The Principal Direction of Mediation Supplementary Materials

Oliver Chén, Elizabeth Ogburn, Ciprian Crainiceanu,
Brian Caffo, Martin Lindquist

1 Proofs

Lemma 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:

(1) $Q_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$;

(2) $\boldsymbol{\Theta}$ is compact;

(3) $Q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$;

(4) $Q(\boldsymbol{\theta}; \mathbf{Z}_n)$ 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 1:

See Theorem 2.1 in (Newey and McFadden, 1994). ■

Lemma 2

Consider a compact space $\boldsymbol{\Theta}$. Consider:

$$L(z, \mathbf{w}, \boldsymbol{\theta}, \lambda) = f^{\text{obj}}(z, \mathbf{w}, \boldsymbol{\theta}) + f^{\text{pen}}(z, \mathbf{w}, \lambda),$$

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.

If:

(1) both the objective function and the penalization function can be profiled by $\boldsymbol{\theta}$, defined as $f^{\text{obj}}(z, \boldsymbol{\theta})$ and $f^{\text{pen}}(\boldsymbol{\theta}) := \dfrac{\lambda(\boldsymbol{\theta})\{f^{\text{cons}}(\boldsymbol{\theta}) - c\}}{n}$;

(2) the objective function is a $\log$ likelihood function;

(3) both $f^{\text{obj}}(z, \boldsymbol{\theta})$ and $f^{\text{pen}}(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$;

(4) if 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 $q(\boldsymbol{\theta}; \mathbf{Z}_n) := \dfrac{1}{n} L(Z_i, \boldsymbol{\theta})$.

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 2:

Consider the regularity conditions in the Appendix.

Part i:

$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, 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), then $q_0(\boldsymbol{\theta}_k) := \mathbb{E}_{\boldsymbol{\theta}_0}(L(Z, \boldsymbol{\theta}_k)) \to \mathbb{E}_{\boldsymbol{\theta}_0}(L(Z, \boldsymbol{\theta})) = q_0(\boldsymbol{\theta})$, by the dominated convergence theorem (DCT henceforth). Hence, $q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$.

Part ii:

We need to show that $\forall \epsilon, \xi$, $\exists N(\epsilon, \xi)$ such that $\forall n > N(\epsilon, \xi)$,

$$P[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mid q(\boldsymbol{\theta}; \mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid] < \xi.$$

Since both $f^{\text{obj}}(z, \boldsymbol{\theta})$ and $f^{\text{prof}}(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$, then $L(z, \boldsymbol{\theta})$ is uniformly continuous. 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 $\bigcup_{j=1}^{J<\infty} B(\boldsymbol{\theta}_j,\delta) \supset \boldsymbol{\Theta}$.

By the triangle inequality, we have:

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

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,

$$(2) < \max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid$$

Moreover, choosing δ to be small,

$$(3) \leq \sup_{\{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2):\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|<\delta\}} \mid q_0(\boldsymbol{\theta}_1) - q_0(\boldsymbol{\theta}_2) \mid \leq \epsilon^*(\delta)$$

where $\epsilon(\delta) \to 0$ as $\delta \to 0$.

Combine (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 δ such that $\epsilon^*(\delta) < \frac{\epsilon}{3}$. Call this δ_1. So for any $\delta < \delta_1$, we have:

$$P_{\boldsymbol{\theta}_0}[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mid q(\boldsymbol{\theta};\mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid > \epsilon]$$

$$\leq P_{\boldsymbol{\theta}_0}[\frac{1}{n}\textstyle\sum_{i=1}^{n}\Delta(Z_i,\delta) + \max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid > \frac{2\epsilon}{3}]$$

$$\leq P_{\boldsymbol{\theta}_0}[\frac{1}{n}\sum_{i=1}^{n}\Delta(Z_i,\delta) > \frac{\epsilon}{3}] + \tag{4}$$

$$P_{\boldsymbol{\theta}_0}[\max_{j\in\{1,\ldots,J\}} \mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid > \frac{\epsilon}{3}] \tag{5}$$

Note (4) $= P_{\boldsymbol{\theta}_0}[\frac{1}{n}\sum_{i=1}^{n}\{\Delta(Z_i,\delta) - \mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)]\} + \mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)] > \frac{\epsilon}{3}]$, where $\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)] \to 0$ as $\delta \to 0$.

Choose δ small enough such that $\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta(Z;\delta)] < \frac{\epsilon}{6}$. Call this δ_2.

Take $\delta < \min(\delta_1, \delta_2)$. Then:

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

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

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

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

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 $N_2(\epsilon,\xi) = \max_{j\in\{1,\ldots,J\}}\{N_{2j}\}$. Then, $\forall n > N_2(\epsilon,\xi)$, we have:

$$\textstyle\sum_{j=1}^{J} P_{\boldsymbol{\theta}_0}[\mid q(\boldsymbol{\theta}_j;\mathbf{Z}_n) - q_0(\boldsymbol{\theta}_j) \mid > \frac{\epsilon}{3}] < \frac{\xi}{2}.$$

Hence, (5) $< \frac{\xi}{2}$.

The result for (4) and (5) show that $\exists$ an $N(\epsilon,\xi) = \max\big(N_1(\epsilon,\xi), N_2(\epsilon,\xi)\big)$ such that $\forall n > N(\epsilon,\xi)$,

$$P_{\boldsymbol{\theta}_0}[\sup_{\boldsymbol{\theta}\in H} \mid q(\boldsymbol{\theta};\mathbf{Z}_n) - q_0(\boldsymbol{\theta}) \mid] < \xi.\blacksquare$$

Proof of Theorem 1:

Define $Q(\boldsymbol{\theta};\mathbf{Z}_n) := \dfrac{1}{n}L(\boldsymbol{\theta};\mathbf{Z}_n)$ and $Q_0(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}_0}(L(\boldsymbol{\theta}_0;\mathbf{Z}_n))$.

Recall $L(\boldsymbol{\theta};z) := f^{\text{obj}}\big(\boldsymbol{\theta};z\big) + \dfrac{\lambda(\boldsymbol{\theta})(f^{\text{cons}} - c)}{n}$.

(i. Consistency)

We want to show $\hat{\boldsymbol{\theta}}(\mathbf{Z}_n) := \text{argmax}_{\boldsymbol{\theta}}\{L(\mathbf{Z}_n;\boldsymbol{\theta})\}$ converges to $\boldsymbol{\theta}_0$ in probability, i.e.

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0. \tag{6}$$

By Lemma 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 $\boldsymbol{\theta}$;

(c) $\sup_{\boldsymbol{\theta}\in H} \mid Q(\boldsymbol{\theta};\mathbf{Z}_n) - Q_0(\boldsymbol{\theta}) \mid \xrightarrow{p} 0$.

(a) is implied by (N-3);

Due to (N-3) and (N-4), by Lemma 2, we have:

(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 $\hat{q}(\boldsymbol{\theta};\mathbf{Z}_n) = \dfrac{1}{n}\sum_{i=1}^{n} L(Z_i,\boldsymbol{\theta})$

(d) implies (b);

(e) implies (c).$\blacksquare$

(ii. Asymptotic Normality)

The estimator $\hat{\boldsymbol{\theta}}(\mathbf{Z}_n)$ (henceforth $\hat{\boldsymbol{\theta}}$) satisfies:

$$\dot{L}(\mathbf{Z}_n, \hat{\boldsymbol{\theta}}) = \frac{\partial L(\mathbf{Z}_n, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big/_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$
$$= \sum_{i=1}^{n} \left\{ \frac{\partial f^{\text{obj}}(Z_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\nabla^{\boldsymbol{\theta}} \lambda(\boldsymbol{\theta})(f^{\text{cons}}(\boldsymbol{\theta}) - c) + \lambda(\boldsymbol{\theta}) \nabla^{\boldsymbol{\theta}} f^{\text{cons}}(\boldsymbol{\theta})}{n} \right\} = 0.$$

Recall that we have defined:

$$\ell(z; \boldsymbol{\theta}) = \frac{\partial f^{\text{obj}}(z; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\nabla^{\boldsymbol{\theta}} \lambda(\boldsymbol{\theta})(f^{\text{cons}}(\boldsymbol{\theta}) - c) + \lambda(\boldsymbol{\theta}) \nabla^{\boldsymbol{\theta}} f^{\text{cons}}(\boldsymbol{\theta})}{n} \Bigg\}.$$

Define:

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

Hence, $\hat{\boldsymbol{\theta}}$ satisfies

$$\hat{\dot{q}}(\mathbf{Z}_n; \hat{\boldsymbol{\theta}}) = 0. \tag{7}$$

By (N-2) and (N-6),

$$\dot{q}_0(\boldsymbol{\theta}) = O_p(n^{-1/2}). \tag{8}$$

Expanding $\hat{\dot{q}}(\mathbf{Z}_n; \hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$, we have:

$$0 = \hat{\dot{q}}(\mathbf{Z}_n; \hat{\boldsymbol{\theta}}) = \hat{\dot{q}}(\mathbf{Z}_n; \boldsymbol{\eta}_0) + D_n^*(\mathbf{Z}_n)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{9}$$

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

Assume (N-9) and (N-10), by Lemma 2, we have:

$$\sup \parallel \hat{D}(\mathbf{Z}_n; \boldsymbol{\theta}) - D_0(\boldsymbol{\theta}) \parallel \xrightarrow{p} 0 \tag{10}$$

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.,

$$D_n^*(\mathbf{Z}_n) \xrightarrow{p} D_0(\boldsymbol{\theta}_0). \tag{11}$$

Hence, from (9), we have:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\{D_n^*(\mathbf{Z}_n)\}^{-1}\big\{\sqrt{n}\hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta}_0)\big\}. \tag{12}$$

By the C.L.T., the result from (8), and Slutsky's theorem, we have:

$$\sqrt{n}\hat{\dot{q}}(\mathbf{Z}_n;\boldsymbol{\theta}_0) = \sqrt{n}\left(\frac{\sum_{i=1}^n \ell(Z_i;\boldsymbol{\theta}_0)}{n} - \dot{q}_0(\boldsymbol{\theta})\right) + \sqrt{n}\dot{q}_0(\boldsymbol{\theta}) \to N\big(m(\boldsymbol{\theta}_0), V(\boldsymbol{\theta}_0)\big) \tag{13}$$

where $V(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\ell(z;\boldsymbol{\theta})\ell^\intercal(z;\boldsymbol{\theta})\right) - m(\boldsymbol{\theta}_0)[m(\boldsymbol{\theta}_0)]^\intercal$ and $m(\boldsymbol{\theta}_0) = O_p(1)$.

By (11) - (13), we have:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(\mu(\boldsymbol{\theta}_0), \Sigma(\boldsymbol{\theta}_0)) \tag{14}$$

where $\mu(\boldsymbol{\theta}_0) = D_0^{-1}(\boldsymbol{\theta}_0)m(\boldsymbol{\theta}_0)$, $\Sigma(\boldsymbol{\theta}_0) = D_0^{-1}(\boldsymbol{\theta}_0)V(\boldsymbol{\theta}_0)\big)[D_0^{-1}(\boldsymbol{\theta}_0)]^\intercal$, and $m(\boldsymbol{\theta}_0) = O_p(1)$.

■

Lemma 3 (Multivariate Delta Method When $\overset{\star}{J} > \overset{\circ}{J}$, where $g(\boldsymbol{\theta}) : \mathbb{R}^{\overset{\circ}{J}} \mapsto \mathbb{R}^{\overset{\star}{J}}$)

Many a time the Multivariate Delta Method is used without sepcifying dimension differences between the domain space and codomain space, or when it does, it is implied when the dimension of the domain space, $\overset{\circ}{J}$, is larger than or equal to it of the codomain space, $\overset{\star}{J}$. Since our method and theory apply for both cases when $\overset{\circ}{J} \geq \overset{\star}{J}$ and $\overset{\circ}{J} < \overset{\star}{J}$, for completeness purpose, we include a proof of the Multivariate Delta Method with a particular emphasis that the method is applicable for cases when $\overset{\circ}{J} < \overset{\star}{J}$. The $\overset{\circ}{J} \geq \overset{\star}{J}$ case can be proven using the same technique.

Proof of Lemma 3:

We consider the mapping $\mathbf{w}(\boldsymbol{\theta}) : \mathbb{R}^{\mathring{J}} \mapsto \mathbb{R}^{J}$, where $\mathring{J} < J$, as an example.

Assume:

$\mathbf{w}(\boldsymbol{\theta})$ is continuously differentiable in $\mathcal{N}_r(\boldsymbol{\theta}_0)$.

If:

$$a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} V$$

where $\boldsymbol{\theta}$ and V are some random vectors in $\mathbb{R}^{\mathring{J}}$, $\boldsymbol{\theta}_0$ is fixed in $\mathbb{R}^{\mathring{J}}$, and $a_n \to \infty$.

then:

$$a_n\big(\mathbf{w}(\hat{\boldsymbol{\theta}}) - \mathbf{w}(\boldsymbol{\theta}_0)\big) \xrightarrow{d} \nabla \mathbf{w}(\boldsymbol{\theta}) V$$

Proof of Lemma 3:

$\forall k \in \{1, \ldots, J\}$, extend $\hat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$, we have:

$$\left\{\mathbf{w}_k(\hat{\boldsymbol{\theta}})\right\}_{1\times 1} = \left\{\mathbf{w}_k(\boldsymbol{\theta}_0)\right\}_{1\times 1} + \left\{\nabla^{\boldsymbol{\theta}} \mathbf{w}_k(\boldsymbol{\theta}_k^*)\right\}_{1\times \mathring{J}} \left\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\}_{\mathring{J}\times 1}$$

where $\boldsymbol{\theta}_k^* \in \overline{(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})}$, and $\overline{(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})}$ denotes the open ball bounded by $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$.

Define $\left\{\nabla^{\boldsymbol{\theta}} \mathbf{w}\right\}_{J\times \mathring{J}}$ as the partial derivative of $\mathbf{w}_i$ with respect to $\boldsymbol{\theta}$.

Define $P_n^* = \{\nabla^{\boldsymbol{\theta}} \mathbf{w}_1(\boldsymbol{\theta}_1^*), \ldots, \nabla^{\boldsymbol{\theta}} \mathbf{w}_J(\boldsymbol{\theta}_J^*)\}$.

We have:

$$\mathbf{w}(\boldsymbol{\theta}) = \mathbf{w}(\boldsymbol{\theta}_0) + P_n^*(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Hence, $\forall c \in \mathbb{R}^{J_1}$:

$$c^\intercal a_n\big(\mathbf{w}(\hat{\boldsymbol{\theta}}) - \mathbf{w}(\boldsymbol{\theta}_0)\big) = c^\intercal \big(P_n^* - \nabla^{\boldsymbol{\theta}} \mathbf{w}(\boldsymbol{\theta}_0)\big)\big(a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) + c^\intercal \nabla^{\boldsymbol{\theta}} \mathbf{w}(\boldsymbol{\theta}_0) a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

While the first term converges to 0 in probability, the second term converges to $c^\intercal \nabla^{\boldsymbol{\theta}} \mathbf{w}(\boldsymbol{\theta}_0) V$ in distribution, by Cramér-Wold Theorem.

Therefore, $\forall c \in \mathbb{R}^{J_1}$, where $J_1 > \mathring{J}$:

$$c^\intercal a_n \big(\mathbf{w}(\boldsymbol{\theta}) - \mathbf{w}(\boldsymbol{\theta}_0)\big) \xrightarrow{d} c^\intercal \nabla^{\boldsymbol{\theta}} \mathbf{w}(\boldsymbol{\theta}_0) V.$$

The result follows by applying the Cramér-Wold Theorem again. ■

Proof of Theorem 2:

(i. Consistency)

Under Regularity Condition (N-0), it can be shown:

$$\hat{\mathbf{w}} | \boldsymbol{\theta}, \lambda = (\lambda \mathbf{I} + \boldsymbol{\psi}(\boldsymbol{\theta}))^{-1} \boldsymbol{\phi}(\boldsymbol{\theta}) \tag{15}$$

$$\hat{\lambda} | \boldsymbol{\theta} = \arg_\lambda \left\{ [(\lambda \mathbf{I} + \boldsymbol{\psi}(\boldsymbol{\theta}))^{-1} \boldsymbol{\phi}(\boldsymbol{\theta})]^\intercal [(\lambda \mathbf{I} + \boldsymbol{\psi}(\boldsymbol{\theta})^{-1} \boldsymbol{\phi}(\boldsymbol{\theta})] = 1 \right\} \tag{16}$$

$$\hat{\boldsymbol{\theta}} | \hat{\mathbf{w}}, \hat{\lambda} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\mathbf{D}; \hat{\mathbf{w}}, \boldsymbol{\theta}, \hat{\lambda}) \tag{17}$$

By (15) and (16), we have:

$$\hat{\mathbf{w}} | \boldsymbol{\theta} = \hat{\mathbf{w}}(\boldsymbol{\theta}). \tag{18}$$

Hence, for every fixed $\hat{\boldsymbol{\theta}}$ inputting to (18), there is an unique output of $\mathbf{w} : \mathbb{R}^{\mathring{J}} \longmapsto \mathbb{R}^J$, associated with that particular $\mathbf{w}$.

Specifically, for every fixed $\mathbf{w}$ in (6), $\exists\ \mathbf{w}$, such that

$$\mathbf{w} = \mathbf{w}(\hat{\boldsymbol{\theta}}). \tag{19}$$

By the consistency proof of Theorem 1, under Regularity Condition (N-12), that $\mathbf{w}(\boldsymbol{\theta})$ is continuously differentiable in $\mathcal{N}_r(\boldsymbol{\theta}_0)$, we have:

$$\mathbf{w}(\hat{\boldsymbol{\theta}}) := \hat{\mathbf{w}} \xrightarrow{p} \mathbf{w}(\boldsymbol{\theta}_0) := \mathbf{w}_0 \tag{20}$$

$\forall i \in \{1, \ldots, q_1\}$, by the continuous mapping theorem. ■

(ii. Asymptotic Normality)

Under Regularity Condition (N-12) and (N-13), and by Lemma 3,

$$\sqrt{n}(\mathbf{w}(\hat{\boldsymbol{\theta}}) - \mathbf{w}(\boldsymbol{\theta}_0)) \to N\bigg(\nabla\mathbf{w}(\boldsymbol{\theta})\mu(\boldsymbol{\theta}_0), [\nabla\mathbf{w}(\boldsymbol{\theta})]^\intercal\Sigma(\boldsymbol{\theta}_0)\nabla\mathbf{w}(\boldsymbol{\theta})\bigg). \tag{21}$$

■

1.1 Obtain Variance Estimates of the first PDM

Recall for the first Principal Direction of Mediation, the lagrangian can be written as:

$$G(\mathbf{D}; \mathbf{w}, \boldsymbol{\theta}, \lambda) = g_1(\mathbf{D}; \mathbf{w}, \boldsymbol{\theta}) - \lambda(\mathbf{w}^\intercal\mathbf{w} - 1), \tag{22}$$

where g_1 is the joint $\log$-likelihood function.

Conditioning on $\hat{\lambda}$ and $\hat{\mathbf{w}}$, the estimates of the multiplier and the first PDM, we have:

$$G(\mathbf{D}, \hat{\mathbf{w}}, \hat{\lambda}; \boldsymbol{\theta}) = g_1(\mathbf{D}, \hat{\mathbf{w}}; \boldsymbol{\theta}). \tag{23}$$

Define $\dot{G}(\mathbf{D}, \hat{\mathbf{w}}, \hat{\lambda}; \boldsymbol{\theta}) = \dfrac{\partial g_1(\mathbf{D}, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n} \dfrac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$, where $f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})$ is the $\log$-likelihood function for $\mathbf{D}_i = (X_i, Y_i, \mathbf{M}_i)$ conditioning on $\mathbf{w} = \hat{\mathbf{w}}$, i.e.,

$$f_i(D_i, \hat{\mathbf{w}}; \boldsymbol{\theta}) = \frac{1}{2}\bigg\{\frac{(Y_i - \beta_0 - X_i\gamma_1 - \mathbf{M}_i\hat{\mathbf{w}}\beta_1)^2}{\sigma_\epsilon} + \frac{(\mathbf{M}_i\hat{\mathbf{w}} - \alpha_0 - X_i\alpha_1)^2}{\sigma_\eta}\bigg\}.$$

The normalizing term $\log\dfrac{1}{2\pi\sigma_\epsilon\sigma_\eta}$ and the minus sign in the $\log$-likelihood function are dropped because they do not affect the final result.

Define $\ell(D, \hat{\mathbf{w}}; \boldsymbol{\theta}) = \dfrac{\partial f(D, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$.

1.1.1 Estimation of $D_0(\boldsymbol{\theta}_0)$

Recall $D_0(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\frac{\partial \ell(D; \mathbf{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$.

It is easy to see:

$$\ell(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta}) = \frac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \alpha_0} \\ \frac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \alpha_1} \\ \frac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \beta_0} \\ \frac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \beta_1} \\ \frac{\partial f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \gamma_1} \end{pmatrix} = \begin{pmatrix} -\frac{(\mathbf{M}_i\mathbf{w} - \alpha_0 - X_i\alpha_1)}{\sigma_\eta^2} \\ -\frac{(\mathbf{M}_i\mathbf{w} - \alpha_0 - X_i\alpha_1)}{\sigma_\eta^2} X_i \\ -\frac{(Y_i - \beta_0 - X_i\gamma_1 - \mathbf{M}_i\mathbf{w}\beta_1)}{\sigma_\epsilon^2} \\ -\frac{(Y_i - \beta_0 - X_i\gamma_1 - \mathbf{M}_i\mathbf{w}\beta_1)}{\sigma_\epsilon^2} \mathbf{M}_i\mathbf{w} \\ -\frac{(Y_i - \beta_0 - X_i\gamma_1 - \mathbf{M}_i\mathbf{w}\beta_1)}{\sigma_\epsilon^2} X_i \end{pmatrix}. \tag{24}$$

Moreover,

$$\frac{\ell(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial f_i}{\partial \alpha_0 \partial \alpha_0} & \frac{\partial f_i}{\partial \alpha_0 \partial \alpha_1} & \frac{\partial f_i}{\partial \alpha_0 \partial \beta_0} & \frac{\partial f_i}{\partial \alpha_0 \partial \beta_1} & \frac{\partial f_i}{\partial \alpha_0 \partial \gamma_1} \\ \frac{\partial f_i}{\partial \alpha_1 \partial \alpha_0} & \frac{\partial f_i}{\partial \alpha_1 \partial \alpha_1} & \frac{\partial f_i}{\partial \alpha_1 \partial \beta_0} & \frac{\partial f_i}{\partial \alpha_1 \partial \beta_1} & \frac{\partial f_i}{\partial \alpha_1 \partial \gamma_1} \\ \frac{\partial f_i}{\partial \beta_0 \partial \alpha_0} & \frac{\partial f_i}{\partial \beta_0 \partial \alpha_1} & \frac{\partial f_i}{\partial \beta_0 \partial \beta_0} & \frac{\partial f_i}{\partial \beta_0 \partial \beta_1} & \frac{\partial f_i}{\partial \beta_0 \partial \gamma_1} \\ \frac{\partial f_i}{\partial \beta_1 \partial \alpha_0} & \frac{\partial f_i}{\partial \beta_1 \partial \alpha_1} & \frac{\partial f_i}{\partial \beta_1 \partial \beta_0} & \frac{\partial f_i}{\partial \beta_1 \partial \beta_1} & \frac{\partial f_i}{\partial \beta_1 \partial \gamma_1} \\ \frac{\partial f_i}{\partial \gamma_1 \partial \alpha_0} & \frac{\partial f_i}{\partial \gamma_1 \partial \alpha_1} & \frac{\partial f_i}{\partial \gamma_1 \partial \beta_0} & \frac{\partial f_i}{\partial \gamma_1 \partial \beta_1} & \frac{\partial f_i}{\partial \gamma_1 \partial \gamma_1} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_\eta^2} & \frac{X_i}{\sigma_\eta^2} & 0 & 0 & 0 \\ \frac{X_i}{\sigma_\eta^2} & \frac{X_i^2}{\sigma_\eta^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma_\epsilon^2} & \frac{\mathbf{M}_i\mathbf{w}}{\sigma_\epsilon^2} & \frac{X_i}{\sigma_\epsilon^2} \\ 0 & 0 & \frac{\mathbf{M}_i\mathbf{w}}{\sigma_\epsilon^2} & \frac{(\mathbf{M}_i\mathbf{w})^2}{\sigma_\epsilon^2} & \frac{X_i\mathbf{M}_i\mathbf{w}}{\sigma_\epsilon^2} \\ 0 & 0 & \frac{X_i}{\sigma_\epsilon^2} & \frac{X_i\mathbf{M}_i\mathbf{w}}{\sigma_\epsilon^2} & \frac{X_i^2}{\sigma_\epsilon^2} \end{pmatrix}. \tag{25}$$

Therefore, $D_0(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\frac{\partial \ell(D; \mathbf{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$ can be estimated by

$$\hat{D}_0 = \begin{pmatrix} \frac{1}{\hat{\sigma}_\eta^2} & \frac{\bar{X}}{\hat{\sigma}_\eta^2} & 0 & 0 & 0 \\ \frac{\bar{X}}{\hat{\sigma}_\eta^2} & \frac{\sum_{i=1}^n X_i^2}{n}/\hat{\sigma}_\eta^2 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\hat{\sigma}_\epsilon^2} & \frac{\sum_{i=1}^n \mathbf{M}_i\hat{\mathbf{w}}}{n}/\hat{\sigma}_\epsilon^2 & \frac{\bar{X}}{\hat{\sigma}_\epsilon^2} \\ 0 & 0 & \frac{\sum_{i=1}^n \mathbf{M}_i\hat{\mathbf{w}}}{n}/\hat{\sigma}_\epsilon^2 & \frac{\sum_{i=1}^n (\mathbf{M}_i\hat{\mathbf{w}})^2}{n}/\hat{\sigma}_\epsilon^2 & \frac{\sum_{i=1}^n X_i\mathbf{M}_i\hat{\mathbf{w}}}{n}/\hat{\sigma}_\epsilon^2 \\ 0 & 0 & \frac{\bar{X}}{\hat{\sigma}_\epsilon^2} & \frac{\sum_{i=1}^n X_i\mathbf{M}_i\hat{\mathbf{w}}}{n}/\hat{\sigma}_\epsilon^2 & \frac{\sum_{i=1}^2 X_i^2}{n}/\hat{\sigma}_\epsilon^2 \end{pmatrix}. \quad (26)$$

1.1.2 Estimation of $V(\boldsymbol{\theta}_0)$

Recall $V(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\left(\ell(D; \mathbf{w}, \boldsymbol{\theta})\ell^\intercal(D; \mathbf{w}, \boldsymbol{\theta})\right) - m(\boldsymbol{\theta}_0)m^\intercal(\boldsymbol{\theta}_0)$.

For simplicity, denote $f_i = f_i(\mathbf{D}_i, \hat{\mathbf{w}}; \boldsymbol{\theta})$. It is easy to see:

$$\ell(\mathbf{D}_i,\hat{\mathbf{w}};\boldsymbol{\theta})\ell^{\intercal}(\mathbf{D}_i,\hat{\mathbf{w}};\boldsymbol{\theta}) = \begin{pmatrix} \left[\frac{\partial f_i}{\partial \alpha_0}\right]^2 & \frac{\partial f_i}{\partial \alpha_0}\frac{\partial f_i}{\partial \alpha_1} & \frac{\partial f_i}{\partial \alpha_0}\frac{\partial f_i}{\partial \beta_0} & \frac{\partial f_i}{\partial \alpha_0}\frac{\partial f_i}{\partial \beta_1} & \frac{\partial f_i}{\partial \alpha_0}\frac{\partial f_i}{\partial \gamma 1} \\ \frac{\partial f_i}{\partial \alpha_1}\frac{\partial f_i}{\partial \alpha_0} & \left[\frac{\partial f_i}{\partial \alpha_1}\right]^2 & \frac{\partial f_i}{\partial \alpha_1}\frac{\partial f_i}{\partial \beta_0} & \frac{\partial f_i}{\partial \alpha_1}\frac{\partial f_i}{\partial \beta_1} & \frac{\partial f_i}{\partial \alpha_1}\frac{\partial f_i}{\partial \gamma 1} \\ \frac{\partial f_i}{\partial \beta_0}\frac{\partial f_i}{\partial \alpha_0} & \frac{\partial f_i}{\partial \beta_0}\frac{\partial f_i}{\partial \alpha_1} & \left[\frac{\partial f_i}{\partial \beta_0}\right]^2 & \frac{\partial f_i}{\partial \beta_0}\frac{\partial f_i}{\partial \beta_1} & \frac{\partial f_i}{\partial \beta_0}\frac{\partial f_i}{\partial \gamma 1} \\ \frac{\partial f_i}{\partial \beta_1}\frac{\partial f_i}{\partial \alpha_0} & \frac{\partial f_i}{\partial \beta_1}\frac{\partial f_i}{\partial \alpha_1} & \frac{\partial f_i}{\partial \beta_1}\frac{\partial f_i}{\partial \beta_0} & \left[\frac{\partial f_i}{\partial \beta_1}\right]^2 & \frac{\partial f_i}{\partial \beta_1}\frac{\partial f_i}{\partial \gamma 1} \\ \frac{\partial f_i}{\partial \gamma_1}\frac{\partial f_i}{\partial \alpha_0} & \frac{\partial f_i}{\partial \gamma_1}\frac{\partial f_i}{\partial \alpha_1} & \frac{\partial f_i}{\partial \gamma_1}\frac{\partial f_i}{\partial \beta_0} & \frac{\partial f_i}{\partial \gamma_1}\frac{\partial f_i}{\partial \beta_1} & \left[\frac{\partial f_i}{\partial \gamma_1}\right]^2 \end{pmatrix} \tag{27}$$

$$= \begin{pmatrix} R_i^2 & R_i^2 X_i & R_i S_i & R_i S_i \mathbf{M}_i\mathbf{w} & R_i S_i X_i \\ R_i^2 X_i & R_i^2 X_i^2 & R_i S_i X_i & R_i S_i \mathbf{M}_i\mathbf{w} X_i & R_i S_i X_i^2 \\ R_i S_i & R_i S_i X_i & S_i^2 & S_i^2 \mathbf{M}_i\mathbf{w} & S_i^2 X_i \\ R_i S_i \mathbf{M}_i\mathbf{w} & R_i S_i \mathbf{M}_i\mathbf{w} X_i & S_i \mathbf{M}_i\mathbf{w} & S_i^2 (\mathbf{M}_i\mathbf{w})^2 & S_i^2 \mathbf{M}_i\mathbf{w} X_i \\ R_i S_i X_i & R_i S_i X_i^2 & S_i^2 X_i & S_i^2 \mathbf{M}_i\mathbf{w} X_i & S_i^2 X_i^2 \end{pmatrix}$$

where $R_i = -\dfrac{(\mathbf{M}_i\mathbf{w} - \alpha_0 - X_i\alpha_1)}{\sigma_\eta^2}$ and $S_i = -\dfrac{(Y_i - \beta_0 - X_i\gamma_1 - \mathbf{M}_i\mathbf{w}\beta_1)}{\sigma_\epsilon^2}$.

Since $m(\boldsymbol{\theta}_0) = O_p(1)$, $V(\boldsymbol{\theta}_0)$ can be estimated by

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \frac{\sum_{i=1}^n \hat{R}_i^2}{n} & \frac{\sum_{i=1}^n \hat{R}_i^2 X_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i \mathbf{M}_i\mathbf{w}}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i X_i}{n} \\ \frac{\sum_{i=1}^n \hat{R}_i^2 X_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i^2 X_i^2}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i X_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i \mathbf{M}_i\hat{\mathbf{w}} X_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i X_i^2}{n} \\ \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i X_i}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 \mathbf{M}_i\hat{\mathbf{w}}}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 X_i}{n} \\ \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i \mathbf{M}_i\hat{\mathbf{w}}}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i \mathbf{M}_i\hat{\mathbf{w}} X_i}{n} & \frac{\sum_{i=1}^n \hat{S}_i \mathbf{M}_i\hat{\mathbf{w}}}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 (\mathbf{M}_i\hat{\mathbf{w}})^2}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 \mathbf{M}_i\hat{\mathbf{w}} X_i}{n} \\ \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i X_i}{n} & \frac{\sum_{i=1}^n \hat{R}_i \hat{S}_i X_i^2}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 X_i}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 \mathbf{M}_i\hat{\mathbf{w}} X_i}{n} & \frac{\sum_{i=1}^n \hat{S}_i^2 X_i^2}{n} \end{pmatrix}. \tag{28}$$

where $\hat{\mathbf{w}}$ is the estimates of the first PDM, and $\hat{R}_i = -\dfrac{(\mathbf{M}_i\hat{\mathbf{w}} - \hat{\alpha}_0 - X_i\hat{\alpha}_1)}{\hat{\sigma}_\eta^2}$ and $\hat{S}_i =$

$$-\frac{(Y_i - \hat{\beta}_0 - X_i\hat{\gamma}_1 - \mathbf{M}_i\hat{\mathbf{w}}\hat{\beta}_1)}{\hat{\sigma}^2_\epsilon}.$$

1.1.3 Estimation of Asymptotic Variance for Pathway Coefficients

Hence, we can estimate the asymptotic variance for pathway coefficients, $\Sigma(\boldsymbol{\theta}_0)$, by

$$\hat{\Sigma}(\hat{\boldsymbol{\theta}}) = \hat{D}_0^{-1}\hat{V}(\hat{\boldsymbol{\theta}})[\hat{D}_0^{-1}]^\intercal \tag{29}$$

where $\hat{D}_0$ and $\hat{V}(\hat{\boldsymbol{\theta}})$ are based upon (26) and (28).

1.1.4 Estimation of Asymptotic Variance for the first PDM

First, we need to find the first partial derivative of $\mathbf{w}$ with respect to $\boldsymbol{\theta}$.

Due to (11) in the paper, we have:

$$\mathbf{w}|\boldsymbol{\theta}, \hat{\lambda} = \left(\hat{\lambda}\mathbf{I} + \psi(\boldsymbol{\theta})\right)^{-1}\phi(\theta) \tag{30}$$

where

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \frac{\mathbf{M}^\intercal\mathbf{M}\beta_1^2}{\sigma^2_{\epsilon_1}} + \frac{\mathbf{M}^\intercal\mathbf{M}}{\sigma^2_{\eta_1}} \tag{31}$$

and

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = \frac{\mathbf{M}^\intercal(\alpha_{0,1} + \alpha_1\mathbf{X})}{\sigma^2_{\eta_1}} + \frac{\mathbf{M}^\intercal(\mathbf{Y} - \beta_{0,1} - \mathbf{X}\gamma_1)\beta_1}{\sigma^2_{\epsilon_1}}. \tag{32}$$

.

Lemma 4 For a non-singular square matrix $\mathbf{A}$, where elements are a function of $\boldsymbol{\theta}$, then:

$$\frac{\partial\mathbf{A}^{-1}}{\partial\boldsymbol{\theta}} = -\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\boldsymbol{\theta}}\,\mathbf{A}^{-1}.$$

Due to Lemma 4, we have $\nabla^{\mathbf{w}}(\boldsymbol{\theta}) = \dfrac{\partial\mathbf{w}}{\partial\boldsymbol{\theta}} = -\mathbf{A}^{-1}\dfrac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\,\mathbf{A}^{-1}\phi(\boldsymbol{\theta}) + \mathbf{A}^{-1}\dfrac{\partial\phi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$, where $\mathbf{A} = \hat{\lambda}\mathbf{I} + \psi(\boldsymbol{\theta})$.

With some algebra we can show that:

$$\nabla^{\mathbf{w}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \mathbf{w}}{\partial \alpha_0} \\ \frac{\partial \mathbf{w}}{\partial \alpha_1} \\ \frac{\partial \mathbf{w}}{\partial \beta_0} \\ \frac{\partial \mathbf{w}}{\partial \beta_1} \\ \frac{\partial \mathbf{w}}{\partial \gamma_1} \end{pmatrix} = \begin{pmatrix} \mathbf{A}^{-1}\frac{\mathbf{M}^\intercal \mathbf{1}}{\sigma_\eta^2} \\ \mathbf{A}^{-1}\frac{\mathbf{M}^\intercal \mathbf{X}}{\sigma_\eta^2} \\ \mathbf{A}^{-1}\big(-\frac{\mathbf{M}^\intercal \mathbf{1}\beta_1}{\sigma_\epsilon^2}\big) \\ -\mathbf{A}^{-1}\big[\frac{2\beta_1 \mathbf{M}^\intercal \mathbf{M}}{\sigma_\epsilon^2}\big]\mathbf{A}^{-1}\phi(\boldsymbol{\theta}) + \mathbf{A}^{-1}\frac{\mathbf{M}^\intercal(\mathbf{Y}-\beta_0-\mathbf{X}\gamma_1)}{\sigma_\epsilon^2} \\ \mathbf{A}^{-1}\big(-\frac{\mathbf{M}^\intercal \mathbf{X}\beta_1}{\sigma_\epsilon^2}\big) \end{pmatrix} \tag{33}$$

where $\mathbf{1} = \text{vec}_{n\times 1}(1)$.

Hence, $\nabla^{\mathbf{w}}(\boldsymbol{\theta}_0)$ can be estimated by

$$\hat{\nabla}^{\mathbf{w}}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \hat{\mathbf{A}}^{-1}\frac{\mathbf{M}^\intercal \mathbf{1}}{\hat{\sigma}_\eta^2} \\ \hat{\mathbf{A}}^{-1}\frac{\mathbf{M}^\intercal \mathbf{X}}{\hat{\sigma}_\eta^2} \\ \hat{\mathbf{A}}^{-1}\big(-\frac{\mathbf{M}^\intercal \mathbf{1}\hat{\beta}_1}{\hat{\sigma}_\epsilon^2}\big) \\ -\hat{\mathbf{A}}^{-1}\big[\frac{2\hat{\beta}_1 \mathbf{M}^\intercal \mathbf{M}}{\hat{\sigma}_\epsilon^2}\big]\hat{\mathbf{A}}^{-1}\hat{\phi}(\hat{\boldsymbol{\theta}}) + \hat{\mathbf{A}}^{-1}\frac{\mathbf{M}^\intercal(\mathbf{Y}-\hat{\beta}_0-\mathbf{X}\hat{\gamma}_1)}{\hat{\sigma}_\epsilon^2} \\ \hat{\mathbf{A}}^{-1}\big(-\frac{\mathbf{M}^\intercal \mathbf{X}\hat{\beta}_1}{\hat{\sigma}_\epsilon^2}\big) \end{pmatrix} \tag{34}$$

where $\hat{\mathbf{A}} = \hat{\lambda}\mathbf{I} + \hat{\psi}(\hat{\boldsymbol{\theta}})$, and $\hat{\boldsymbol{\theta}} = \{\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_1\}$.

Therefore, the asymptotic variance for the first PDM, $\Sigma^{\mathbf{w}}(\boldsymbol{\theta}_0)$ can be estimated by:

$$\hat{\Sigma}^{\mathbf{w}}(\hat{\boldsymbol{\theta}}) = [\hat{\nabla}^{\mathbf{w}}(\hat{\boldsymbol{\theta}})]^\intercal \hat{D}_0^{-1}\hat{V}(\hat{\boldsymbol{\theta}})[\hat{D}_0^{-1}]^\intercal \hat{\nabla}^{\mathbf{w}}(\hat{\boldsymbol{\theta}}) \tag{35}$$

with each term on the right side based upon (26), (28), and (34).

1.1.5 Estimation of the Conditional Asymptotic Variance for the first PDM in a Ultrahigh Dimensional Setting

When the mediation matrix $\mathbf{M}$ is ultrahigh dimensional, we first conduct Population Value Decomposition (PVD), as explained in the paper. We obtain $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{w}}$, where $\mathbf{M}\mathbf{w} = \tilde{\mathbf{M}}\tilde{\mathbf{w}}$ and $\tilde{\mathbf{w}} = \mathbf{D}\mathbf{w}$, where $\mathbf{D}$ is the matrix carrying population level information, obtained via the PVD in the paper. We then obtain the asymptotic variance for $\tilde{\mathbf{w}}$ using the aforementioned method, say, $\hat{\Sigma}^{\tilde{\mathbf{w}}}(\hat{\boldsymbol{\theta}})$, and it follows that estimate of the conditional asymptotic variance for $\mathbf{w}$ is

$$\hat{\Sigma}^{\mathbf{w}}(\hat{\boldsymbol{\theta}}|\mathbf{D}) = \mathbf{D}^{-}\frac{\hat{\Sigma}^{\tilde{\mathbf{w}}}(\hat{\boldsymbol{\theta}})}{n}[\mathbf{D}^{\intercal}]^{-}. \tag{36}$$

It is necessary to mention that, $\mathbf{D}$ obtained from PVD is random. If it were not conditioned, it is difficulty to obtain $\hat{\Sigma}^{\mathbf{w}}(\hat{\boldsymbol{\theta}})$ in that when $\mathbf{D}$ is of ultrahigh dimensional, it is computationaly costly to estimate the variance-covariance matrix of $\mathbf{D}$.

References and Notes

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