

# Multilevel Mediation Analysis with Structured Unmeasured Mediator-Outcome Confounding

Yi Zhao<sup>1</sup> and Xi Luo<sup>2\*</sup>

<sup>1</sup>Department of Biostatistics and Health Data Science, Indiana University School of Medicine

<sup>2</sup>Department of Biostatistics and Data Science,  
The University of Texas Health Science Center at Houston

## Abstract

Mediation analysis usually requires the assumption that there is no unmeasured mediator-outcome confounder. However, this may not hold in many social and scientific studies. Though various parametric and nonparametric mediation methods have been developed, this assumption remains instrumental, without which the causal effects are not identifiable unless alternative assumptions are imposed. To circumvent this, a multilevel parametric structural equation modeling framework is proposed to relax this no unmeasured mediator-outcome confounding assumption under a specific data setting inspired by a real experiment. Using the proposed framework, it is shown that the causal effects are identifiable and consistently estimated. Likelihood-based approaches are proposed with efficient optimization algorithms to estimate the parameters, including the unmeasured confounding effect, instead of performing sensitivity analysis. The asymptotic consistency is established. Using extensive simulations and a functional magnetic resonance imaging dataset, the improvement of the approaches over existing methods is demonstrated. The R package `macc` for implementation is available on CRAN.

## 1 Introduction

In diverse fields of empirical research, including many in the biological and social science, scientists are often interested in identifying causal mechanisms through which treatment affects an outcome. Mediation analysis is widely used to quantify the effect that is mediated by a third variable (called a mediator) in the causal pathways. Structural equation modeling (SEM), such as the Baron-Kenny

---

\*To whom correspondence should be addressed. Email: [rossi.stat@gmail.com](mailto:rossi.stat@gmail.com).

method (Baron and Kenny, 1986) and its extensions (MacKinnon et al., 2007), can be implemented for complex data. However, the resulting SEM coefficients do not have causal interpretations unless certain causal assumptions are met. A critical and hardly testable assumption is the no unmeasured mediator-outcome confounding assumption. Imai et al. (2010) considers this assumption to be “too strong for the typical situations”, which is also unlikely to hold in our motivating functional magnetic resonance imaging (fMRI) experiment. In this paper, a novel multilevel SEM framework is proposed, which parameterizes and estimates the unmeasured mediator-outcome confounding effect. It will then adjust the estimated confounding yielding unbiased and consistent estimate of the causal effects.

The assumptions involved to perform mediation analysis have been widely studied in the literature, usually, for one-level experimental studies. See Imai et al. (2010) and VanderWeele (2016) for a review and many references therein. Three critical assumptions in mediation analysis are the no unmeasured exposure-outcome confounding, no unmeasured exposure-mediator confounding, and no unmeasured mediator-outcome confounding assumptions. Existing literature aims to seek unbiased estimate of the causal effects by relaxing one or multiple of these assumptions. In a recent study, Fulcher et al. (2020) introduced a nonparametric approach for estimating a new type of indirect effect, the population intervention indirect effect, that allows the presence of an unmeasured common cause of the exposure and outcome, but still assumes no unmeasured confounding between the mediator and outcome. This paper aims to investigate the settings under which the causal effects remain identifiable without the no unmeasured mediator-outcome confounding assumption.

The motivating example of the existence of unmeasured mediator-outcome confounding and multilevel data comes from an fMRI experiment. Scientists are interested in how the stimulus is processed in the human brain across the population. In particular, based on the prior findings (Dunn et al., 2009; Obeso et al., 2013), an important scientific question is to quantify the mediating role of a brain region called the presupplementary motor area (preSMA) when the final motor responses are programmed by another brain region called the primary motor cortex (PMC). In the experiment, the assumption of no unmeasured mediator-outcome confounder is violated as the task-unrelated activity poses significant influences on both the mediator and outcome regions. To address this complexity, one may adopt the sensitivity analysis (Imai et al., 2010) or apply the instrumental variable (IV) approach under additional structural assumptions (Ten Have et al., 2007; Small, 2011). Lindquist (2012) applied the IV approach for an fMRI dataset when the outcome variable is outside the brain, where the structural assumptions are more likely to hold.

Recently, mediation analysis has been extended to hierarchically organized experiments, partly due to the increasing popularity of such experiments. These approaches are generally developed for two-level data in practical settings. The first-level data are usually modeled following the Baron-Kenny method, and various second-level models are introduced for the parameters (see Krull and MacKinnon, 1999; Kenny et al., 2003, for example). It has not been rigorously studied if the resulting parameters have causal interpretation and suffers from similar limitations of assuming no unmeasured mediator-outcome confounding.

In this manuscript, we propose an optimization-based multilevel mediation framework, which we call it Correlated-error Mediation Analysis (CMA). We make the following specific contributions. First, parametric SEMs are introduced that allow modeling the mediator and outcome variables jointly with correlated errors to parameterize the effect of unmeasured mediator-outcome confounding. An integrated approach is introduced for modeling two/three-level data. Second, causal assumptions are studied. It is proved that parameters associated with causal effects and the unmeasured mediator-outcome confounding are identifiable from multilevel data, overcoming the nonidentifiability under single-level mediation models (Imai et al., 2010). Asymptotic analyses show that the estimators are consistent with the parametric rates. Third, efficient computational algorithms are developed to compute for a large number of parameters.

This paper is organized as follows. We introduce the multilevel SEMs for mediation analysis with correlated errors in Section 2. In Section 3, we propose likelihood-based methods to estimate the coefficients, study the identifiability of model parameters, and the asymptotic properties of the estimators. We compare these methods and demonstrate the improvement using extensive simulations in Section 4 (and Section E of the supplementary materials) and compare the analysis results of different methods in the fMRI data application in Section 5. Section 6 summarizes the paper with discussions.

## 2 Model

For simplicity, we refer to the three levels of data in the fMRI experiment by trial, session, and participant. Let  $Z_{ikl}$  be the treatment assignment for the  $l$ th trial of the  $k$ th session of the  $i$ th participant, where  $i = 1, \dots, N$ ,  $k = 1, \dots, K_i$ , and  $l = 1, \dots, n_{ik}$ . Similarly, define  $M_{ikl}$  and  $R_{ikl}$  for the observed mediator and outcome values from the same multilevel unit. In the experiment,  $Z_{ikl}$  is a binary stimulus assignment;  $M_{ikl}$  and  $R_{ikl}$  are the preSMA and PMC activations, respectively, of

the same trial. It is straightforward to adapt the proposed framework for varying  $K_i$ , and thus we will focus on  $K_i = K$ , as in the experiment, to fix the idea. Unless noted otherwise, this three-level notation will also be used for two-level data, by setting  $K = 1$ .

The proposed multilevel mediation model contains two components. At the first level, the following matrix-format mediation model is proposed, for every  $i$  and  $k$ ,

$$\begin{pmatrix} \mathbf{M}_{ik} & \mathbf{R}_{ik} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{ik} & \mathbf{M}_{ik} \end{pmatrix} \begin{pmatrix} A_{ik} & C_{ik} \\ 0 & B_{ik} \end{pmatrix} + \begin{pmatrix} \mathbf{E}_{1_{ik}} & \mathbf{E}_{2_{ik}} \end{pmatrix}, \quad (2.1)$$

$$\text{and } \text{vec} \left[ \begin{pmatrix} \mathbf{E}_{1_{ik}} & \mathbf{E}_{2_{ik}} \end{pmatrix} \right] \sim \mathcal{N} \left( \mathbf{0}, \begin{pmatrix} \sigma_{1_{ik}}^2 & \delta_{ik} \sigma_{1_{ik}} \sigma_{2_{ik}} \\ \delta_{ik} \sigma_{1_{ik}} \sigma_{2_{ik}} & \sigma_{2_{ik}}^2 \end{pmatrix} \otimes \mathbf{I}_{n_{ik}} \right), \quad (2.2)$$

where  $\text{vec}[\cdot]$  is the vectorization operator by stacking columns of a matrix,  $\otimes$  is the Kronecker product operator, and  $\mathbf{I}_{n_{ik}}$  is the  $n_{ik}$ -dimensional identity matrix.  $A_{ik}$ ,  $B_{ik}$ , and  $C_{ik}$  are the SEM coefficients in session  $k$  of participant  $i$ . Without loss of generality,  $\mathbf{R}_{ik}$  and  $\mathbf{M}_{ik}$  are assumed to be centered around 0, and thus no intercepts are included in the model. This can be achieved via subtracting the sample means from each variable. The effect of other covariates can be removed using regression (Rosenbaum, 2002). Standard parametric mediation models (Baron and Kenny, 1986; Imai et al., 2010) without unmeasured mediator-outcome confounding are special cases of models (2.1)-(2.2) by assuming  $\delta_{ik} = 0$  for every  $i$  and  $k$ , see Section 3.1 for a more detailed comparison. In Sections 3.2 and 3.3, an approach to estimate  $\delta_{ik}$  will be introduced.

For higher levels, it is proposed to pool information across the first-level coefficients in (2.1) using the following model,

$$\mathbf{b}_{ik} = \begin{pmatrix} A_{ik} \\ B_{ik} \\ C_{ik} \end{pmatrix} = \begin{pmatrix} A \\ B \\ C \end{pmatrix} + \begin{pmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{pmatrix} + \begin{pmatrix} \epsilon_{ik}^A \\ \epsilon_{ik}^B \\ \epsilon_{ik}^C \end{pmatrix} = \mathbf{b} + \mathbf{u}_i + \boldsymbol{\eta}_{ik}. \quad (2.3)$$

For three-level data, model (2.3) in essence is a mixed effects model, where  $A$ ,  $B$  and  $C$  are the fixed effects;  $\alpha_i$ ,  $\beta_i$ , and  $\gamma_i$  are the random effects of  $A_{ik}$ ,  $B_{ik}$  and  $C_{ik}$ , respectively. Similar to many mixed effects models (see Penny et al., 2003, for a review on mixed effects models in fMRI analysis), it is assumed that the random effect,  $\mathbf{u}_i$ , follows a trivariate normal distribution with mean zero and covariance matrix  $\boldsymbol{\Psi}$ ; and  $\epsilon_{ik}^A$ ,  $\epsilon_{ik}^B$ , and  $\epsilon_{ik}^C$  are the random errors in session  $k$  of participant  $i$  following a trivariate normal distribution with mean zero and covariance matrix  $\boldsymbol{\Lambda}$ .  $\mathbf{u}_i$  and  $\boldsymbol{\eta}_{ik}$ , for all  $i$  and  $k$ , are mutually independent. For simplicity, this paper will focus on diagonal matrices,  $\boldsymbol{\Psi} = \text{diag}\{\psi_\alpha^2, \psi_\beta^2, \psi_\gamma^2\}$  and  $\boldsymbol{\Lambda} = \text{diag}\{\lambda_\alpha^2, \lambda_\beta^2, \lambda_\gamma^2\}$ , in the numerical studies

for fast and stable covariance estimation. Though the method can be extended to non-diagonal matrices, the computational complexity will increase accordingly. For two-level data or  $K = 1$ , one cannot estimate the random effects  $\mathbf{u}_i$ . Thus, we set  $\mathbf{u}_i = \mathbf{0}$  and  $\Psi = \mathbf{0}$  in model (2.3) without changing all other modeling assumptions. This essentially reduces to an ANOVA model for the first-level SEM coefficients. Figure 1 depicts a conceptual diagram of the proposed multilevel model and the relationship between the parameters across multiple levels. We next describe in detail each modeling component and the causal assumptions.

## 2.1 First-level Model

Keeping the notations uncluttered, when discussing the first-level model in this section, we denote the vector  $\mathbf{Z} = (Z_{ikj}, j = 1, \dots, n_{ik})$  by omitting the subscripts  $i$  and  $k$ , and similarly omit the subscripts in  $\mathbf{M}$ ,  $\mathbf{R}$ ,  $n$ , and so on. Using the simplified notations, the model is written as

$$\begin{pmatrix} \mathbf{M} & \mathbf{R} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} & \mathbf{M} \end{pmatrix} \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} + \begin{pmatrix} \mathbf{E}_1 & \mathbf{E}_2 \end{pmatrix}. \quad (2.4)$$

where the errors  $\mathbf{E}_1$  and  $\mathbf{E}_2$  have correlation  $\delta$  as in model (2.2). Baron and Kenny (1986) also considered a third but redundant equation

$$\mathbf{R} = \mathbf{Z}C' + \mathbf{E}', \quad (2.5)$$

where  $C'$  is the coefficient of interest and  $\mathbf{E}'$  is the noise term. Under model (2.4), the average total effect ( $C'$ ) can be decomposed into the indirect effect ( $AB$  or  $C' - C$ ) and direct effect ( $C$ , see Imai et al., 2010, for a review).

The standard mediation models assume independent  $\mathbf{E}_1$  and  $\mathbf{E}_2$ , and thus models in (2.4) are fitted separately, or equivalently, setting  $\delta = 0$ . Because it is expected that  $\delta \neq 0$  in many practical settings, these standard analyses are usually followed by sensitivity analysis, where the users change the  $\delta$  values (sometimes within a hypothesized range) to check its influence on the direct and indirect effect estimates (Imai et al., 2010). We do not impose this assumption on  $\delta$ , but rather treat it as a modeling parameter to be estimated from data. A hypothetical example of the existence of the unmeasured mediator-outcome confounding is described as follows. Suppose

$$\mathbf{M} = \mathbf{Z}A + g_1\mathbf{U} + \tilde{\mathbf{E}}_1, \quad \mathbf{R} = \mathbf{Z}C + \mathbf{M}B + g_2\mathbf{U} + \tilde{\mathbf{E}}_2,$$

where  $\mathbf{U}$ ,  $\tilde{\mathbf{E}}_1$ , and  $\tilde{\mathbf{E}}_2$  are mutually independent normal variables, and  $g_1$  and  $g_2$  are fixed and unknown scalars.  $\mathbf{U}$  represents the overall effect from all unmeasured mediator-outcome confounding

factors. Under this model, the errors in (2.4) are correlated when  $g_1g_2 \neq 0$ . The proposed approach will remove the influence of  $\mathbf{U}$  by accounting the estimation bias due to  $\delta \neq 0$ .

### 2.1.1 Assumptions and causal interpretation

We use Rubin’s potential outcome framework (Rubin, 2005) to assess the causal interpretation of model (2.4). As Rubin conjectured but not analyzed in Section 7 of his paper, when  $Z$  is randomized, a valid approach for mediation analysis is to consider “a bivariate outcome variable”,  $(M, R)$ , in order to infer the direct and indirect effects. We analyze and extend this conjecture in this paper. Under the potential outcome framework,

$$\begin{pmatrix} \mathbf{M}(\mathbf{z}) & \mathbf{R}(\mathbf{z}, \mathbf{m}) \end{pmatrix} = \begin{pmatrix} \mathbf{z} & \mathbf{M}(\mathbf{z}) \end{pmatrix} \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} + \begin{pmatrix} \mathbf{E}_1(\mathbf{z}) & \mathbf{E}_2(\mathbf{z}) \end{pmatrix}, \quad (2.6)$$

where all the potential outcomes are vectors of length  $n$ . Based on the potential outcomes, the average total effect (ATE) can be decomposed as the sum of the average indirect effect (AIE) and average direct effect (ADE):

$$\begin{aligned} \text{ATE}(z_i, z'_i) &= \mathbb{E} \{ R_i(z_i, m_i) - R_i(z'_i, m'_i) \} \\ &= \mathbb{E} \{ R_i(z_i, m_i) - R_i(z_i, m'_i) \} + \mathbb{E} \{ R_i(z_i, m'_i) - R_i(z'_i, m'_i) \} \\ &= \text{AIE}(z_i, z'_i) + \text{ADE}(z_i, z'_i), \end{aligned}$$

In order to estimate the AIE and ADE from the data, the following assumptions are imposed:

- (A1) stable unit treatment value assumption (SUTVA);
- (A2) model (2.4) is correctly specified and there is no treatment-mediator interaction;
- (A3) the observed bivariate outcome is one realization of the potential outcomes with the observed treatment assignment vector  $\mathbf{Z} = \mathbf{z}$ ;
- (A4) randomized treatment  $\mathbf{Z}$  with  $0 < \mathbb{P}(\mathbf{Z} = \mathbf{z}) < 1$  for every  $\mathbf{z}$ , i.e.,  $\mathbf{Z} \perp \{ \mathbf{R}(\mathbf{z}', \mathbf{m}), \mathbf{M}(\mathbf{z}) \}$  for all  $\mathbf{z}$  and  $\mathbf{z}'$ ; similarly  $\mathbf{Z} \perp \{ \mathbf{E}_1(\mathbf{z}), \mathbf{E}_2(\mathbf{z}') \}$ ;
- (A5) the covariance matrix of the Gaussian errors in model (2.4) is not affected by the treatment assignments, that is

$$\text{Cov} \{ \mathbf{E}_1(\mathbf{z}), \mathbf{E}_2(\mathbf{z}') \} = \text{Cov} \{ \mathbf{E}_1(\mathbf{z}), \mathbf{E}_2(\mathbf{z}) \} = \begin{pmatrix} \sigma_1^2 & \delta\sigma_1\sigma_2 \\ \delta\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \otimes \mathbf{I}_n.$$

Assumptions (A1)-(A3) are standard regularity assumptions in causal inference (see for example Rubin, 1978; Imai et al., 2010). Petersen et al. (2006) considered a weaker version of the first part of assumption (A4) and these two versions are equivalent in randomized experiments. For randomized trials, assumption (A4) is automatically satisfied. Assumption (A5) replaces the no unmeasured mediator-outcome confounding assumption with a Gaussian covariance assumption. Under Gaussianity, when  $\delta \neq 0$ , it is equivalent to the “cross-world” independence assumption (VanderWeele et al., 2014).

A version of the no unmeasured mediator-outcome confounding assumption is written as,

$$R_i(z'_i, m_i) \perp M_i(z_i) \mid Z_i = z_i, \quad (2.7)$$

for all  $z'_i$  and  $z_i, i = 1, \dots, n$ . Under the finest fully randomized causally interpreted structured tree graph (FRCISTG) model, Robins (2003) assumed the following,

$$R_i(z_i, m_i) \perp M_i(z_i) \mid Z_i = z_i, \quad (2.8)$$

together with the first part of assumption (A4), in order to identify the causal effects. As discussed in Imai et al. (2010), assumption (2.8) allows for conditioning on observed post-treatment confounders, but requires an additional no-interaction assumption; while assumption (2.7) does not depend on post-treatment confounders and the no-interaction assumption is not required. Neither of these two versions holds in (2.6) when there exists unmeasured mediator-outcome confounding or when the errors,  $\mathbf{E}_1$  and  $\mathbf{E}_2$ , are correlated. Instead, assumption (A5) is imposed considering a correlation assumption between the Gaussian errors. Setting  $\delta = 0$ , it implies the so-called “cross-world” independence assumption (VanderWeele et al., 2014). Imai et al. (2010) treated  $\delta$  as a sensitivity parameter. By imposing this assumption, it is assumed that the effect of unmeasured mediator-outcome confounding is fully characterized by the error correlation,  $\delta$ .

Under the proposed parametric model (2.4),

$$\text{AIE}(z_i, z'_i) = (z_i - z'_i)AB, \quad \text{ADE} = (z_i - z'_i)C.$$

As we show in the proof in the supplementary materials (Section A.1), parameters  $(AB, C)$  are consistently estimated from observed data only if  $\delta$  is given within an asymptotically shrinking neighborhood near the true value. With single-level data, it is impossible to estimate  $\delta$  as shown in Theorem 3.2. However, one can leverage the existence of higher level data which provide additional information about  $\delta$ . In the next section, we list the assumptions needed to estimate  $\delta$  consistently,

which in turn enable consistent estimation of the causal parameters. Thus, multilevel modeling is critical for the identification of the unmeasured mediator-outcome confounding effect ( $\delta$ ).

## 2.2 Higher-Level models

Motivated by the fMRI experiment, the primary interest is to infer the *population* parameters,  $A$ ,  $B$  and  $C$ , in model (2.3), for either two-level or three-level data. However, these parameters are not causally identifiable if one allows  $\delta_{ik}$  to vary across  $i$  and  $k$ . In order to make causal interpretation of the estimates, the following assumption is required for the two-level model:

(A6)  $\delta_{ik}$  is constant across participants, i.e.,  $\delta_{ik} = \delta$  for all  $i$  and  $k = 1$ ;

or the following for the three-level model:

(A6')  $\delta_{ik}$  is constant across participants and sessions, i.e.,  $\delta_{ik} = \delta$ , for all  $i$  and  $k$ .

Intuitively, either assumption allows us to pool data across levels to improve the estimation of  $\delta_{ik}$ . As it will be proved in Section 3.1.1,  $\delta_{ik}$ , if allowed to vary with  $i$  and  $k$ , is not identifiable in the likelihood sense. These assumptions are minimal for the purpose of model identifiability. Both assumptions (A6) and (A6') are weaker than a multilevel mediation model proposed by Kenny et al. (2003), where  $\delta_{ik} = \delta = 0$  to fit models in (2.4) separately for the first-level data followed by the same higher-level model formulation as ours. Their approach thus suffers from the estimation bias due to nonzero  $\delta$  or unmeasured mediator-outcome confounding.

The proposed framework can also address the following relaxed assumption of (A6'):

(A6'')  $\delta_{ik}$  is constant across participants in session  $k$ , i.e.,  $\delta_{ik} = \delta_k$ , for all  $i$  and  $\forall k \in \{1, \dots, K\}$ . This assumption allows  $\delta$  to vary across sessions and is equivalent to having  $K$  versions of (A6). To model three-level data under (A6''), one only needs to fit  $K$  two-level models. Using the estimates of  $\delta_k$ 's, one can check empirically if (A6') holds. The validity of (A6') of the fMRI dataset is illustrated in Section F.6 of the supplementary materials. Since (A6'') introduces a minor modification methodologically, we will focus on assumptions (A6) and (A6') in this paper.

Under either (A6) or (A6'), it is proved in Sections 3.2 and 3.3, respectively, that  $\delta_{ik}$  is identifiable and estimated consistently. The identifiability and estimation consistency of parameters,  $(A_{ik}, B_{ik}, C_{ik})$  and  $(A, B, C)$ , then follow. Following the term ‘‘population inference’’ (Penny et al., 2003),  $AB_d = C' - C$  and  $C$  are called the *population* indirect and direct effects, respectively, as they represent the causal effect estimates for a population after accounting for individual variability.

The proposed method also produces the estimate for  $AB_p = A \times B$ , which is equivalent to  $AB_d$  under certain conditions (Kenny et al., 2003).

### 3 Method

A multilevel likelihood framework is proposed to estimate the parameters:

$$\begin{aligned} \ell &= \sum_{i=1}^N \sum_{k=1}^K \log \mathbb{P}(\mathbf{R}_{ik}, \mathbf{M}_{ik} | \mathbf{Z}_{ik}, \mathbf{b}_{ik}, \delta_{ik}, \sigma_{1_{ik}}, \sigma_{2_{ik}}) + \sum_{i=1}^N \log \mathbb{P}(\mathbf{b}_{i1}, \dots, \mathbf{b}_{iK} | \mathbf{b}, \mathbf{\Lambda}, \mathbf{\Psi}) \\ &= \ell^{(1)} + \ell^{(2)}, \end{aligned} \quad (3.1)$$

where  $\ell^{(1)}$  is the log-likelihood of the first-level model and  $\ell^{(2)}$  is of the higher levels. The specific formulations of  $\ell^{(1)}$  and  $\ell^{(2)}$  will be introduced in the following sections.

#### 3.1 Method for the first-level model

Though the integrated method is to maximize  $\ell$  (asymptotically), it is worthwhile to discuss the methodological and theoretical issues related to maximizing  $\ell^{(1)}$ . As  $\ell^{(1)}$  is the sum of the conditional log-likelihood,  $\ell_{ik}^{(1)}$ , of model (2.1), we will focus on  $\ell_{ik}^{(1)}$  in this section. The likelihood  $\ell_{ik}^{(1)}$  has six parameters,  $(A_{ik}, B_{ik}, C_{ik}, \delta_{ik}, \sigma_{1_{ik}}, \sigma_{2_{ik}})$ , for each  $i$  and  $k$ . The exact formulation of  $\ell_{ik}^{(1)}$  is given in Section A.2 of the supplementary materials. To characterize the changes in the estimates due to  $\delta$  in the model, it is firstly considered the case when the covariance parameters  $(\delta_{ik}, \sigma_{1_{ik}}, \sigma_{2_{ik}})$  are given. The unknown covariance case will be discussed in Section 3.1.1.

**Theorem 3.1.** *Given  $(\delta_{ik}, \sigma_{1_{ik}}, \sigma_{2_{ik}})$ , the solution that maximizes  $\ell_{ik}^{(1)}$  is given by*

$$\begin{aligned} \hat{A}_{ik} &= (\mathbf{Z}_{ik}^\top \mathbf{Z}_{ik})^{-1} \mathbf{Z}_{ik}^\top \mathbf{M}_{ik}, \\ \hat{C}_{ik} &= (\mathbf{Z}_{ik}^\top \mathbf{H}_{\mathbf{M}_{ik}} \mathbf{Z}_{ik})^{-1} \mathbf{Z}_{ik}^\top \mathbf{H}_{\mathbf{M}_{ik}} \mathbf{R}_{ik} + \delta_{ik} \frac{\sigma_{2_{ik}}}{\sigma_{1_{ik}}} (\mathbf{Z}_{ik}^\top \mathbf{Z}_{ik})^{-1} \mathbf{Z}_{ik}^\top \mathbf{M}_{ik}, \\ \hat{B}_{ik} &= (\mathbf{M}_{ik}^\top \mathbf{M}_{ik})^{-1} \mathbf{M}_{ik}^\top \left( \mathbf{I}_{n_{ik}} - \mathbf{Z}_{ik} (\mathbf{Z}_{ik}^\top \mathbf{H}_{\mathbf{M}_{ik}} \mathbf{Z}_{ik})^{-1} \mathbf{Z}_{ik}^\top \mathbf{H}_{\mathbf{M}_{ik}} \right) \mathbf{R}_{ik} - \delta_{ik} \frac{\sigma_{2_{ik}}}{\sigma_{1_{ik}}}, \end{aligned}$$

where  $\mathbf{I}_{n_{ik}}$  is the  $n_{ik}$ -dimensional identity matrix;  $\mathbf{H}_{\mathbf{M}_{ik}} = \mathbf{I}_{n_{ik}} - \mathbf{P}_{\mathbf{M}_{ik}}$ , and  $\mathbf{P}_{\mathbf{M}_{ik}} = \mathbf{M}_{ik} (\mathbf{M}_{ik}^\top \mathbf{M}_{ik})^{-1} \mathbf{M}_{ik}^\top$  is the projection matrix of  $\mathbf{M}_{ik}$ .

This theorem shows how  $\delta_{ik}$  and the variance parameters affect the maximum likelihood estimates (MLEs) of  $B_{ik}$  and  $C_{ik}$  by two different additive terms related to  $\delta_{ik}$ . The standard Baron-Kenny estimates (Baron and Kenny, 1986) are special cases by setting  $\delta_{ik} = 0$  in Theorem 3.1.

The differences between the Baron-Kenny estimates and the proposed are the biases due to the unmeasured mediator-outcome confounding, which are corrected in Theorem 3.1. Moreover, the biases of the Baron-Kenny estimates increase when  $\delta_{ik}$  moves away from 0 and are proportional to the variance ratio  $\sigma_{2_{ik}} / \sigma_{1_{ik}}$ . The biases are asymptotically independent of the Baron-Kenny estimates (see the proof of Theorem 3.3 in Section A.7 of the supplementary materials). Finally, the estimate of  $A_{ik}$  is the same as the Baron-Kenny estimate since  $\mathbf{Z}_{ik}$  is randomized.

The asymptotic properties are demonstrated in Theorem A.1 (Section A.3 of the supplementary materials). Briefly, the estimator is not only consistent but also achieves the Fisher efficiency, and thus those estimators without accounting for  $\delta_{ik} \neq 0$  are asymptotically biased. By the Delta method (Sobel, 1982), the asymptotic standard errors of the product estimator,  $\widehat{AB}_{p_{ik}} = \hat{A}_{ik} \hat{B}_{ik}$ , and the difference estimator,  $\widehat{AB}_{d_{ik}} = \hat{C}'_{ik} - \hat{C}_{ik}$ , are calculated using the asymptotic distributions of  $(\hat{A}_{ik}, \hat{B}_{ik}, \hat{C}_{ik})$  and  $(\hat{C}_{ik}, \hat{C}'_{ik})$ , respectively. Importantly, all these asymptotic standard errors depend on  $\delta_{ik}$ , see the explicit formulas in Section A.3 of the supplementary materials.

### 3.1.1 Estimation and identifiability under unknown variances

It is known that the parameters in model (2.1) are in general not all identifiable without additional assumptions. The non-identifiability issue can be verified using either a rank condition (Hausman, 1983) or comparing the numbers of parameters and equations (Imai et al., 2010). For completeness, we prove the non-identifiability issue in the first-level model from the likelihood perspective.

**Theorem 3.2.** *For every fixed  $\delta_{ik} \in (-1, +1)$ ,  $i = 1, \dots, N$  and  $k = 1, \dots, K$ ,  $\ell_{ik}^{(1)}$  achieves the same maximum (profile) likelihood value  $\ell_{ik}^{(1)}(\delta_{ik})$ , where the maximum is taken over all the remaining parameters  $(A_{ik}, B_{ik}, C_{ik}, \sigma_{1_{ik}}, \sigma_{2_{ik}})$ . For the variance components, with a given  $\delta_{ik}$ , the following estimates maximize  $\ell_{ik}^{(1)}$ ,*

$$\hat{\sigma}_{1_{ik}}^2 = \frac{1}{n_{ik}} \mathbf{M}_{ik}^\top (\mathbf{I}_{n_{ik}} - \mathbf{P}_{\mathbf{Z}_{ik}}) \mathbf{M}_{ik}, \quad \hat{\sigma}_{2_{ik}}^2 = \frac{1}{n_{ik}(1 - \delta_{ik}^2)} \mathbf{R}_{ik}^\top (\mathbf{I}_{n_{ik}} - \mathbf{P}_{\mathbf{M}_{ik}\mathbf{Z}_{ik}} - \mathbf{P}_{\mathbf{M}_{ik}}) \mathbf{R}_{ik}, \quad (3.2)$$

where  $\mathbf{P}_{\mathbf{Z}_{ik}} = \mathbf{Z}_{ik}(\mathbf{Z}_{ik}^\top \mathbf{Z}_{ik})^{-1} \mathbf{Z}_{ik}^\top$  and  $\mathbf{P}_{\mathbf{M}_{ik}\mathbf{Z}_{ik}} = \mathbf{H}_{\mathbf{M}_{ik}} \mathbf{Z}_{ik}(\mathbf{Z}_{ik}^\top \mathbf{H}_{\mathbf{M}_{ik}} \mathbf{Z}_{ik})^{-1} \mathbf{Z}_{ik}^\top \mathbf{H}_{\mathbf{M}_{ik}}$  are projection matrices. The estimate of  $A_{ik}$ ,  $B_{ik}$ , and  $C_{ik}$  are obtained by plugging in the variance estimates above into Theorem 3.1.

This theorem shows that  $\ell_{ik}^{(1)}(\delta_{ik})$  achieves the same maximum value regardless of  $\delta_{ik}$ . It holds for both two-level and three-level data. Figure 2a demonstrates it using a simulated two-level dataset (see Section E.2 for the simulation setup), where the computed maximum value is constant

over  $\delta$ . Thus, one cannot estimate  $\delta$  by maximizing  $\ell^{(1)}$ . Section B of the supplementary materials derives the relationship between the parameters and presents an example where two generative models with zero or nonzero indirect effects will yield the same data distribution or likelihood.

Imai et al. (2010) considered only the one-level setting and derived the same estimates as in (3.2) with  $\delta_{ik}$  as a sensitivity parameter. As shown in Theorem 3.1, the impact of  $\delta_{ik}$  on the estimates can be large. Under such a situation, it is challenging to employ sensitivity analysis as the resulting estimates depend heavily on the choice of  $\delta_{ik}$ . We illustrate this limitation using the fMRI dataset in Figure F.16, where opposite conclusions are drawn with a moderate change in  $\delta_{ik}$ . The sensitivity analysis in Imai et al. (2010) fails to account for individual variability, which is an important issue in datasets with multiple nested levels (Kenny et al., 2003).

Theorem 3.2 also suggests that it is not easy to avoid assumption (A6) or (A6'). Suppose  $\delta_{ik}$  is different across  $i$  and  $k$ . For any population parameter set  $(B, C)$ , one can pick different  $\delta_{ik}$  such that  $\hat{B}_{ik} = B$  and  $\hat{C}_{ik} = C$  yielding the same likelihood  $\ell^{(1)}$ . The term  $\ell^{(2)}$  is also the same as it depends on  $(A, B, C)$  and their first-level counterparts,  $(A_{ik}, B_{ik}, C_{ik})$ . Thus, there exists multiple estimates such that  $\ell$  is the same and the model is not identifiable in the likelihood sense. The next two sections prove the identifiability if either (A6) or (A6') is satisfied.

### 3.2 Method for the two-level model

For the two-level model, the second term in the likelihood criterion (3.1) becomes

$$\ell^{(2)} = \sum_{i=1}^N \log \mathbb{P}(\mathbf{b}_i | \mathbf{b}, \mathbf{\Lambda}), \quad (3.3)$$

where the above is the log-likelihood of regression model (2.3).

To estimate the parameters, it is proposed to maximize the likelihood via a two-stage algorithm. In the algorithm, the first step optimizes  $(\mathbf{b}_i, \sigma_{1_i}, \sigma_{2_i})$  in  $\ell^{(1)}$  for a given  $\delta_i = \delta$ . The second step optimizes  $(\mathbf{b}, \mathbf{\Lambda})$  in  $\ell^{(2)}$  by plugging in the estimated  $\mathbf{b}_i$ 's from the first step. Since  $\delta$  is usually unknown, we further optimize over  $\delta$  to yield the maximum likelihood. This idea is summarized in Algorithm 1. The following theorem demonstrates that this algorithm maximizes  $\ell$  and the resulting estimate of  $\delta$  is consistent asymptotically.

**Theorem 3.3.** *Assume assumptions (A1)-(A6) are satisfied. Let  $\mathbf{Z}_i^\top \mathbf{Z}_i / n_i \rightarrow q_i < \infty$  as  $n = \min_i n_i \rightarrow \infty$ , for  $i = 1, \dots, N$ .*

1. *If  $\mathbf{\Lambda}$  is known, the two-stage estimator  $\hat{\delta}$  maximizes the profile likelihood of model (2.3) asymptotically, and  $\hat{\delta}$  is  $\sqrt{Nn}$ -consistent.*

---

**Algorithm 1** Estimating the parameters in the two-level mediation model sequentially.

---

With a given  $\delta$ :

1. Estimate  $(\mathbf{b}_i, \sigma_{1_i}, \sigma_{2_i})$  for each  $i$  using Theorem 3.1 and (3.2).
2. Fit model (2.3) on the estimated  $\hat{\mathbf{b}}_i$ 's, and estimate  $\mathbf{b}$  and  $\mathbf{\Lambda}$  via maximum likelihood.
3. Return the maximum log-likelihood value of the regression model.

When  $\delta$  is unknown, apply an optimization algorithm (e.g., Newton's method) to maximize over  $\delta$  using the maximum log-likelihood value at Step 3 above.

---

2. If  $\mathbf{\Lambda}$  is unknown, the profile likelihood of model (2.3) has a unique maximizer  $\hat{\delta}$  asymptotically, and  $\hat{\delta}$  is  $\sqrt{Nn}$ -consistent provided that  $1/\varpi = \mathcal{O}_p(1/\sqrt{Nn})$ , where  $\varpi = \bar{\kappa}^2/\varrho^2$ ,  $\kappa_i = \sigma_{2_i}/\sigma_{1_i}$ ,  $\bar{\kappa} = \sum \kappa_i/N$ , and  $\varrho^2 = \sum(\kappa_i - \bar{\kappa})^2/N$ .

Compared to Theorem 3.2, this theorem shows that  $\delta$  is identifiable under the multilevel model. The intuition is that multilevel observations provide additional information to help avoid the over-parameterization issue in the first-level model. The likelihood term  $\ell^{(2)}$  has a unique maximizer converging to the true  $\delta$ . In practice, the shape of  $\ell$  with varying  $\delta$  depends on the observed data. We propose to check empirically the maximum log-likelihood value as a function of  $\delta$  using a simulated dataset, see Figure 2b. In this example,  $\ell^{(2)}(\delta)$  is unimodal achieving its maximum at  $\hat{\delta} = 0.476$  (the truth is 0.5), while the first-level likelihood is flat (Figure 2a).

### 3.2.1 An alternative coordinate-descent algorithm

In general, it is challenging to find the global optimum of a function like  $\ell$ , especially when it contains many parameters. Though Algorithm 1 is simple and consistent, it may not optimize  $\ell$  as a whole in finite samples. To address this issue, we propose an alternative algorithm, which draws on the following properties of  $\ell$ .

**Theorem 3.4.** *Assume  $\delta_{ik} = \delta$  is given. The negative of log-likelihood function (3.1) is conditional convex in the parameter sets  $(\sigma_{1_i}^{-1}, \sigma_{2_i}^{-1})$ ,  $(\mathbf{b}_i)$ ,  $\mathbf{b}$ ,  $\mathbf{\Lambda}^{-1}$ , respectively. The conditional optimizer of each parameter set is given in explicit forms in Section A.6 of the supplementary materials.*

Based on this theorem, a block coordinate-descent algorithm is proposed (Algorithm 2). Each descent step is computed efficiently using explicit updates (see Section A.6). When  $\delta$  is unknown, it is estimated with the value that yields the highest maximum profile likelihood. In practice, this approach yields estimates closer to the truth in the simulation studies (see the numerical results in

---

**Algorithm 2** Estimating the parameters in the two-level mediation model using (3.1).

---

**With a given  $\delta$ :**

1. Estimate  $(\sigma_{1_i}, \sigma_{2_i})$ ,  $\mathbf{b}_i$ ,  $\mathbf{b}$ , and  $\mathbf{\Lambda}$  by maximizing the log-likelihood function (3.1) over these remaining parameters using block coordinate descent.
2. Return the maximum log-likelihood value.

**When  $\delta$  is unknown, apply an optimization algorithm (e.g., Newton's method) to maximize over  $\delta$  using the profile log-likelihood value at Step 2 above.**

---

Section 4 and Section E). In the toy example, the numerical value of  $\ell$  is a unimodal function of  $\delta$  (Figure 2c). It yields a better estimate ( $\hat{\delta} = 0.519$ ) with lower bias than Algorithm 1 does.

### 3.3 Method for the three-level model

For the three-level model, the second term of the likelihood criterion becomes

$$\ell^{(2)} = \sum_{i=1}^N \log \mathbb{P}(\mathbf{b}_{i1}, \dots, \mathbf{b}_{iK} | \mathbf{b}, \mathbf{\Psi}, \mathbf{\Lambda}), \quad (3.4)$$

where the above is now the log-likelihood of the mixed effects model (2.3).

A two-stage algorithm is proposed to optimize the likelihood criterion, very similar to Algorithm 1. The first-level term  $\ell^{(1)}$  is firstly optimized with a given  $\delta$  and then optimize the mixed effects likelihood  $\ell^{(2)}$  using the estimated coefficients. The modification to Algorithm 1 is to re-place the model and likelihood with the mixed effects model and its likelihood, respectively. Thus, the description of this algorithm is omitted here.  $\delta$  is again estimated via maximizing the profile likelihood  $\ell^{(2)}$  (after maximizing over  $\mathbf{\Psi}$  and  $\mathbf{\Lambda}$  if unknown). The following two theorems show that this algorithm identifies  $\delta$  consistently when  $\mathbf{\Psi}$  and  $\mathbf{\Lambda}$  are either known or estimated.

**Theorem 3.5.** *Assume (A1)-(A6') are satisfied. Let  $\mathbf{Z}_{ik}^\top = \mathbf{Z}_{ik}/n_{ik} \rightarrow q_{ik} < \infty$  as  $n = \min_{i,k} n_{ik} \rightarrow \infty$ , for  $\forall i, k$ . Suppose  $\mathbf{\Psi}$  and  $\mathbf{\Lambda}$  in model (2.3) are known. For fixed  $K$ , the two-stage estimator  $\hat{\delta}$  maximizes the profile likelihood of the mixed effects model and  $\hat{\delta}$  is  $\sqrt{Nn}$ -consistent.*

**Theorem 3.6.** *Assume (A1)-(A6') are satisfied. Suppose that  $\mathbf{\Psi}$  and  $\mathbf{\Lambda}$  are estimated via maximizing the likelihood of the mixed model. Assume  $1/\varpi = \mathcal{O}_p(1/\sqrt{Nn})$ , where  $\varpi = \bar{\kappa}^2/\varrho^2$ ,  $\kappa_{ik} = \sigma_{2_{ik}}/\sigma_{1_{ik}}$ ,  $\bar{\kappa} = (nN)^{-1} \sum \kappa_{ik}$ , and  $\varrho^2 = (nN)^{-1} \sum (\kappa_{ik} - \bar{\kappa})^2$ . For fixed  $K$ , the two-stage estimator  $\hat{\delta}$  maximizes the profile likelihood of the mixed effects model and  $\hat{\delta}$  is  $\sqrt{Nn}$ -consistent provided that either one of the following conditions holds: (a)  $\lambda_\alpha^2 \geq \psi_\alpha^2$ ,  $K \geq 2$ ; or (b)  $\lambda_\alpha^2 < \psi_\alpha^2$ ,  $K \geq \{(\lambda_\gamma^2 \lambda_\alpha^2)/(\psi_\lambda^2(\psi_\alpha^2 - \lambda_\alpha^2)) + 1\}$ .*

These two theorems show that  $\delta$  is identifiable and estimated consistently under regularity conditions. When the covariance matrices are unknown, additional regularity conditions are required, such as the minimal  $K$  condition, to estimate the variance components well. Theoretically, the minimal  $K$  condition holds automatically if one sets  $K \rightarrow \infty$  in a typical asymptotic setting. Here,  $K$  is assumed to be fixed because it is less restrictive for practical examples. When  $K$  is fixed, the convergence rates depend on the number of participants and the number of trials. These rates are consistent with those of standard linear mixed effects models (Nie, 2007). In practice, the following approach is suggested to check the condition on  $K$ . Since  $\lambda_\alpha^2$  and  $\psi_\alpha^2$  depend only on  $A_{ik}$ 's independent of  $\delta$  as shown by Theorem 3.1, they are estimated unbiasedly without knowing  $\delta$ . If these two estimates satisfy condition (a) in Theorem 3.6, it only needs to verify  $K > 2$ . If condition (b) is satisfied instead, the lower bounds of  $K$  in the theorem is computed by varying  $\delta$  values in a range and verify if  $K$  is greater than the maximum of the lower bound.

### 3.3.1 An alternative coordinate-descent algorithm

Similar to Section 3.2.1, an alternative coordinate descent algorithm is proposed to explore its improvement in finite samples. Analogous to Theorem 3.4, the resulting likelihood, named as m-likelihood (marginal-likelihood), is shown to be conditional convex and the iterative updates are given in explicit forms. As these results are parallel to those in Section 3.2.1, we include them in Section A.6.2 of the supplementary materials.

### 3.3.2 An alternative likelihood formulation

In the mixed effects modeling, alternative likelihood-based criteria have been proposed to improve finite sample performance. Because the framework is flexible enough, various criteria can be incorporated. Following Lee and Nelder (1996), a hierarchical-likelihood (h-likelihood) criterion is considered, which replaces  $\ell^{(2)}$  in (3.4) with

$$\ell^{(2)} = \sum_{i=1}^N \sum_{k=1}^K \log \mathbb{P}(\mathbf{b}_{ik} | \mathbf{u}_i, \mathbf{b}, \mathbf{\Lambda}) + \sum_{i=1}^N \log \mathbb{P}(\mathbf{u}_i | \mathbf{\Psi}). \quad (3.5)$$

This criterion includes the random effects  $\mathbf{u}_i$ . The introduction of the random effects term is sometimes viewed as an computational convenience to simplify the mixed effects likelihood, especially when it is not straightforward to integrate over this term explicitly. Lee and Nelder (1996) proved that maximizing the hierarchical likelihood is asymptotically equivalent to maximizing the standard likelihood, so it is expected that it will yield estimates close to those from the other two algorithms

when the sample size is reasonably large. Since the last term in (3.5) is sometimes viewed as a penalty term to stabilize the estimates in finite samples, it may introduce estimation bias, though it is usually negligible in practice (see Commenges, 2009, for a review). The hierarchical likelihood is conditional convex and a block coordinate descent algorithm is proposed. These results are included in Section A.6.1 of the supplementary materials.

### 3.4 Inference

Due to the complexity of the multilevel model, the distributions of the estimators, especially the indirect effect and  $\delta$  estimators, may deviate from normal in finite samples. The wild bootstrap (Wu, 1986) is considered to compute the confidence intervals.

## 4 Simulation study

In this section, the proposed estimators are compared with others under the three-level model when  $\delta$  is unknown. The simulation results of the first- and two-level models are presented in Section E of the supplementary materials. The methods include the two-stage mixed effects algorithm (CMA-ts) from Section 3.3, the coordinate descent algorithm for the marginal likelihood (CMA-m) from Section 3.3.1, the coordinate descent algorithm for the hierarchical likelihood (CMA-h) from Section 3.3.2, the first-level method (CMA- $\delta$ ) from Section 3.1, the linear mixed effects SEM (KKB) method (Kenny et al., 2003), and the Baron-Kenny (BK) method (Baron and Kenny, 1986). Neither KKB or BK can estimate  $\delta$  as no unmeasured confounding (or  $\delta = 0$ ) is assumed. Because the CMA- $\delta$  method allows  $\delta$  as input, the true  $\delta$  value will be used to assess the oracle performance of this single level method in multilevel data. Since both BK and CMA- $\delta$  are developed for one-level data, they are implemented by concatenating the multilevel data from all sessions and all participants. The CMA methods are implemented using the developed R package `macc`, KKB using the `lme4` package, and BK via standard regression.

The total number of participants is set to be  $N = 50$  and the number of sessions  $K = 4$ . Under each session and for each participant, the number of trials is a random draw from the Poisson distribution with mean 100. The main objective is to identify the population direct effect (denoted by  $C$ ) and the population indirect effect (denoted by  $C' - C$  or  $AB$ ). Using the product definition of the indirect effect, the null hypothesis is  $H_0: AB = 0$ , which includes three scenarios for  $A$  and  $B$ . That is, a)  $A = 0, B \neq 0$ ; b)  $A \neq 0, B = 0$ ; and c)  $A = B = 0$ . Since all methods yield

unbiased estimate for  $A$  (independent of  $\delta$ ), only the scenario of b)  $A \neq 0, B = 0$  for the null of  $AB$  is presented. Under the alternatives, the population level parameters are set as  $A = 0.5, B = -1$ , and  $C = 0.5$ . Both  $\Psi$  and  $\Lambda$  are set to be diagonal, and the variance components are  $\psi_\alpha^2 = \psi_\beta^2 = \psi_\gamma^2 = 0.5$  and  $\lambda_\alpha^2 = \lambda_\beta^2 = \lambda_\gamma^2 = 0.5$ . For each participant in each session, the variances of the errors in the first level mediation model are  $\sigma_{1_{ik}} = 1$  and  $\sigma_{2_{ik}} = 2$ , for  $i = 1, \dots, N$  and  $k = 1, \dots, K$ . The correlation between the errors (denoted by  $\delta$ ) is either 0.5 or 0, to simulate the settings with and without unmeasured confounding, respectively. Simulations are repeated 200 times.

In Table E.3 of the supplementary materials, the computation time of the CMA approaches is compared. The CMA-h approach computes the fastest. Table 1 presents the point estimates from all the methods considered. From the table, CMA-ts, CMA-h, and CMA-m have small biases in estimating  $\delta$ . CMA-h has slightly lower biases than CMA-ts and CMA-m on average. The KKB, BK and CMA- $\delta$  estimates yield large biases in  $B$  and  $C$ , as well as in the indirect effect when the true  $\delta$  is nonzero.

To test whether the methods are robust to the magnitude of unmeasured confounding, the first case in Table 1 are simulated with varying  $\delta \in (-1, 1)$ . Figure 3 shows that our methods yield lower estimation biases of  $\delta$ , and also lower biases than other competing methods in terms of estimating  $AB$ . Across different  $\delta$  values, CMA-h has the lowest biases among all methods. This is also consistent with the simulation results earlier. The biases in KKB, BK and CMA- $\delta$  increase dramatically as  $|\delta|$  approaches to one, while our multilevel methods have numerically negligible biases across all  $\delta$  values. KKB has the largest bias, followed by BK and CMA- $\delta$ , and the biases can be as large as 200%. The biases in estimating  $B, C$  and  $AB$  of BK and KKB are approximately a linear function of  $\delta$ , as predicted by our theory, see Section D of the supplementary materials.

To validate the consistency theory, the simulation of the first case in Table 1 is expanded with  $N = n_{ik} = 50, 200, 500, 1000$  and  $K = 4, 10$ . Figure 4 shows that the estimates of  $\delta$  (from CMA-ts, CMA-h, and CMA-m) approach the true values as the number of trials and the number of participants increase. This convergence does not depend on the increase of the number of sessions, as predicted by the theory. CMA-h has the lowest bias in finite samples, probably due to the regularization term as discussed in Section 3.3.2. Figure E.5 in the supplementary materials compares the biases in estimating the direct and indirect effects. CMA-h again achieves the lowest biases. In Section E.4 of the supplementary materials, a scenario of  $\sigma_{1_{ik}} = 2$  and  $\sigma_{2_{ik}} = 1$  is considered. The performance of the CMA-h and CMA-ts approaches generalizes, while the CMA-m

approach attains estimates with a slightly greater bias. Section E.5 of the supplementary materials discusses the robustness of the proposed approaches under various settings.

## 5 Application

The proposed model and methods are applied to an fMRI dataset. In the experiment,  $N = 96$  participants consented to fMRI scanning while performing a response conflict task, where the conflict occurs between the GO trial (pressing a button when seeing a “circle” on the screen) and the STOP trial (withholding the press when seeing a “cross”). Each participant  $i$  was scanned in  $K_i = K = 4$  sessions. Each session  $k$  is about ten minutes in length with  $n_{ik}$  (median 90) randomized STOP/GO trials. For all sessions, with probability  $3/4$ , the  $j$ th trial is a GO trial ( $Z_{ikj} = 0$ ); and with probability  $1/4$ , it is a STOP trial ( $Z_{ikj} = 1$ ). The experiment paradigm was described in details in Duann et al. (2009) and Luo et al. (2012). Prior modeling efforts have identified various brain pathways in this experiment (Duann et al., 2009). In this study, a brain pathway from the presupplementary motor area (preSMA) to the primary motor cortex (PMC) is investigated. The latter is a region that carries out the primary function of movements and the former is a primary region of motor response prohibition (Duann et al., 2009). The existence of this pathway has been confirmed using transcranial magnetic stimulation (Obeso et al., 2013). Brain activations of preSMA ( $M_{ik}$ ) and PMC ( $R_{ik}$ ) in each trial  $Z_{ikj}$  are extracted through a widely used single trial analysis (Lindquist, 2008). More details of data processing are in Section F of the supplementary materials.

Neuroimaging analysis usually concerns about population estimates rather than individual variability. Both the two-level and three-level models are capable of estimating the population effects. For the sake of space, only the three-level model is considered. Additional data analyses and the validation of modeling assumptions are included in Section E. In fMRI experiments, the mediator-outcome confounding factor may come from several sources. First, systematic errors, such as head motions, influence brain activities in both regions (see the discussion in Sobel and Lindquist, 2014). Second, brain activity under a task can be reliably modeled by a linear superposition of task-related activity and (spontaneous) task-unrelated activity. The task-unrelated activity is shown to account for a significant fraction of the variation (Cole et al., 2014; Fox et al., 2006). Third, other brain regions that are not included in the model may have impact on the two brain regions considered, for example, a third region inferior frontal gyrus may influence both preSMA and PMC (Obeso et al.,

2013). Since head motions are robustly estimated in standard neuroimaging processing pipelines, we will treat head motions as a measurable source of confounding. Two analyses, with and without adjusting for head motions, are conducted to assess the robustness of the methods.

In the first analysis, head motions are intentionally not adjusted (Rosenbaum, 2002) and thus, contribute to unmeasured confounding. All proposed methods (denoted as CMA) yield similar estimates of  $\delta$  far from zero. The differences between them are consistent with the simulation results, where CMA-ts (two-stage approach in Section 3.3) and CMA-m (marginal likelihood in Section 3.3.1) slightly underestimate and overestimate  $\delta$ , respectively. Table 2 (and Table F.5) compares the inference results of different one-level and three-level methods. CMA-h (hierarchical likelihood in Section 3.3.2) estimates  $\delta = -0.465$  ( $-0.578, -0.325$ ),  $AB_p = 0.293$  ( $0.237, 0.350$ ) (Table F.5), and  $C = -0.177$  ( $-0.247, -0.108$ ) with 500 wild bootstrap samples. Because both KKB (linear mixed effects SEM by Kenny et al., 2003) and BK (Baron and Kenny, 1986) ignored the unmeasured confounding effect, the indirect effect estimates are about 50% less. The direct effect estimates of CMA-ts, CMA-h, and CMA-m are significant at the 95% level, while the KKB and BK estimates are not. The CMA methods lead to important and interpretable scientific conclusions suggesting that the STOP stimulus increases the preSMA activity which then increases the PMC activity via the preSMA-PMC pathway, while directly decreases the PMC activity. The indirect effect estimates are about two folds larger than the direct effect, which quantifies the important role of preSMA in motion prohibition. The direct effect estimates are negative and significant, which is consistent with the expectation of the participants withholding motor movements during the STOP trials. KKB and BK, without accounting for the unmeasured confounding, underestimate the role of preSMA and miss the significant direct prohibition effect.

With the second analysis, the robustness of the methods is validated. Head motions are adjusted in the data processing step, which is expected to reduce the magnitude of unmeasured confounding. Both CMA-h and CMA-m yield a slightly smaller estimate of  $\delta$  in magnitude, though this change is not significant. The estimates of other parameters are similar (Tables 2 and F.5). This demonstrates that the multilevel method is stable under varying magnitude of confounding. In contrast, the indirect effect estimates by KKB and BK are outside the corresponding confidence intervals after motion correction suggesting that the estimates are sensitive to confounding.

## 6 Discussion

In this study, a multilevel mediation modeling framework is proposed for data with a hierarchically nested structure. This framework simultaneously addresses the unmeasured mediator-outcome confounding issue and individual variation in causal mediation analysis. For the one-, two-, and three-level mediation models, optimization-based methods are introduced along with efficient algorithms for computing, especially for a large number of parameters in the multilevel models. We prove theoretically these methods consistently estimate the magnitude of unmeasured confounding, which is mathematically impossible under single-level mediation models. Using extensive simulations and an fMRI dataset, it is shown that the resulting estimates correct for biases effectively in finite samples, and are robust to the magnitude of the unmeasured confounding effect. Though the method is motivated by an fMRI experiment, the methodology can be generalized to many other studies when the correlated errors are present due to confounding.

The proposed framework can be extended to other types of mixed effects models for multilevel data. One example is that the mediator is a constant at the second level and does not change over sessions/time. One can impose the corresponding model parameters to be fixed in the estimation. Given that there are many variants in the literature on specifying the random effects and covariance structures, we leave to future research the analysis of more complex multilevel SEMs, for example, fitting  $\delta_{ik}$  in a mixed effects model. Moreover, it is also of interest to test whether the finite sample estimation accuracy can be improved by using a global optimization algorithm.

Though the proposed framework is derived under binary treatment assignment, it can be used in other applications when the treatment is continuous. For studies with covariates, one can conduct covariate adjustment on both the mediator and the outcome before applying the proposed method, similar to what have been done to the fMRI data on motion correction. Another option is to include the covariates in the first-level model. However, it makes the computation more challenging as the number of parameters increases. The optimization problem becomes more complex when the covariates differ at different levels, see a discussion in Section C of the supplementary materials. Both approaches for covariate adjustment can be applied to observational studies under certain assumptions (see Rosenbaum, 2002, and Section C of the supplementary materials).

For studies with multiple mediators, when the causal ordering of the mediators is known, one can formulate the first-level SEMs accordingly based on the causal diagram. Hyperparameters at higher levels can be then modeled analogously. However, one should note that as the number

of mediators increases, the decomposition of the mediation effect becomes more complex (Daniel et al., 2015). When the ordering of the mediators is unknown, a marginal SEM can be employed. However, the interpretation of the effects may differ (Imai and Yamamoto, 2013; VanderWeele and Vansteelandt, 2014). If such a marginal model is considered, similarly, hyperparameters at higher levels can be modeled. Investigation of other complex scenarios will be future research.

Via simulation studies, it suggests that a correctly specified model is essential (Section E.5 of the supplementary materials). However, this assumption is intrinsically difficult to examine. For the ordering assumption of the variables, it is mainly determined based on domain knowledge or the temporal ordering of the measurements. As our proposed framework accounts for unmeasured confounding, missing measured confounding factors in the model will not impact the estimate of the indirect effect, but the estimation of the correlation parameter (as demonstrated in the fMRI data application). The proposed estimation approaches are likelihood based, thus one can consider comparing different models using existing tools, such as AIC/BIC if models are nested. One may also consider prediction performance on out-of-sample data to compare different models and select the best one.

We leave to future work on various extensions of the proposed framework, such as including interactions (MacKinnon et al., 2007; Valeri and VanderWeele, 2013). It is also possible to consider mediator and outcome from other distributions, as the framework is based on maximum likelihood. However, quantifying the unmeasured confounding effect using correlations is not straightforward. It is also of interest to study nonlinear mediation models as well.

## References

- 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(6):1173.
- Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., and Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron*, 83(1):238–251.
- Commenges, D. (2009). Statistical models: Conventional, penalized and hierarchical likelihood. *Statistics Surveys*, 3:1–17.

- Daniel, R., De Stavola, B., Cousens, S., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14.
- Duann, J.-R., Ide, J. S., Luo, X., and Li, C.-s. R. (2009). Functional connectivity delineates distinct roles of the inferior frontal cortex and presupplementary motor area in stop signal inhibition. *The Journal of Neuroscience*, 29(32):10171–10179.
- Fox, M. D., Snyder, A. Z., Zacks, J. M., and Raichle, M. E. (2006). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience*, 9(1):23–25.
- Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. (2020). Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Hausman, J. A. (1983). Specification and estimation of simultaneous equation models. *Handbook of econometrics*, 1:391–448.
- Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.
- Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171.
- Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8(2):115.
- Krull, J. L. and MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, 23(4):418–444.
- Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678.
- Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.
- Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309.

- Luo, X., Small, D., Li, C., and Rosenbaum, P. (2012). Inference with interference between units in an fMRI experiment of motor inhibition. *Journal of the American Statistical Association*, 107(498):530–541.
- MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annual review of psychology*, 58:593.
- Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *Journal of Statistical Planning and Inference*, 137(6):1787–1804.
- Obeso, I., Cho, S. S., Antonelli, F., Houle, S., Jahanshahi, M., Ko, J. H., and Strafella, A. P. (2013). Stimulation of the pre-SMA influences cerebral blood flow in frontal areas involved with inhibitory control of action. *Brain Stimulation*, 6(5):769–776.
- Penny, W. D., Holmes, A., and Friston, K. (2003). Random effects analysis. *Human brain function*, 2:843–850.
- Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, 17(3):276–284.
- Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. *Highly structured stochastic systems*, pages 70–81.
- Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469).
- Small, D. S. (2011). Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables. *arXiv preprint arXiv:1109.1070*.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13(1982):290–312.

- Sobel, M. E. and Lindquist, M. A. (2014). Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *Journal of the American Statistical Association*, 109(507):967–976.
- Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, 63(3):926–934.
- Valeri, L. and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological methods*, 18(2):137.
- VanderWeele, T. J. (2016). Mediation analysis: a practitioner’s guide. *Annual review of public health*, 37:17–32.
- VanderWeele, T. J. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115.
- VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306.
- Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, pages 1261–1295.

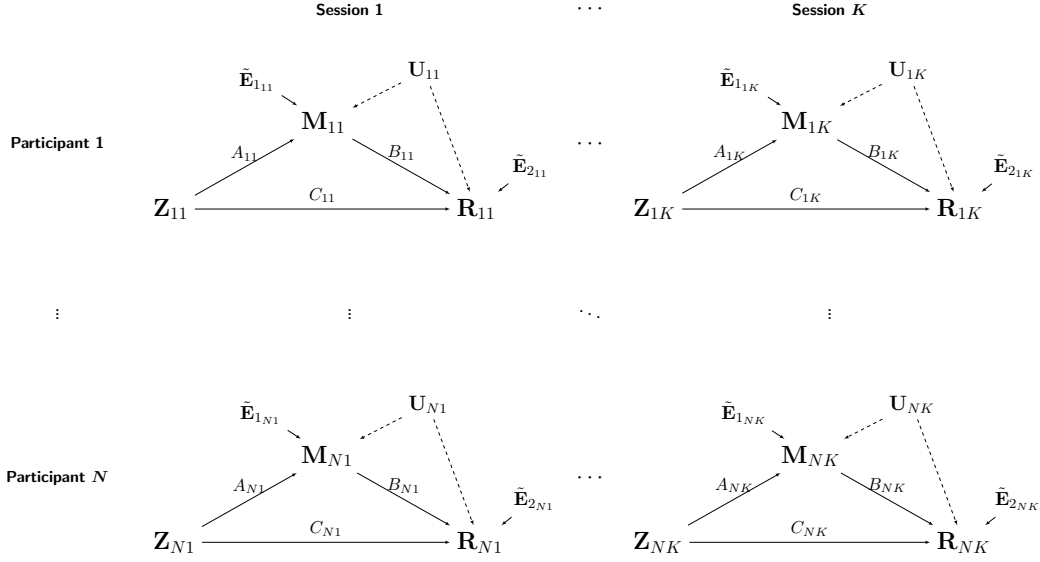


Figure 1: Conceptual causal diagram of the multilevel model.  $\mathbf{Z}_{ik}$ ,  $\mathbf{M}_{ik}$  and  $\mathbf{R}_{ik}$  are the randomized treatment, mediator and outcome vectors, respectively, in session  $k$  of participant  $i$ , for  $i = 1, \dots, N$  and  $k = 1, \dots, K$ .  $\mathbf{U}_{ik}$ 's are the unmeasured confounders, and  $\tilde{\mathbf{E}}_{1ik}$  and  $\tilde{\mathbf{E}}_{2ik}$  are the independent model errors.  $A_{ik}$ ,  $B_{ik}$  and  $C_{ik}$ 's are the SEM coefficients.

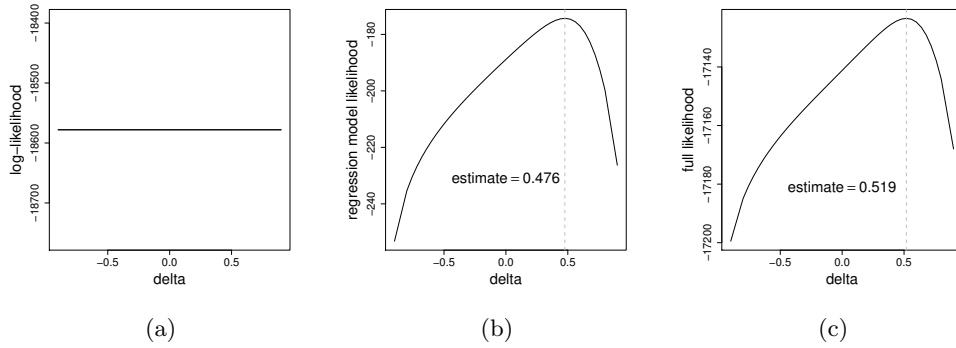


Figure 2: The log-likelihood functions of (a) the first-level model ( $\ell^{(1)}$ ), (b) the higher-level model ( $\ell^{(2)}$ ), and (c) the two-level model ( $\ell$ ) of a simulated two-level dataset. The true  $\delta$  value is 0.5. The dashed lines in (b) and (c) are the estimates from the two-stage algorithm and the block coordinate-descent algorithm, respectively.

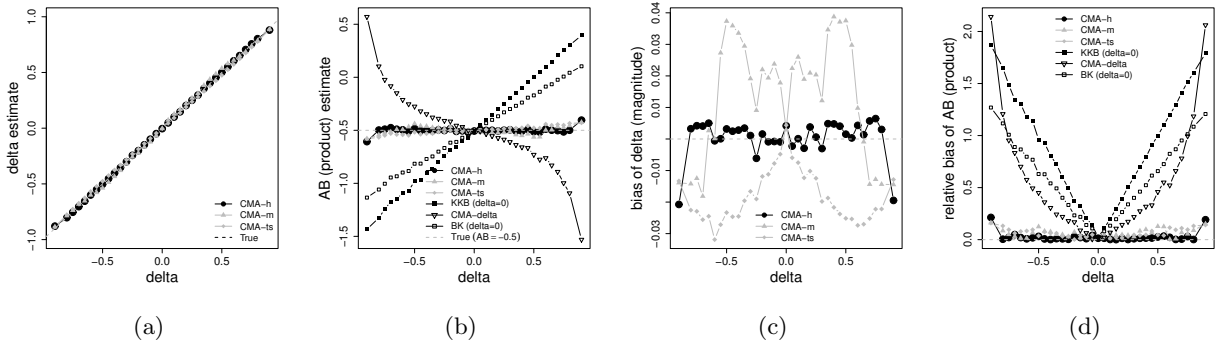


Figure 3: Point estimates of (a)  $\delta$  and (b)  $AB$ , (c) the bias of  $\hat{\delta}$ , and (d) the relative bias of  $\widehat{AB}_p$  with varying  $\delta$ . Solid circles are from CMA-h, solid triangles are from CMA-m, solid diamonds are from CMA-ts, solid squares are from KKB with  $\delta = 0$ , triangles are from CMA- $\delta$  with the true  $\delta$  value, and squares are from BK with  $\delta = 0$ . Dashed lines are the true parameter values in (a) and (b), and zeros in (c) and (d).

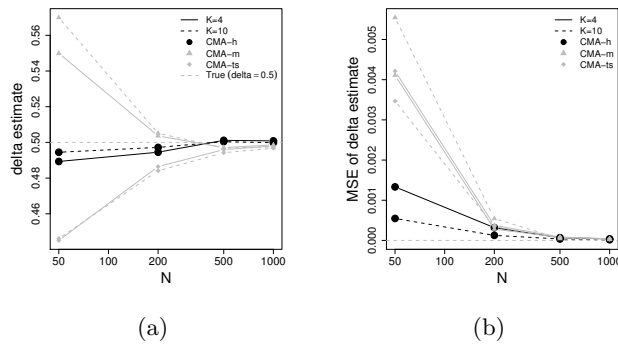


Figure 4: (a) Average point estimates of  $\delta$  and (b) the mean squared errors (MSEs) of  $\hat{\delta}$  by CMA-ts, CMA-h and CMA-m. Solid circles are from CMA-h, solid triangles are from CMA-m, and solid diamonds are from CMA-ts. Dashed lines show the true value of  $\delta$  in (a) and zero in (b).

Table 1: Average point estimates and empirical standard errors (in brackets) of CMA-ts, CMA-h, CMA-m, CMA- $\delta$ , KKB and BK.  $\delta$  is estimated in the multilevel CMA methods. It is set to zero in KKB and BK, and set to the true  $\delta$  in CMA- $\delta$ .

<i>Method</i>	$\delta$	$C$	$B$	$AB_p$	$AB_d$
True value	0.5	0.5	-1	-0.5	-0.5
CMA-ts	0.476 (0.029)	0.473 (0.121)	-0.940 (0.129)	-0.455 (0.119)	-0.450 (0.147)
CMA-h	0.502 (0.031)	0.504 (0.121)	-1.006 (0.136)	-0.488 (0.129)	-0.482 (0.153)
CMA-m	0.541 (0.037)	0.557 (0.131)	-1.117 (0.155)	-0.541 (0.143)	-0.535 (0.165)
CMA- $\delta$	-	0.719 (0.163)	-1.436 (0.127)	-0.699 (0.174)	-0.699 (0.174)
KKB	-	0.016 (0.171)	0.000 (0.105)	0.001 (0.053)	0.006 (0.110)
BK	-	0.183 (0.173)	-0.337 (0.122)	-0.163 (0.068)	-0.163 (0.068)
True value	0.5	0.5	0	0	0
CMA-ts	0.474 (0.029)	0.474 (0.117)	0.052 (0.132)	0.025 (0.069)	0.021 (0.117)
CMA-h	0.499 (0.030)	0.506 (0.117)	-0.014 (0.134)	-0.007 (0.070)	-0.011 (0.116)
CMA-m	0.538 (0.040)	0.560 (0.129)	-0.121 (0.170)	-0.062 (0.090)	-0.065 (0.132)
CMA- $\delta$	-	0.717 (0.160)	-0.442 (0.129)	-0.221 (0.091)	-0.221 (0.091)
KKB	-	0.013 (0.155)	0.985 (0.112)	0.485 (0.125)	0.482 (0.155)
BK	-	0.174 (0.163)	0.656 (0.124)	0.322 (0.093)	0.322 (0.093)
True value	0	0.5	-1	-0.5	-0.5
CMA-ts	-0.001 (0.040)	0.490 (0.125)	-0.988 (0.137)	-0.485 (0.132)	-0.479 (0.158)
CMA-h	-0.001 (0.044)	0.490 (0.126)	-0.988 (0.141)	-0.485 (0.133)	-0.478 (0.158)
CMA-m	0.000 (0.086)	0.493 (0.151)	-0.991 (0.197)	-0.488 (0.154)	-0.481 (0.176)
CMA- $\delta$	-	0.498 (0.144)	-0.991 (0.120)	-0.485 (0.126)	-0.485 (0.126)
KKB	-	0.491 (0.122)	-0.991 (0.108)	-0.485 (0.119)	-0.479 (0.151)
BK	-	0.498 (0.144)	-0.991 (0.120)	-0.485 (0.126)	-0.485 (0.126)

Table 2: Average estimates and 95% confidence intervals from the multilevel CMA methods, CMA- $\delta$  method with  $\delta$  estimated from CMA-h, KKB and BK on the fMRI dataset with and without motion correction (MC), using 500 bootstrap samples.

<i>Data</i>	<i>Method</i>	$\delta$	<i>C</i>	<i>AB<sub>p</sub></i>
Without MC	CMA-ts	-0.127 (-0.201, -0.056)	-0.051 (-0.099, -0.002)	0.166 (0.142, 0.191)
	CMA-h	-0.465 (-0.578, -0.325)	-0.177 (-0.247, -0.108)	0.293 (0.237, 0.350)
	CMA-m	-0.652 (-0.794,-0.456)	-0.288 (-0.412, -0.164)	0.406 (0.289, 0.523)
	CMA- $\delta$	-	-0.214 (-0.289, -0.138)	0.295 (0.231, 0.359)
	KKB	-	-0.007 (-0.050, 0.036)	0.123 (0.119, 0.127)
	BK	-	-0.029 (-0.071, 0.012)	0.111 (0.107, 0.115)
With MC	CMA-ts	-0.149 (-0.212, -0.04)	-0.074 (-0.122, -0.025)	0.155 (0.134, 0.176)
	CMA-h	-0.439 (-0.532, -0.322)	-0.178 (-0.239, -0.117)	0.260 (0.215, 0.306)
	CMA-m	-0.586 (-0.730, -0.410)	-0.253 (-0.357, -0.150)	0.336 (0.241, 0.432)
	CMA- $\delta$	-	-0.225 (-0.290, -0.160)	0.261 (0.210, 0.312)
	KKB	-	-0.023 (-0.069, 0.022)	0.106 (0.102, 0.110)
	BK	-	-0.061 (-0.105, 0.017)	0.097 (0.093, 0.101)