Identifiability of TIDEs and TEs in multivariate mediation model

Eli Sun

2025-04-04

A note: These proofs follow directly by application of Theorems 1 and 2 from Imai et al (2010). We work out the integrals for completeness and their structure closely resembles those of Sohn and Li (2019).

The model

Let there be $\boldsymbol{X}$ be an $n \times q$ matrix of (binary) treatments, $\boldsymbol{U}$ an $n \times \ell$ matrix of confounders, $\boldsymbol{M}$ an $n \times p$ matrix of candidate mediators, and $\boldsymbol{Y}$ an $n \times k$ matrix of responses. Then we assume the data are from an LSEM of the form (equivalently, that the generative model is the DAG given):

$$\boldsymbol{M} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{E}_{\boldsymbol{M}}$$

$$\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{U}\boldsymbol{\eta} + \boldsymbol{E}_{\boldsymbol{Y}}$$

where $\boldsymbol{E}_{\boldsymbol{M}}$ and $\boldsymbol{E}_{\boldsymbol{Y}}$ are appropriately-dimensioned normally distributed error matrices:

$$\text{vec}(\boldsymbol{E}_{\boldsymbol{M}}) \sim \mathcal{N}(\boldsymbol{0}_{np}, \boldsymbol{\Sigma}_{\boldsymbol{M}} \otimes I_n)$$

$$\text{vec}(\boldsymbol{E}_{\boldsymbol{Y}}) \sim \mathcal{N}(\boldsymbol{0}_{nk}, \boldsymbol{\Sigma}_{\boldsymbol{Y}} \otimes I_n)$$

where $\boldsymbol{\Sigma}_{\boldsymbol{M}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ are unspecified covariance matrices. The LSEM can be rewritten in terms of potential outcomes as

$$\boldsymbol{M}_i(\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{E}_{\boldsymbol{M}}(\boldsymbol{X}_i)$$

$$\boldsymbol{Y}_i(\boldsymbol{X}_i, \boldsymbol{M}_i(\boldsymbol{X}_i)) = \boldsymbol{M}_i(\boldsymbol{X}_i)\boldsymbol{\beta} + \boldsymbol{X}_i\boldsymbol{\tau} + \boldsymbol{U}\boldsymbol{\eta} + \boldsymbol{E}_{\boldsymbol{Y}}(\boldsymbol{X}_i, \boldsymbol{M}_i(\boldsymbol{X}_i))$$

While we will let $\boldsymbol{X}_i$ denote the treatment(s) for subject i, it will generally be the case that $q = 1$ and we restrict attention to a single treatment at a time. Additional assumptions beyond those considered below may be needed for identifiability in the cases of multiple treatments (or, at the least, certain assumptions are rendered substantially stronger).

Necessary assumptions

Before the total indirect effects (TIDEs) and total effects (TEs) can be identified, we need to make four assumptions. Unfortunately, none of these assumptions are directly testable and can only be empirically evaluated via sensitivity analyses and subject-matter knowledge.

Let $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{U}}$ be the supports for the treatment, mediator, and confounders, respectively. Then we make the following assumptions:

(A1) SUTVA: The potential outcomes for subject i do not depend on the treatment assignments for the other subjects and for each subject i, there are no hidden variations of treatment.

(A2) Positivity:

$$P(\boldsymbol{X}_i = \boldsymbol{x}|\boldsymbol{U}_i) > 0 \text{ and } P(\boldsymbol{M}_i(X_i) = \boldsymbol{m}|\boldsymbol{X}_i, \boldsymbol{U}_i) > 0$$

(A3) No interaction: There is no interaction between the treatment and the mediators on the responses

(A4) Conditional independence: For all $\boldsymbol{x}, \boldsymbol{x'} \in \boldsymbol{\mathcal{X}}$, $\boldsymbol{m} \in \boldsymbol{\mathcal{M}}$, and $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$:

(a) The treatment is ignorable: $\{\boldsymbol{Y}_i(\boldsymbol{x'}, \boldsymbol{m}), \boldsymbol{M}_i(\boldsymbol{x})\} \perp\!\!\!\perp \boldsymbol{X}_i|\boldsymbol{U}_i = \boldsymbol{u}$

(b) The mediators are ignorable: $\boldsymbol{Y}_i(\boldsymbol{x'}, \boldsymbol{m}) \perp\!\!\!\perp \boldsymbol{M}_i(\boldsymbol{x})|\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{U}_i = \boldsymbol{u}$

Assumptions (A2) and (A4), together, are often called "sequential ignorability" (Imai et al. (2010)) and they imply several of the usual causal assumptions (e.g., no unmeasured confounding).

For modeling purposes, we assume that each of the error terms are normally distributed with mean 0 and covariance matrices as given above. However, for the purpose of proving identifiability, we can relax this assumption to simply assume that the errors have mean 0.

The first conditional independence assumption implies independence between the mediator error matrix and treatment for the i^{th} individual, conditional on the confounders

$$\boldsymbol{E}_{\boldsymbol{M},i}(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{X_i}|\boldsymbol{U}_i = \boldsymbol{u}$$

from which we have the conditional expectation

$$E(\boldsymbol{E}_{\boldsymbol{M},i}(\boldsymbol{X}_i)|\boldsymbol{X}_i) = E(\boldsymbol{E}_{\boldsymbol{M},i}(\boldsymbol{X}_i)) = \boldsymbol{0}.$$

Likewise, by the second conditional independence assumption, we have the conditional independence relationship

$$\boldsymbol{E}_{\boldsymbol{Y},i}(\boldsymbol{x}, \boldsymbol{m}) \perp\!\!\!\perp \boldsymbol{M}_i|\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{U}_i = \boldsymbol{u}$$

which, in turn, gives the following conditional expectation

$$E(\boldsymbol{E}_{\boldsymbol{Y},i}(\boldsymbol{X}_i, \boldsymbol{M}_i)|\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{M}_i = \boldsymbol{m}) = E(\boldsymbol{E}_{\boldsymbol{Y},i}(\boldsymbol{x}, \boldsymbol{m})|\boldsymbol{X}_i = \boldsymbol{x}) = E(\boldsymbol{E}_{\boldsymbol{Y},i}(\boldsymbol{x}, \boldsymbol{m})) = \boldsymbol{0}.$$

Indentifiability of the indirect effects

Let $\boldsymbol{x}$ be the observed treatment for individual i, $\boldsymbol{x}_0$ a reference value for the treatment, and $\chi \in \{\boldsymbol{x}, \boldsymbol{x}_0\}$. Then the causal total indirect effect matrix is identifiable as

$$\begin{aligned}
\boldsymbol{\delta}(\chi) &= E\left\{\boldsymbol{Y}_i(\chi, \boldsymbol{M}_i(\boldsymbol{x})) - \boldsymbol{Y}_i(\chi, \boldsymbol{M}_i(\boldsymbol{x}_0)|\boldsymbol{U}_i = \boldsymbol{u}\right\} \\
&= \int \cdots \int E\left\{\boldsymbol{Y}_i|\boldsymbol{M}_i = \boldsymbol{m}, \boldsymbol{X}_i = \chi, \boldsymbol{U}_i = \boldsymbol{u}\right\} \left(dF_{\boldsymbol{M}_i|\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{U}_i = \boldsymbol{u}}(\boldsymbol{m}) - dF_{\boldsymbol{M}_i|\boldsymbol{X}_i = \boldsymbol{x}_0, \boldsymbol{U}_i = \boldsymbol{u}}(\boldsymbol{m})\right) dF_{\boldsymbol{U_i}}(\boldsymbol{u}) \\
&= \int \cdots \int (\boldsymbol{m}\boldsymbol{\beta} + \boldsymbol{x}\boldsymbol{\tau} + \boldsymbol{u}\boldsymbol{\eta}) \left(dF_{\boldsymbol{M}_i|\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{U}_i = \boldsymbol{u}}(\boldsymbol{m}) - dF_{\boldsymbol{M}_i|\boldsymbol{X}_i = \boldsymbol{x}_0, \boldsymbol{U}_i = \boldsymbol{u}}(\boldsymbol{m})\right) dF_{\boldsymbol{U_i}}(\boldsymbol{u}) \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}(\boldsymbol{x} - \boldsymbol{x}_0)
\end{aligned}$$

where the fourth equality holds by substituting the first model equation into the second . And so the total indirect effect is identifiable as $\boldsymbol{\alpha}\boldsymbol{\beta}$ per unit change in treatment. For our purposes, with $q = 1$ binary treatment, $\boldsymbol{x} - \boldsymbol{x}_0 = 1$ or 0 when the subject receives or does not receive treatment, respectively, and so the indirect effect is the comparison of treatment and control, as expected.

Identifiability of the direct effect

Similarly to the TIDEs, the direct effect, ζ, can be identified for $\chi \in \{\boldsymbol{x}, \boldsymbol{x}_0\}$ as:

$$
\begin{aligned}
\zeta(\chi) &= E\{\boldsymbol{Y}_i(\boldsymbol{x}, \boldsymbol{M}_i(\chi)) - \boldsymbol{Y}_i(\boldsymbol{x}_0, \boldsymbol{M}_i(\chi)) | \boldsymbol{U}_i = \boldsymbol{u}\} \\
&= \int \cdots \int \{E\left[\boldsymbol{Y}_i(\boldsymbol{x}, \boldsymbol{m}) | \boldsymbol{M}_i(\chi) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}\right] \\
&\qquad - E\left[\boldsymbol{Y}_i(\boldsymbol{x}_0, \boldsymbol{m}) | \boldsymbol{M}_i(\chi) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}\right]\} dF_{\boldsymbol{M}_i(\chi)|\boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \int \cdots \int \{E\left[\boldsymbol{Y}_i(\boldsymbol{x}, \boldsymbol{m}) | \boldsymbol{M}_i(\chi) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \chi\right] \\
&\qquad - E\left[\boldsymbol{Y}_i(\boldsymbol{x}_0, \boldsymbol{m}) | \boldsymbol{M}_i(\chi) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \chi\right]\} dF_{\boldsymbol{M}_i(\chi)|\boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \int \cdots \int \{E\left[\boldsymbol{Y}_i(\boldsymbol{x}, \boldsymbol{m}) | \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \chi\right] \\
&\qquad - E\left[\boldsymbol{Y}_i(\boldsymbol{x}_0, \boldsymbol{m}) | \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \chi\right]\} dF_{\boldsymbol{M}_i(\chi)|\boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \int \cdots \int \{E\left[\boldsymbol{Y}_i(\boldsymbol{x}, \boldsymbol{m}) | \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \boldsymbol{x}\right] \\
&\qquad - E\left[\boldsymbol{Y}_i(\boldsymbol{x}_0, \boldsymbol{m}) | \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \boldsymbol{x}_0\right]\} dF_{\boldsymbol{M}_i|\boldsymbol{X}_i=\chi, \boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \int \cdots \int \{E\left[\boldsymbol{Y}_i(\boldsymbol{x}, \boldsymbol{m}) | \boldsymbol{M}_i(\boldsymbol{x}) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \boldsymbol{x}\right] \\
&\qquad - E\left[\boldsymbol{Y}_i(\boldsymbol{x}_0, \boldsymbol{m}) | \boldsymbol{M}_i(\boldsymbol{x}_0) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \boldsymbol{x}_0\right]\} dF_{\boldsymbol{M}_i|\boldsymbol{X}_i=\chi, \boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \int \cdots \int \{E\left[\boldsymbol{Y}_i | \boldsymbol{M}_i(\boldsymbol{x}) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \boldsymbol{x}\right] \\
&\qquad - E\left[\boldsymbol{Y}_i | \boldsymbol{M}_i(\boldsymbol{x}_0) = \boldsymbol{m}, \boldsymbol{U}_i = \boldsymbol{u}, \boldsymbol{X}_i = \boldsymbol{x}_0\right]\} dF_{\boldsymbol{M}_i|\boldsymbol{X}_i=\chi, \boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \int \cdots \int \boldsymbol{\tau}(\boldsymbol{x} - \boldsymbol{x}_0) dF_{\boldsymbol{M}_i|\boldsymbol{X}_i=\chi, \boldsymbol{U}_i=\boldsymbol{u}}(\boldsymbol{m}) dF_{\boldsymbol{U}_i}(\boldsymbol{u}) \\
&= \boldsymbol{\tau}(\boldsymbol{x} - \boldsymbol{x}_0)
\end{aligned}
$$

In order, we have used (A4a), (A4b), (A4a), and (A4b) for the third through sixth equalities, respectively. The eighth equality follows by substituting the model and cancelling terms accordingly.

Calculating response-wise TIDEs and PIDEs

Suppressing the factor of $(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{0}})$, we have the matrix of TIDEs as

$$
\delta(\boldsymbol{x}) = \boldsymbol{\alpha}\boldsymbol{\beta} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1p} \\ \vdots & \ddots & \vdots \\ \alpha_{q1} & \cdots & \alpha_{qp} \end{pmatrix} \begin{pmatrix} \beta_{11} & \cdots & \beta_{1k} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pk} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p} \alpha_{1j}\beta_{j1} & \cdots & \sum_{j=1}^{p} \alpha_{1j}\beta_{jk} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{p} \alpha_{qj}\beta_{j1} & \cdots & \sum_{j=1}^{p} \alpha_{qj}\beta_{jk} \end{pmatrix}
$$

The indirect effects matrix will be $(q \times p)(p \times k) = (q \times k)$. The $(q'k')^{\text{th}}$, for $q' = 1, \ldots, q$ and $k' = 1, \ldots, k$, element of the matrix is the total indirect effect from treatment q' to response k'

Further, this allows for computation of partial indirect effects. Let $\boldsymbol{\alpha}_{q'}$ be the row p-vectors of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{k'}$ be the column p-vectors of $\boldsymbol{\beta}$. Then the partial indirect effects for treatment q' on outcome k' through mediator $p' = 1, \ldots, p$ are found as

$$\left(\boldsymbol{\alpha}_{q'} \odot \boldsymbol{\beta}_{k'}^{\top}\right) = \begin{pmatrix} \alpha_{11}\beta_{11} & \cdots & \alpha_{1p}\beta_{p1} \end{pmatrix}$$

where $\odot$ is the element-wise (Hadamard) product. The total indirect effect from treatment q' to all responses, as well as the total indirect effect from all treatments to response k' can be similarly calculated as the row and column sums, respectively, of $\boldsymbol{\alpha}\boldsymbol{\beta}$, should they be of interest.