assuming $\eta_{it} \sim N(0, 1)$, we simulated X_{it}, Y_{it}, and $\mathbf{M}_{it}$, with $n_i = 1\ \forall i$ and $n = 100$, and $1,000$. This procedure is repeated $1,000$ times, and each time the parameter estimates associated with the first DM are recorded.

Simulation 3. Data are generated under the null hypothesis that α_k and β_k are both equal to $0\ \forall k$, i.e., assuming no indirect effect. Consider $\mathbf{X}$, a vector of length $1,149$, that takes values between $[44.3, 49.3]$. The values of $\mathbf{X}$ are chosen to be equivalent to those in the fMRI data studied in the next section. Thus, in this simulation $n = 33$ and n_i values range from 58 to 75. Assuming $(\beta_0, \gamma) = (-15, 0.5)$ and $\epsilon_{it} \sim N(0, 0.5)$, we generate $\mathbf{Y}_{it}$ according to (3.17), and let $M_{it}^{(j)} \sim N(m_i, s_i)$ for $j = 1 \ldots p$, where $p = 10,000$, $m_i \sim N(2, 5)$ and $s_i \sim N(20, 5)$. Here $M_{it}^{(j)}$ represents the simulated value of the j^{th} voxel for subject i on trial t. We perform GPVD on $\mathbf{M}$ assuming $B = 35$. Next, we perform the bootstrap procedure ($R = 500$), and create confidence intervals for $\mathbf{w}_k$ and $\alpha_k\beta_k$ for the first three DMs.

Simulation 4. Again, consider $\mathbf{X}$, a vector of length $1,149$, that ranges between $[44.3, 49.3]$, with values chosen to be equivalent to the fMRI data studied in the next section. Let $p = 12,000$. The values of $\mathbf{M}$ are generated under the assumption that the center $4,000$ voxels are active, while the remaining $8,000$ are non-active. This is achieved by simulating data from active voxels using a $N(\alpha\mathbf{X}, 5)$ distribution, and from non-active voxels using a $N(0, 5)$ distribution. Entries of $\mathbf{w}_1$ are set to a boxcar that weighs active voxels by $1/4000$ and non-active voxels by 0. Values of $\mathbf{Y}$ are simulated according to (3.17), where $(\beta_0, \gamma, \beta_1) = (0, 0.12, 0.5)$ and $\eta_{it} \sim N(0, 0.5)$. We perform GPVD on $\mathbf{M}$ assuming $B = 35$. Next, we perform the bootstrap procedure ($R = 500$), and create confidence intervals for $\mathbf{w}_k$ and $\alpha_k\beta_k$ for the first three DMs.

D.2 Simulation Results

Fig. 1 displays results for the case when $p = 3$, and the sample size n is equal to 10, 100, 500, and $1,000$, respectively. Clearly, the method provides good estimates of the parameters of interest, with all bootstrap distributions centered around the true values (illustrated using a red dashed line). In addition, the bootstrap distributions narrow around the true values as the sample size increases, as would be expected. Note that $\mathbf{w}_k$, α_k, and β_k are all bimodal due to the sign ambiguity described in Section 3.2. Fig. 2 displays similar results for the case when $p = 10$, and the sample size n is equal to 100 and $1,000$. Again, as n increases, the estimates become increasingly accurate with a smaller standard deviation. Together, Simulations 1 and 2 indicate that the estimation algorithm is able to effectively estimate both the first DM ($\mathbf{w}_1$) and the terms needed to estimate the direct and indirect effects ($\boldsymbol{\theta}_1$).

Simulations 3 and 4 were designed to evaluate the GPVD approach towards data reduction, and the bootstrap procedure for evaluating the voxel weights and parameters corresponding to the k^{th} DM. Fig. 3 shows the results of Simulation 3. Recall that data is generated assuming no indirect effect. The first row shows results corresponding to the first DM. The histogram of the p-values for each element of $\mathbf{w}_1$ are shown to the left. These clearly follow a uniform distribution, as one would expect in the null setting. Similarly, the bootstrap distributions of α_1, β_1, $\alpha_1\beta_1$, and γ are all centered around zero; again consistent with the null setting. The bottom two rows illustrate equivalent results for the second and third DM, which show the expected behavior under the null hypothesis.

Fig. 4 shows results for Simulation 4. Here, data is generated assuming that the center $4,000$ voxels are active, while the remaining $8,000$ are non-active. The top row shows the 500 bootstrap samples for $|\mathbf{w}_1|$ overlayed on one another. Clearly, there is a shift in the distribution for active voxels. The bottom row shows histograms of the bootstrap distribution for an exemplar non-active (voxel 3000) and active (voxel 7000) element. Here it is important to note that the active voxel follows an approximate normal distribution, while the non-active voxels is roughly half-normal. The latter result motivates our choice of the half-normal for computing p-values for $|\mathbf{w}_k|$; see Section 4.

Appendix E. Data Acquisition and Preprocessing

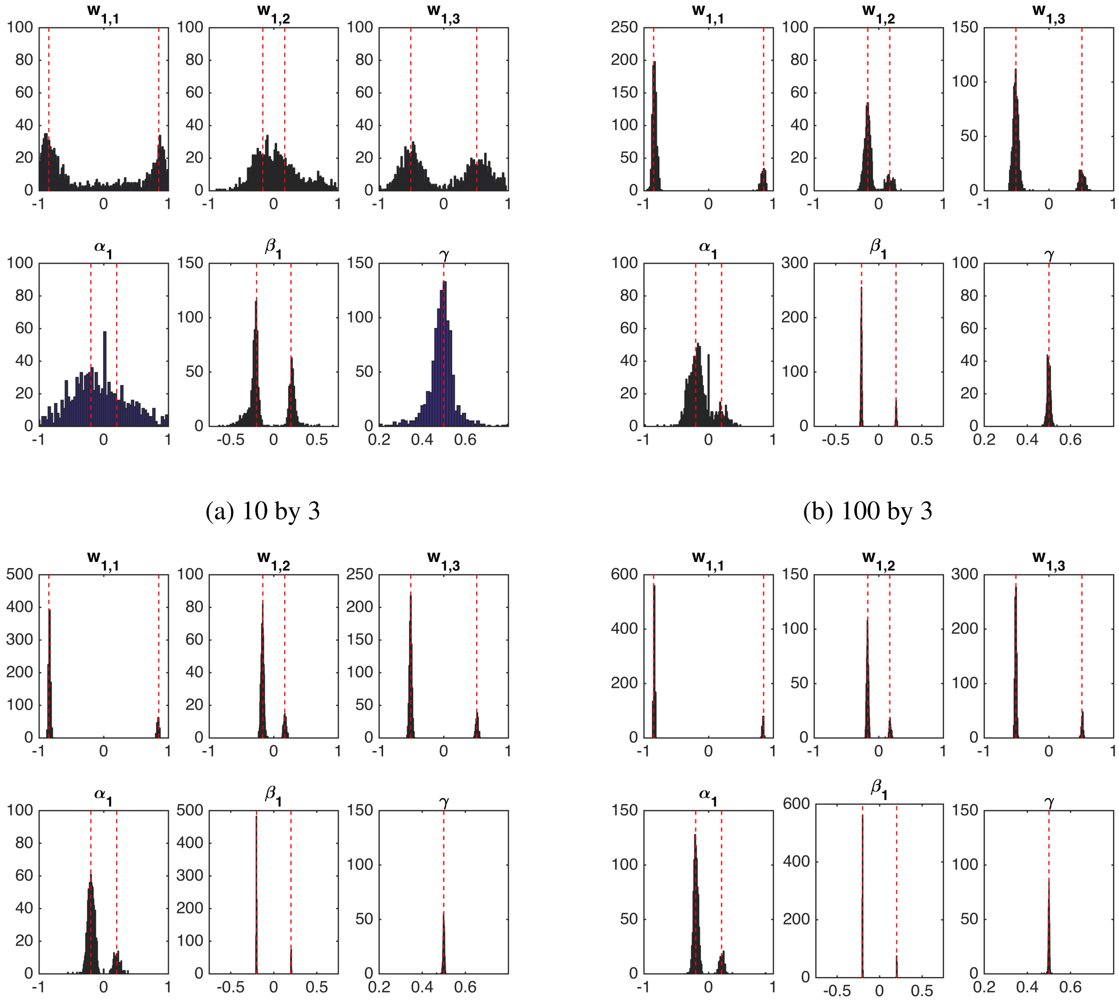
Whole-brain fMRI data was acquired on a 3T Philips Achieva TX scanner at Columbia University. Structural images were acquired using high-resolution T1 spoiled gradient recall (SPGR) images with the intention of using them for anatomical localization and warping to a standard space. Functional EPI images were acquired with TR = 2000ms, TE = 20ms, field of view = 224mm, 64×64 matrix, $3 \times 3 \times 3$mm^3 voxels, 42 interleaved slices, parallel imaging, SENSE factor 1.5. For each subject, structural images were co-registered to the mean functional image using the iterative mutual information-based algorithm implemented in SPM8[1]. Subsequently, structural images were normalized to MNI space using SPM8's generative segment-and-normalize algorithm. Prior to preprocessing of functional images, the first four volumes were removed to allow for image intensity stabilization. Outliers were identified using the Mahalanobis distance for the matrix of slice-wise mean and the standard deviation values. The functional images were corrected for differences in slice-timing, and were motion corrected using SPM8. The functional images were warped to SPMs normative atlas using warping parameters estimated from coregistered, high resolution structural images, and smoothed with an 8mm FWHM Gaussian kernel. A high-pass filter of 180s was applied to the time series data.

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

References and Notes

NEWEY, W.K. AND MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics* **4**, 2111–2245.

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).



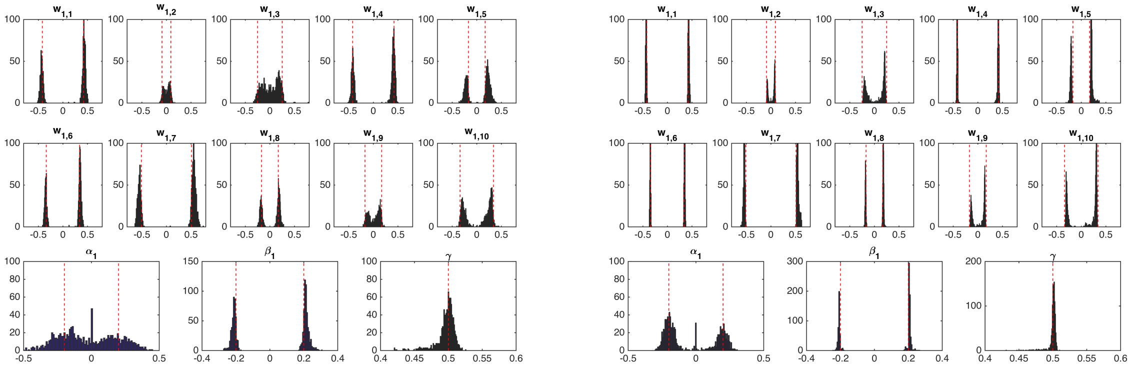
(a) 10 by 3

(b) 100 by 3

(c) 500 by 3

(d) 1000 by 3

Figure 1: Results for Simulation 1. Here $p = 3$ and values of n range from 10 to 1000 while keeping the ground truth values of $\mathbf{w}_1$ and $\boldsymbol{\theta}_1 = (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma)$ fixed. The top row of each panel shows the distribution across repetitions for each element of $\mathbf{w}_1$. The bottom row shows distributions for α_1, β_1, and γ. The red lines indicate the true values. Note for $\mathbf{w}_1$, α_1 and β_1 we show both plus and minus the absolute value of the true parameter values due to the sign ambiguity.



(a) 100 by 10

(b) 1000 by 10

Figure 2: Results for Simulation 2. Here $p = 10$ and values of n range from 100 to 1000 while keeping the ground truth values of $\mathbf{w}_1$ and $\boldsymbol{\theta}_1 = (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma)$ fixed. The top two rows of each panel shows the distribution across repetitions for each element of $\mathbf{w}_1$. The bottom row shows distributions for α_1, β_1, and γ. The red lines indicate the true values. Note for $\mathbf{w}_1$, α_1 and β_1 we show both plus and minus the absolute value of the true parameter values due to the sign ambiguity.

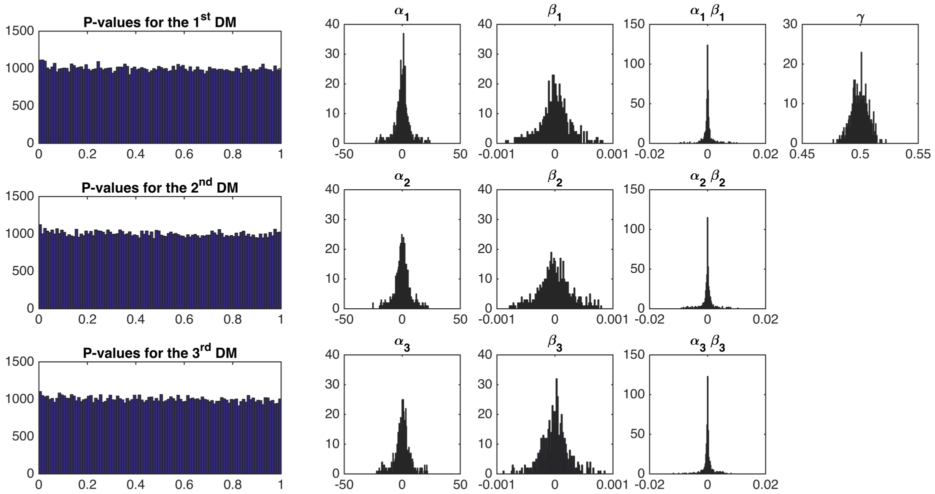


Figure 3: Results for Simulation 3. Data is generated under the null hypothesis of no mediation effect. The first row shows results corresponding to the first DM. From left to right is the histogram of the p-values for each element of $\mathbf{w}_1$, and bootstrap distributions for α_1, β_1, $\alpha_1\beta_1$, and γ. The second and third rows show similar results for the second and third DM. Note γ is the same throughout and is only shown once.

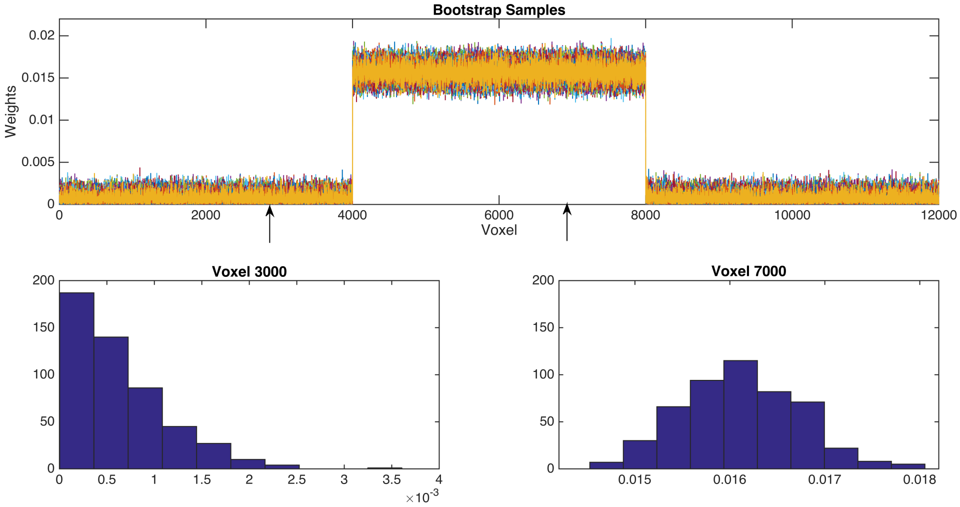


Figure 4: Results for Simulation 4. Data is generated assuming voxels $4001 - 8000$ are active, while the remaining 8000 are non-active. The top row shows the 500 bootstrap samples for $|\mathbf{w}_1|$ overlayed on one another. The bottom row shows histograms of the bootstrap distribution for an exemplar non-active (voxel 3000) and active (voxel 7000) element. The active voxel follows an approximate normal distribution, while the non-active voxels is roughly half-normal.