Time-varying mediation analysis for incomplete data with application to DNA methylation study for PTSD

Kecheng Wei[a], Fei Xue[b], Qi Xu[c], Yubai Yuan[d], Yuexia Zhang[e], Guoyou Qin[a], Agaz H. Wani[f], Allison E. Aiello[g], Derek E. Wildman[f], Monica Uddin[f] and Annie Qu[h*]

[a]Department of Biostatistics, Fudan University;

[b]Department of Statistics, Purdue University;

[c]Department of Statistics and Data Science, Carnegie Mellon University;

[d]Department of Statistics, Penn State University;

[e]Department of Management Science and Statistics, The University of Texas at San Antonio;

[f]College of Public Health, University of South Florida;

[g]Department of Epidemiology and Robert N. Butler Columbia Aging Center, Columbia University

[h]Department of Statistics, University of California Irvine;

Abstract

DNA methylation (DNAm) has been shown to mediate causal effects from traumatic experiences to post-traumatic stress disorder (PTSD). However, the scientific question about whether the mediation effect changes over time remains unclear. In this paper, we develop time-varying structural equation models to identify cytosine-phosphate-guanine (CpG) sites where DNAm mediates the effect of trauma exposure on PTSD, and to capture dynamic changes in mediation effects. The proposed methodology is motivated by the Detroit Neighborhood Health Study (DNHS) with high-dimensional and longitudinal DNAm measurements. To handle the non-monotone missing DNAm in the dataset, we propose a novel Longitudinal Multiple Imputation (LMI) method utilizing dependency among repeated measurements, and employ the generalized method of moments to integrate the multiple imputations. Simulations confirm that the proposed method outperforms existing approaches in various longitudinal settings. In DNHS data analysis, our method identifies several CpG sites where DNAm exhibits dynamic mediation effects. Some of the corresponding genes have been shown to be associated with PTSD in the existing literature, and our findings

*CONTACT Annie Qu. Email: aqu2@uci.edu

on their time-varying effects could deepen the understanding of the mediation role of DNAm on the causal path from trauma exposure to PTSD risk.

Keywords: generalized method of moments, high dimensional mediators, longitudinal data, missing data, multiple imputation, structural equation modeling, variable selection.

1 Introduction

Post-traumatic stress disorder (PTSD) is a serious mental health disorder that arises after an individual witnesses or experiences traumatic events, such as severe accidents, natural disasters, or warfare. Although the causality of trauma on PTSD has been established from the biological perspective, the trauma-PTSD relation has not been elucidated on the population level of public health data. World Health Organization World Mental Health Surveys show that, although 69.7% of the surveyed population were exposed to at least one traumatic event, the cross-national lifetime prevalence of PTSD was only 5.6% among the trauma-exposed subpopulation (Koenen et al. 2017). This significant disparity of prevalence between trauma exposure and PTSD implies that there are potential intermediate factors intervening in the trauma-PTSD pathway to explain individual vulnerability to trauma and tendency to develop PTSD. On the other hand, extensive meta-analyses show that the variability in PTSD on the population level is not sufficiently explained by incorporating phenotype variables as predictors (Tortella-Feliu et al. 2019). These limitations in previous PTSD studies lead us to further explore genotype information, such as genetic and relevant molecular variation, as intermediate variables, and identify causal mediation pathways from trauma exposure to PTSD via the transmission of genetic variation.

DNA methylation (DNAm) is a molecular modification that alters gene expression without changing DNA sequence, and can be influenced by stress and socio-environmental factors (Turecki and Meaney 2016, Lussier et al. 2023). It has been found that aberrant DNAm is associated with a variety of stress-related neuropsychiatric disorders such as depression, schizophrenia, and PTSD (Wani et al. 2021, Wilker et al. 2023). Thus, DNAm may be a pathway by which traumatic events become biologically embedded and contribute to mental illness. Given the intermediate role of DNAm in mental disorder formulation, studies have suggested that DNAm may also mediate the trauma-PTSD mechanism: trauma increases or reduces DNAm levels, which in turn exacerbate or alleviate PTSD symptoms.

For example, Rutten et al. (2018) found that DNAm mediates the relationship between combat trauma and PTSD. Occean et al. (2022) showed that DNAm in the nuclear factor of activated T cells 1 mediates the exposure to different types of traumatic events and the risk for PTSD. Xue et al. (2022) identified the heterogeneous mediation effects of DNAm in different subpopulations.

From the perspective of mediation methodology, most of the existing mediation analyses on DNAm are cross-sectional based, i.e., only estimating mediation pathways at a single time point. However, DNAm levels and PTSD severity typically progress over time, therefore the relationships among trauma, DNAm, and PTSD may temporally change. For example, Wu et al. (2021) found that different pre-migration and post-migration stressors significantly affect refugees' mental health at different resettlement stages. Lussier et al. (2023) demonstrated the temporal effects of adversity exposure on genetic variation across childhood and adolescence. Wilker et al. (2023) showed that the association between glucocorticoid receptor gene methylation status and PTSD symptoms is only significant in the early stages of study. Nevertheless, dynamic mediation analysis tools are still not well developed. By performing time-varying mediation analysis on longitudinal data, we can achieve a more comprehensive understanding of how causal pathways between variables in the disease process evolve over time: when the effects occur, diminish, strengthen, weaken, or change direction. Such insights could enrich understanding of the dynamic mediation effects of DNAm on pathways from trauma exposure to PTSD risk, and help to develop more targeted and time-sensitive interventions.

There have been some pioneering works about time-varying mediation analysis where exposure, mediator, and outcome are observed over multiple time points. For example, Rijnhart et al. (2022) and Cai et al. (2022) included time-interaction terms in their model to account for temporal effects. Zeng et al. (2021) and Zeng et al. (2022) extended the classical causal mediation method from a functional data analysis perspective. Ge et al. (2023) and Luo et al. (2023) studied dynamic mediation effects under a reinforcement learning framework. Nevertheless, most focus on cases where there is only one or a small number of mediators, with completely observed data. Few studies consider scenarios of high-dimensional mediators with incomplete data.

Missing data is ubiquitous in longitudinal study. In the statistical literature, two types of missingness structures are generally considered. One is called "monotone missingness"

or "dropout", where a subject may leave the study at a certain time point and never return. On the other hand, "non-monotone missingness" or "intermittent missingness" refers to when a subject may miss certain visits during follow-up and return at later visits. For our motivating Detroit Neighborhood Health Study (DNHS) which will be introduced later, subjects are supposed to have the methylation status of a large number of cytosine-phosphate-guanine (CpG) sites measured at each follow-up visit. Due to variability of compliance, the DNAm data exhibits the non-monotone missingness mechanism, and high-dimensional DNAm values are entirely missing at some visits by the individual. Classical approaches, such as weighting (Chen et al. 2021), imputation (Jahangiri et al. 2023), and likelihood-based methods (Tseng et al. 2016), have been developed to address non-monotone missingness in longitudinal data. However, these approaches typically consider non-monotone missingness of the outcome or with a few covariates, and may not be directly applicable to our setting where high-dimensional DNAm values are entirely unobserved at some visits.

In this paper, we develop a new time-varying mediation method to investigate the temporal mediation effects of DNAm transmission from trauma to PTSD. We propose time-varying structural equation models, and adopt a regularization approach to select relevant mediators and identify time-varying effects, which yields several advantages. First, we can identify the important mediators from high-dimensional DNAm values at different time points, in contrast to existing dynamic mediation approaches considering only one or a small number of mediators (Rijnhart et al. 2022). Second, we can capture the trajectory of mediation effects even with a small number of time points, while existing approaches require dense time grids (Cai et al. 2022).

To handle non-monotone missing DNAm values, we propose a new imputation method called Longitudinal Multiple Imputation (LMI), which has the following major advantages. First, it makes full use of the dependence among repeated measurements, which exploits the inherent correlation structure of longitudinal data, not just based on low-rank structures or trajectory means (Mazumder, Hastie and Tibshirani 2010, Jahangiri et al. 2023). Second, it can handle high-dimensional mediators being entirely unobserved at some time points, in contrast to existing approaches which only consider missingness of the outcome or with a few covariates (Chen et al. 2021). Third, it does not require some individuals with complete repeated measurements in the dataset, whereas existing approaches may require that there

must be some subjects with complete data, or that the data at the first time visit is always observed for all subjects (Jahangiri et al. 2023).

In simulation studies, the proposed method outperforms existing approaches under different missingness structures and different dimensions of mediators, and its performance is robust to different missingness mechanisms. In DNHS data analysis, our method identifies some DNAm CpG sites which exhibit dynamic mediation effects. Specifically, certain CpG sites initially have nonzero mediation effects, but the effects disappear over time. Certain CpG sites do not show mediation effects in early stages, but the effects emerge over time. Some genes (e.g., *PTPRK*) where selected CpG sites (e.g., cg09247979) are located have been reported as differentially expressed genes in previous PTSD meta-analyses (Chitrala, Nagarkatti and Nagarkatti 2016), and our further findings on their time-varying effects elucidate trauma-PTSD dynamic causal mechanisms.

The rest of the paper is organized as follows. Section 2 describes the background of the DNHS dataset. Section 3 introduces the proposed time-varying mediation models and the non-monotone missingness structure of mediators. Section 4 presents the proposed approaches for multiple imputation, data integration, and regularization. Section 5 illustrates the detailed implementation and algorithm. Section 6 includes the results of simulation studies. Section 7 applies the proposed method to the DNHS dataset. Section 8 concludes the research with some discussion.

2 Motivating application: Longitudinal trauma, PTSD, and DNAm

2.1 DNHS data

The Detroit Neighborhood Health Study (DNHS) is a prospective and representative longitudinal cohort study of predominantly African American adults living in Detroit, Michigan (Uddin et al. 2010, Goldmann et al. 2011). Participants completed a 40-minute structured telephone interview annually between 2008-2012 (i.e., waves 1-5), to assess demographic characteristics, exposure to traumatic events, and mental health status. All participants were given the opportunity to provide a specimen (venipuncture, blood spot, or saliva) for immune and inflammatory marker testing as well as genetic testing of DNA. We analyze

the DNAm data with 526 subjects from waves 1, 2, 4, and 5 as there was no full blood draw in wave 3 as in the other waves. We refer to waves 1, 2, 4, and 5 as time points 1-4 throughout the paper.

The exposure variable of interest in our study is the number of traumatic events. In the initial interview (time point 1), participants were asked about their lifetime traumatic experiences using a predefined list of 19 traumatic events (Breslau et al. 1998). In subsequent interviews (time points 2-4), they were additionally queried about traumatic experiences since their last interview within the same list. At time points 2-4, we incorporate the number of traumatic events experienced since the previous interview to capture the cumulative severity of trauma exposure. The first plot in the left panel of Figure 1 shows the longitudinal trajectories of trauma from 15 subjects.

The outcome variable of interest is PTSD symptom severity, which is measured through the widely used self-report PTSD Checklist (PCL, Ruggiero et al. 2003). The PCL contains 17 items corresponding to key symptoms of PTSD. In the initial interview (time point 1), participants were asked how much they had been bothered by each symptom using a 5-point scale (1-5) in reference to their worst lifetime traumatic experience, and the summary score for the 17 items were used to characterize the PTSD symptom severity. In subsequent interviews (time points 2-4), they were additionally queried about the worst traumatic experience since their last interview, and the summary score corresponding to it was similarly calculated. At time points 2-4, we choose the largest score calculated in previous and current interviews to capture the cumulative severity of PTSD symptoms. The second plot in the left panel of Figure 1 shows the longitudinal trajectories of PTSD from 15 subjects. We additionally incorporate demographic information such as age and gender, and blood work such as CD4 T cells and B cells, measured at baseline.

Existing studies suggest that epigenomic variations can be induced by trauma exposure and accompany the development of stress-related disorders (Rutten et al. 2018, Occean et al. 2022, Xue et al. 2022). However, few studies have reported brain-related epigenomic profiles that associate with PTSD risk. In DNHS, the methylation levels of 1879 CpG sites from paired blood and brain tissue (Braun et al. 2019) were assessed. We treat the DNAm levels of these 1879 CpG sites as potential mediators. The methylation level of a CpG site is quantified using the "Beta-value", which ranges from 0 to 1 and is calculated as the proportion $M/(M+U+100)$, where M and U represent the signal intensities of methylated

and unmethylated probes (Mou et al. 2022). The third and fourth plots in the left panel of Figure 1 show the longitudinal trajectories of DNAm levels for two CpG sites from 15 subjects. We observe that the DNAm levels fluctuate within a specific range over the follow-up period, without exhibiting a clear trend or consistent pattern. The average rate of change in DNAm levels for all 1879 CpG sites across the 15 subjects between time points t and t' can be calculated as $\mathrm{CR}_{tt'} = \frac{1}{15}\sum_{i=1}^{15}\sum_{j=1}^{1879}(M_{it}^{j} - M_{it'}^{j})/M_{it'}^{j}$. The corresponding results for CR_{21}, CR_{31}, CR_{41}, CR_{32}, CR_{42}, and CR_{43} are 2.6%, 0%, 1.9%, -0.1%, 1.2%, and 3.2%, respectively.

The traumatic events and PTSD severity are easily assessed, so that we have complete repeated measurements for exposure and outcome. However, collecting and testing biospecimens is difficult and costly. Due to variability of compliance, loss to follow-up, and budget restriction, DNAm data could suffer from missingness. In DNHS, DNAm data of some subjects may be available only at certain visits, but are unobserved at other visits. The right panel of Figure 1 shows the non-monotone missingness structure of DNAm. Each blank column represents unobserved DNAm, while the colored ones represent observed DNAm. Participants can be divided into 15 groups, where group 1 has 4 repeated observations at all time points, while group 2 only has observations at time points 1-3, with high-dimensional DNAm data entirely missing at time point 4. The sample sizes of groups 1-15 are (15, 29, 7, 28, 21, 45, 27, 12, 23, 9, 21, 147, 24, 102, 16), respectively. There are 511 out of 526 subjects with missing values and the missingness rate of DNAm is 58.3%.

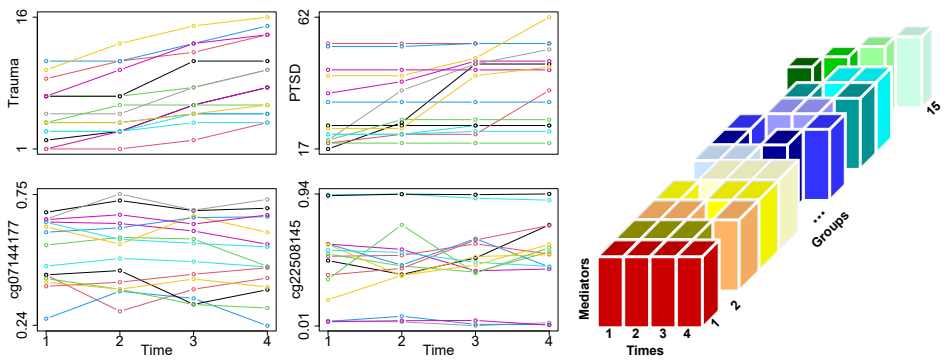

Figure 1: Left: Longitudinal trajectories of trauma, PTSD, and DNAm of two CpG sites from 15 subjects with complete repeated measurements of DNAm. Right: Non-monotone missingness structure of DNAm across all time points. Each blank column represents unobserved DNAm, while the colored ones represent observed DNAm.

2.2 Novelty of our analysis

Several studies have suggested that DNAm plays a critical role as a mediator in the causal relationship between trauma and PTSD. For instance, Xue et al. (2022) leveraged DNHS data to identify multiple DNAm CpG sites mediating the effect of trauma on PTSD, highlighting that these epigenetic factors can play different mediation roles across different subpopulations. However, their analysis relied on baseline wave to estimate static mediation effects, which is not applicable for discovering potential time-varying genetic contributions to the disease process.

Time-varying effects are a common phenomenon in the genetic regulation of disease progression. Dynamic factors such as environmental changes and the natural course of the disease interact intricately with gene regulation mechanisms, suggesting that genetic variants may adapt to changing environmental conditions and show different effects at different stages of the disease (Lussier et al. 2023). Investigating time-varying genetic effects provides a more nuanced understanding of how genetic factors influence disease processes, helps identify critical time windows for intervention, and facilitates the design of time-sensitive therapeutic strategies.

However, studying time-varying effects poses great challenges. As illustrated in Figure 1, while trauma exposure and PTSD severity tend to increase monotonically over time, the dynamics of DNAm appear irregular, making it difficult to capture their patterns of variation. Moreover, the issues of missingness and high dimensionality within the DNHS dataset further complicate the analysis. These challenges underscore the necessity of developing innovative statistical models and estimation methods to characterize time-varying effects and accommodate complex data structures.

3 Models

In this section, we propose time-varying structural equation models to study dynamic mediation effects, and introduce the non-monotone missingness structure of mediators.

3.1 Time-varying structural equation models

Let $\boldsymbol{M}_{it} = (M_{it}^1, ..., M_{it}^p)^T$ be a $p \times 1$ vector of mediators, where M_{it}^j is the j-th ($j = 1, ..., p$) mediator of the i-th ($i = 1, ..., N$) subject, measured at the t-th ($t = 1, ..., T$) time point.

Let $X_{it} \in \mathbb{R}$, $Y_{it} \in \mathbb{R}$, and $\boldsymbol{Z}_i \in \mathbb{R}^q$ be the longitudinal exposure, the longitudinal outcome, and the demographic variables measured at baseline, respectively. We propose the following structural equation models:

$$Y_{it} = \alpha_t X_{it} + \boldsymbol{\beta}_t^T \boldsymbol{M}_{it} + \boldsymbol{\delta}^T \boldsymbol{Z}_i + e_{it}, \tag{1}$$

$$\boldsymbol{M}_{it} = \boldsymbol{\gamma}_t X_{it} + \boldsymbol{\eta}_t Y_{i,t-1} + \boldsymbol{\Gamma} \boldsymbol{Z}_i + \boldsymbol{\varepsilon}_{it}, \tag{2}$$

with $Y_{i0} = 0$, where the coefficients α_t, $\boldsymbol{\beta}_t = (\beta_t^1, ..., \beta_t^p)^T$, $\boldsymbol{\gamma}_t = (\gamma_t^1, ..., \gamma_t^p)^T$, and $\boldsymbol{\eta}_t = (\eta_t^1, ..., \eta_t^p)^T$ measure the time-varying effects at the t-th time point, $\boldsymbol{\delta} \in \mathbb{R}^q$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times q}$ measure the time-invariant effects. The random errors e_{it} and $\boldsymbol{\varepsilon}_{it} = (\varepsilon_{it}^1, ..., \varepsilon_{it}^p)^T$ are zero-mean, where e_{it} is independent of X_{it}, $\boldsymbol{M}_{it}$, and $\boldsymbol{Z}_i$, and $\boldsymbol{\varepsilon}_{it}$ is independent of X_{it}, $Y_{i,t-1}$, and $\boldsymbol{Z}_i$. We define $\boldsymbol{\Theta} = (\boldsymbol{\theta}^T, \boldsymbol{\vartheta}^{1T}, ..., \boldsymbol{\vartheta}^{pT})^T$ with $\boldsymbol{\theta} = (\alpha_1, ..., \alpha_T, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_T^T, \boldsymbol{\delta}^T)^T$ and $\boldsymbol{\vartheta}^j = (\gamma_1^j, ..., \gamma_T^j, \eta_2^j, ..., \eta_T^j, \boldsymbol{\Gamma}^{jT})^T$, where $\boldsymbol{\Gamma}^j$ is the j-th row of $\boldsymbol{\Gamma}$ for $j = 1, ..., p$.

Figure 2 illustrates the relationships between variables in models (1) and (2) without presenting $\boldsymbol{Z}_i$, where solid unidirectional arrows represent causal relationships, while dashed bidirectional arrows indicate correlations. For a given time point t, α_t represents the direct effect from the exposure X_{it} to the outcome Y_{it}, and $\gamma_t^j \beta_t^j$ represents the indirect effect through the mediator M_{it}^j. There are correlations among the mediators $(M_{it}^1, ..., M_{it}^p)$. In the context of our real data, the DNAm levels of multiple CpG sites exhibit correlations, reflecting their shared regulatory roles or proximity to similar biological functions (Mou et al. 2022).

For adjacent time points $t-1$ and t, we assume both correlation and causation exist among the exposure, mediators, and outcome. We postulate correlations among repeated measurements of outcome Y_{it}. And similarly, we assume correlations among repeated measurements of mediator M_{it}^j. Regarding the causal relationships across time points, since X_{it} reflects the accumulated level of trauma exposure, it is influenced mainly by the exposure at the previous time point. Based on existing literature, DNAm is primarily influenced by external adverse environmental exposures or disease conditions (Turecki and Meaney 2016), we specify the causal effect of PTSD condition $Y_{i,t-1}$ on $(M_{it}^1, ..., M_{it}^p)$ to be $(\eta_t^1, ..., \eta_t^p)$. Following the Markov assumption (Luo et al. 2023), we assmue that spillover effects among variables occur exclusively at adjacent time points.

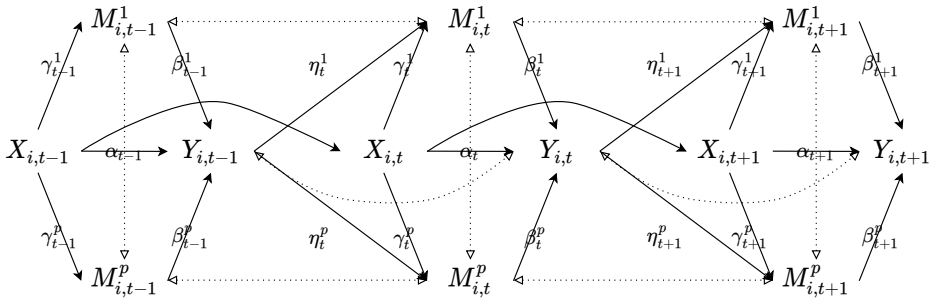

Figure 2: Dynamic mediation structure among time points $t-1$, t, and $t+1$. Solid unidirectional arrows represent causal relationships, while dashed bidirectional arrows indicate correlations.

Regression coefficients in models (1) and (2) can be linked to causal estimands. Let $Y^{x,\boldsymbol{m}}$ be the potential outcome that would have been observed if X was set to x and $\boldsymbol{M}$ was set to $\boldsymbol{m}$, and $\boldsymbol{M}^x$ be the potential mediators that would have been observed if X was set to x. Following Bind et al. (2016), we formalize the following assumptions:

$$\text{(i)}\ \ Y_{it}^{x,\boldsymbol{m}} \perp\!\!\!\perp X_{it} | \boldsymbol{Z}_i = \boldsymbol{z}, \qquad \text{(ii)}\ \ \boldsymbol{M}_{it}^{x} \perp\!\!\!\perp X_{it} | \boldsymbol{Z}_i = \boldsymbol{z},$$

$$\text{(iii)}\ \ Y_{it}^{x,\boldsymbol{m}} \perp\!\!\!\perp \boldsymbol{M}_{it} | X_{it} = x, \boldsymbol{Z}_i = \boldsymbol{z}, \qquad \text{(iv)}\ \ Y_{it}^{x,\boldsymbol{m}} \perp\!\!\!\perp \boldsymbol{M}_{it}^{\tilde{x}} | \boldsymbol{Z}_i = \boldsymbol{z},$$

which indicate that there is no (i) unmeasured exposure-outcome confounding, (ii) exposure-mediator confounding, (iii) mediator-outcome confounding, or (iv) exposure-induced confounding, where $\tilde{x}$ is the realization of exposure at a different value from x. Corresponding to our DNHS data, to estimate the direct effect of trauma exposure X_{it} on PTSD outcome Y_{it}, as well as the indirect effect mediated through DNAm $\boldsymbol{M}_{it}$, it is essential to control for confounding bias. Assumption (i) requires that all causes of trauma and PTSD are observed and adjusted for. These causes may include demographic information such as age and gender (Wani et al. 2021), which have been accounted for in $\boldsymbol{Z}_i$. Assumption (ii) requires that all causes of trauma and DNAm are observed and adjusted for. These may include demographic information like age and gender (Wani et al. 2021), which have been accounted for in $\boldsymbol{Z}_i$. Assumption (iii) requires that all causes of DNAm and PTSD are observed and adjusted for. These may include demographic information such as age and gender, as well as blood work (Occean et al. 2022), which have been accounted for in $\boldsymbol{Z}_i$. Assumption (iv) requires that the causes (except for trauma) of DNAm and PTSD are not induced by trauma. This assumption is reasonable, as the demographic information and blood work we control for are intrinsic attributes of individuals. Note that although

$Y_{i,t-1}$ may have a spillover effect on $(M_{it}^1, ..., M_{it}^p)$, it does not directly affect X_{it} and Y_{it}. Consequently, there is no time-varying confounding in our study, similar to the arguments by VanderWeele and Tchetgen Tchetgen (2017), and thus following Bind et al. (2016), the natural direct effect at time point t for x versus $\tilde{x}$ is $\mathbb{E}(Y_{it}^{x,\boldsymbol{m}_{it}^{\tilde{x}}} - Y_{it}^{\tilde{x},\boldsymbol{m}_{it}^{\tilde{x}}}) = \alpha_t(x - \tilde{x})$, and the natural indirect effect at time point t for x versus $\tilde{x}$ is $\mathbb{E}(Y_{it}^{x,\boldsymbol{m}_{it}^{x}} - Y_{it}^{x,\boldsymbol{m}_{it}^{\tilde{x}}}) = \boldsymbol{\beta}_t^T \boldsymbol{\gamma}_t(x - \tilde{x})$.

3.2 Non-monotone missing mediators

In our DNHS data, exposure X_{it}, outcome Y_{it}, and confounders $\boldsymbol{Z}_i$ are always observed, but the mediators $\boldsymbol{M}_{it}$ suffer from non-monotone missingness due to the following reasons: $\boldsymbol{M}_i = (\boldsymbol{M}_{i1}, \ldots, \boldsymbol{M}_{iT})^T$ of some subjects may be available only at certain visits, but missing at the next time point, and measured again at later visits. Given the non-monotone missingness, subjects can be divided into R disjoint groups based on different missingness patterns across all time points. Figure 3 illustrates an example where we have $T = 3$ time points and $R = 7$ groups. Each blank column represents unobserved mediators, while the colored ones represent observed mediators. Notice that subjects in group 1 have observed $\boldsymbol{M}_i$ at all time points, but subjects in group 2 have observed $\boldsymbol{M}_i$ only at time points 1 and 2.

Define the missing indicator $\boldsymbol{R}_{it} = (R_{it}^1, ..., R_{it}^p)^T$, where $R_{it}^j = 1$ for $j = 1, ..., p$ if $\boldsymbol{M}_{it}$ is observed at time point t, and $R_{it}^j = 0$ for $j = 1, ..., p$ if $\boldsymbol{M}_{it}$ is missing at time point t. In this paper, we assume that the missingness mechanism is missing at random (MAR, Little and Rubin 2019), meaning that the conditional distribution of the missing indicator $\boldsymbol{R}_i = (\boldsymbol{R}_{i1}, \ldots, \boldsymbol{R}_{iT})^T$, denoted as $f_{\boldsymbol{R}_i}(\boldsymbol{r}_i | \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{M}_i)$, depends only on the exposure $\boldsymbol{X}_i = (X_{i1}, ..., X_{iT})^T$, the outcome $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iT})^T$, the confounders $\boldsymbol{Z}_i$, and the observed mediator component $\boldsymbol{R}_i \boldsymbol{M}_i$ (Wei et al. 2022).

Let $\mathscr{H}(r)$ be the index set of subjects in group r and $n_r = |\mathscr{H}(r)|$ be the sample size $(r = 1, ..., R)$. Let $o(r)$ and $u(r)$ be the index sets of time points corresponding to observed and unobserved mediators in group r. Consequently, $\boldsymbol{M}_{o(r)}$ and $\boldsymbol{M}_{u(r)}$ represent observed and unobserved mediators in group r. In addition, for each $t \in u(r)$, we define $\mathscr{T}(r, t)$ as the collection of time sets, and each element in $\mathscr{T}(r, t)$ is a subset of $o(r)$. The $\mathscr{T}(r, t)$ can be obtained by the following steps: (i) let $\mathscr{G}(r, t)$ be the index set of the groups where mediators are observed not only at time point t, but also at least one time point in $o(r)$; (ii) for each $r' \in \mathscr{G}(r, t)$, let $\mathscr{t}_{r'} = o(r) \cap o(r')$; (iii) $\mathscr{T}(r, t) = \bigcup_{r' \in \mathscr{G}(r,t)} \mathscr{t}_{r'}$. Note that the

definition of $\mathscr{T}(r,t)$ is to prepare for the imputation method later. That is, we can use observed values at each time set $\boldsymbol{t} \in \mathscr{T}(r,t)$ to impute the unobserved values at time point $t \in u(r)$.

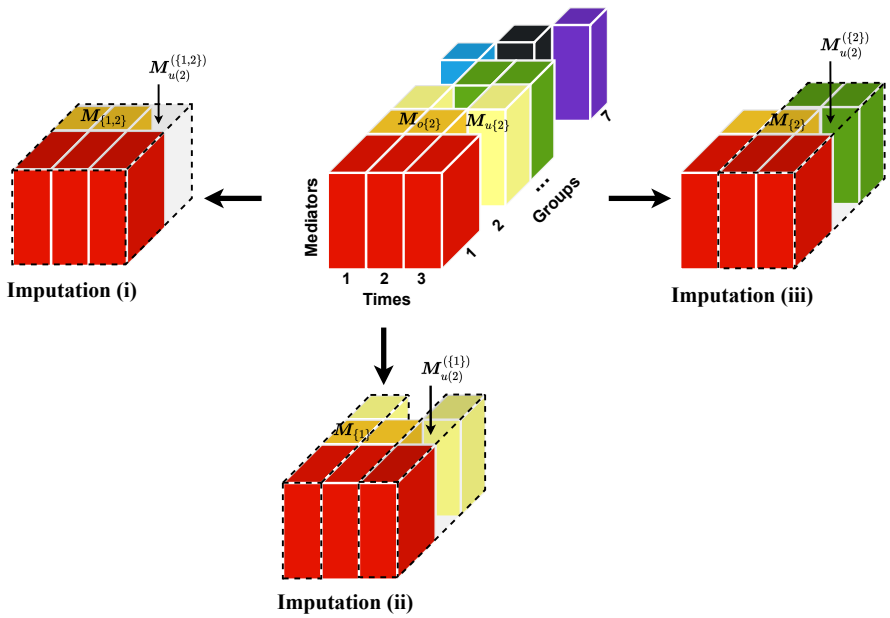

Figure 3: Middle: Non-monotone missingness structure of mediators across all time points. Each blank column represents unobserved mediators, while the colored ones represent observed mediators. Surroundings: Multiple imputations for the missing column $\boldsymbol{M}_{u(2)}$ in group 2.

4 Methods

In this section, we first propose a new imputation method called Longitudinal Multiple Imputation (LMI) to handle non-monotone missing data. Second, we adopt the generalized method of moments (GMM, Hansen 1982) to combine the multiple imputations. Third, we propose a regularization approach to identify relevant mediators and estimate time-varying effects. Finally, we employ principal component analysis to solve the singularity issue of the weighting matrix in GMM.

4.1 Longitudinal multiple imputation (LMI)

In this subsection, we propose the LMI approach to handle non-monotone missing longitudinal data. The main idea is that, due to the correlation between repeated measurements $\boldsymbol{M}_t$ $(t = 1, ..., T)$ from the same subject, we can use the observed values at some time points to predict the unobserved values at other time points. For example, for subjects with observed $\boldsymbol{M}_t$ at time points 1 and 2, we can use some statistical methods, such as regression, to learn the correlation of variables between these two time points. Thus, for some other individuals with observed $\boldsymbol{M}_t$ at time point 1 but missing $\boldsymbol{M}_t$ at time point 2, we can impute the data at time point 2 based on the learned correlation.

Given a group r with a time point $t \in u(r)$, subjects have unobserved M_t^j at this time point, but have observed $\boldsymbol{M}_{\mathscr{t}}$ at each time set $\mathscr{t} \in \mathscr{T}(r, t)$. We can use observed values at each time set $\mathscr{t} \in \mathscr{T}(r, t)$ to impute the unobserved values at time point $t \in u(r)$. Specifically, given a time set $\mathscr{t} \in \mathscr{T}(r, t)$, we first estimate $\mathbb{E}(M_t^j | \boldsymbol{M}_{\mathscr{t}}, \boldsymbol{D})$ using other groups containing observed M_t^j and $\boldsymbol{M}_{\mathscr{t}}$, where $\boldsymbol{D}$ contains exposure $\boldsymbol{X} = (X_1, ..., X_T)^T$ at all time points, outcome Y_{t-1} at time point $t - 1$, and confounders $\boldsymbol{Z}$. Based on the learned correlation, we then impute unobserved M_t^j using $\boldsymbol{M}_{\mathscr{t}}$ and $\boldsymbol{D}$. Let $\boldsymbol{M}_t^{(\mathscr{t})} = \{\mathbb{E}(M_t^j | \boldsymbol{M}_{\mathscr{t}}, \boldsymbol{D}), j = 1, .., p\}$ represent the imputed values based on the time set $\mathscr{t} \in \mathscr{T}(r, t)$, then we can obtain $|\mathscr{T}(r, t)|$ imputations.

We illustrate the LMI with an example in Figure 3. Specifically, $\mathscr{H}(2)$ refers to the index set of subjects in group 2. Since subjects in group 2 have observed mediators at time points 1 and 2, but have unobserved mediators at time point 3, $o(2) = \{1, 2\}$ and $\boldsymbol{M}_{o(2)}$ refers to observed mediators, $u(2) = \{3\}$ and $\boldsymbol{M}_{u(2)}$ refers to unobserved mediators. The collection of time sets $\mathscr{T}(2, 3) = \{\{1, 2\}, \{1\}, \{2\}\}$ since: (i) $\mathscr{G}(2, 3) = \{1, 3, 4\}$; (ii) $\mathscr{t}_1 = o(2) \cap o(1) = \{1, 2\}$, $\mathscr{t}_3 = o(2) \cap o(3) = \{1\}$, and $\mathscr{t}_4 = o(2) \cap o(4) = \{2\}$; (iii) $\mathscr{T}(2, 3) = \bigcup_{r' \in \mathscr{G}(2,3)} \mathscr{t}_{r'} = \{\{1, 2\}, \{1\}, \{2\}\}$.

Given group 2 with unobserved $\boldsymbol{M}_t$ at time point $t \in u(r) = 3$, we can impute the unobserved values in $|\mathscr{T}(r, t)| = |\{\{1, 2\}, \{1\}, \{2\}\}| = 3$ ways: for $j = 1, .., p$, (i) for $\mathscr{t} = \{1, 2\}$, we estimate $\mathbb{E}(M_3^j | \boldsymbol{M}_{\{1,2\}}, \boldsymbol{D})$ using group 1, and then obtain the imputed values $\boldsymbol{M}_3^{(\{1,2\})}$; (ii) for $\mathscr{t} = \{1\}$, we estimate $\mathbb{E}(M_3^j | \boldsymbol{M}_{\{1\}}, \boldsymbol{D})$ using groups 1 and 3, and then obtain the imputed values $\boldsymbol{M}_3^{(\{1\})}$; (iii) for $\mathscr{t} = \{2\}$, we estimate $\mathbb{E}(M_3^j | \boldsymbol{M}_{\{2\}}, \boldsymbol{D})$ using groups 1 and 4, and then obtain the imputed values $\boldsymbol{M}_3^{(\{2\})}$.

Remark 1. One novelty of our LMI method is to make full use of the dependence among

repeated measurements, which exploits the inherent correlation structure of longitudinal data, not based on stringent low-rank assumptions (Mazumder, Hastie and Tibshirani 2010) or information-poor trajectory means (Jahangiri et al. 2023). The LMI method allows high-dimensional variables to be entirely unobserved at some time points, while existing approaches only consider missingness of the outcome or with a few covariates (Chen et al. 2021). Furthermore, the LMI method allows for the absence of individuals with complete repeated measurements in the dataset, whereas existing approaches may require that there must be some subjects with complete data, or that the data at the first time visit is always observed for all subjects (Jahangiri et al. 2023).

4.2 Combining multiple imputations

In this subsection, we propose to combine the information from all LMIs. Since different imputations correspond to different estimating functions, based on moment conditions (Hansen 1982), we integrate all information through the GMM.

First, we construct the moment conditions corresponding to model (1). For a given group r, $j = 1, .., p$, and each time set $\boldsymbol{t} \in \mathscr{T}(r, t)$, since unobserved M_t^j at time point $t \in u(r)$ is imputed through $M_t^{j(\boldsymbol{t})} = \mathbb{E}(M_t^j | \boldsymbol{M}_{\boldsymbol{t}}, \boldsymbol{D})$, the covariates $\boldsymbol{M}_{\boldsymbol{t}}$ and $\boldsymbol{D}$ are uncorrelated with residuals of the projection $M_t^j - M_t^{j(\boldsymbol{t})}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left\{\left(X_t, \boldsymbol{M}_t^{(\boldsymbol{t})T}\right)^T \left(Y_t - \alpha_{0t} X_t - \boldsymbol{\beta}_{0t}^T \boldsymbol{M}_t^{(\boldsymbol{t})} - \boldsymbol{\delta}_0^T \boldsymbol{Z}\right)\right\} \\
=&\mathbb{E}\left\{\left(X_t, \boldsymbol{M}_t^{(\boldsymbol{t})T}\right)^T \left(e_t + \boldsymbol{\beta}_{0t}^T \left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(\boldsymbol{t})}\right)\right)\right\} = \boldsymbol{0},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{t=1}^{T} \boldsymbol{Z} \left(Y_t - \alpha_{0t} X_t - \boldsymbol{\beta}_{0t}^T \boldsymbol{M}_t^{(\boldsymbol{t})} - \boldsymbol{\delta}_0^T \boldsymbol{Z}\right)\right\} \\
=&\mathbb{E}\left\{\sum_{t=1}^{T} \boldsymbol{Z} \left(e_t + \boldsymbol{\beta}_{0t}^T \left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(\boldsymbol{t})}\right)\right)\right\} = \boldsymbol{0}.
\end{aligned}$$

The moment conditions are satisfied for the true parameter $\boldsymbol{\theta}_0 = (\alpha_{01}, ..., \alpha_{0T}, \boldsymbol{\beta}_{01}^T, ..., \boldsymbol{\beta}_{0T}^T, \boldsymbol{\delta}_0^T)^T$.

We then construct estimating functions based on these moment conditions. For each $i \in \mathscr{H}(r)$, let $\boldsymbol{M}_{it}^{(\boldsymbol{t})}$ be the imputed values based on time set $\boldsymbol{t} \in \mathscr{T}(r, t)$, where $\boldsymbol{M}_{it}^{(\boldsymbol{t})} = \boldsymbol{M}_{it}$ for $t \in o(r)$, and $M_{it}^{j(\boldsymbol{t})} = \mathbb{E}(M_{it}^j | \boldsymbol{M}_{i\boldsymbol{t}}, \boldsymbol{D}_i)$ for $t \in u(r)$ and $j = 1, .., p$. The estimating

functions for X_{it} and $\boldsymbol{M}_{it}^{(t)}$ at time point $t \in o(r)$ are $\boldsymbol{g}_{it}^{(r)}(\boldsymbol{\theta}) = (X_{it}, \boldsymbol{M}_{it}^{(t)T})^T(Y_{it} - \alpha_t X_{it} - \boldsymbol{\beta}_t^T \boldsymbol{M}_{it}^{(t)} - \boldsymbol{\delta}^T \boldsymbol{Z}_i)$ in which $\boldsymbol{M}_{it}^{(t)} = \boldsymbol{M}_{it}$, and the estimating functions for X_{it} and $\boldsymbol{M}_{it}^{(t)}$ at time point $t \in u(r)$ are $\boldsymbol{g}_{it}^{(r)}(\boldsymbol{\theta}) = \{\boldsymbol{g}_{it}^{(r,\acute{t})}(\boldsymbol{\theta}), \acute{t} \in \mathscr{T}(r,t)\}$, where

$$\boldsymbol{g}_{it}^{(r,\acute{t})}(\boldsymbol{\theta}) = \left(X_{it}, \boldsymbol{M}_{it}^{(t)T}\right)^T \left(Y_{it} - \alpha_t X_{it} - \boldsymbol{\beta}_t^T \boldsymbol{M}_{it}^{(t)} - \boldsymbol{\delta}^T \boldsymbol{Z}_i\right). \tag{3}$$

The estimating functions for $\boldsymbol{Z}_i$ are $\mathfrak{g}_i^{(r)}(\boldsymbol{\theta}) = \{\sum_{t=1}^T \mathfrak{g}_{it}^{(r)}(\boldsymbol{\theta})\}$, where $\mathfrak{g}_{it}^{(r)}(\boldsymbol{\theta}) = \boldsymbol{Z}_i(Y_{it} - \alpha_t X_{it} - \boldsymbol{\beta}_t^T \boldsymbol{M}_{it}^{(t)} - \boldsymbol{\delta}^T \boldsymbol{Z}_i)$ at time point $t \in o(r)$ in which $\boldsymbol{M}_{it}^{(t)} = \boldsymbol{M}_{it}$, and $\mathfrak{g}_{it}^{(r)}(\boldsymbol{\theta}) \in \{\mathfrak{g}_{it}^{(r,\acute{t})}(\boldsymbol{\theta}), \acute{t} \in \mathscr{T}(r,t)\}$ at time point $t \in u(r)$ in which

$$\mathfrak{g}_{it}^{(r,\acute{t})}(\boldsymbol{\theta}) = \boldsymbol{Z}_i \left(Y_{it} - \alpha_t X_{it} - \boldsymbol{\beta}_t^T \boldsymbol{M}_{it}^{(t)} - \boldsymbol{\delta}^T \boldsymbol{Z}_i\right). \tag{4}$$

To integrate information from all groups, we propose an aggregated vector of estimating functions:

$$\boldsymbol{g}(\boldsymbol{\theta}) = \left\{\left(\boldsymbol{g}^{(1)}(\boldsymbol{\theta})\right)^T, ..., \left(\boldsymbol{g}^{(R)}(\boldsymbol{\theta})\right)^T\right\}^T, \tag{5}$$

where $\boldsymbol{g}^{(r)}(\boldsymbol{\theta}) = \frac{1}{n_r}\sum_{i \in \mathscr{K}(r)}\{(\boldsymbol{g}_{i1}^{(r)}(\boldsymbol{\theta}))^T, ..., (\boldsymbol{g}_{iT}^{(r)}(\boldsymbol{\theta}))^T, (\mathfrak{g}_i^{(r)}(\boldsymbol{\theta}))^T\}^T$. Note that the total number of estimating functions exceeds the dimension of parameters, and the estimating functions from groups with fewer missingness or more precise imputations tend to have smaller variance. To combine all estimating functions in $\boldsymbol{g}(\boldsymbol{\theta})$ and put more weights on high-quality moments, we propose the following quadratic loss function:

$$G(\boldsymbol{\theta}) = \sum_{r=1}^{R} \left(\boldsymbol{g}^{(r)}(\boldsymbol{\theta})\right)^T \left(\boldsymbol{W}^{(r)}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{g}^{(r)}(\boldsymbol{\theta}), \tag{6}$$

with weighting matrix $\boldsymbol{W}^{(r)}(\boldsymbol{\theta}) = \frac{1}{n_r}\sum_{i \in \mathscr{K}(r)} \boldsymbol{g}^{(r)}(\boldsymbol{\theta})(\boldsymbol{g}^{(r)}(\boldsymbol{\theta}))^T$.

For each mediator M^j $(j = 1, ..., p)$ in model (2), we can get the aggregated vector of estimating functions $\boldsymbol{h}^j(\boldsymbol{\vartheta}^j)$ similar to (5). Detailed formulations are shown in the Supplementary Material. We propose the following quadratic loss function:

$$H^j\left(\boldsymbol{\vartheta}^j\right) = \sum_{r=1}^{R} \left(\boldsymbol{h}^{j(r)}\left(\boldsymbol{\vartheta}^j\right)\right)^T \left(\boldsymbol{W}^{j(r)}\left(\boldsymbol{\vartheta}^j\right)\right)^{-1} \boldsymbol{h}^{j(r)}\left(\boldsymbol{\vartheta}^j\right), \tag{7}$$

with weighting matrix $\boldsymbol{W}^{j(r)}(\boldsymbol{\vartheta}^j) = \frac{1}{n_r}\sum_{i \in \mathscr{K}(r)} \boldsymbol{h}^{j(r)}(\boldsymbol{\vartheta}^j)(\boldsymbol{h}^{j(r)}(\boldsymbol{\vartheta}^j))^T$. The loss function incorporating models (1) and (2) for all subjects is $G(\boldsymbol{\theta}) + \sum_{j=1}^p H^j(\boldsymbol{\vartheta}^j)$.

Remark 2. For multiple imputed data, different imputations correspond to different estimating functions, all of which satisfy the moment conditions. We use GMM to combine all estimating functions, which adaptively puts more weights to imputations with smaller

variance, while still consider the information contained in imputations with larger variance. In contrast, if we only use the imputation with the smallest variance while discarding the others, it may lead to efficiency losses due to not use of available information.

4.3 Regularization

To select the relevant mediators from high-dimensional variables and to estimate the time-varying effects, we adopt the regularization approach. Specifically, we propose to incorporate three penalty terms to: (i) identify mediators with large mediation effects at each time point, (ii) shrink similar effects at adjacent time points, and (iii) shrink effects of demographic variables. Consequently, the objective function of the proposed method is

$$L\left(\boldsymbol{\Theta}\right) = G\left(\boldsymbol{\theta}\right) + \sum_{j=1}^{p} H^{j}\left(\boldsymbol{\vartheta}^{j}\right) + P_{1}\left(\boldsymbol{\Theta}\right) + P_{2}\left(\boldsymbol{\Theta}\right) + P_{3}\left(\boldsymbol{\Theta}\right), \tag{8}$$

where

$$P_{1}\left(\boldsymbol{\Theta}\right) = \sum_{t=1}^{T} P_{\mathrm{SCAD},\lambda_1,a}\left(\left|\alpha_t\right|\right) + \lambda_1 \sum_{t=1}^{T}\sum_{j=1}^{p}\left\{1 - \frac{1}{\left(1+b\left|\beta_t^j\right|\right)\left(1+b\left|\gamma_t^j\right|\right)}\right\}, \tag{9}$$

is a penalty for direct effects and indirect effects, with $P'_{\mathrm{SCAD},\lambda_1,a}(\alpha) = \lambda_1\{\mathbb{I}(\alpha \leq \lambda_1) + \frac{(a\lambda_1-\alpha)_+}{(a-1)\lambda_1}\mathbb{I}(\alpha > \lambda_1)\}$ for some $\alpha > 0$, $a > 2$, and $b > 0$,

$$P_{2}\left(\boldsymbol{\Theta}\right) = \lambda_2 \sum_{t=2}^{T}\left|\alpha_t - \alpha_{t-1}\right| + \lambda_2 \sum_{t=2}^{T}\sum_{j=1}^{p}\left(\left|\beta_t^j - \beta_{t-1}^j\right| + \left|\gamma_t^j - \gamma_{t-1}^j\right|\right), \tag{10}$$

is a fusion penalty for time-varying effects, and

$$P_{3}\left(\boldsymbol{\Theta}\right) = \lambda_3\left(\left\|\boldsymbol{\delta}\right\|_1 + \sum_{j=1}^{p}\left\|\boldsymbol{\Gamma}^j\right\|_1\right), \tag{11}$$

is a penalty for demographic effects. The regularization parameters λ_1, λ_2, and λ_3 control the amount of shrinkage.

The second term in $P_1(\boldsymbol{\Theta})$ is a bi-level mediation penalty for indirect effects, which links models (1) and (2) by penalizing $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ jointly rather than separately, and achieves bi-level variable selection (Huang et al. 2009). This estimator not only obtains the sparse solution for β_t^j and γ_t^j when $\beta_t^j = 0$ and $\gamma_t^j = 0$, but also obtains the sparse solution for β_t^j or γ_t^j when $\beta_t^j = 0$ or $\gamma_t^j = 0$. On the other hand, it inherits the advantage of the smoothly clipped absolute deviation (SCAD, Fan and Li 2001) penalty where the function gradually

levels off as β_t^j and γ_t^j increase, which avoids over-shrinkage. In general, it is possible that not every mediator has different effects at different time points. We use the fused lasso penalty (Tibshirani et al. 2005) in $P_2(\boldsymbol{\Theta})$ to shrink similar effects at adjacent time points, which links $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ across all time points. $P_3(\boldsymbol{\Theta})$ is a lasso penalty (Friedman, Hastie and Tibshirani 2010) for selecting demographic effects to avoid over-fitting.

4.4 Principal components of weighting matrix

The GMM estimators in (6) and (7) may have poor finite sample performance in highly overidentified models, due to imprecise estimation of the weighting matrix $\boldsymbol{W}^{(r)}(\boldsymbol{\theta})$ or $\boldsymbol{W}^{j(r)}(\boldsymbol{\vartheta}^j)$, or under the singularity of the weighting matrix (Doran and Schmidt 2006). This could be attributed to the large number of estimating functions compared to the relatively small sample size, or due to the overlap of information between LMIs. For example, as illustrated in Figure 3, the observed values at time point 1 in group 2 are used for the estimation of both $\boldsymbol{M}_{u(2)}^{(\{1,2\})}$ and $\boldsymbol{M}_{u(2)}^{(\{1\})}$, and the observed values at time point 2 in group 2 are used for the estimation of both $\boldsymbol{M}_{u(2)}^{(\{1,2\})}$ and $\boldsymbol{M}_{u(2)}^{(\{2\})}$.

To address the singularity issue, we reduce the dimension of $\boldsymbol{g}^{(r)}$ and $\boldsymbol{h}^{j(r)}$ ($r = 1, ..., R$ and $j = 1, ..., p$), by combining informative estimating functions, for example, using the first several largest principal components. If $\boldsymbol{W}^{(r)} = \frac{1}{n_r}\sum_{i\in\mathscr{G}(r)} \boldsymbol{g}_i^{(r)}\boldsymbol{g}_i^{(r)T}$ is singular or close to singular, we extract the first $u^{(r)}$ principal components $\tilde{\boldsymbol{g}}^{(r)} = \boldsymbol{U}^{(r)}\boldsymbol{g}^{(r)}$ from $\boldsymbol{g}^{(r)}$, where $\boldsymbol{U}^{(r)}$ contains $u^{(r)}$ eigenvectors of $\boldsymbol{W}^{(r)}$ corresponding to the largest $u^{(r)}$ nonzero eigenvalues. Similarly, if $\boldsymbol{W}^{j(r)} = \frac{1}{n_r}\sum_{i\in\mathscr{G}(r)} \boldsymbol{h}_i^{j(r)}\boldsymbol{h}_i^{j(r)T}$ is singular or close to singular, we extract the first $u^{j(r)}$ principal components $\tilde{\boldsymbol{h}}^{j(r)} = \boldsymbol{U}^{j(r)}\boldsymbol{h}^{j(r)}$ from $\boldsymbol{h}^{j(r)}$, where $\boldsymbol{U}^{j(r)}$ contains $u^{j(r)}$ eigenvectors of $\boldsymbol{W}^{j(r)}$ corresponding to the largest $u^{j(r)}$ nonzero eigenvalues. The numbers of principal components $u^{(r)}$ and $u^{j(r)}$ ($r = 1, ..., R$ and $j = 1, ..., p$) can be selected based on the information criterion proposed by Cho and Qu (2015) to capture sufficient information from the estimating functions. Note that if $\boldsymbol{W}^{(r)}$ or $\boldsymbol{W}^{j(r)}$ is not singular, $\boldsymbol{U}^{(r)}$ or $\boldsymbol{U}^{j(r)}$ is an identity matrix. Consequently, the final objective function is

$$\tilde{L}\left(\boldsymbol{\Theta}\right) = \tilde{G}\left(\boldsymbol{\theta}\right) + \sum_{j=1}^{p} \tilde{H}^j\left(\boldsymbol{\vartheta}^j\right) + P_1\left(\boldsymbol{\Theta}\right) + P_2\left(\boldsymbol{\Theta}\right) + P_3\left(\boldsymbol{\Theta}\right), \tag{12}$$

where the quadratic form $\tilde{G}(\boldsymbol{\theta}) = \sum_{r=1}^{R}(\tilde{\boldsymbol{g}}^{(r)}(\boldsymbol{\theta}))^T(\tilde{\boldsymbol{W}}^{(r)}(\boldsymbol{\theta}))^{-1}\tilde{\boldsymbol{g}}^{(r)}(\boldsymbol{\theta})$ with transformed weighting matrix $\tilde{\boldsymbol{W}}^{(r)}(\boldsymbol{\theta}) = \frac{1}{n_r}\sum_{i\in\mathscr{H}(r)}\tilde{\boldsymbol{g}}^{(r)}(\boldsymbol{\theta})(\tilde{\boldsymbol{g}}^{(r)}(\boldsymbol{\theta}))^T$, and the quadratic form $\tilde{H}^j(\boldsymbol{\vartheta}^j) =$

$\sum_{r=1}^{R}(\tilde{\boldsymbol{h}}^{j(r)}(\boldsymbol{\vartheta}^j))^T(\tilde{\boldsymbol{W}}^{j(r)}(\boldsymbol{\vartheta}^j))^{-1}\tilde{\boldsymbol{h}}^{j(r)}(\boldsymbol{\vartheta}^j)$ with transformed weighting matrix $\tilde{\boldsymbol{W}}^{j(r)}(\boldsymbol{\vartheta}^j) = \frac{1}{n_r}\sum_{i\in\mathscr{H}(r)}\tilde{\boldsymbol{h}}^{j(r)}(\boldsymbol{\vartheta}^j)(\tilde{\boldsymbol{h}}^{j(r)}(\boldsymbol{\vartheta}^j))^T$.

5 Computation

In this section, we provide the computational details of the proposed method. We first describe the implementation of LMI from Section 4.1. Then we introduce an information criterion to select the numbers of principal components from Section 4.4. Next, we present a smoothing technique to handle the non-separable fusion penalty from Section 4.3. Afterwards, we introduce the overall optimization algorithm of the proposed method for a given set of regularization parameters $(\lambda_1, \lambda_2, \lambda_3)^T$ from Section 4.3. Finally, we propose an information criterion to select these regularization parameters.

When conducting LMI, the conditional expectation $\mathbb{E}(M_t^j|\boldsymbol{M}_t, \boldsymbol{D})$ is required to be estimated for each time set $\mathscr{t} \in \mathscr{T}(r,t)$, time point $t \in u(r)$, group $r = 1, ..., R$, and dimension $j = 1, ..., p$. In this paper, we adopt the lasso-regularized linear model (Friedman, Hastie and Tibshirani 2010) to account for high-dimensionality, where the regularization parameter is selected by cross-validation.

To capture most of the information from the estimating functions $\boldsymbol{g}^{(r)}$ and $\boldsymbol{h}^{j(r)}$ ($r = 1, ..., R$ and $j = 1, ..., p$), we select the numbers of principal components $u^{(r)}$ and $u^{j(r)}$, by minimizing the following Bayesian-type information criterion (Cho and Qu 2015):

$$\Psi(u) = \frac{\operatorname{tr}\left\{\boldsymbol{\Omega} - \tilde{\boldsymbol{\Omega}}(u)\right\}}{\operatorname{tr}\{\boldsymbol{\Omega}\}} + u\frac{\log(n_r d)}{n_r d}, \tag{13}$$

where $\operatorname{tr}\{\cdot\}$ is the trace of a matrix, $\boldsymbol{\Omega}$ represents the weighting matrix $\boldsymbol{W}^{(r)}$ or $\boldsymbol{W}^{j(r)}$ with dimension d, $\tilde{\boldsymbol{\Omega}}(u) = \sum_{k=1}^{u}\xi_k\boldsymbol{\nu}_k\boldsymbol{\nu}_k^T$ is an approximation of $\boldsymbol{\Omega}$ based on the largest u eigenvectors, and ξ_k is the k-th largest eigenvalue of $\boldsymbol{\Omega}$ corresponding to the eigenvector $\boldsymbol{\nu}_k$. Since $\operatorname{tr}\{\boldsymbol{\Omega} - \tilde{\boldsymbol{\Omega}}(u)\} = \sum_{k=u+1}^{d}\xi_k$, the minimizer of $\Psi(u)$ is indeed the number of eigenvalues greater than $\operatorname{tr}\{\boldsymbol{\Omega}\}\log(n_r d)/(n_r d)$.

Since the non-smooth and non-separable of fusion penalty $P_2(\boldsymbol{\Theta})$ in (10) makes the optimization challenging, we use the smoothing technique from Chen et al. (2012). Specifically, we reformulate the fused lasso as $P_2(\boldsymbol{\Theta}) = \lambda_2\|\mathbb{D}\boldsymbol{\Theta}\|_1 = \lambda_2 \max_{\|\boldsymbol{\eta}\|_\infty\le 1}(\boldsymbol{\eta}^T\mathbb{D}\boldsymbol{\Theta})$, where $\mathbb{D}$ is a difference operator corresponding to the differences in $P_2(\boldsymbol{\Theta})$. Let $\tilde{P}_2(\boldsymbol{\Theta};\rho) =$

$\lambda_2 \max_{\|\boldsymbol{\eta}\|_\infty \le 1} (\boldsymbol{\eta}^T \mathbb{D}\boldsymbol{\Theta} - \rho\|\boldsymbol{\eta}\|_2^2/2)$, where ρ is a positive smoothing parameter and $\tilde{P}_2(\boldsymbol{\Theta};\rho)$ approximates $P_2(\boldsymbol{\Theta})$ as $\rho \to 0$. Define a projection operator $S(x) = -\mathbb{I}(x < -1) + x\mathbb{I}(-1 \le x \le 1) + \mathbb{I}(x > 1)$, we can deduce that $\boldsymbol{\eta}^* = S(\mathbb{D}\boldsymbol{\Theta}/\rho)$ is indeed the optimal solution in $\tilde{P}_2(\boldsymbol{\Theta};\rho)$. Thus we have

$$\tilde{P}_2\left(\boldsymbol{\Theta},\rho\right) = \lambda_2\left(\boldsymbol{\eta}^{*T}\mathbb{D}\boldsymbol{\Theta} - \frac{\rho}{2}\left\|\boldsymbol{\eta}^*\right\|_2^2\right), \tag{14}$$

which is convex and differentiable in $\boldsymbol{\Theta}$. We set $\rho = 10^{-4}$ following Chen et al. (2012).

We propose an algorithm based on the iterated GMM with principal component analysis. Specifically, given $\boldsymbol{\Theta}^{(s-1)}$ from the $(s-1)$-th iteration, for $r = 1, ..., R$ and $j = 1, ..., p$, if the weighting matrix $\boldsymbol{W}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ or $\boldsymbol{W}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$ is singular or close to singular, we first select the number of principal components $u^{(r)(s-1)}$ and $u^{j(r)(s-1)}$ by minimizing the information criterion in (13), then construct principal component matrices $\boldsymbol{U}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ and $\boldsymbol{U}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$, next get the reduced moment conditions $\tilde{\boldsymbol{g}}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ and $\tilde{\boldsymbol{h}}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$, and finally obtain the transformed weighting matrices $\tilde{\boldsymbol{W}}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ and $\tilde{\boldsymbol{W}}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$.

At the s-th iteration, we solve

$$\boldsymbol{\Theta}^{(s)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \tilde{L}^{(s)}\left(\boldsymbol{\Theta}\right), \tag{15}$$

where $\tilde{L}^{(s)}(\boldsymbol{\Theta}) = \tilde{G}^{(s-1)}(\boldsymbol{\theta}) + \sum_{j=1}^{p} \tilde{H}^{j(s-1)}(\boldsymbol{\vartheta}^j) + P_1(\boldsymbol{\Theta}) + \tilde{P}_2(\boldsymbol{\Theta};\rho) + P_3(\boldsymbol{\Theta})$. The quadratic form $\tilde{G}^{(s-1)}(\boldsymbol{\theta}) = \sum_{r=1}^{R} (\tilde{\boldsymbol{g}}^{(r)(s-1)}(\boldsymbol{\theta}))^T (\tilde{\boldsymbol{W}}^{(r)}(\boldsymbol{\theta}^{(s-1)}))^{-1} \tilde{\boldsymbol{g}}^{(r)(s-1)}(\boldsymbol{\theta})$ with $\tilde{\boldsymbol{g}}^{(r)(s-1)}(\boldsymbol{\theta}) = \boldsymbol{U}^{(r)}(\boldsymbol{\theta}^{(s-1)})\boldsymbol{g}^{(r)}(\boldsymbol{\theta})$ uses principal component matrix $\boldsymbol{U}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ and transformed weighting matrix $\tilde{\boldsymbol{W}}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ obtained at the $(s-1)$-th iteration. The quadratic form $\tilde{H}^{j(s-1)}(\boldsymbol{\vartheta}^j) = \sum_{r=1}^{R} (\tilde{\boldsymbol{h}}^{j(r)(s-1)}(\boldsymbol{\vartheta}^j))^T (\tilde{\boldsymbol{W}}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)}))^{-1} \tilde{\boldsymbol{h}}^{j(r)(s-1)}(\boldsymbol{\vartheta}^j)$ uses principal component matrix $\boldsymbol{U}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$ and transformed weighting matrix $\tilde{\boldsymbol{W}}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$ obtained at the $(s-1)$-th iteration. The above calculation process can be summarized in Algorithm 1, and an expanded version of this algorithm is shown in the Supplementary Material.

The turning parameter a in (9) is set to be 3.7 following Fan and Li (2001) and b in (9) is set to be 10 following Xue et al. (2022). To determine the regularization parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$, we propose the following Bayesian-type information criterion (Xue and Qu 2021, Xue et al. 2022):

$$\Phi\left(\lambda_1,\lambda_2,\lambda_3\right) = \log\left\{\frac{\mathrm{RSS}_Y\left(\hat{\boldsymbol{\Theta}}_{\boldsymbol{\lambda}}\right)}{NT}\right\} + \log\left\{\sum_{j=1}^{p}\frac{\mathrm{RSS}_{M^j}\left(\hat{\boldsymbol{\Theta}}_{\boldsymbol{\lambda}}\right)}{NT}\right\} + \frac{\log\left(NT\right)}{NT}\mathrm{df}_{\boldsymbol{\lambda}}, \tag{16}$$

Algorithm 1

1: LMI. Conduct longitudinal multiple imputation.

2: Initialization. Obtain initial values $\boldsymbol{\Theta}^{(0)}$ based on the averaged imputed data. Set the regularization parameters λ_1, λ_2, and λ_3.

3: At the s-th iteration, given $\boldsymbol{\Theta}^{(s-1)}$ from the $(s-1)$-th iteration:

(a) Principal component analysis. For $r = 1, ..., R$ and $j = 1, ..., p$, select the number of principal components using (13) if $\boldsymbol{W}^{(r)}(\boldsymbol{\theta}^{(s-1)})$ or $\boldsymbol{W}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$ is singular or close to singular, and obtain $\boldsymbol{U}^{(r)}(\boldsymbol{\theta}^{(s-1)})$, $\boldsymbol{U}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$, $\tilde{\boldsymbol{W}}^{(r)}(\boldsymbol{\theta}^{(s-1)})$, and $\tilde{\boldsymbol{W}}^{j(r)}(\boldsymbol{\vartheta}^{j(s-1)})$.

(b) Optimization. Solve (15) through the gradient descent to obtain $\boldsymbol{\Theta}^{(s)}$.

4: Iterate Step 3 until $\|\boldsymbol{\Theta}^{(s)} - \boldsymbol{\Theta}^{(s-1)}\|_1 < 10^{-6}$. Return $\boldsymbol{\Theta}^{(s)}$.

where $\mathrm{RSS}_Y(\hat{\boldsymbol{\Theta}}_{\boldsymbol{\lambda}}) = \sum_{r=1}^{R} \sum_{i \in \mathscr{H}(r)} \sum_{t=1}^{T} \frac{1}{|\mathscr{T}(r,t)|} \sum_{t \in \mathscr{T}(r,t)} (Y_{it} - \hat{\alpha}_t X_{it} - \hat{\boldsymbol{\beta}}_t^T \boldsymbol{M}_{it}^{(t)} - \hat{\boldsymbol{\delta}}^T \boldsymbol{Z}_i)^2$ and $\mathrm{RSS}_{M^j}(\hat{\boldsymbol{\Theta}}_{\boldsymbol{\lambda}}) = \sum_{r=1}^{R} \sum_{i \in \mathscr{H}(r)} \sum_{t=1}^{T} \frac{1}{|\mathscr{T}(r,t)|} \sum_{t \in \mathscr{T}(r,t)} (M_{it}^{j(t)} - \hat{\gamma}_t^j X_{it} - \hat{\boldsymbol{\Gamma}}^{jT} \boldsymbol{Z}_i)^2$ are the residual sum of squares based on imputed data corresponding to models (1) and (2). Degrees of freedom $\mathrm{df}_{\boldsymbol{\lambda}} = \mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}) + \mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}) + \mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\gamma}}) + \mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\Gamma}})$, where $\mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\Gamma}})$ is the number of nonzero estimated values in $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Gamma}}$, and

$$\mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}) = T - \sum_{t=1}^{T} \mathbb{I}(\hat{\alpha}_t = 0) - \sum_{t=2}^{T} \mathbb{I}(\hat{\alpha}_t = \hat{\alpha}_{t-1} \text{ and } \hat{\alpha}_t \hat{\alpha}_{t-1} \neq 0),$$

$$\mathrm{df}_{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\beta}}\right) = \sum_{j=1}^{p} \left\{ T - \sum_{t=1}^{T} \mathbb{I}\left(\hat{\beta}_t^j = 0\right) - \sum_{t=2}^{T} \mathbb{I}\left(\hat{\beta}_t^j = \hat{\beta}_{t-1}^j \text{ and } \hat{\beta}_t^j \hat{\beta}_{t-1}^j \neq 0\right) \right\},$$

$$\mathrm{df}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\gamma}}) = \sum_{j=1}^{p} \left\{ T - \sum_{t=1}^{T} \mathbb{I}\left(\hat{\gamma}_t^j = 0\right) - \sum_{t=2}^{T} \mathbb{I}\left(\hat{\gamma}_t^j = \hat{\gamma}_{t-1}^j \text{ and } \hat{\gamma}_t^j \hat{\gamma}_{t-1}^j \neq 0\right) \right\},$$

are numbers of nonzero and time-varying estimated values in $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, ..., \hat{\alpha}_T)^T$, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, ..., \hat{\boldsymbol{\beta}}_T^T)^T$, and $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1^T, ..., \hat{\boldsymbol{\gamma}}_T^T)^T$, respectively.

Remark 3. There are three regularization parameters $(\lambda_1, \lambda_2, \lambda_3)$ that need to be selected during computation, and three-dimensional searches may be computationally costly. In practical applications, we can adjust for confounders $\boldsymbol{Z}_i$ in models (1) and (2) during the data preprocessing stage (Schuurmans et al. 2024), and then use the residuals for subsequent analyses. This approach removes the need to include parameters $\boldsymbol{\delta}$ and $\boldsymbol{\Gamma}$ in the model, thereby eliminating the necessity of selecting λ_3. For the remaining parameters (λ_1, λ_2), we perform a two-dimensional grid search to identify the combinations which minimize the criterion (16).

6 Simulations

In this section, we conduct comprehensive simulation studies to compare the performance of proposed method with some existing approaches. We consider the following time-varying structural equation models:

$$Y_{it} = \alpha_t X_{it} + \boldsymbol{\beta}_t^T \boldsymbol{M}_{it} + e_{it}, \tag{17}$$

$$\boldsymbol{M}_{it} = \boldsymbol{\gamma}_t X_{it} + \boldsymbol{\eta}_t Y_{i,t-1} + \boldsymbol{\varepsilon}_{it}, \tag{18}$$

with $q = 0$ and $T = 4$. We generate the exposure $X_{i1} \sim \text{Uniform}(0,1)$ and $X_{it} = X_{i,t-1} + \text{Uniform}(0,1)$ for $t = 2, ..., T$. The random error vector $(e_{i1}, ..., e_{iT})^T$ is generated from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\rho_1))$, where $\boldsymbol{\Sigma}(\rho_1)$ is the exchangeable correlation structure with parameter ρ_1. The random error vector $(\boldsymbol{\varepsilon}_{i1}^T ..., \boldsymbol{\varepsilon}_{iT}^T)^T = (\varepsilon_{i1}^1, ..., \varepsilon_{i1}^p, ..., \varepsilon_{iT}^1, ..., \varepsilon_{iT}^p)^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\rho_2))$, where $\boldsymbol{\Sigma}(\rho_2)$ is the (block-wise) compound symmetry correlation structure with parameter ρ_2. We set $\rho_1 = 0.1$, $\rho_2 = 0.5$ and $\boldsymbol{\eta}_t = \mathbf{0}$.

We consider the following settings with different sample sizes, different dimensions of mediators, different non-monotone missingness structures, and different missingness mechanisms:

Setting I. We consider the missingness structure (A) illustrated in Figure 4. We set $N = 350$, $p = 50$, and the sample sizes in group 1-4 are $(50, 100, 100, 100)$. The true parameters $\boldsymbol{\alpha} = (0.5, 0.5, 1, 1)^T$, β_t^j and γ_t^j for $j = 1, ..., 14$ and $t = 1, ..., 4$ are shown in Figure 5. The β_t^j and γ_t^j are both zero for $j = 15, ..., 50$ and all t.

Setting II. We consider the missingness structure (B) illustrated in Figure 4, where there is no complete case group. We set $N = 375$, $p = 50$, and the sample sizes in group 1-6 are $(75, 75, 75, 50, 50, 50)$. The true parameters $\boldsymbol{\alpha} = (0.5, 0.5, 1, 1)^T$, β_t^j and γ_t^j for $j = 1, ..., 14$ and $t = 1, ..., 4$ are shown in Figure 5. The β_t^j and γ_t^j are both zero for $j = 15, ..., 50$ and all t.

Setting III. We consider the missingness structure (A) illustrated in Figure 4. We set $N = 550$, $p = 200$, and the sample sizes in group 1-4 are $(100, 150, 150, 150)$. The true parameters $\boldsymbol{\alpha} = (1, 1, 1, 1)^T$, β_t^j and γ_t^j for $j = 1, ..., 22$ and $t = 1, ..., 4$ are shown in Figure 6. The β_t^j and γ_t^j are both zero for $j = 23, ..., 200$ and all t.

Setting IV. We consider the missingness structure (B) illustrated in Figure 4, where there is no complete case group. We set $N = 675$, $p = 200$, and the sample sizes in group 1-6 are $(125, 125, 125, 100, 100, 100)$. The true parameters $\boldsymbol{\alpha} = (1, 1, 1, 1)^T$, β_t^j and γ_t^j for

$j = 1, ..., 22$ and $t = 1, ..., 4$ are shown in Figure 6. The β_t^j and γ_t^j are both zero for $j = 23, ..., 200$ and all t.

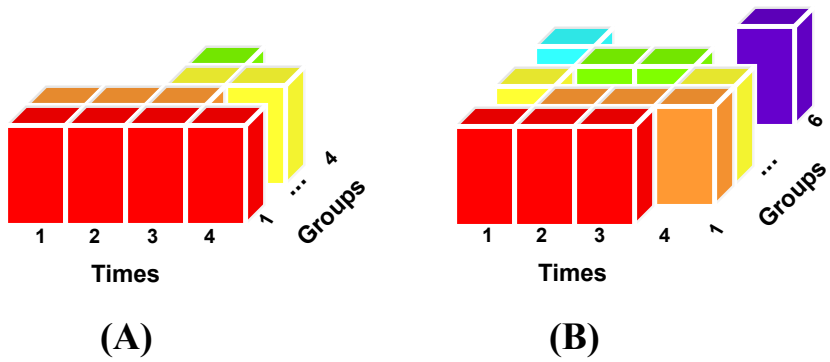

Figure 4: Missingness structures in simulations. Each blank column represents unobserved mediators, while the colored ones represent observed mediators.

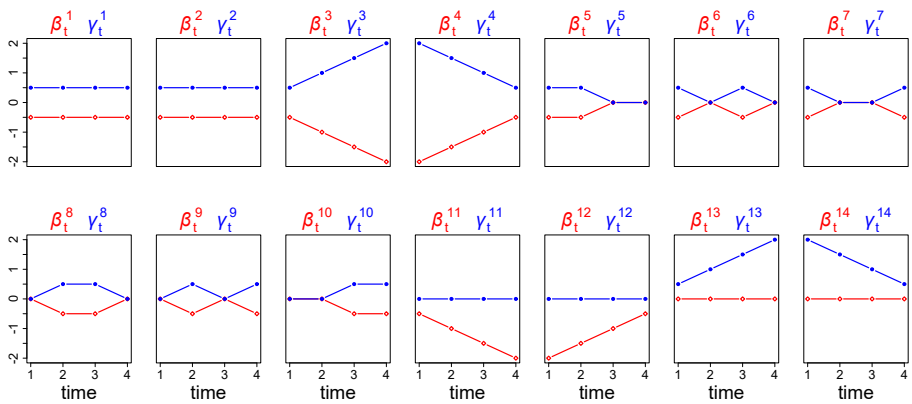

Figure 5: True parameters of β_t^j (red colored) and γ_t^j (blue colored) for $j = 1, ..., 14$ and $t = 1, ..., 4$ in simulation settings I and II. The β_t^j and γ_t^j are both zero for $j = 15, ..., 50$ and all t.

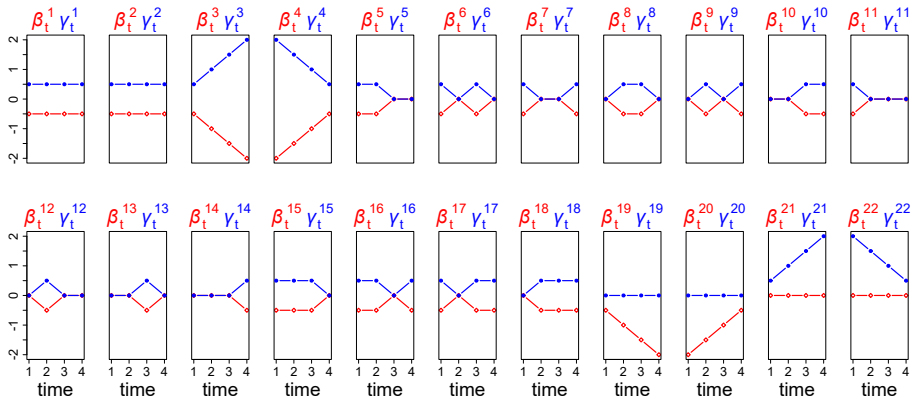

Figure 6: True parameters of β_t^j (red colored) and γ_t^j (blue colored) for $j = 1, ..., 22$ and $t = 1, ..., 4$ in simulation settings III and IV. The β_t^j and γ_t^j are both zero for $j = 23, ..., 200$ and all t.

In each of the settings I-IV, we consider three different missingness mechanisms (Little and Rubin 2019): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In the MCAR scenario, all subjects are completely randomly assigned to the groups. In the MAR scenario, subjects $i = 1, ..., N$ are sequentially randomly assigned to group 1 with probability proportional to $\sum_{t=1}^{T} |X_{it}|$. The remaining subjects $i \notin \mathscr{H}(1)$ are sequentially randomly assigned to group 2 with probability proportional to $\sum_{t=1}^{T} |X_{it}|$, and so on until all subjects are grouped. In the MNAR scenario, the grouping process is the same as that in the MAR scenario, except that the probability is proportional to $\sum_{j=1}^{p} \sum_{t=1}^{T} |M_{it}^j|$.

Our method, using varying-coefficient (VC) models (17) and (18) with LMI-data, is compared with other mediator selection approaches, including: (i) Mix: mixed-effects models with time interactions and lasso penalty (Rijnhart et al. 2022); (ii) Path: linear structural equation models with pathway lasso penalty (Zhao and Luo 2022); (iii) HIMA: linear structural equation models with de-biased lasso and false discovery rate control (Perera et al. 2022); (iv) Bayes: Bayesian sparse models with continuous shrinkage (Song et al. 2020). Note that for VC and Mix, the parameters are estimated using data from all time points simultaneously, while for Path, HIMA, and Bayes, the parameters are estimated using data from each time point separately. We also consider different approaches for handling missing data, including: (i) CC: complete-case analysis which retains only subjects in group 1 with fully observed repeated measurements; (ii) MC: matrix completion via iterative soft-thresholded singular value decomposition (Mazumder, Hastie and Tibshirani 2010); (iii)

SI: single imputation based on trajectory means (Jahangiri et al. 2023).

In each replication, to evaluate the accuracy of the mediator selection for each method, we calculate the false negative rate (FNR) $\frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{j=1}^{p}\mathbb{I}(\beta_t^j\gamma_t^j\neq 0,\hat{\beta}_t^j\hat{\gamma}_t^j=0)}{\sum_{j=1}^{p}\mathbb{I}(\beta_t^j\gamma_t^j\neq 0)}$ and the false positive rate (FPR) $\frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{j=1}^{p}\mathbb{I}(\beta_t^j\gamma_t^j=0,\hat{\beta}_t^j\hat{\gamma}_t^j\neq 0)}{\sum_{j=1}^{p}\mathbb{I}(\beta_t^j\gamma_t^j=0)}$, across all time points, where $\hat{\beta}_t^j$ and $\hat{\gamma}_t^j$ are estimated values of β_t^j and γ_t^j for $t = 1, ..., T$ and $j = 1, ..., p$. We also calculate the mean squared error (MSE) $\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}(\beta_t^j\gamma_t^j - \hat{\beta}_t^j\hat{\gamma}_t^j)^2$ to evaluate the precision of the mediation effect estimation for each method. We say that a method has better performance if it has a smaller FNR+FPR and a smaller MSE.

Tables 1 and 2 summarize the averaged estimation results across 50 replications of different methods for different missingness mechanisms under settings I and II. Tables 1 shows that the proposed VC-LMI method has the smallest FNR+FPR and MSE among all missingness mechanisms, which suggests that the proposed VC-LMI method has the highest accuracy of mediator selection and precision of mediation effect estimation, and its performance is robust to different missingness mechanisms. The VC-CC has a relatively large FNR+FPR, due to the fact that only 50/350 of the subjects with complete repeated measurements are used for analysis. The VC-MC has large FNR+FPR and MSE, possibly because the MC assumes the low-rank structure and noisy entries of the variable matrix, which may not be suitable for our longitudinal data. The VC-SI also has large FNR+FPR and MSE, possibly because the SI utilizes the trajectory mean for imputation, which erases some of the longitudinal changes in variables. Our VC-LMI method makes full use of the correlation among repeated measurements for imputation, and effectively integrates the multiple imputations, which therefore exhibits powerful performance compared with candidate approaches.

For other mediator selection methods in Tables 1, both FNR+FPR and MSE are large. Specifically, the Mix shows large FPR, while the Path provides large FNR. The reason for their poor estimations may be because the mediator selection methods with some shortcomings (too many interactions in Mix and over-shrinkage in Path) are used on the datasets with poor quality. The HIMA and Bayes provide a smaller FPR but a larger FNR, possibly due to the multiple-testing procedure for HIMA and the posterior inclusion probability for Bayes being overly conservative.

Table 1: False negative rate (FNR), false positive rate (FPR), FNR+FPR, and mean squared error (MSE) of different methods for different missingness mechanisms under setting I.

	missing completely at random				missing at random				missing not at random			
	FNR	FPR	FNR+FPR	MSE	FNR	FPR	FNR+FPR	MSE	FNR	FPR	FNR+FPR	MSE
VC-LMI	0.059	0.086	**0.145**	**0.719**	0.049	0.090	**0.139**	**0.774**	0.064	0.088	**0.152**	**0.715**
VC-CC	0.174	0.054	0.228	1.032	0.180	0.057	0.237	1.018	0.150	0.058	0.208	0.860
VC-MC	0.419	0.072	0.491	8.332	0.333	0.074	0.407	8.590	0.401	0.071	0.471	8.176
VC-SI	0.203	0.036	0.239	5.065	0.180	0.034	0.214	5.228	0.196	0.036	0.232	5.208
Mix-CC	0.111	0.105	0.215	2.638	0.109	0.114	0.222	2.525	0.116	0.099	0.215	2.319
Mix-MC	0.086	0.327	0.414	8.815	0.077	0.320	0.397	9.214	0.092	0.314	0.407	8.772
Mix-SI	0.057	0.167	0.224	5.839	0.043	0.163	0.205	6.016	0.049	0.157	0.206	5.965
Path-CC	0.272	0.065	0.337	7.215	0.288	0.064	0.351	7.125	0.258	0.065	0.323	7.082
Path-MC	0.560	0.032	0.592	10.420	0.470	0.028	0.498	10.364	0.541	0.033	0.574	10.382
Path-SI	0.248	0.020	0.268	9.198	0.227	0.022	0.250	9.319	0.249	0.021	0.270	9.285
HIMA-CC	0.679	0.005	0.685	1.715	0.702	0.004	0.707	2.198	0.669	0.004	0.673	1.708
HIMA-MC	0.574	0.035	0.609	8.742	0.489	0.035	0.524	9.173	0.547	0.040	0.587	8.579
HIMA-SI	0.376	0.014	0.390	4.895	0.361	0.015	0.377	5.066	0.370	0.015	0.385	5.158
Bayes-CC	0.916	0.000	0.916	5.577	0.916	0.000	0.916	5.508	0.901	0.000	0.901	4.797
Bayes-MC	0.786	0.019	0.805	10.043	0.754	0.021	0.775	10.840	0.760	0.020	0.780	9.711
Bayes-SI	0.689	0.001	0.690	6.129	0.675	0.001	0.676	6.129	0.659	0.001	0.660	6.115

Note: Each estimator is denoted as "mediator selection method - missing data handling method". Results are averaged across all time points and 50 replications.

Table 2 summarizes the estimation results under setting II corresponding to the missingness structure (B) in Figure 4. Since no subject has complete repeated measurements, the CC is not applicable. Our proposed VC-LMI method has the smallest FNR+FPR and MSE among all candidate approaches. The estimation results under settings III and IV in the high-dimensional context are provided in Tables S1 and S2 in the Supplementary Material, which show that our proposed VC-LMI method also beats other candidate approaches, based on its smallest FNR+FPR and MSE. The Supplementary Material includes additional simulations, where Tables S3 and S4 demonstrate that our method continues to outperform other candidate approaches in higher-dimensional settings ($p = 2000$) and scenarios with more time points ($T = 5$), respectively. Table S5 shows that combining multiple imputations through the GMM outperforms approach that selecting imputations with the smallest variance.

Table 2: False negative rate (FNR), false positive rate (FPR), FNR+FPR, and mean squared error (MSE) of different methods for different missingness mechanisms under setting II.

	missing completely at random				missing at random				missing not at random			
	FNR	FPR	FNR+FPR	MSE	FNR	FPR	FNR+FPR	MSE	FNR	FPR	FNR+FPR	MSE
VC-LMI	0.046	0.105	**0.151**	**0.790**	0.048	0.113	**0.160**	**0.855**	0.056	0.107	**0.163**	**0.755**
VC-MC	0.439	0.076	0.514	9.195	0.369	0.070	0.439	9.082	0.426	0.072	0.498	8.962
VC-SI	0.254	0.043	0.297	5.426	0.276	0.043	0.319	5.477	0.246	0.043	0.290	5.374
Mix-MC	0.105	0.330	0.435	10.002	0.091	0.291	0.382	10.131	0.114	0.297	0.411	9.880
Mix-SI	0.057	0.140	0.198	6.707	0.060	0.143	0.203	6.606	0.056	0.151	0.207	6.581
Path-MC	0.646	0.022	0.668	10.908	0.559	0.017	0.577	10.770	0.611	0.018	0.630	10.851
Path-SI	0.286	0.019	0.305	9.609	0.276	0.016	0.292	9.534	0.264	0.019	0.283	9.544
HIMA-MC	0.654	0.045	0.698	9.667	0.571	0.046	0.618	9.837	0.631	0.048	0.680	9.564
HIMA-SI	0.457	0.021	0.478	5.321	0.461	0.018	0.479	5.287	0.429	0.021	0.450	5.360
Bayes-MC	0.915	0.028	0.943	10.791	0.876	0.031	0.907	11.405	0.902	0.030	0.933	10.800
Bayes-SI	0.756	0.008	0.764	5.960	0.774	0.005	0.780	6.075	0.764	0.006	0.770	6.010

Note: Each estimator is denoted as "mediator selection method - missing data handling method". Results are averaged across all time points and 50 replications.

7 DNHS data analysis

In this section, we investigate how DNAm dynamically mediates the effect of traumatic experiences on development of PTSD, utilizing the DNHS data. Specifically, we treat the number of traumatic events as exposure X_{it}, the logarithm transformed PTS symptom severity score as outcome Y_{it}, and the M-value transformed (Kruppa et al. 2021) DNAm CpG sites as potential mediators $\boldsymbol{M}_{it}$. The study includes 526 subjects in waves 1, 2, 4, and 5, which are referred to as time points 1-4. In our data, exposure X_{it} and outcome Y_{it} are always observed, but the mediators $\boldsymbol{M}_{it}$ suffer from non-monotone missingness as shown in Figure 1: $\boldsymbol{M}_i$ of some subjects may be available at certain visits, but missing at the next time point, and measured again at later visits.

There are two steps in the data preprocessing phase. First, we adjust some demographic information and blood work measured at baseline. For the outcome corresponding to model (1), we adjust age, gender, race, education, income, cigarette use, depression, and social cohesion score (Johns et al. 2012, Wani et al. 2021). For the mediators corresponding to model (2), we adjust age, gender, cigarette use, CD4 T cells, CD8 T cells, natural killer cells, B cells, and monocytes (Occean et al. 2022). The residuals after adjustment and the

models (17) and (18) are used for subsequent analyses. Second, given the limited number of subjects and the ultra-high dimensional potential mediators (1879 DNAm CpG sites), we carry out a screening process to reduce the dimension to a moderate scale below the sample size (Fan and Lv 2008). Specifically, for $t = 1, ..., T$ and $j = 1, ..., p$, we consider a series of marginal models $Y_{it} = \alpha_t X_{it} + \beta_t^j M_{it}^j + e_{it}$ and $M_{it}^j = \gamma_t^j X_{it} + \eta_t^j Y_{i,t-1} + \varepsilon_{it}^j$. Along the lines of the sure independence mediator screening (Perera et al. 2022), for each time point $t = 1, ..., T$, we use the subjects with observed mediators to fit the marginal models, then obtain a subset $\mathscr{D}_t = \{j\colon M^j \text{ is among the top } [n_t/\log(n_t)] \text{ largest } \beta_t^j \gamma_t^j \text{ effect}\}$, where n_t is the sample size at time point t and $[\cdot]$ denotes the integer part. Then we take $\bigcup_{t=1,...,T} \mathscr{D}_t$ as the index set of potential mediators after screening. There are 153 DNAm CpG sites remaining for the next step of analysis.

We use the proposed VC-LMI method to fit the preprocessed data. The estimated direct effect is $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4) = (0.50, 0.54, 0.54, 0.54)$, which means that trauma exposure has a sustained direct effect on PTSD. Figures 7 and 8 show the 13 selected CpG sites, each of which has nonzero $\hat{\beta}_t^j \hat{\gamma}_t^j$ at least at one time point. We can find that certain CpG sites (e.g., cg22564046) initially have nonzero mediation effects, but the effects disappear over time. Certain CpG sites (e.g., cg17318247) do not show mediation effects in the early stages, but the effects emerge over time. This indicates that the effects at the CpG level may vary in different stages of the disease process.

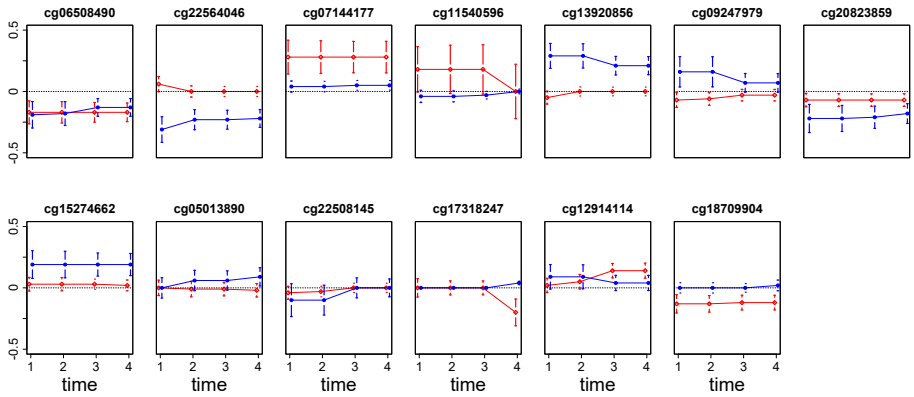

Figure 7: Estimated values $\hat{\beta}_t^j$ (red colored) and $\hat{\gamma}_t^j$ (blue colored) with Bootstrap 95% confidence intervals for the selected CpG sites and $t = 1, ..., 4$, based on the proposed VC-LMI method.

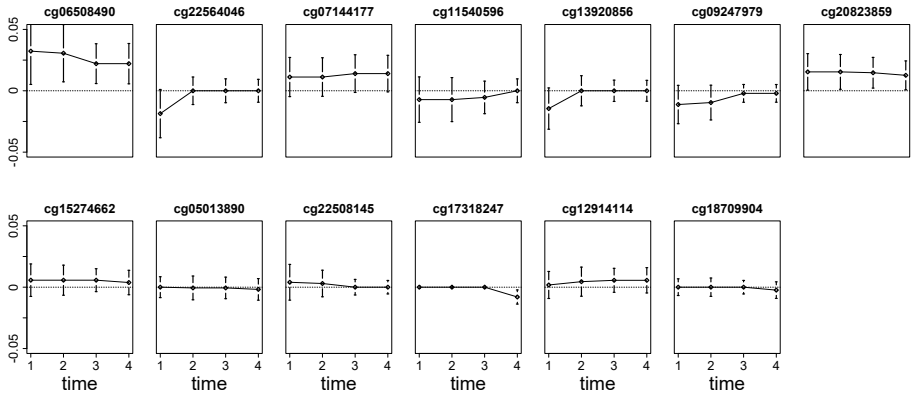

Figure 8: Estimated mediation effects $\hat{\beta}_t^j \hat{\gamma}_t^j$ with Bootstrap 95% confidence intervals for the selected CpG sites and $t = 1, ..., 4$, based on the proposed VC-LMI method.

As shown in Figure 8, the estimated mediation effects for certain CpG sites at specific time points, such as cg22508145 at $t = 1, 2$ and cg09247979 at $t = 3, 4$, appear small or close to zero. These seemingly negligible effects should be interpreted within the broader context of longitudinal mediation analysis and the cumulative role of DNAm across multiple time points. First, mediation effects can vary over time due to the dynamic nature of DNAm and its interactions with environmental exposures and disease progression. Small mediation effects at individual time points do not necessarily indicate that the CpG site lacks relevance as a mediator. Instead, they may reflect temporal fluctuations or periods during which the indirect effect is less pronounced. Second, while mediation effects may appear negligible at specific time points, the cumulative contribution of a CpG site across all time points could still be meaningful. Longitudinal analyses are designed to capture these aggregate patterns, which may not be evident when focusing solely on isolated single-time-point effects. Finally, even small mediation effects can hold biological significance, especially in high-dimensional settings where individual mediators contribute subtly but collectively to disease mechanisms (Xue et al. 2022).

Table 3 lists the gene information corresponding to the selected CpG sites. Eleven of these sites are annotated to human genes according to the Illumina annotation files (*CAT*, *PFKP*, *EEF1E1*, *TANC1*, *PTPRK*, *OVGP1*, *RPS6KA2*, *CPAMD8*, *CDK16*, *FAM120B*, and *C14orf182*). Of these eleven genes, *PTPRK* has been shown to be associated with mental health disorders such as PTSD (Chitrala, Nagarkatti and Nagarkatti 2016), *RPS6KA2* has been reported as significant KEGG pathways for PTSD in epigenome-wide DNA methylation study (Kuan et al. 2017). Other CpG sites while not located near any protein-coding

genes, fall within ENCODE candidate Cis-Regulatory Elements that are adjacent to highly conserved non-coding genomic regions (Moore et al. 2020).

Table 3: Gene information corresponding to the selected CpG sites.

CpG	chromosome	Gene
cg06508490	11	CAT
cg22564046	10	PFKP
cg07144177	6	EEF1E1
cg11540596	6	NA/Desert
cg13920856	2	TANC1
cg09247979	6	PTPRK
cg20823859	1	OVGP1
cg15274662	7	NA/Desert
cg05013890	6	RPS6KA2
cg22508145	19	CPAMD8
cg17318247	X	CDK16
cg12914114	6	FAM120B
cg18709904	14	C14orf182

Beyond the time-invariant associations, our results further show that the effects of genes change over time, which may reveal how the biological embedding of trauma exposure through DNAm contributes to PTSD risk over development. Our findings could have important clinical implications for time-sensitive risk prediction in trauma-exposed populations.

To compare the performance of the proposed VC-LMI method with other approaches, we randomly split the preprocessed data into a training set (80% of subjects) and a testing set (20% of subjects; the index set is denoted as $\mathfrak{T}$) for 50 replications. In each replication, for each method, we train the model on the training set and obtain the estimated values of parameters. Then we calculate the mean number of selected mediators (No.-mediator) as $\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}\mathbb{I}(\hat{\beta}_t^j\hat{\gamma}_t^j \neq 0)$, the prediction mean squared error (PMSE-β) for the outcome model (17) as $\frac{1}{|\mathfrak{T}|}\sum_{i\in\mathfrak{T}}\sum_{r=1}^{R}\mathbb{I}(i\in\mathscr{H}(r))\frac{1}{|o(r)|}\sum_{t\in o(r)}(Y_{it}-\hat{Y}_{it})^2$, and the prediction mean squared error (PMSE-γ) for the mediator model (18) as $\frac{1}{|\mathfrak{T}|p}\sum_{i\in\mathfrak{T}}\sum_{r=1}^{R}\mathbb{I}(i\in\mathscr{H}(r))\frac{1}{|o(r)|}\sum_{t\in o(r)}(\boldsymbol{M}_{it}-\hat{\boldsymbol{M}}_{it})^T(\boldsymbol{M}_{it}-\hat{\boldsymbol{M}}_{it})$, on the testing set across

all time points, where $\hat{Y}_{it}$ and $\hat{\boldsymbol{M}}_{it}$ are the fitted values of Y_{it} and $\boldsymbol{M}_{it}$ for $i \in \mathfrak{T}$ and $t \in o(r)$.

Table 4 summarizes the averaged estimation results across 50 replications. We observe that our VC-LMI method has the smallest PMSE both in the outcome model and the mediator model, indicating that the proposed method is more accurate in terms of prediction. For the Mix and the Path, despite more mediators being selected based on the imputed dataset, the PMSE is larger, possibly due to overfitting. For the HIMA and the Bayesian approaches, no mediator passes their overly conservative significance tests, which is consistent with the simulation results in Section 6.

Table 4: Mean number of selected mediators (No.-mediator), prediction mean squared error (PMSE-β) for the outcome model (17), and prediction mean squared error (PMSE-γ) for the mediator model (18) in DNHS data analysis.

	No.-mediator	PMSE-β	PMSE-γ
VC-LMI	10.350	**0.087**	**1.240**
VC-CC	8.900	0.286	1.754
VC-MC	14.500	0.097	1.249
VC-SI	22.250	0.104	1.254
Mix-CC	2.400	0.133	1.396
Mix-MC	6.350	0.101	1.249
Mix-SI	32.200	0.097	1.251
Path-CC	12.300	0.125	1.385
Path-MC	43.900	0.077	1.248
Path-SI	45.700	0.092	1.251
HIMA-CC	0.000	0.129	1.250
HIMA-MC	0.150	0.111	1.250
HIMA-SI	0.000	0.113	1.250
Bayes-CC	0.000	0.110	1.250
Bayes-MC	0.000	0.113	1.250
Bayes-SI	0.000	0.113	1.251

Note: Each estimator is denoted as "mediator selection method - missing data handling method". Results are averaged across all time points and 50 replications.

8 Discussion

In this paper, to investigate the time-varying mediation effects of DNAm on the relationship between trauma and PTSD, we propose time-varying structural equation models. The LMI approach is proposed to handle non-monotone missing DNAm and the regularization approach is incorporated to select relevant mediators and capture time-varying effects. In simulations, the proposed method compares favorably against existing competitors in various scenarios. In DNHS data analysis, we successfully identify potential DNAm CpG sites which show dynamic mediation effects.

We assume the mediators $(M_{it}^1, ..., M_{it}^p)$ are correlated but we do not specify causal ordering. The main reason for this assumption is that, based on the existing literature, external adverse environmental exposures or disease conditions are likely to influence DNAm across multiple CpG sites (Turecki and Meaney 2016), whereas the relationships between DNAm levels across CpG sites are primarily correlational (Mou et al. 2022). Such assumptions of parallel relationships among mediators are commonly made in high-dimensional DNAm-related mediation analysis (e.g., Song et al. 2020, Perera et al. 2022, Xue et al. 2022). Furthermore, even if causal ordering existed among high-dimensional DNAm CpG sites, establishing the correct causal order is infeasible in complex biological systems, instead they are characterized by intricate interactions and feedback loops (Tai et al. 2022). In our study, we also lack sufficient genetic knowledge to reliably support causal ordering. A more refined understanding of the causal relationships among high-dimensional mediators may require the application of directed acyclic graph learning techniques (Shi and Li 2022), which might be able to identify the complex causal structures among variables. However, such analyses extend beyond the scope of this study and could be explored in future research.

Unmeasured confounding is an inherent challenge in population-based studies. In our DNHS data analysis, we incorporate a comprehensive set of demographic information and blood work measured at baseline to control for confounding bias as much as possible. Nonetheless, some unmeasured confounders, such as participants' other medical treatments or psychological interventions for PTSD not captured in the DNHS investigation, may still exist. To address these remaining unmeasured confounders, methods such as instrumental variable approaches (Chen et al. 2023), proxy causal learning frameworks (Dukes et al. 2023), or de-confounding techniques (Yuan and Qu 2024) could be employed. Exploring

these methodologies could be a promising direction for future research.

In our imputation step, we estimate $\mathbb{E}(M_t^j|\boldsymbol{M}_{\mathcal{t}}, \boldsymbol{D})$ using specific groups that include observed M_t^j and $\boldsymbol{M}_{\mathcal{t}}$, then impute the unobserved M_t^j in other groups that contain observed $\boldsymbol{M}_{\mathcal{t}}$ but missing M_t^j. The rationale for our imputation method relies on the premise that the conditional distributions of mediators in the observed data are the same as those in the missing data, which holds only under the assumptions of MCAR or MAR. For mediators that are MNAR, the direct and indirect effects corresponding to the coefficients in models (1) and (2) are not identifiable without additional assumptions. Potential strategies include modeling the missing indicator $f_{\boldsymbol{R}_i}(\boldsymbol{r}_i|\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{M}_i)$ based on the completeness assumption (Zuo et al. 2024), or introducing instrumental or shadow variables to impose specific structures on the missingness mechanism (Shan et al. 2024). Extending our method to accommodate MNAR mediators using these strategies, as well as further addressing missing outcomes (Qin et al. 2019), are important directions for future research.

For the statistical inference of the mediation effect, there are two main challenges including the limiting distributions of the estimators for parameters in models (1) and (2), and the composite p-value of the mediation effect. Specifically, the penalty functions in (9) and (10) may present challenges in deriving the asymptotic distribution. Some existing studies used the bi-level penalty similar to that of $P_1(\boldsymbol{\Theta})$ in (9) and the corresponding theoretical results show that the estimator of nonzero coefficients in general converges to nonnormal distribution (Huang et al. 2009). Our estimator includes an additional fusion penalty $P_2(\boldsymbol{\Theta})$ in (10) to serve the purpose of shrinking similar effects at adjacent time points, which may lead to more complex limiting distributions. On the other hand, to test whether M^j mediates the effect of X on Y, the null and alternative hypotheses can be formulated as $H_0^j : \beta^j\gamma^j = 0$ versus $H_1^j : \beta^j\gamma^j \neq 0$. The null hypothesis encompasses a product of parameters, which is composite and consists of three cases as case 1: $\beta^j = 0$ and $\gamma^j \neq 0$, case 2: $\beta^j \neq 0$ and $\gamma^j = 0$, and case 3: $\beta^j = 0$ and $\gamma^j = 0$. Constructing the test statistic and deriving the reference distribution and composite p-value under the composite null hypothesis are still open problems in the field of high-dimensional mediation analysis, and relevant studies can be found in the review by Du et al. (2023). Considering the complex limiting distributions of $\hat{\beta}^j$ and $\hat{\gamma}^j$, it is more challenge to conduct statistical inference for the mediation effect $\hat{\beta}^j\hat{\gamma}^j$. We will develop theories and inference tools of the proposed estimator in our future research.

Funding

The research is supported by NSF Grant DMS-2210640 and 1952406, and in part by NIH Grant R01MD011728.

Supplement

Supplement to "Time-varying mediation analysis for incomplete data with application to DNA methylation study for PTSD". The Supplementary Material includes three sections. Section 1 contains the GMM estimator corresponding to each mediator M^j $(j = 1, ..., p)$ in model (2). Section 2 contains the expanded version of Algorithm 1. Section 3 contains the additional simulation results.

Supplemental code. We have implemented the proposed longitudinal multiple imputation method as an R function named "LMI" and the generalized method of moments with principal component analysis as another R function named "GMM_PCA". Simulation code, data, and application code, along with detailed instructions, are provided.

References

Bind, M. A. C., Vanderweele, T. J., Coull, B. A. and Schwartz, J. D. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics* **17** 122–134.

Braun, P. R., Han, S., Hing, B., Nagahama, Y., Gaul, L. N., Heinzman, J. T., Grossbach, A. J., Close, L., Dlouhy, B. J., Howard III, M. A., Kawasaki, H., Potash, J. B. and Shinozaki, G. (2019). Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl. Psychiatry* **9** 47.

Breslau, N., Kessler, R. C., Chilcoat, H. D., Schultz, L. R., Davis, G. C. and Andreski, P. (1998). Trauma and posttraumatic stress disorder in the community: The 1996 Detroit Area Survey of Trauma. *Arch. gen. psychiatry* **55** 626–632.

Cai, X., Coffman, D. L., Piper, M. E. and Li, R. (2022). Estimation and inference

for the mediation effect in a time-varying mediation model. *BMC Medical Res. Methodol.* **22** 113.

Chen, C., Shen, B., Liu, A., Wu, R. and Wang, M. (2021). A multiple robust propensity score method for longitudinal analysis with intermittent missing data. *Biometrics* **77** 519–532.

Chen, F., Hu, W., Cai, J., Chen, S., Si, A., Zhang, Y. and Liu, W. (2023). Instrumental variable-based high-dimensional mediation analysis with unmeasured confounders for survival data in the observational epigenetic study. *Front. Genet.* **14** 1092489.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G. and Xing, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Stat.* **6** 719–752.

Chitrala, K. N., Nagarkatti, P. and Nagarkatti, M. (2016). Prediction of possible biomarkers and novel pathways conferring risk to post-traumatic stress disorder. *PLOS ONE* **11** e0168404.

Cho, H. and Qu, A. (2015). Efficient estimation for longitudinal data by combining large-dimensional moment conditions. *Electron. J. Stat.* **9** 1315–1334.

Doran, H. E. and Schmidt, P. (2006). GMM estimators with improved finite sample properties using principal components of the weighting matrix, with an application to the dynamic panel data model. *J. Econom.* **133** 387–409.

Du, J., Zhou, X., Clark-Boucher, D., Hao, W., Liu, J., Smith, J. A. and Mukherjee, B. (2023). Methods for large-scale single mediator hypothesis testing: Possible choices and comparisons. *Genet. Epidemiol.* **47** 167–184.

Dukes, O., Shpitser, I. and Tchetgen Tchetgen, E. J. (2023). Proximal mediation analysis. *Biometrika* **110** 973–987.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

Ge, L., Wang, J., Shi, C., Wu, Z. and Song, R. (2023). A reinforcement learning framework for dynamic mediation analysis. *PMLR* **202** 11050-11097.

Goldmann, E., Aiello, A. E., Uddin, M., Delva, J., Koenen, K., Gant, L. M. and Galea, S. (2011). Pervasive exposure to violence and posttraumatic stress disorder in a predominantly African American Urban Community: The Detroit Neighborhood Health Study. *J. Trauma. Stress* **24** 747–751.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.

Huang, J., Ma, S., Xie, H. and Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355.

Jahangiri, M., Kazemnejad, A., Goldfeld, K. S., Daneshpour, M. S., Mostafaei, S., Khalili, D., Moghadas, M. R. and Akbarzadeh, M. (2023). A wide range of missing imputation approaches in longitudinal data: A simulation study and real data analysis. *BMC Medical Res. Methodol.* **23** 161.

Johns, L. E., Aiello, A. E., Cheng, C., Galea, S., Koenen, K. C. and Uddin, M. (2012). Neighborhood social cohesion and posttraumatic stress disorder in a community-based sample: Findings from the Detroit Neighborhood Health Study. *Soc. Psychiatry Psychiatr. Epidemiol.* **47** 1899–1906.

Koenen, K. C., Ratanatharathorn, A., Ng, L., McLaughlin, K. A., Bromet, E. J., Stein, D. J., Karam, E. G., Ruscio, A. M., Benjet, C., Scott, K., Atwoli, L., Petukhova, M., Lim, C. C. W., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Bunting, B., Ciutan. M., de Girolamo, G., Degenhardt, L., Gureje, O., Haro, J. M., Huang, Y., Kawakami, N., Lee, S., Navarro-Mateu, F., Pennell, B. E., Piazza, M., Sampson, N., ten Have, M., Torres, Y., Viana, M. C., Xavier, M. and Kessler, R. C. (2017). Posttraumatic stress disorder in the World Mental Health Surveys. *Psychol. Med.* **47** 2260–2274.

Kruppa, J., Sieg, M., Richter, G. and Pohrt, A. (2021). Estimands in epigenome-wide association studies. *Clin. Epigenet.* **13** 98.

Kuan, P., Waszczuk, M. A., Kotov, R., Marsit, C. J., Guffanti, G., Gonzalez, A., Yang, X., Koenen, K., Bromet, E. and Luft, B. J. (2017). An epigenome-wide DNA methylation study of PTSD and depression in World Trade Center responders. *Transl. Psychiatry* **7** e1158.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data, 3rd ed.* Wiley, New York.

Luo, L., Shi, C., Wang, J., Wu, Z. and Li, L. (2023). Multivariate dynamic mediation analysis under a reinforcement learning framework. *arXiv preprint arXiv:2310.16203.*

Lussier, A. A., Zhu, Y., Smith, B. J., Cerutti, J., Fisher, J., Melton, P. E., Wood, N. M., Cohen-Woods, S., Huang, R. C., Mitchell, C., Schneper, L., Notterman, D. A., Simpkin, A. J., Smith, A. D. A. C., Suderman, M. J., Walton, E., Relton, C. L., Ressler. K. and Dunn, E. C. (2023). Association between the timing of childhood adversity and epigenetic patterns across childhood and adolescence: Findings from the Avon Longitudinal Study of Parents and Children (ALSPAC) prospective cohort. *Lancet Child Adolesc. Health* **7** 532–543.

Mazumder, R., Hastie, T. and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322.

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang XO, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lécuyer E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guigó R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison

RC, Gingeras TR, Stamatoyannopoulos JA and Weng Z (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583** 699–710.

Mou, X., Zhang, H. and Arshad, S. H. (2022). Identifying intergenerational patterns of correlated methylation sites. *Ann. Appl. Stat.* **16** 521-536.

Occean, J. R., Wani, A. H., Donglasan, J., Aiello, A. E., Galea, S., Koenen, K. C., Qu, A., Wildman, D. and Uddin, M. (2022). DNA methylation of Nuclear Factor of Activated T Cells 1 mediates the prospective relation between exposure to different traumatic event types and post-traumatic stress disorder. *Psychiatry Res.* **311** 114510.

Perera, C., Zhang, H., Zheng, Y., Hou, L., Qu, A., Zheng, C., Xie, K. and Liu, L. (2022). HIMA2: High-dimensional mediation analysis and its application in epigenome-wide DNA methylation data. *BMC Bioinform.* **23** 296.

Qin, X., Hong, G., Deutsch, J. and Bein, E. (20219). Multisite causal mediation analysis in the presence of complex sample and survey designs and non-random non-response. *J. R. Stat. Soc., A: Stat. Soc.* **182** 1343–1370.

Rijnhart, J. J. M., Twisk, J. W. R., Valente, M. J. and Heymans, M. W. (2022). Time lags and time interactions in mixed effects models impacted longitudinal mediation effect estimates. *J. Clin. Epidemiol.* **151** 143–150.

Ruggiero, K. J., Ben, K. D., Scotti, J. R. and Rabalais, A. E. (2003). Psychometric properties of the PTSD Checklist - Civilian version. *J. Trauma. Stress* **16** 495–502.

Rutten, B. P. F., Vermetten, E., Vinkers, C. H., Ursini, G., Daskalakis, N. P., Pishva, E., de Nijs, L., Houtepen, L. C., Eijssen, L., Jaffe, A. E., Kenis, G., Viechtbauer, W., van den Hove, D., Schraut, K. G., Lesch, K. P., Kleinman, J. E., Hyde, T. M., Weinberger. D. R., Schalkwyk, L., Lunnon, K., Mill, J., Cohen, H., Yehuda, R., Baker, D. G., Maihofer, A. X., Nievergelt, C. M., Geuze, E. and Boks, M. P. M. (2018). Longitudinal analyses of the DNA methylome in deployed military servicemen identify susceptibility loci for post-traumatic stress disorder. *Mol. Psychiatry* **23** 1145–1156.

Schuurmans, I. K., Dunn, E. C. and Lussier, A. A. (2024). DNA methylation as a possible causal mechanism linking childhood adversity and health: Results from two-sample mendelian randomization study. *Am. J. Epidemiol.* **193** 1541–1552.

Shan, J., Li, W. and Ai, C. (2024). Efficient nonparametric inference of causal mediation effects with nonignorable missing confounders. *arXiv preprint arXiv:2402.05384*.

Shi, C. and Li, L. (2022). Testing mediation effects using logic of boolean matrices. *J. Am. Stat. Assoc.* **117** 2014–2027.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L. R., Diez Roux, A. V., Needham, B. L., Smith, J. A. and Mukherjee, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics* **76** 700–710.

Tai, A. S., Lin, P. H., Huang, Y. T. and Lin, S. H. (2022). Path-specific effects in the presence of a survival outcome and causally ordered multiple mediators with application to genomic data. *Stat. Methods Med. Res.* **31** 1916–1933.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108.

Tortella-Feliu, M., Fullana, M. A., Pérez-Vigil, A., Torres, X., Chamorro, J., Littarelli, S. A., Solanes, A., Ramella-Cravaro, V., Vilar, A., González-Parra, J. A., Andero, R., Reichenberg, A., Mataix-Cols, D., Vieta, E., Fusar-Poli, P., Ioannidis, J. P. A., Stein, M. B., Radua. J. and Fernández de la Cruz, L. (2019). Risk factors for posttraumatic stress disorder: An umbrella review of systematic reviews and meta-analyses. *Neurosci. Biobehav. Rev.* **107** 154–165.

Tseng, C. H., Elashoff, R., Li, N. and Li, G. (2016). Longitudinal data analysis with non-ignorable missing data. *Stat. Methods Med. Res.* **25** 205–220.

Turecki, G. and Meaney, M. J. (2016). Effects of the social environment and stress on glucocorticoid receptor gene methylation: A systematic review. *Biol. Psychiatry* **79** 87–96.

Uddin, M., Aiello, A. E., Wildman, D. E., Koenen, K. C., Pawelec, G., de los Santos, R., Goldmann, E. and Galea, S. (2010). Epigenetic and immune function profiles associated with posttraumatic stress disorder. *PNAS* **107** 9470–9475.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 917–938.

Wani, A. H., Aiello, A. E., Kim, G. S., Xue, F., Martin, C. L., Ratanatharathorn, A., Qu, A., Koenen, K., Galea, S., Wildman, D. and Uddin, M. (2021). The impact of psychopathology, social adversity and stress-relevant DNA methylation on prospective risk for post-traumatic stress: A machine learning approach. *J. Affect. Disord.* **282** 894–905.

Wei, K., Zhu, H., Qin, G., Zhu, Z. and Tu, D. (2022). Multiply robust subgroup analysis based on a single-index threshold linear marginal model for longitudinal data with dropouts. *Stat. Med.* **41** 2822–2839.

Wilker, S., Vukojevic, V., Schneider, A., Pfeiffer, A., Inerle, S., Pauly, M., Elbert, T., Papassotiropoulos, A., de Quervain, D. and Kolassa, I. T. (2023). Epigenetics of traumatic stress: The association of NR3C1 methylation and posttraumatic stress disorder symptom changes in response to narrative exposure therapy. *Transl. Psychiatry* **13** 14.

Wu, S., Renzaho, A. M. N., Hall, B. J., Shi, L., Ling, L. and Chen, W. (2021). Time-varying associations of pre-migration and post-migration stressors in refugees' mental health during resettlement: A longitudinal study in Australia. *Lancet Psychiatry* **8** 36–47.

Xue, F. and Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *J. Am. Stat. Assoc.* **116** 1914–1927.

Xue, F., Tang, X., Kim, G., Koenen, K. C., Martin, C. L., Galea, S., Wildman, D., Uddin, M. and Qu, A. (2022). Heterogeneous mediation analysis on epigenomic ptsd and traumatic stress in a predominantly African American cohort. *J. Am. Stat. Assoc.* **117** 1669–1683.

Yuan, Y. and Qu, A. (2024). De-confounding causal inference using latent multiple-mediator pathways. *J. Am. Stat. Assoc.* **119** 2051–2065.

Zeng, S., Lange, E. C., Archie, E. A., Campos, F. A., Alberts, S. C. and Li, F. (2022). A causal mediation model for longitudinal mediators and survival outcomes with an application to animal behavior. *J. Agric. Biol. Environ. Stat.* **28** 197–218.

Zeng, S., Rosenbaum, S., Alberts, S. C., Archie, E. A. and Li, F. (2021). Causal mediation analysis for sparse and irregular longitudinal data. *Ann. Appl. Stat.* **15** 747–767.

Zhao, Y. and Luo, X. (2022). Pathway lasso: Pathway estimation and selection with high-dimensional mediators. *Stat. Its Interface* **15** 39–50.

Zuo, S., Ghosh, D., Ding, P. and Yang, F. (2024). Mediation analysis with the mediator and outcome missing not at random. *J. Am. Stat. Assoc.* 1–12.