Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Statistical Inference for Mediation Models with High Dimensional Exposures and Mediators

Xinyu Zhang, Wei Zhou, Jingyuan Liu & Jian Kang

To cite this article: Xinyu Zhang, Wei Zhou, Jingyuan Liu & Jian Kang (09 Feb 2026): Statistical Inference for Mediation Models with High Dimensional Exposures and Mediators, Journal of the American Statistical Association, DOI: 10.1080/01621459.2026.2621518

To link to this article: https://doi.org/10.1080/01621459.2026.2621518

Statistical Inference for Mediation Models with High Dimensional Exposures and Mediators

Xinyu Zhang[a,*], Wei Zhou[b,*], Jingyuan Liu[c,†], Jian Kang[d]

[a]Paula and Gregory Chow Institute for Studies in Economics, Xiamen University

[b]Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics

[c]MOE Key Laboratory of Econometrics, Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics, Fujian Key Lab of Statistics, Xiamen University

[d]Department of Biostatistics, University of Michigan

*These two authors are co-first authors

[†]Corresponding author: jingyuan@xmu.edu.cn

Abstract

High-dimensional mediation analysis has gained increasing interest in various fields, particularly in genetic and medical research. Compared with existing works that focus mainly on high-dimensional mediators, this paper advocates a new framework of Partial Regularization-based Inference for Mediation Effects (PRIME) when both exposures and mediators are high-dimensional. Estimated direct and indirect effects are established using a group-wise partially penalized least squares method, incorporating a double-layer latent factor structure. F-type and Wald tests for the high-dimensional direct and indirect effects, respectively, are advocated based on the proposed estimators. Both theoretical and numerical performance of PRIME have been carefully studied. PRIME is also applied to investigating direct effects of genetic variants on Alzheimer's disease (AD) and indirect effects of them mediated by changes in brain activity intensity.

Keywords: High-dimensional mediation analysis; Partially penalized least squares; Latent factor model; Multiple test

1 Introduction

Mediation analysis is an important tool for analyzing the relationship between exposures and outcomes through mediators, with application in a wide spectrum of scientific domains, including genetic studies (Chernozhukov et al., 2021), psychology (Rucker et al., 2011), and economics (Guo et al., 2022). The most widely used mediation analysis framework is based on a two-layer linear structural equation model (Baron and Kenny, 1986), which represents causal relationships between variables through associated regression coefficients. For a comprehensive overview, see VanderWeele (2016).

A considerable amount of work has been devoted to statistical estimation and inference for mediation analysis with low-dimensional exposures and mediators; see Preacher and Hayes (2008); VanderWeele and Vansteelandt (2014), among others. In terms of high-dimensional mediation analysis, the recent development of the methodology mainly focuses on purely high-dimensional mediators, keeping exposures low-dimensional, such as investigating the effect of early life trauma on cortisol stress reactivity in adulthood through DNA methylation levels (Guo et al., 2022), or identifying activity levels of brain regions that act as mediators between a thermal stimulus and certain self-reported pain (Chén et al., 2018). To this end, Zhou et al. (2020) provides asymptotic inference guarantees for the indirect effect of low-dimensional exposure through a debiasing approach. Guo et al. (2023, 2022) develop a two-stage procedure to account for both direct and indirect effects. Another branch of addressing high-dimensional mediators typically examines one potential mediator at a time, based on the estimated structures of directed acyclic graphs; see Shi and Li (2022); Chakrabortty et al. (2018). Another research stream to handling high-dimensional mediators involves the use of latent variables. For example, Derkach et al. (2019) proposes a model where exposures influence a set of latent factors, which in turn affect both mediators and the outcome. The authors employ a LASSO-penalized likelihood and an expectation-maximization algorithm to jointly estimate model parameters. Building on this work, Cai et al. (2023) extends the framework by incorporating penalized regularization to select both latent factors and candidate mediators. Meanwhile, Zhang et al. (2024) introduces a latent factor model on the coefficients of a linear mediation model to capture time-varying and heterogeneous mediation effects, enabling the estimation of individualized causal effects. Although these studies establish solid theoretical properties for mediation analysis with high-dimensional mediators, they restrict the exposure dimensions to be fixed and finite.

However, the increasing prevalence of high-dimensional data often necessitates the consideration of high-dimensional exposures in mediation analysis. For instance, in the imaging proteomics study by Zhao et al. (2022), the exposures are taken as the intensity measures of the peptides, while the mediators are the brain imaging data obtained by anatomical magnetic resonance imaging (MRI). To address the high-dimensionality of both exposures and mediators, Zhang (2022) introduces two regularization methods for mediator selection under the assumption of an independence structure among mediators. Zhao et al. (2022) employs a penalized principal component regression method to reduce dimensionality, which can alter the original structure of exposures, potentially complicating the interpretation of the causal pathway. With regard to statistical inference, few existing works have been found in the literature. One notable method to test direct and path-specific indirect effects is a decorrelating-debiasing approach proposed by Wang et al. (2025). However, it ignores the

interplay among mediators by considering each path separately, and the debiasing step leads to a loss of efficiency and high computational costs.

In this paper, we propose a new inference framework, called Partial Regularization-based Inference for Mediation Effects (PRIME), for mediation models with both high-dimensional exposures and mediators. In contrast to previous studies that typically focus on testing each individual path through each mediator, PRIME provides a more comprehensive assessment of the indirect effect of each exposure in all mediators. The group-wise multiple testing setup renders the validity of partial regularization, which automatically avoids introducing estimation bias of the primary exposures and hence further avoid efficiency loss due to debiasing. To address the high correlations among exposures and between exposures and mediators, as well as to relax the irrepresentable conditions, we assume that latent factors influence the exposures and mediators simultaneously, and thus impose a double-layer latent factor structure on the high-dimensional mediation models. Then an F-type and Wald test for the high-dimensional direct and indirect effects are advocated, respectively, under a partially multiple testing framework. Theoretically, we establish the rates of convergence and asymptotic distributions of the proposed estimators after factor adjustment and derive the size and power of the proposed test for direct and indirect effects under both null and local alternative hypotheses.

Our work is motivated by the needs of analyzing imaging genetics data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which provides rich, longitudinal multi-modal data aimed at uncovering the biological mechanisms underlying Alzheimer's disease (AD). Our goal is to disentangle the direct genetic effects on AD symptoms from those that are mediated through changes in brain activity, thus enhancing our understanding of the genotype-phenotype pathway and identifying potential biomarkers for early detection and intervention. In our study, the exposures are high-dimensional genetic variants (SNPs), the mediators are regional brain activity intensities measured by PET imaging across 116 anatomical regions, and the outcome is the neuropsychiatric inventory score (NPISCORE), a clinical measure of psychological symptoms related to AD severity. The high dimensionality and complex interdependencies in both genetic and imaging data pose significant challenges for traditional mediation analysis.

The contributions of the proposed PRIME are threefold. First, to our best knowledge, PRIME is the first inference method, with rigorous theoretical guarantee, for both high-dimensional direct and indirect effects. Second, we advocate a double-layer latent factor structure to mediation models, to address estimation bias caused by the high correlations among the high-dimensional exposures and mediators, respectively, and between exposures and mediators. Crucially, the proposed factor structure is carefully designed in order to preserve the identifiability of the original direct and indirect effects. Building on the factor-adjusted models and the partial regularization, we derive the size and power of the subsequent tests for the direct and indirect effects of interest. Third, we analyze the direct and indirect effects of genetic variants on Alzheimer's disease through the intensity of brain activity in an imaging genetics study. We discover single nucleotide polymorphisms (SNPs) with significant direct and indirect effects, some of which are consistent with existing findings in biology and medicine, while others are newly identified and have not been previously reported in the literature. These results may contribute to gene localization and offer insights for disease diagnosis and treatment.

The remainder of the paper is organized as follows. In Section 2, we introduce the problem setup, including the linear structural equation mediation model with both high-dimensional exposures and mediators, the hypotheses of interest, and the double-layer latent factor structure. In Section 3, we propose the estimation and inference procedures and establish the theoretical guarantee of the methods. Section 4 presents simulation studies. The real data analysis to test the direct and indirect effects of SNPs on Alzheimer's disease is provided in Section 5. The last section is dedicated to the conclusion and discussion. Technical proofs and additional results are deferred to a separate supplementary file.

2 Problem Setup

Notations. For a vector $\boldsymbol{v} = (v_1, \ldots, v_h)^\top \in \mathbb{R}^h$ and an index set $A \subset \{1, 2, \ldots, h\}$ with cardinality $|A|$, denote $\boldsymbol{v}_A$ to be the sub-vector of $\boldsymbol{v}$ corresponding to the indices in A, and $\boldsymbol{v}_{-A}$ to be its complement. Further, let $\boldsymbol{v}_{-A,S}$ denote the sub-vector of $\boldsymbol{v}_{-A}$ corresponding to the set $S \subseteq \{1, \ldots, h-|\mathcal{A}|\}$. Similarly, for a matrix $\boldsymbol{V} = (V_{ij})_{n\times h}$, denote $\boldsymbol{V}_A = (V_{ij})_{1\le i\le n,\, j\in A}$, $\boldsymbol{V}_{-A} = (V_{ij})_{1\le i\le n,\, j\notin A}$, and $\boldsymbol{V}_{-A,S} = (V_{ij})_{1\le i\le n,\, j\in S\cap\{1,\ldots,h-|\mathcal{A}|\}}$ as the sub-matrix of $\boldsymbol{V}$. We define $\|\boldsymbol{V}\|_\infty = \max_i \sum_j |V_{ij}|$, $1 \le i \le n$, $\|\boldsymbol{V}\|_{\max} = \max_{ij} |V_{ij}|$, $\|\boldsymbol{V}\|_{2,\infty} = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \|\boldsymbol{V}\boldsymbol{x}\|_\infty$, and $\|\boldsymbol{V}\|_{\mathbb{F}} = \sqrt{\sum_{ij} V_{ij}^2}$. Let $\lambda_{\min}(\boldsymbol{U})$ and $\lambda_{\max}(\boldsymbol{U})$ denote the minimum and maximum eigenvalues of a square matrix $\boldsymbol{U}$, respectively. For two positive sequences $\{a_n\}_{n\ge1}$ and $\{b_n\}_{n\ge1}$, we write $a_n \gg b_n$ if $\lim a_n / b_n = \infty$. The sub-Gaussian norm for a scalar random variable Z is defined as $\|Z\|_{\psi_2} = \inf\left\{t > 0 : \mathrm{E}\exp\left(Z^2/t^2\right) \leqslant 2\right\}$. For a random vector $\boldsymbol{x} \in \mathbb{R}^m$, we use $\|\boldsymbol{x}\|_{\psi_2} = \sup_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{v}^\top \boldsymbol{x}\|_{\psi_2}$ to denote its sub-Gaussian norm.

2.1 High-dimensional Linear Structural Equation Models

In this paper, we consider the following high-dimensional linear structural equation model (HD-LSEM) for mediation analysis:

$$y = \boldsymbol{\alpha}_0^\top \boldsymbol{m} + \boldsymbol{\alpha}_1^\top \boldsymbol{x} + \varepsilon_1, \quad (1)$$

$$\boldsymbol{m} = \Gamma^\top \boldsymbol{x} + \boldsymbol{\varepsilon}, \quad (2)$$

where y is the univariate outcome variable, $\boldsymbol{m} = (m_1, \ldots, m_p)^\top$ is the p-dimensional mediator vector, and $\boldsymbol{x} = (x_1, \ldots, x_q)^\top$ is the q-dimensional exposure vector. We assume that both p and q are high-dimensional compared with the sample size n. Correspondingly, $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0p})^\top \in \mathbb{R}^p$, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1q})^\top \in \mathbb{R}^q$, and $\Gamma \in \mathbb{R}^{q\times p}$ are coefficient vectors and matrix. For parsimony and to avoid redundant parameters, all variables are standardized and the intercept terms are omitted from the model. We also impose the sparsity assumption that only a small proportion of entries in $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are non-zero. We assume that there are no unmeasured confounders to avoid endogeneity, which implies independence between ε_1 and

$\boldsymbol{m}$ and $\boldsymbol{x}$, between $\boldsymbol{\varepsilon}$ and $\boldsymbol{x}$, and between ε_1 and $\boldsymbol{\varepsilon}$. Furthermore, suppose that $\mathrm{var}(\varepsilon_1) = \sigma_1^2$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \Sigma^\star$. Note that the above HD-LSEM can be directly extended to accommodate observed confounders with no specific technical challenges, thus we choose not to include such confounders for notation simplicity.

Plugging (2) into (1) yields

$$y = (\boldsymbol{\beta} + \boldsymbol{\alpha}_1)^\top \boldsymbol{x} + \varepsilon_1 + \varepsilon_2 \coloneqq \boldsymbol{\gamma}^\top \boldsymbol{x} + \varepsilon_3, \quad (3)$$

where $\boldsymbol{\beta} = \Gamma \boldsymbol{\alpha}_0 = (\beta_1, \ldots, \beta_q)^\top \in \mathbb{R}^q$, $\varepsilon_2 \coloneqq \boldsymbol{\alpha}_0^\top \boldsymbol{\varepsilon}$ with $\mathrm{var}(\varepsilon_2) = \sigma_2^2 = \boldsymbol{\alpha}_0^\top \Sigma^\star \boldsymbol{\alpha}_0$, $\boldsymbol{\gamma} \coloneqq \boldsymbol{\beta} + \boldsymbol{\alpha}_1 = (\gamma_1, \ldots, \gamma_q)^\top \in \mathbb{R}^q$, and $\varepsilon_3 \coloneqq \varepsilon_1 + \varepsilon_2$ is the total random error. Following the literature, we refer $\boldsymbol{\beta}$ to the indirect effect of $\boldsymbol{x}$ on *y* mediated by $\boldsymbol{m}$, $\boldsymbol{\alpha}_1$ to the direct effect of $\boldsymbol{x}$, and $\boldsymbol{\gamma} = \boldsymbol{\beta} + \boldsymbol{\alpha}_1$ to the total effect of $\boldsymbol{x}$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_1$ possess their own interpretations as natural indirect effect and natural direct effect in the causality regime. See Section S1 for the causal interpretation of such effects.

2.2 Double-layer Latent Factor Structure for HD-LSEM

For high-dimensional regression models, such as those used in genetic studies, it is common to encounter high correlations among predictors, resulting in potential estimation bias and efficiency loss. In the double-layer model HD-LSEM, such issue caused by correlations between predictors (including exposures and mediators) can be more severe. For example, as illustrated in Figure 1, if x_1 and x_2 are highly correlated, then regressing *y* against $\boldsymbol{x}$ can lead to a non-zero estimate of γ_1, although the true value of γ_1 is 0. To address the correlation within the exposure $\boldsymbol{x}$, we introduce the following latent factor structure in the first-layer model:

$$\boldsymbol{x} = \boldsymbol{B}_1 \boldsymbol{f}_1 + \boldsymbol{v}_1, \quad (4)$$

where $\boldsymbol{f}_1 \in \mathbb{R}^{K_1}$ is the latent K_1-dimensional factor capturing the correlation within $\boldsymbol{x}$, $\boldsymbol{B}_1 \in \mathbb{R}^{q \times K_1}$ is the corresponding factor loading matrix, and $\boldsymbol{v}_1 \in \mathbb{R}^q$ is the idiosyncratic component of $\boldsymbol{x}$, with independent components. Plugging (4) into the model (3), we obtain

$$y = \boldsymbol{\gamma}^\top (\boldsymbol{B}_1 \boldsymbol{f}_1 + \boldsymbol{v}_1) + \varepsilon_3 \coloneqq \boldsymbol{\phi}_1^\top \boldsymbol{f}_1 + \boldsymbol{\gamma}^\top \boldsymbol{v}_1 + \varepsilon_3, \quad (5)$$

where $\boldsymbol{\phi}_1 \coloneqq \boldsymbol{B}_1^\top \boldsymbol{\gamma}$. Then (5) is, in fact, the factor-augmented model to estimate the total effect $\boldsymbol{\gamma}$. By this means, the correlation structure of original exposure $\boldsymbol{x}$ has been captured by $\boldsymbol{f}_1$, and the effect of $\boldsymbol{x}$ on the outcome is migrated to the effect of the idiosyncratic component $\boldsymbol{v}_1$, where the total effect $\boldsymbol{\gamma}$ is preserved.

In addition, the correlation between $\boldsymbol{m}$ and $\boldsymbol{x}$ tends to bring estimation bias for the direct effect $\boldsymbol{\alpha}_1$, and hence the indirect effect $\boldsymbol{\beta}$. In the example illustrated by Figure 1, x_3 has a non-zero indirect effect on *y* through m_2, while the true direct effect of x_3 on *y* is assumed to

be zero. However, regressing y on $\boldsymbol{m}$ and $\boldsymbol{x}$ leads to estimation bias of $\alpha_{1,3}$ due to the correlation between x_3 and m_2, which in turn yields potential large bias when estimating β_3.

To address this, we employ another layer of latent factor structure to capture the relationship between $\boldsymbol{m}$ and $\boldsymbol{x}$ via

$$(\boldsymbol{m}^\top, \boldsymbol{x}^\top)^\top = (\boldsymbol{B}_m^\top, \boldsymbol{B}_x^\top)^\top \boldsymbol{f}_2 + (\boldsymbol{v}_{2m}^\top, \boldsymbol{v}_{2x}^\top)^\top = \boldsymbol{B}_2 \boldsymbol{f}_2 + \boldsymbol{v}_2, \quad (6)$$

where $\boldsymbol{f}_2 \in \mathbb{R}^{K_2}$ is the K_2-dimensional latent factor, $\boldsymbol{B}_m \in \mathbb{R}^{p \times K_2}$ and $\boldsymbol{B}_x \in \mathbb{R}^{q \times K_2}$ are the factor loading matrices corresponding to $\boldsymbol{m}$ and $\boldsymbol{x}$ with $\boldsymbol{B}_2 := (\boldsymbol{B}_{\boldsymbol{m}}^\top, \boldsymbol{B}_{\boldsymbol{x}}^\top)^\top$, and $\boldsymbol{v}_{2m} \in \mathbb{R}^p$ and $\boldsymbol{v}_{2x} \in \mathbb{R}^q$ are the respective idiosyncratic components with $\boldsymbol{v}_2 := (\boldsymbol{v}_{2m}^\top, \boldsymbol{v}_{2x}^\top)^\top$. We refer to (4) and (6) as the double-layer latent factor model. Plugging (6) into (1) yields that

$$\begin{aligned} y &= \boldsymbol{\alpha}_0^\top (\boldsymbol{B}_m \boldsymbol{f}_2 + \boldsymbol{v}_{2m}) + \boldsymbol{\alpha}_1^\top (\boldsymbol{B}_x \boldsymbol{f}_2 + \boldsymbol{v}_{2x}) + \varepsilon_1 \\ &= (\boldsymbol{\alpha}_0^\top \boldsymbol{B}_m + \boldsymbol{\alpha}_1^\top \boldsymbol{B}_x) \boldsymbol{f}_2 + \boldsymbol{\alpha}_0^\top \boldsymbol{v}_{2m} + \boldsymbol{\alpha}_1^\top \boldsymbol{v}_{2x} + \varepsilon_1 \quad (7) \\ &:= \boldsymbol{\phi}_2^\top \boldsymbol{f}_2 + \boldsymbol{\alpha}_0^\top \boldsymbol{v}_{2m} + \boldsymbol{\alpha}_1^\top \boldsymbol{v}_{2x} + \varepsilon_1, \end{aligned}$$

where $\boldsymbol{\phi}_2 := \boldsymbol{B}_m^\top \boldsymbol{\alpha}_0 + \boldsymbol{B}_x^\top \boldsymbol{\alpha}_1$. A byproduct of imposing this second factor layer (6) is that arbitrary correlation between $\boldsymbol{m}$ and $\boldsymbol{x}$ is now allowed in the original HD-LSEM, unlike most existing approaches, which typically prohibit highly correlated mediators and exposures.

While there are alternative factor models to capture correlations among predictors, the proposed double-layer latent factor structure in (4) and (6) offers two distinct advantages. The first-layer factor model (4) ensures estimation efficiency of the total effect. An alternative approach involves directly substituting the decomposition of $\boldsymbol{x}$, expressed in terms of $\boldsymbol{f}_2$ and $\boldsymbol{v}_{2x}$ from model (6), into the total effect model (3). However, $\boldsymbol{f}_2$ inherently contains information from $\boldsymbol{m}$, which compromises the efficiency of estimating the total effect. Second, the factor structure in (6) guarantees that the predictors of interest in the resulting augmented model (2.2), typically elements in $\boldsymbol{v}_{2x}$, are orthogonal to the nuisance predictors. This, as elaborated in the next section, leads to an unbiased and analytically tractable solution for estimating the parameters of interest through partial regularization.

2.3 Test Hypotheses for Direct and Indirect Effects

Our interest is to test the direct and indirect effects of each individual exposure variable in $\boldsymbol{x}$ on the outcome; that is,

$$H_{0j,\text{dir}} : \alpha_{1j} = 0 \textit{versus}\ \ H_{1j,\text{dir}} : \alpha_{1j} \neq 0,\ j = 1, \ldots, q, \text{ and } \quad (8)$$

$$H_{0j,\text{ind}} : \beta_j = 0 \textit{versus}\ \ H_{1j,\text{ind}} : \beta_j \neq 0,\ j = 1, \ldots, q. \quad (9)$$

This is essentially the multiple hypothesis testing problems.

However, the high-dimensional exposure variable $\boldsymbol{x}$ may be classified as *G* groups from prior knowledge. For instance, Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003) classifies SNPs into different functional categories based on gene functions, biological pathways, and disease associations, using annotations such as gene ontology and KEGG pathways; Functional Mapping and Annotation of GWAS (FUMA) (Watanabe et al., 2017) offers a more comprehensive grouping strategy, incorporating gene mapping, regulatory region analysis, eQTL data and chromatin conformation. Based on properly grouped SNPs, researchers can better investigate the joint mechanisms of association between SNPs and relevant diseases or phenotypes. Therefore, in genetic practice, it is more interesting to map genes or to evaluate the effects of individual or certain small groups of genes on the phenotype. Rather than testing the direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta}$ of exposures (e.g., genes or SNPs) in HD-LSEM as whole high-dimensional vectors or individual one-dimensional scalars, multiple test hypotheses involving subsets of exposures are more relevant. Then for the direct effect $\boldsymbol{\alpha}_1$, the null and alternative hypotheses are

$$H_{0,\text{dir}}^{(g)}:\boldsymbol{\alpha}_{1,g}=\mathbf{0} \text{ versus } H_{1,\text{dir}}^{(g)}:\boldsymbol{\alpha}_{1,g}\neq\mathbf{0}, \quad (10)$$

for $g=1,\ldots,G$. Similarly, for the indirect effect $\boldsymbol{\beta}$, we have

$$H_{0,\text{ind}}^{(g)}:\boldsymbol{\beta}_{g}=\mathbf{0} \text{ versus } H_{1,\text{ ind}}^{(g)}:\boldsymbol{\beta}_{g}\neq\mathbf{0}, \quad (11)$$

for $g=1,\ldots,G$. When $G=q$, (10) and (11) reduce to the multiple hypothesis testing problems for the individual direct and indirect effects in (8) and (9).

However, if the prior information about the group strategy is unavailable, we might conduct individual tests in (8) and (9). Therefore, our proposal is based on the grouped hypotheses (10) and (11), which contain the individual hypotheses (8) and (9) as special cases.

3 PRIME: Partial Regularization-based Inference for Mediation Effects in HD-LSEM

In this section, we introduce the estimation and testing procedure for the high dimensional direct and indirect effects of the exposure $\boldsymbol{x}$ on the outcome *y*, respectively. We name the proposed method Partial Regularization-based Inference for Mediation Effects (PRIME), which consists of estimating and testing procedures as well as FDR control for the multiple testing problem.

Observe that directly estimating the indirect effect $\boldsymbol{\beta}$ via $\boldsymbol{\beta}=\Gamma\boldsymbol{\alpha}_0$ introduces non-negligible errors and heavy computational burden through the estimation of the high-dimensional matrix Γ and the high-dimensional vector $\boldsymbol{\alpha}_0$. However, inspired by Guo et al. (2023), the indirect effect $\boldsymbol{\beta}$ is in fact the difference between the total effect $\boldsymbol{\gamma}$ and the direct effect $\boldsymbol{\alpha}_1$, that is, $\boldsymbol{\beta}=\boldsymbol{\gamma}-\boldsymbol{\alpha}_1$. Thus, we propose to first estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}_1$, and the indirect effect $\boldsymbol{\beta}$ can then naturally be estimated by taking the difference.

For the gth group of exposure $\boldsymbol{x}_g$, $g = 1,\ldots,G$, we first analyze the total effect $\boldsymbol{\gamma}_g$ corresponding to exposures from the gth group on the outcome y. Thus, we rewrite the factor-augmented total effect model (5) as

$$y = \boldsymbol{\gamma}_{-g}^{\top}\boldsymbol{u}_{-g} + \boldsymbol{\gamma}_{g}^{\top}\boldsymbol{v}_{1,g} + \boldsymbol{\phi}_1^{\top}\boldsymbol{f}_1 + \varepsilon_3, \quad (12)$$

where $\boldsymbol{v}_{1,g}$ is the sub-vector of the idiosyncratic component $\boldsymbol{v}_1$ corresponding to the gth group of exposures, $\boldsymbol{u}_{-g} := \boldsymbol{v}_{1,-g}$ is the idiosyncratic component except for the gth group, and $\boldsymbol{\gamma}_{-g}$ is the corresponding augmented coefficients.

To analyze the gth group of direct effect $\boldsymbol{\alpha}_{1,g}$, we reparameterize model (2.2) as

$$y = \boldsymbol{\theta}_{-g}^{T}\boldsymbol{w}_{-g} + \boldsymbol{\alpha}_{1,g}^{T}\boldsymbol{v}_{2x,g} + \boldsymbol{\phi}_2^{\top}\boldsymbol{f}_2 + \varepsilon_1, \quad (13)$$

where $\boldsymbol{v}_{2x,g}$ is the sub-vector of the idiosyncratic component $\boldsymbol{v}_{2x}$ corresponding to the gth group exposures, $\boldsymbol{w}_{-g} = (\boldsymbol{v}_{2m}^{\top}, \boldsymbol{v}_{2x,-g}^{\top})^{\top}$ is the idiosyncratic components $\boldsymbol{v}_{2m}$ and the sub-vector of $\boldsymbol{v}_{2x}$ that excludes the gth group, and $\boldsymbol{\theta}_{-g} = (\boldsymbol{\alpha}_0^{\top}, \boldsymbol{\alpha}_{1,-g}^{\top})^{\top}$ is the corresponding augmented coefficients. Naturally, the indirect effect for the gth group is $\boldsymbol{\beta}_g = \boldsymbol{\gamma}_g - \boldsymbol{\alpha}_{1,g}$.

3.1 Estimation of Direct and Indirect Effects

Denote $\{\boldsymbol{m}_i, \boldsymbol{x}_i, y_i\}_{i=1}^{n}$ as the random sample generated from HD-LSEM and $\{\boldsymbol{f}_{1,i}, \boldsymbol{f}_{2,i}, \boldsymbol{v}_{1,i}, \boldsymbol{v}_{2m,i}, \boldsymbol{v}_{2x,i}\}_{i=1}^{n}$ as the corresponding factors and idiosyncratic components in (5) and (2.2). Define $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\top} \in \mathbb{R}^{n\times q}$, $\boldsymbol{M} = (\boldsymbol{m}_1,\ldots,\boldsymbol{m}_n)^{\top} \in \mathbb{R}^{n\times p}$, $\boldsymbol{F}_1 = (\boldsymbol{f}_{1,1},\ldots,\boldsymbol{f}_{1,n})^{\top} \in \mathbb{R}^{n\times K_1}$, $\boldsymbol{F}_2 = (\boldsymbol{f}_{2,1},\ldots,\boldsymbol{f}_{2,n})^{\top} \in \mathbb{R}^{n\times K_2}$, $\boldsymbol{V}_1 = (\boldsymbol{v}_{1,1},\ldots,\boldsymbol{v}_{1,n})^{\top} \in \mathbb{R}^{n\times q}$, $\boldsymbol{V}_{2m} = (\boldsymbol{v}_{2m,1},\ldots,\boldsymbol{v}_{2m,n})^{\top} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{V}_{2x} = (\boldsymbol{v}_{2x,1},\ldots,\boldsymbol{v}_{2x,n})^{\top} \in \mathbb{R}^{n\times q}$ with $\boldsymbol{V}_2 = (\boldsymbol{V}_{2m}, \boldsymbol{V}_{2x})$. Then we obtain the double-layer factor structure (4) and (6) in matrix form:

$$\begin{aligned}\boldsymbol{X} &= \boldsymbol{F}_1\boldsymbol{B}_1^{\top} + \boldsymbol{V}_1,\\ (\boldsymbol{M}, \boldsymbol{X}) &= \boldsymbol{F}_2\boldsymbol{B}_2^{\top} + \boldsymbol{V}_2,\end{aligned}$$

In this paper, we employ an efficient constrained least squares method motivated by Fan et al. (2024) to estimate the factor scores and their loading matrices. For the first layer, we solve for

$$\begin{aligned}&(\boldsymbol{F}_1, \boldsymbol{B}_1) = \underset{\boldsymbol{F}_1\in\mathbb{R}^{n\times\hat{K}_1}, \boldsymbol{B}_1\in\mathbb{R}^{q\times\hat{K}_1}}{\arg\min} \| \boldsymbol{X} - \mathbf{F}_1\boldsymbol{B}_1^{\top} \|_{\mathbb{F}}^2 \\ &\text{subject to } \frac{1}{n}\boldsymbol{F}_1^{\top}\boldsymbol{F}_1 = \boldsymbol{I}_{\hat{K}_1} \text{ and } \boldsymbol{B}_1^{\top}\boldsymbol{B}_1 \text{ is diagonal},\end{aligned} \quad (14)$$

where the constraints are designed for the identifiability of factors and $\hat{K}_1$ is the estimator of the number of factors, estimated by the eigenvalue ratio approach (Fan et al., 2024). Specifically, the columns of $\boldsymbol{F}_1/\sqrt{n}$ are eigenvectors corresponding to the largest $\hat{K}_1$ eigenvalues of the matrix $\boldsymbol{X}\boldsymbol{X}^\top$, and thus the least squares estimator of $\boldsymbol{B}_1$ is given by $\boldsymbol{B}_1^\top = (\boldsymbol{F}_1^\top \boldsymbol{F}_1)^{-1}\boldsymbol{F}_1^\top \boldsymbol{X} = n^{-1}\boldsymbol{F}_1^\top \boldsymbol{X}$. Then $\boldsymbol{V}_1 = \boldsymbol{X} - \boldsymbol{F}_1\boldsymbol{B}_1^\top = (\boldsymbol{I}_n - n^{-1}\boldsymbol{F}_1\boldsymbol{F}_1^\top)\boldsymbol{X}$. In the same fashion, $\hat{K}_2$, $\boldsymbol{F}_2$, $\boldsymbol{B}_m$, $\boldsymbol{B}_x$, $\boldsymbol{V}_{2m}$ and $\boldsymbol{V}_{2x}$ can also be obtained. Furthermore, we can also replace the constrained least squares method with alternatives, including maximum likelihood estimation (Bai and Li, 2012), principal component method (Bai, 2003), and random projection method (Fan and Liao, 2022), since the proposed method is not restricted to any specific estimation approach.

Based on estimated factors and associated loading matrices, we can further estimate direct and indirect effects. Define $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\boldsymbol{W}_{-g} = (\boldsymbol{V}_{2m}, \boldsymbol{V}_{2x,-g})$, where $\boldsymbol{V}_{2x,-g}$ is the sub-matrix of $\boldsymbol{V}_{2x}$ except for the *g*-th group. Next, we solve the following partially penalized least squares problem to obtain the estimator of the direct effect $\hat{\boldsymbol{\alpha}}_{1,g}$ corresponding to the *g*th group of exposures,

$$(\hat{\boldsymbol{\theta}}_{-g}, \hat{\boldsymbol{\alpha}}_{1,g}, \hat{\phi}_2) = \underset{\boldsymbol{\theta}_{-g}, \boldsymbol{\alpha}_{1,g}, \phi_2}{\arg\min} \frac{1}{2n} \| \boldsymbol{y} - \hat{\boldsymbol{W}}_{-g}\boldsymbol{\theta}_{-g} - \boldsymbol{V}_{2x,g}\boldsymbol{\alpha}_{1,g} - \boldsymbol{F}_2\phi_2 \|^2 + \sum_{j=1}^{p+q-|\mathcal{I}_g|} p_{\lambda_1}(|\theta_{-g,j}|), \quad (15)$$

where $p_{\lambda_1}(\cdot)$ is a penalty function with a tuning parameter λ_1, and $\boldsymbol{\theta}_{-g} = (\theta_{-g,j})_{j=1}^{p+q-|\mathcal{I}_g|}$ with $\theta_{-g,j}$ being the *j*th element of $\boldsymbol{\theta}_{-g}$. Note that we only penalize the reunified high-dimensional sparse vector $\boldsymbol{\theta}_{-g}$ defined in (13), excluding $\boldsymbol{\alpha}_{1,g}$ to avoid introducing additional estimation bias and achieve local power for direct effects.

We further propose another partially penalized least squares procedure to estimate the total effect $\hat{\boldsymbol{\gamma}}_g$ for the *g*th group,

$$(\hat{\boldsymbol{\gamma}}_{-g}, \hat{\boldsymbol{\gamma}}_g, \hat{\phi}_1) = \underset{\boldsymbol{\gamma}_{-g}, \boldsymbol{\gamma}_g, \phi_1}{\arg\min} \frac{1}{2n} \| \boldsymbol{y} - \hat{\boldsymbol{U}}_{-g}\boldsymbol{\gamma}_{-g} - \boldsymbol{V}_{1,g}\boldsymbol{\gamma}_g - \hat{\boldsymbol{F}}_1\phi_1 \|^2 + \sum_{j=1}^{q-|\mathcal{I}_g|} p_{\lambda_2}(|\gamma_{-g,j}|), \quad (16)$$

where $\gamma_{-g,j}$ is the *j*th element of $\boldsymbol{\gamma}_{-g}$. Similarly, we only penalize the high-dimensional vector $\boldsymbol{\gamma}_{-g}$ while not for $\boldsymbol{\gamma}_g$. To compute $\hat{\boldsymbol{\alpha}}_{1,g}$ and $\hat{\boldsymbol{\gamma}}_g$ and select the tuning parameters λ_1 and λ_2 in (15)-(16), we provide more details explicitly in Section S2.

Therefore, the indirect effect $\boldsymbol{\beta}_g$ is calculated as $\boldsymbol{\beta}_g = \hat{\boldsymbol{\gamma}}_g - \hat{\boldsymbol{\alpha}}_{1,g}$. Repeating these procedures above for each group with $g = 1, \ldots, G$, we obtain the estimators for direct effects $\hat{\boldsymbol{\alpha}}_1$ with $\{\hat{\boldsymbol{\alpha}}_{1,g}\}_{g=1}^G$ and indirect effects $\hat{\boldsymbol{\beta}}$ with $\{\hat{\boldsymbol{\beta}}_g\}_{g=1}^G$ in the group-wise manner, given the group structure $I_1, \ldots, I_G$. It should be noted that high-dimensional direct and total effects estimation problems given in (15) and (16) can be simplified to the one-dimensional estimation

problems for q times if $|I_g|=1$ for each g with $G=q$. This factor-adjusted regularization method proposed for HD-LSEM is summarized in Algorithm 1.

Algorithm 1: The factor-adjusted regularization algorithm for HD-LSEM

Input : $\boldsymbol{X}\in\mathbb{R}^{n\times q}$, $\boldsymbol{M}\in\mathbb{R}^{n\times p}$, $\boldsymbol{Y}\in\mathbb{R}^{n}$.

1: Use the eigenvalue ratio method to estimate the number of factors $\hat{K}_1$ and $\hat{K}_2$.

2: Estimate factors and idiosyncratic components $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{V}_1$, $\boldsymbol{V}_{2m}$ and $\boldsymbol{V}_{2x}$.

3: For $g=1,\ldots,G$, estimate direct effect $\boldsymbol{\alpha}_{1,g}$ and total effect $\hat{\boldsymbol{\gamma}}_g$ in the gth group through (15) and (16).

4: The indirect effect in the gth group is $\boldsymbol{\beta}_g=\hat{\boldsymbol{\gamma}}_g-\boldsymbol{\alpha}_{1,g}$.

Output : The estimators for direct effects $\{\hat{\boldsymbol{\alpha}}_{1,g}\}_{g=1}^{G}$ and indirect effects $\{\hat{\boldsymbol{\beta}}_g\}_{g=1}^{G}$.

3.2 Asymptotic Distributions of Estimated Direct and Indirect Effects

In this subsection, we explore the asymptotic behaviors of the proposed estimators, which serve as a basis for subsequent inference. To start with, denote $\boldsymbol{\alpha}_0^\star$, $\boldsymbol{\alpha}_1^\star$, $\boldsymbol{\beta}^\star$, and $\boldsymbol{\gamma}^\star$ as the true values of $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. Additionally, $\boldsymbol{\theta}_0^\star=(\boldsymbol{\alpha}_0^{\star\top},\boldsymbol{\alpha}_1^{\star\top})^\top$. Correspondingly, we have $\boldsymbol{\alpha}_{1,g}^\star$, $\boldsymbol{\theta}_{-g}^\star$, $\boldsymbol{\gamma}_g^\star$, $\boldsymbol{\gamma}_{-g}^\star$, $\phi_1^\star$, and $\phi_2^\star$. Further, we define $\boldsymbol{\psi}_2^\star=\phi_2^\star-\boldsymbol{B}_2^\top\boldsymbol{\theta}_0^\star$ and $\boldsymbol{\psi}_1^\star=\phi_1^\star-\boldsymbol{B}_1^\top\boldsymbol{\gamma}^\star$.

Denote $\mathcal{A}_1=\{j:\theta_{-g,j}^\star\neq 0\}$, $s_1=|\mathcal{A}_1|$, $\mathcal{N}_{0,1}=\{\boldsymbol{\delta}_1\in\mathbb{R}^{s_1}:\|\boldsymbol{\delta}_1-\boldsymbol{\theta}_{-g,\mathcal{A}_1}^\star\|_2\le d_{1n}\}$ where $d_{1n}=\min_{j\in A_1}|\theta_{-g,j}^\star|/2$ is the half minimum signal of $\boldsymbol{\theta}_{-g,\mathcal{A}_1}^\star$. In addition, $\mathcal{A}_2=\{j:\gamma_{-g,j}^\star\neq 0\}$, $s_2=|\mathcal{A}_2|$, and $\mathcal{N}_{0,2}=\{\boldsymbol{\delta}_2\in\mathbb{R}^{s_2}:\|\boldsymbol{\delta}_2-\boldsymbol{\gamma}_{-g,\mathcal{A}_2}^\star\|_2\le d_{2n}\}$ where $d_{2n}=\min_{j\in A_2}|\gamma_{-g,j}^\star|/2$ is the half minimum signal of $\boldsymbol{\gamma}_{-g,\mathcal{A}_2}^\star$. Let

$$\begin{aligned}
&\boldsymbol{\Sigma}_{V_{1g}V_{1g}}=\mathrm{E}(\boldsymbol{v}_{1,g}\boldsymbol{v}_{1,g}^\top),\qquad \boldsymbol{\Sigma}_{V_{2g}V_{2g}}=\mathrm{E}(\boldsymbol{v}_{2x,g}\boldsymbol{v}_{2x,g}^\top),\\
&\boldsymbol{\Sigma}_{V_{2g}W_{-g}}=\mathrm{E}(\boldsymbol{v}_{2x,g}\boldsymbol{w}_{-g,\mathcal{A}_1}^\top),\quad \boldsymbol{\Sigma}_{V_{1g}U_{-g}}=\mathrm{E}(\boldsymbol{v}_{1,g}\boldsymbol{u}_{-g,\mathcal{A}_2}^\top),\\
&\boldsymbol{\Sigma}_{W_{-g}W_{-g}}=\mathrm{E}(\boldsymbol{w}_{-g,\mathcal{A}_1}\boldsymbol{w}_{-g,\mathcal{A}_1}^\top),\quad \boldsymbol{\Sigma}_{W_{-g}U_{-g}}=\mathrm{E}(\boldsymbol{w}_{-g,\mathcal{A}_1}\boldsymbol{u}_{-g,\mathcal{A}_2}^\top),\quad \boldsymbol{\Sigma}_{U_{-g}U_{-g}}=\mathrm{E}(\boldsymbol{U}_{-g,\mathcal{A}_2}\boldsymbol{u}_{-g,\mathcal{A}_2}^\top).
\end{aligned}$$

Denote

$$\Sigma_{1g}=\begin{pmatrix}\Sigma_{V_{2g}V_{2g}} & \Sigma_{V_{2g}W_{-g}}\\ \Sigma_{W_{-g}V_{2g}} & \Sigma_{W_{-g}W_{-g}}\end{pmatrix},\ \Sigma_{2g}=\begin{pmatrix}\Sigma_{V_{1g}U_{-g}} & \Sigma_{V_{1g}U_{-g}}\\ \Sigma_{U_{-g}V_{1g}} & \Sigma_{U_{-g}U_{-g}}\end{pmatrix}.$$

Assumption 1 .

There exists some positive constants $c_2 > c_1 > 0$, *such that* $c_1 \le \lambda_{\min}(\Sigma_{1g}) \le \lambda_{\max}(\Sigma_{1g}) \le c_2$, *and* $c_1 \le \lambda_{\min}(\Sigma_{2g}) \le \lambda_{\max}(\Sigma_{2g}) \le c_2$. *Furthermore, assume* $\|(\boldsymbol{M}, \boldsymbol{X})\|_2 = O_P(\sqrt{n})$, $\|\boldsymbol{W}_{-g,\mathcal{A}_1^c}^\top(\boldsymbol{M}, \boldsymbol{X})\|_{2,\infty} = O_P(n)$ *and* $\|\boldsymbol{U}_{-g,\mathcal{A}_2^c}^\top \boldsymbol{X}\|_{2,\infty} = O_P(n)$.

Assumption 2 .

The penalty function $p_\lambda(t_0)$ *is increasing and concave in* $t_0 \in [0,\infty)$, *and has a continuous derivative* $p_\lambda'(t_0)$ *with* $p_\lambda'(0+) > 0$. *Denote* $\rho(t_0,\lambda) = p_\lambda(t_0)/\lambda$ *for* $\lambda > 0$. *Assume that* $\rho'(t_0,\lambda)$ *increases in* $\lambda \in (0,\infty)$ *and* $\rho'(0+,\lambda)$ *do not depend on* λ.

Assumption 3 .

Assume that $d_{1n} \gg \lambda_{1n} \gg \|\boldsymbol{\psi}_2^\star\|_2 \cdot \max\left\{\sqrt{\frac{\log(p+q)}{n}}, \frac{1}{p+q-s_1}, \sqrt{A}, \sqrt{(p+q-s_1)BA}\right\}$ *and* $d_{2n} \gg \lambda_{2n} \gg \|\boldsymbol{\psi}_1^\star\|_2 \cdot \max\{\sqrt{\log q/n}, 1/(q-s_2)\sqrt{C}\ \sqrt{(q-s_2)DC}\}$, *where* λ_{ln}, $l = 1,2$, *are tuning parameters in the objective functions (15) and (16), to emphasize their dependence on n, and* $A = \frac{s_1^3 \log s_1}{n} + \frac{s_1^3}{p+q}$, $B = \frac{\log(p+q-s_1)}{n} + \frac{1}{p+q}$, $C = \frac{s_2^3 \log s_2}{n} + \frac{s_2^3}{q}$, $D = \frac{\log(q-s_2)}{n} + \frac{1}{q}$. *In addition, for a vector* $\boldsymbol{v} = (v_1,...,v_h)^\top$, *define* $\bar{\rho}(\boldsymbol{v},\lambda) = \{\mathrm{sgn}(v_1)\rho'(|v_1|,\lambda),...,\mathrm{sgn}(v_h)\rho'(|v_h|,\lambda)\}^\top$, *where* $\mathrm{sgn}(\cdot)$ *is the sign function. Further define the local concavity of* $\rho(\cdot,\lambda)$ *at* $\boldsymbol{v}$ *as*

$$\kappa(\rho,\boldsymbol{v},\lambda) = \lim_{\epsilon\to 0^+} \max_{1\le j\le h} \sup_{t_1<t_2\in(|v_j|-\epsilon,|v_j|+\epsilon)} \frac{\rho'(t_2,\lambda)-\rho'(t_1,\lambda)}{t_2-t_1}.$$

We assume that $p_{\lambda_{ln}}'(d_{ln}) = o((ns_l)^{-1/2})$ *and* $\lambda_{ln}\omega_l = o(1)$ *where* $\omega_l = \max_{\delta\in\mathcal{N}_{0,l}} \kappa(\rho,\boldsymbol{\delta},\lambda_{ln})$ *for* $l = 1,2$.

Assumption 4 .

For some $\varpi_1, \varpi_2 > 2$, *there exist positive sequences* R_{1n}, R_{2n} *such that* $\mathrm{E}[\|\boldsymbol{w}_{-g,\mathcal{A}_1^c}\varepsilon_1\|_\infty^{\varpi_1}] \le R_{1n}^{\varpi_1}$, $\mathrm{E}[\|\boldsymbol{U}_{-g,\mathcal{A}_2^c}\varepsilon_3\|_\infty^{\varpi_2}] \le R_{2n}^{\varpi_2}$, $R_{1n}^2 \log(p+q-|\mathcal{I}_g|)/n^{1-2/\varpi_1-\zeta_1} \to 0$, *and* $R_{2n}^2 \log(q-|\mathcal{I}_g|)/n^{1-2/\varpi_2-\zeta_2} \to 0$ *for some arbitrary small* ζ_1 *and* ζ_2. *Further we assume that there exists a constant* c_3, *such that*

$$\max_{1\le j\le p+q} \mathrm{E}(z_{1j}^4) < c_3 < \infty, \max_{1\le j\le q} \mathrm{E}(z_{2j}^4) < c_3 < \infty, \max_{1\le j\le s_1} E[\boldsymbol{\theta}_{-g,\mathcal{A}_1}^\star]_j^2 < c_3 < \infty, \max_{1\le j\le s_2} E[\boldsymbol{\gamma}_{-g,\mathcal{A}_2}^\star]_j^2 < c_3 < \infty,$$

where z_{1j}, z_{2j}, $[\boldsymbol{\theta}^{\star}_{-g,\mathcal{A}_1}]_j$, $[\boldsymbol{\gamma}^{\star}_{-g,\mathcal{A}_2}]_j$ *are the jth components of* $\boldsymbol{z}_1 = (\boldsymbol{w}^{\top}_{-g}, \boldsymbol{v}^{\top}_{2x,g})^{\top}$, $\boldsymbol{z}_2 = (\boldsymbol{u}^{\top}_{-g}, \boldsymbol{v}^{\top}_{1,g})^{\top}$, $\boldsymbol{\theta}^{\star}_{-g,\mathcal{A}_1}$ *and* $\boldsymbol{\gamma}^{\star}_{-g,\mathcal{A}_2}$, *respectively.*

Assumptions 1 and 2 are mild and commonly imposed in high-dimensional modeling, and more details are referred to Fan and Lv (2011). Assumption 3 imposes minimal signal conditions on nonzero elements in $\boldsymbol{\theta}^{\star}_{-g}$ and $\boldsymbol{\gamma}^{\star}_{-g}$. It should be noted that estimating the direct effect $\boldsymbol{\alpha}_{1,g}$ and the indirect effect $\boldsymbol{\beta}_g$ is of our interest, and minimal signal conditions on these effects themselves are not required since they are imposed only on the nuisance parameters $\boldsymbol{\theta}^{\star}_{-g}$ and $\boldsymbol{\gamma}^{\star}_{-g}$. Assumption 4 is needed for establishing consistency and sparsity results. Compared with the existing literature, this assumption is very mild. For more details, see Wang et al. (2012). In addition, we also suppose some mild conditions hold for factor estimation, which are given in Assumptions S1-S4 and deferred to Section S5 in the supplementary file.

Then we establish the asymptotic normality for the estimators of the direct and indirect effects.

Theorem 3.1 .

Suppose Assumptions 1-4 and S1-S4 hold, and $s_1^2 \log s_1 = o(n)$, $s_1 = o(p+q)$. *Then with probability tending to 1,* $\boldsymbol{\theta}_{-g}$ *must satisfy (i)* $\boldsymbol{\theta}_{-g,\mathcal{A}_1^c} = 0$*; (ii)* $\| \boldsymbol{\theta}_{-g,\mathcal{A}_1} - \boldsymbol{\theta}^{\star}_{-g,\mathcal{A}_1} \|_2 = O_P\left(\sqrt{s_1^3 \log s_1 / n + s_1^3/(p+q)}\right)$. *If further* $s_1^3 (\log(p+q))^2 (\log n)^2 = o(n)$ *and* $n s_1^3 (\log n)^2 = o((p+q)^2)$ *hold, we have*

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_{1,g} - \boldsymbol{\alpha}^{\star}_{1,g})$$
$$= \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}_{V_{2g}V_{2g}} \boldsymbol{V}^{\top}_{2x,g} \boldsymbol{\epsilon}_1 + \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}_{V_{2g}V_{2g}} \boldsymbol{\Sigma}_{V_{2g}W_{-g}} \boldsymbol{\Sigma}^{-1}_{W_{-g}W_{-g}\cdot V_{2g}} (\boldsymbol{\Sigma}_{W_{-g}V_{2g}} \boldsymbol{\Sigma}^{-1}_{V_{2g}V_{2g}} \boldsymbol{V}^{\top}_{2x,g} - \mathbf{W}^{\top}_{-g,\mathcal{A}_1}) \boldsymbol{\epsilon}_1 + o_P(1),$$

where $\boldsymbol{\epsilon}_1 = (\varepsilon_{11}, \varepsilon_{21}, ..., \varepsilon_{n1})^{\top}$ *and* $\Sigma_{W_{-g}W_{-g}\cdot V_{2g}} = \Sigma_{W_{-g}W_{-g}} - \Sigma_{W_{-g}V_{2g}} \Sigma^{-1}_{V_{2g}V_{2g}} \Sigma_{V_{2g}W_{-g}}$. *Based on the asymptotic behavior, we can derive the asymptotic distribution of* $\boldsymbol{\alpha}_{1,g}$, *as*

$$\sqrt{n}(\boldsymbol{\alpha}_{1,g} - \boldsymbol{\alpha}_{1,g}) \xrightarrow{d} N\left(0, \sigma_1^2 (\Sigma^{-1}_{V_{2g}V_{2g}} + \Sigma^{-1}_{V_{2g}V_{2g}} \Sigma_{V_{2g}W_{-g}} \Sigma^{-1}_{W_{-g}W_{-g}\cdot V_{2g}} \Sigma_{W_{-g}V_{2g}} \Sigma^{-1}_{V_{2g}V_{2g}})\right).$$

Theorem 3.1 establishes the sparsity of $\boldsymbol{\theta}_{-g}$, the convergence rate of $\boldsymbol{\theta}_{-g,\mathcal{A}_1}$, and the asymptotic distribution of $\boldsymbol{\alpha}_{1,g}$. Similarly, we can derive the sparsity of $\hat{\boldsymbol{\gamma}}_{-g}$, the convergence rate of $\hat{\boldsymbol{\gamma}}_{-g,\mathcal{A}_2}$, and the asymptotic distribution of $\hat{\boldsymbol{\gamma}}_g$ in the following Proposition 3.2.

Proposition 3.2 .

Suppose Assumptions 1-4 and S1-S4 hold, and $s_2^2 \log s_2 = o(n)$, $s_2^2 = o(q)$. *Then with probability tending to 1,* $\hat{\boldsymbol{\gamma}}_{-g}$ *must satisfy (i)* $\hat{\boldsymbol{\gamma}}_{-g,\mathcal{A}_2^c} = 0$*; (ii)* $\| \hat{\boldsymbol{\gamma}}_{-g,\mathcal{A}_2} - \boldsymbol{\gamma}^{\star}_{-g,\mathcal{A}_2} \|_2 = O_P\left(\sqrt{s_2^3 \log s_2 / n + s_2^3 / q}\right)$. *If further* $s_2^3 (\log q)^2 (\log n)^2 = o(n)$ *and* $n s_2^3 (\log n)^2 = o(q^2)$ *hold, we obtain that*

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_g - \boldsymbol{\gamma}_g^{\star}) = \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{V}_{1,g}^{\top} \boldsymbol{\epsilon}_3 + \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{\Sigma}_{V_{1g}U_{-g}} \boldsymbol{\Sigma}^{-1}_{U_{-g}U_{-g}\cdot V_{1g}} (\boldsymbol{\Sigma}_{U_{-g}V_{1g}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{V}_{1,g}^{\top} - \boldsymbol{U}^{\top}_{-g,\mathcal{A}_2}) \boldsymbol{\epsilon}_3 + o_P(1),$$

where $\boldsymbol{\epsilon}_3 = (\varepsilon_{13}, \varepsilon_{23}, ..., \varepsilon_{n3})^{\top}$ *and* $\Sigma_{U_{-g}U_{-g}\cdot V_{1g}} = \Sigma_{U_{-g}U_{-g}} - \Sigma_{U_{-g}V_{1g}} \Sigma^{-1}_{V_{1g}V_{1g}} \Sigma_{V_{1g}U_{-g}}$. *Furthermore, we can also obtain the asymptotic distribution of* $\hat{\boldsymbol{\gamma}}_g$*,*

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_g - \boldsymbol{\gamma}_g^{\star}) \xrightarrow{d} N\left(0, (\sigma_1^2 + \sigma_2^2)\left(\Sigma^{-1}_{V_{1g}V_{1g}} + \Sigma^{-1}_{V_{1g}V_{1g}} \Sigma_{V_{1g}U_{-g}} \Sigma_{U_{-g}U_{-g}\cdot V_{1g}} \Sigma_{U_{-g}V_{1g}} \Sigma^{-1}_{V_{1g}V_{1g}}\right)\right).$$

Given Theorem 3.1 and Proposition 3.2, we further obtain the asymptotic normality of $\hat{\boldsymbol{\beta}}_g$ in Theorem 3.3.

Theorem 3.3 .

Suppose the conditions of Theorem 3.1 and Proposition 3.2 are satisfied, then

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^{\star}) &= \frac{1}{\sqrt{n}} (\boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{V}_{1,g}^{\top} - \boldsymbol{\Sigma}^{-1}_{V_{2g}V_{2g}} \boldsymbol{V}_{2x,g}^{\top}) \boldsymbol{\epsilon}_1 + \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{V}_{1,g}^{\top} \boldsymbol{\epsilon}_2 \\ &+ \frac{1}{\sqrt{n}} (\boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{\Sigma}_{V_{1g}U_{-g}} \boldsymbol{\Sigma}^{-1}_{U_{-g}U_{-g}\cdot V_{1g}} (\boldsymbol{\Sigma}_{U_{-g}V_{1g}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{V}_{1,g}^{\top} - \mathbf{U}^{\top}_{-g,\mathcal{A}_2}) \\ &- \boldsymbol{\Sigma}^{-1}_{V_{2g}V_{2g}} \boldsymbol{\Sigma}_{V_{2g}W_{-g}} \boldsymbol{\Sigma}^{-1}_{W_{-g}W_{-g}\cdot V_{2g}} (\boldsymbol{\Sigma}_{W_{-g}V_{2g}} \boldsymbol{\Sigma}^{-1}_{V_{2g}V_{2g}} \boldsymbol{V}_{2x,g}^{\top} - \mathbf{W}^{\top}_{-g,\mathcal{A}_1})) \boldsymbol{\epsilon}_1 \\ &+ \frac{1}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{\Sigma}_{V_{1g}U_{-g}} \boldsymbol{\Sigma}^{-1}_{U_{-g}U_{-g}\cdot V_{1g}} (\boldsymbol{\Sigma}_{U_{-g}V_{1g}} \boldsymbol{\Sigma}^{-1}_{V_{1g}V_{1g}} \boldsymbol{V}_{1,g}^{\top} - \mathbf{U}^{\top}_{-g,\mathcal{A}_2}) \boldsymbol{\epsilon}_2 + o_P(1).\end{aligned}$$

Furthermore, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_g$*, as* $\sqrt{n}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^{\star}) \xrightarrow{d} N(0, \boldsymbol{V}_{\beta})$*, where* $\boldsymbol{V}_{\beta}$ *is given in Section S6 of the supplementary file.*

3.3 Test for Direct and Indirect Effects

For the direct effect, testing the hypothesis in (10) essentially is to test low-dimensional regression coefficients in the high-dimensional linear regression model (13). We apply the F-type test, and establish the test statistic

$$T_n = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)}{\mathrm{RSS}_1 / (n - |\mathcal{I}_g|)}, \quad (17)$$

where RSS_0 is the residual sum of squares under $H_{0,\mathrm{dir}}^{(g)}$, i.e., $\mathrm{RSS}_0 = \| \boldsymbol{y} - \hat{\boldsymbol{W}}_{-g}\tilde{\boldsymbol{\theta}}_{-g} - \hat{\boldsymbol{F}}_2\phi_2 \|_2^2$, with

$$(\tilde{\boldsymbol{\theta}}_{-g}, \phi_2) = \arg\min_{\boldsymbol{\theta}_{-g},\phi_2} \frac{1}{2n} \| \boldsymbol{y} - \hat{\boldsymbol{W}}_{-g}\boldsymbol{\theta}_{-g} - \boldsymbol{F}_2\phi_2 \|_2^2 + \sum_{j=1}^{p+q-|\mathcal{I}_g|} p_{\lambda_1}(|\theta_{-g,j}|),$$

and $\mathrm{RSS}_1 = \| \boldsymbol{y} - \hat{\boldsymbol{W}}_{-g}\boldsymbol{\theta}_{-g} - \boldsymbol{V}_{2x,g}\hat{\boldsymbol{\alpha}}_{1,g} - \boldsymbol{F}_2\hat{\phi}_2 \|_2^2$ is the residual sum of squares under $H_{1,\mathrm{dir}}^{(g)}$, with

$$(\hat{\boldsymbol{\theta}}_{-g}, \hat{\boldsymbol{\alpha}}_{1,g}, \phi_2) = \arg\min_{\boldsymbol{\theta}_{-g},\boldsymbol{\alpha}_{1,g},\phi_2} \frac{1}{2n} \| \boldsymbol{y} - \hat{\boldsymbol{W}}_{-g}\boldsymbol{\theta}_{-g} - \boldsymbol{V}_{2x,g}\boldsymbol{\alpha}_{1,g} - \boldsymbol{F}_2\phi_2 \|_2^2 + \sum_{j=1}^{p+q-|\mathcal{I}_g|} p_{\lambda_1}(|\theta_{-g,j}|),$$

To evaluate the local power of T_n under the local alternative hypothesis, we consider the local alternative hypothesis $H_{1,\mathrm{dir},n}^{(g)} : \boldsymbol{\alpha}_{1,g} = \boldsymbol{g}_n$, where $\| \boldsymbol{g}_n \|_2 = O(\sqrt{1/n})$. Theorem 3.4 shows that the asymptotical null distribution of T_n in (17) is a chi-square distribution with $|\mathcal{I}_g|$ degrees of freedom.

Theorem 3.4 .

Suppose the conditions in Theorem 3.1 hold, and $s_1^5(\log s_1)^2 = o(n)$, $s_1^5 \log s_1 = o(p+q)$, *and* $s_1^5 n = o((p+q)^2)$, *then it follows that*

$$\sup_x \left| P(T_n \le x) - P\left(\chi^2_{|\mathcal{I}_g|}\left(n\boldsymbol{g}_n^\top \Psi^{-1}\boldsymbol{g}_n / \sigma_1^2\right) \le x\right) \right| \to 0.$$

Here $\Psi = \left(I_{|\mathcal{I}_g|}, 0_{|\mathcal{I}_g|\times s_1}\right)\Sigma_{1g}^{-1}\left(I_{|\mathcal{I}_g|}, 0_{|\mathcal{I}_g|\times s_1}\right)^\top$ *and* $\chi^2_{|\mathcal{I}_g|}\left(n\boldsymbol{g}_n^\top \Psi^{-1}\boldsymbol{g}_n / \sigma_1^2\right)$ *is a chi-squared random variable with* $|\mathcal{I}_g|$ *degrees of freedom and non-centrality parameter* $n\boldsymbol{g}_n^\top \Psi^{-1}\boldsymbol{g}_n / \sigma_1^2$.

Theorem 3.4 implicitly states that under $H_{0,\mathrm{dir},n}^{(g)}$, T_n asymptotically follows $\chi^2_{|\mathcal{I}_g|}$ distribution. It also implies that T_n can detect local alternatives that are distinct from the null hypothesis at the rate of $1/\sqrt{n}$.

For testing the indirect effect in (11), we observe that the indirect effect involves both layers of models in HD-LSEM, making the development of the lack-of-fit type of test like T_n challenging, if not infeasible, to implement. Therefore, we adopt the Wald test statistic,

$$S_n = n\hat{\boldsymbol{\beta}}_g^\top \boldsymbol{V}_\beta^{-1}\hat{\boldsymbol{\beta}}_g, \quad (18)$$

where $\boldsymbol{V}_\beta$ is a consistent estimator of $\boldsymbol{V}_\beta$ derived in Section S6. Under $H_{0,\mathrm{ind}}^{(g)}$, S_n asymptotically follows $\chi^2_{|\mathcal{I}_g|}$. As for the local power of S_n, we consider the local alternative hypothesis $H_{1,\mathrm{ind},n}^{(g)} : \boldsymbol{\beta} = \Delta/\sqrt{n}$, where Δ is a constant vector.

Theorem 3.5 .

Suppose the conditions in Theorem 3.3 hold. Then under the local alternative hypothesis $H_{1,\text{ind},n}^{(g)}$, S_n *converges to* $\chi^2_{|\mathcal{I}_g|}(\Delta^\top(\boldsymbol{V}_{\boldsymbol{\beta}})^{-1}\Delta)$, *a chi-square distribution with* $|\mathcal{I}_g|$ *degrees of freedom and noncentrality parameter* $\Delta^\top(\boldsymbol{V}_{\boldsymbol{\beta}})^{-1}\Delta$.

From Theorem 3.5, S_n can also detect local indirect effects that converge to 0 at the rate of $1/\sqrt{n}$.

Last but not least, in order to control false discovery rate (FDR) for the multiple hypotheses in (10) and (11) for $g = 1, \ldots, G$, we employ the Benjamini–Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001). More details and discussions are provided in Section S4 of the supplementary file.

4 Simulation Studies

In this section, we conduct simulation studies to assess the finite sample performance of PRIME. We compare the proposed method with the oracle one (Oracle), where the true sparsity structure is known, the two-step partially penalized regression method without imposing the factor structure (NoFactor), and the naive two-step SCAD method (NVSCAD) which penalizes all elements of the high-dimensional vectors. We also consider HILAMA proposed by Wang et al. (2025) for testing the direct effects, though it fails to test the indirect effects. The details of implementing NoFactor and NVSCAD are provided in Section S3 of the supplementary file. We consider different cases for generating error distributions when $|\mathcal{I}_g| = 1$, including normal distributions in Examples 1 and 3 and heavy-tailed distribution in Example 2, with different sample size and dimensions of exposures and mediators. In Example 4, we consider the case where $|\mathcal{I}_g| > 1$.

Example 1 Set $n = 300$, $p = 500$, and $q = 400$. Generate $\boldsymbol{x} \sim N(0, \Sigma_x)$, where $\Sigma_x = (\sigma_{ij})_{q\times q}$, with $\sigma_{ij} = 0.3^{|i-j|}$. Randomly select 10 elements from $\boldsymbol{\alpha}_1$ and set them to be $\pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8, \pm 1$, leaving all other elements zero. We also generate a sparse $\boldsymbol{\alpha}_0$ by randomly selecting its 5 nonzero elements and setting them to be 0.1,0.2,0.3,0.4 and 0.5. Denote their position indices in $\boldsymbol{\alpha}_0$ as i_1, i_2, i_3, i_4 and i_5. In the model $\boldsymbol{m} = \Gamma^\top \boldsymbol{x} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0, \Sigma_{\boldsymbol{\varepsilon}})$, $\Sigma_{\boldsymbol{\varepsilon}}$ is generated from the same pattern as $\Sigma_{\boldsymbol{x}}$. To generate a sparse $\Gamma = (\gamma_{ij})_{q\times p}$, we set 10 nonzero elements in the columns indexed by i_1, i_2, i_3, i_4 and i_5. Specifically, we select 10 row indices $j_1, j_2, \ldots, j_{10}$, each two of which correspond to one column mentioned above. We set $\gamma_{j_{2m-1}, i_m} = 2$ and $\gamma_{j_{2m}, i_m} = -2$ for $m = 1, 2, 3, 4, 5$. By this means, we can naturally obtain a sparse $\boldsymbol{\beta}$, whose nonzero elements are $\pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8, \pm 1$ on the columns indexed by $j_1, j_2, \ldots, j_{10}$. Finally, we generate the response *y* from the model $y = \boldsymbol{\alpha}_0^\top \boldsymbol{m} + \boldsymbol{\alpha}_1^\top \boldsymbol{x} + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 0.5^2)$. Without loss of generality and for notation simplicity, we fix $G = q$ and investigate each exposure separately.

Example 2 In this example, we generate the heavy-tailed errors with $\varepsilon_1 \sim t_6/\sqrt{6}$. The multiplier $\sqrt{6}$ ensures that the variance of ε_1 is equal to that of $\varepsilon_1 \sim N(0, 0.5^2)$. All other settings follow Example 1.

Example 3 In this example, we consider the high dimensional case where $n = 600$, $p = 1000$, and $q = 800$. Assume that $\boldsymbol{x} \sim N(0, \Sigma_{\boldsymbol{x}})$ where $\Sigma_{\boldsymbol{x}} = (\sigma_{ij})_{q\times q}$ with $\sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le q$. For the model $\boldsymbol{m} = \Gamma^\top \boldsymbol{x} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0, \Sigma_{\boldsymbol{\varepsilon}})$, $\Sigma_{\boldsymbol{\varepsilon}}$ is generated from the same pattern as $\Sigma_{\boldsymbol{x}}$. That is to say, we increase the correlation within $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ compared with Example 1. We randomly select 10 elements from $\boldsymbol{\alpha}_1$ and set them to be $\pm 0.1, \pm 0.2, \pm 0.3, \pm 0.4, \pm 0.5$; other elements are zero. All other settings are identical to those in Example 1.

Example 4 In this example, we consider the case when prior grouping strategies are available, i.e., $|\mathcal{I}_g| > 1$. Specifically, we set $|\mathcal{I}_g| = 4$ by letting $x_{4(i-1)+1}$, $x_{4(i-1)+2}$, $x_{4(i-1)+3}$, $x_{4(i-1)+4}$ be the ith group, for $i = 1, 2, ..., 100$. Set $n = 300$, $p = 500$, and $q = 400$. Generate $\boldsymbol{x} \sim N(0, \Sigma_{\boldsymbol{x}})$, where $\Sigma_{\boldsymbol{x}} = (\sigma_{ij})_{q\times q}$, with $\sigma_{ij} = 0.3^{|i-j|}$. We randomly select 3 groups of $\boldsymbol{\alpha}_1$ and set their signal strength to be 0.5,1,1.5, whose signs are randomly assigned. We also generate a sparse $\boldsymbol{\alpha}_0$ by randomly selecting its 3 nonzero elements and setting them to be 1. Denote the nonzero indices in $\boldsymbol{\alpha}_0$ as i_1, i_2, i_3. In the model $\boldsymbol{m} = \Gamma^\top \boldsymbol{x} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0, \Sigma_{\boldsymbol{\varepsilon}})$, $\Sigma_{\boldsymbol{\varepsilon}}$ is generated from the same pattern as $\Sigma_{\boldsymbol{x}}$. To generate a sparse $\Gamma = (\gamma_{ij})_{q\times p}$, we set 12 nonzero elements in the columns indexed by i_1, i_2, i_3. Specifically, we select 12 row indices $j_1, j_2, \ldots, j_{12}$, each four of which correspond to one column mentioned above. $\gamma_{j_{4(m-1)+1}, i_m}$, $\gamma_{j_{4(m-1)+2}, i_m}$, $\gamma_{j_{4(m-1)+3}, i_m}$ and $\gamma_{j_{4(m-1)+4}, i_m}$ are randomly set to be $\pm 0.5, \pm 1, \pm 1.5$ for $m = 1, 2, 3$, respectively. By this means, we can naturally obtain a sparse $\boldsymbol{\beta}$, whose signal strength of the three nonzero groups are $0.5, 1, 1.5$ on the columns indexed by $j_1, j_2, \ldots, j_{12}$. And their signs are randomly set. Finally, we generate the response y from the model $y = \boldsymbol{\alpha}_0^\top \boldsymbol{m} + \boldsymbol{\alpha}_1^\top \boldsymbol{x} + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 0.5^2)$. We fix $G = q/|\mathcal{I}_g| = 100$ and investigate each group separately.

Based on 500 simulation runs, we evaluate the mean square error (MSE) and the bias of the estimated direct and indirect effect of exposures in Example 1-3 with signal s in Tables 1-2, S1-S6, and S8-S11. The signal s indeed refers to the true value of the relevant direct or indirect effect. In Examples 1 and 2, $s \in \{0, \pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8, \pm 1\}$ for both direct effects and indirect effects; in Example 3, $s \in \{0, \pm 0.1, \pm 0.2, \pm 0.3, \pm 0.4, \pm 0.5\}$ for direct effects and $s \in \{0, \pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8, \pm 1\}$ for indirect effects. The MSE is given as

$$\mathrm{MSE}(s) = \frac{\sum_{r=1}^{K} (\hat{s}^r - s)^2}{K},$$

where K denotes the number of signals (i.e., direct or indirect effects) with value s across 500 repeated experiments and $\hat{s}^r$ represents the estimated value of s at the r-th run. Furthermore, we assess the power and size as the proportion of rejections among all single tests where the

signal strength (i.e., magnitude of true signal) equals $|s|$ across 500 replications, as depicted in Figures 2 and S1-S3. Additionally, the nominal false discovery rate (FDR) is fixed at 0.1, and the false discovery proportion (FDP) in a single experiment is calculated as the ratio of false discoveries to the total number of rejections among q tests. Then the FDR is estimated by the average of 500 FDPs, as reported in Tables 3, S7, S12, and S13.

Our proposed method PRIME consistently performs well for all the examples in terms of both estimation and inference. Specifically, PRIME is almost as good as the oracle one, while the competitors exhibit much larger MSE as well as significant fluctuation, especially for NoFactor and NVSCAD. This is possibly due to the fact that NoFactor fails to remove the high-correlation and NVSCAD employs the penalization method for all elements thus brings about non-negligible bias. This further leads to a very slow convergence rate along the power curves.

For the inference of direct and indirect effects, we mainly demonstrate the results in Example 1, and Examples 2-3 follow the similar pattern. In Figure 2(a) for testing direct effects, we observe that the power of the proposed PRIME method quickly approaches 1 as the signal strength increases, which is quite similar with the oracle. In particular, the test in NoFactor performs as good as the proposed one in PRIME for direct effects in Examples 1-3. However, the FDR of NoFactor significantly exceeds the nominal level, while PRIME controls FDR well. In Figure 2(b) for testing indirect effects, only PRIME method demonstrates good estimation accuracy and achieves high testing power. It is worth noting that PRIME also achieves better performance in the high-dimensional case, particularly in terms of the power of testing indirect effects. In Example 4, We observe that the power of PRIME is comparable to that of Oracle, and the FDR is also well controlled when the grouping information is known as prior. In conclusion, PRIME not only controls FDR for testing both direct and indirect effects under the nominal level, but also achieves comparable power with the oracle method across all settings.

All simulations were done using R 4.4.3 and the CPU employed is Core(TM) i7-1165G7. We also compare the running time of the proposed PRIME with HILAMA. In Example 1-2 to estimate the direct effect, our method requires approximately 3.7 minutes, while HILAMA takes more than twice of the time. In Example 3, PRIME takes 17.6 minutes, while HILAMA requires almost 1.2 hours. These results show the proposed PRIME is much more computationally efficient than HILAMA, which is possibly due to the fact that HILAMA involves double debiasing estimation procedures and PRIME is feasible with the partially least square estimation method in large-scale scenarios.

5 Real Data Analysis

In this section, we apply the proposed PRIME method to analyze the effect of genes on Alzheimer's disease (AD) through the brain activity intensity in an imaging genetic study. This study helps us to understand the casual effects of genotype on AD, which enables the realization of effective prevention and treatment for the disease.

The data are available at https://adni.loni.usc.edu. This study involves a total of 235 subjects, including 69 normal ones (NORM), 117 mild cognitive impairment ones (MCI), and 49 AD patients. The outcome variable is the neuropsychiatric inventory score (NPISCORE), which is a commonly used self-assessment measure in neurology to evaluate the psychological and

behavior symptoms of AD's patients. The higher the NPISCORE, the more severe the mental condition of the individual. For brain activity intensity, raw data contain PET images registered in a standard brain temple (MNI 2mm) consisting of 116 automated anatomical labeling (AAL) regions with 185,405 voxels. Each voxel in a 3D coordinate has its corresponding brain region. We record the average intensity of each brain region as the measure for 116 mediators. The NPISCORE and PET image data are monitored at three time points: baseline (bl), 6th month (m6), and 12th month (m12). For genotypes, at each time period, we calculate the Pearson correlation coefficients between SNPs and NPISCORE and select the top 500 SNPs with the strongest correlations.

We apply PRIME to test the direct and indirect effects of genetic variants on AD at different time points. Table 4 reports the number of significant SNPs at each time point. We observe that the number of significant SNPs identified by PRIME gradually decreases over time, which suggests that in the early stages of the disease, AD can be prevented or treated well. Treatment methods include stimulating certain regions of the brain, such as the left postcentral gyrus. To illustrate the similarities and difference of SNPs in different time periods, we display the Venn graphs in Figures 3-4. Specifically, it is worthy to note that most SNPs affect AD directly and one SNP(rs4083115) consistently exerts significant direct effects on AD during all the time (Figure 4). Interestingly, we observe that 97 SNPs out of 182 SNPs have significant indirect effects at bl, but they start to significantly affect AD in a direct way at m6. Among these 97 SNPs, only 2 SNPs retain significant direct effects through the m12 period, shown in Figure 3. On the other hand, 4 SNPs out of 76 SNPs have significant indirect effects at m6, but they become significantly to affect AD directly at m12. Therefore, early detection of disease signals is crucial for treatment, since the effect pathways of SNPs can be interrupted in the early stages of the disease, such as electrical stimulation of certain brain regions. However, in later stages, these SNPs start to directly lead to AD, which complicates treatment efforts.

Further, we also provide those SNPs that have been studied in previous literature among all significant ones, as well as their estimated direct and indirect effects (denoted as DE and IE) and the corresponding p-values in Tables 5-7. In these tables, bold text indicates SNPs with significant effects across multiple time periods, while "–" denotes that the corresponding effect is not significant. The scientific evidence of these SNPs are presented in Table S7 and the full list of significant SNPs are provided in Section S7. In Table S7, Chr. refers to the chromosome index for each SNP, and a tick implies the corresponding effect is significant. We also provide the corresponding descriptions of these SNPs and relevant reference in Table S7. To illustrate direct and indirect effects of SNPs on AD mediated by brain regions at different time periods, we only display the summarized mediation pathways in Figure 5, which corresponds to Tables 5-7. The complete pathways shown in Figure 5 are given in Figures S4-S6, respectively. Therefore, for those detected SNPs that have not yet been discovered to be related to AD in the literature, further exploration is worthwhile.

Our findings not only align with previous research results, but also further clarify the way SNPs affect AD and explore how their impact changes over time. From the findings, we select typical SNPs that have a sustained impact on AD for detailed explanation. For example, rs9271170 is found to have significant direct effect on AD during bl and m6 periods, and it also has significant indirect effect at m6. It has been confirmed to exist in many genes of the HLA family. Neill et al. (1999) confirms that HLA-DRB1*03 is associated with an increased risk, while HLA-DRB1*09 may be linked to a decreased risk for the development of late-onset Alzheimer's disease, with the first detectable clinical symptoms

occurring at age 75 or older. Le Guen et al. (2023) discovers that an HLA-DRB1*04-mediated adaptive immune response may reduce the risks of Parkinson's disease (PD) and Alzheimer's disease (AD), potentially by targeting tau, offering promising therapeutic possibilities. Moreover, it is worth noting that rs16994874 exert significant indirect effect bl, while at the m12 stage, it contributes to the onset of Alzheimer's not only by influencing the brain's activity levels, but its variability is also directly associated with AD. This is verified by the fact that PPARA is a key regulator of autophagy in the clearance of $A\beta$, suggesting that gemfibrozil should be evaluated as a potential treatment for Alzheimer's disease (Luo et al., 2020). We also observe that rs267015 (gene GRIA1) has significant direct and indirect effects at bl period simultaneously, while the indirect effect becomes non-significant during the m6 period. In literature, GRIA1 is a key AMPA-type glutamate receptor essential for synaptic transmission and plasticity, and its aberrant expression suggests a disruption in synaptic function, which is a hallmark of AD pathology (Jin et al., 2024). Further, rs12541729 (gene RBPMS/THEM66) has been proved to strongly connected with many neurological disorders (Shi et al., 2023). Our study further shows that it possesses significant direct and indirect effect on AD during the m6 period, however, by the m12 period, only the direct effect remains significant, which is consistent with the previous research.

In summary, the genetic architecture between SNPs and AD changes at different time periods. Our study elucidates the dynamic impact of SNPs on AD, which offers insights for the prevention of the disease. The genetic variants not only directly associate with AD, but also indirectly affect AD via modulating brain activity, which emphasizes the significance of intervention during the mid-stage of the disease. In the late stage a few genes are detected to cause the disease by mediating some brain regions' activity intensity, and thus interventions targeting on specific brain regions may be helpful for preventing the exacerbation of the disease.

6 Conclusion

In this paper, we propose a new statistical inference procedure, called PRIME, for testing the direct and indirect effects in mediation models when both exposures and mediators are high-dimensional under a multiple testing framework. By incorporating a double-layer latent factor structure, we employ the partially penalized least squares approach to estimate the effects. The F-type and Wald type test procedures for the high-dimensional direct and indirect effects are proposed, respectively. Extensive simulation studies support the validity and efficiency of our method with competitors. We also apply the proposed statistical framework to investigate the direct effects of genetic variants on Alzheimer's disease, as well as the indirect effects mediated through changes in brain activity intensity. It is also interesting to extend the proposed method to accommodate the discrete outcome variable by considering double-layer generalized linear models or introducing general dependence measures among exposures and mediators.

Funding Acknowledgments

This research is partially supported by National Key R&D Program of China 2022YFA1003800, NNSFC 12271456, 72495122, 71988101, 12526213, 12501381, 72473114, the Ministry of Education Research in the Humanities and Social Sciences 22YJA910002, and Sichuan Science and Technology Program (2024NSFSC1393). We thank the editors and referees for their constructive comments and suggestions.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement

The data that support the findings of this study are publicly available at https://adni.loni.usc.edu.

References

Bai, J. (2003), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**(1), 135–171.

Bai, J. and Li, K. (2012), 'Statistical analysis of factor models of high dimension', *The Annals of Statistics* **40**(1), 436–465.

Baron, R. M. and Kenny, D. A. (1986), 'The moderator–mediator variable distinction in social psychological research: Conceptual, Strategic, and Statistical Considerations.', *Journal of Personality and Social Psychology* **51**(6), 1173–1182.

Benjamini, Y. and Yekutieli, D. (2001), 'The control of the false discovery rate in multiple testing under dependency', *Annals of Statistics* **29**(4), 1165–1188.

Cai, X., Wang, Q. and Zhu, Y. (2023), 'Mediation analysis with latent factors using simultaneous group-wise and parameter-wise penalization', *Stat* **12**(1), e630.

Chakrabortty, A., Nandy, P. and Li, H. (2018), 'Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models', *arXiv preprint arXiv:1809.10652* .

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D. and Lindquist, M. A. (2018), 'High-dimensional multivariate mediation with application to neuroimaging data', *Biostatistics* **19**(2), 121–136.

Chernozhukov, V., Kasahara, H. and Schrimpf, P. (2021), 'Causal impact of masks, policies, behavior on early covid-19 pandemic in the us', *Journal of Econometrics* **220**(1), 23–62.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003), 'DAVID: Database for annotation, visualization, and integrated discovery', *Genome Biology* **4**, 1–11.

Derkach, A., Pfeiffer, R. M., Chen, T.-H. and Sampson, J. N. (2019), 'High dimensional mediation analysis with latent variables', *Biometrics* **75**(3), 745–756.

Fan, J. and Liao, Y. (2022), 'Learning latent factors from diversified projections and its applications to over-estimated and weak factors', *Journal of the American Statistical Association* **117**(538), 909–924.

Fan, J., Lou, Z. and Yu, M. (2024), 'Are latent factor regression and sparse regression adequate?', *Journal of the American Statistical Association* **119**(546), 1076–1088.

Fan, J. and Lv, J. (2011), 'Nonconcave penalized likelihood with NP-dimensionality', *IEEE Transactions on Information Theory* **57**(8), 5467–5484.

Guo, X., Li, R., Liu, J. and Zeng, M. (2022), 'High-dimensional mediation analysis for selecting DNA methylation loci mediating childhood trauma and cortisol stress reactivity', *Journal of the American Statistical Association* **117**(539), 1110–1121.

Guo, X., Li, R., Liu, J. and Zeng, M. (2023), 'Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to COVID-19 pandemic', *Journal of Econometrics* **235**(1), 166–179.

Jin, K., Lv, Z., Pang, X., Zhu, C., Liu, R., Wei, Y. and Pang, C. (2024), 'Biomarkers in Alzheimer's disease progression: A longitudinal cohort study of NPTX2, GRIA1, and GRIA4', *Aging Advances* **1**(1), 52–59.

Le Guen, Y., Luo, G., Ambati, A., Damotte, V., Jansen, I., Yu, E., Nicolas, A., De Rojas, I., Peixoto Leal, T., Miyashita, A. et al. (2023), 'Multiancestry analysis of the HLA locus in Alzheimer's and Parkinson's diseases uncovers a shared adaptive immune response mediated by HLA-DRB1* 04 subtypes', *Proceedings of the National Academy of Sciences* **120**(36), e2302720120.

Luo, R., Su, L.-Y., Li, G., Yang, J., Liu, Q., Yang, L.-X., Zhang, D.-F., Zhou, H., Xu, M., Fan, Y. et al. (2020), 'Activation of PPARA-mediated autophagy reduces Alzheimer disease-like pathology and cognitive decline in a murine model', *Autophagy* **16**(1), 52–69.

Neill, D., Curran, M., Middleton, D., Mawhinney, H., Edwardson, J., McKeith, I., Ballard, C., Morris, C., Ince, P., Jaros, E. et al. (1999), 'Risk for Alzheimer's disease in older late-onset cases is associated with HLA-DRB1* 03', *Neuroscience Letters* **275**(2), 137–140.

Preacher, K. J. and Hayes, A. F. (2008), 'Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models', *Behavior Research Methods* **40**(3), 879–891.

Rucker, D. D., Preacher, K. J., Tormala, Z. L. and Petty, R. E. (2011), 'Mediation analysis in social psychology: Current practices and new recommendations', *Social and Personality Psychology Compass* **5**(6), 359–371.

Shi, C. and Li, L. (2022), 'Testing mediation effects using logic of boolean matrices', *Journal of the American Statistical Association* **117**(540), 2014–2027.

Shi, Y., Liu, Y., Wu, C., Liu, X., Hu, W., Yang, Z., Li, Z., Li, Y., Deng, C., Wei, K. et al. (2023), 'N, N-Dimethyl-3β-hydroxycholenamide attenuates neuronal death and retinal

inflammation in retinal ischemia/reperfusion injury by inhibiting Ninjurin 1', *Journal of Neuroinflammation* **20**(1), 91.

VanderWeele, T. J. (2016), 'Mediation analysis: A practitioner's guide', *Annual Review of Public Health* **37**(1), 17–32.

VanderWeele, T. and Vansteelandt, S. (2014), 'Mediation analysis with multiple mediators', *Epidemiologic Methods* **2**(1), 95–115.

Wang, L., Wu, Y. and Li, R. (2012), 'Quantile regression for analyzing heterogeneity in ultra-high dimension', *Journal of the American Statistical Association* **107**(497), 214–222.

Wang, X., Liu, J., Hu, S. S., Liu, Z., Lu, H., Liu, L. and for the Alzheimer's Disease Neuroimaging Initiative (2025), 'Hilama: High-dimensional multi-omics mediation analysis with latent confounding', *BMC Medical Research Methodology* **25**(1), 239.

Watanabe, K., Taskesen, E., Van Bochoven, A. and Posthuma, D. (2017), 'Functional mapping and annotation of genetic associations with FUMA', *Nature Communications* **8**(1), 1826.

Zhang, Q. (2022), 'High-dimensional mediation analysis with applications to causal gene identification', *Statistics in Biosciences* **14**(3), 432–451.

Zhang, Y., Yuan, Y., Zhang, Y., Zhu, Z. and Qu, A. (2024), 'Individualized dynamic mediation analysis using latent factor models', *arXiv preprint arXiv:2405.17591* .

Zhao, Y., Li, L. and Initiative, A. D. N. (2022), 'Multimodal data integration via mediation analysis with high-dimensional exposures and mediators', *Human Brain Mapping* **43**(8), 2519–2533.

Zhou, R. R., Wang, L. and Zhao, S. D. (2020), 'Estimation and inference for the indirect effect in high-dimensional linear mediation models', *Biometrika* **107**(3), 573–589.

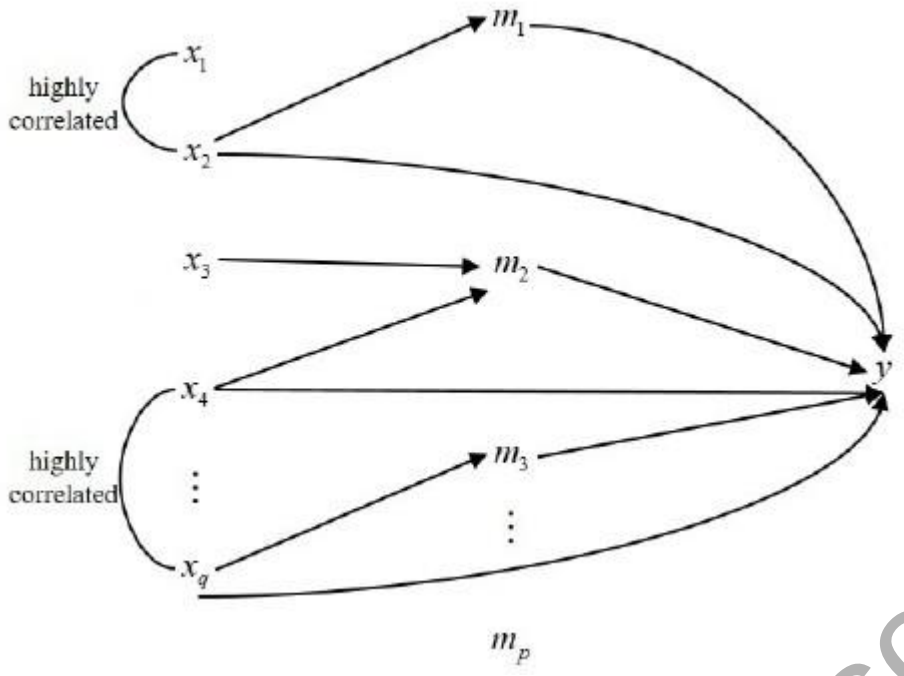
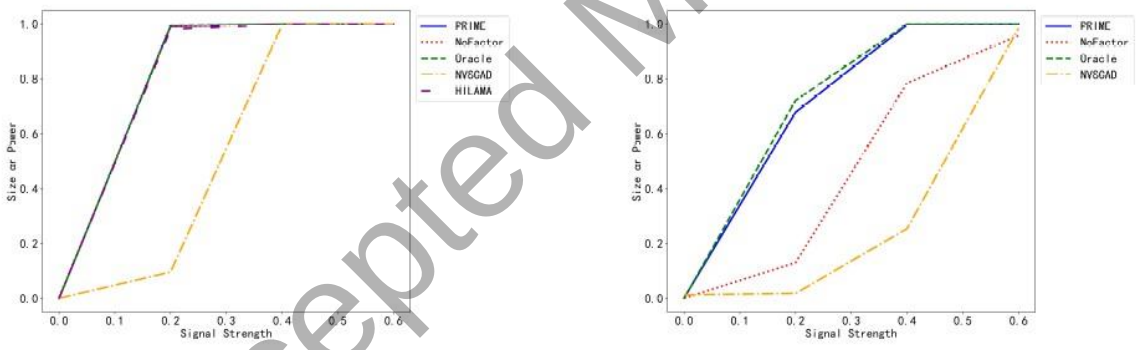


Figure 1: An illustration of the high correlation issue in HD-LSEM, where $x_1,\ldots,x_q$ are high-dimensional exposures, $m_1,\ldots,m_p$ are high-dimensional mediators, and y is the outcome. Directed edges between $\boldsymbol{x}$, $\boldsymbol{m}$, and y represent the potential mediation pathways, undirected edges among $\boldsymbol{x}$ and $\boldsymbol{m}$ indicate high correlations.



(a) Test for direct effects (b) Test for indirect effects

Figure 2: The power of T_n and S_n for the tests of the direct effect and indirect effect, respectively, when signal strength varies from 0 to 0.6 in Example 1.

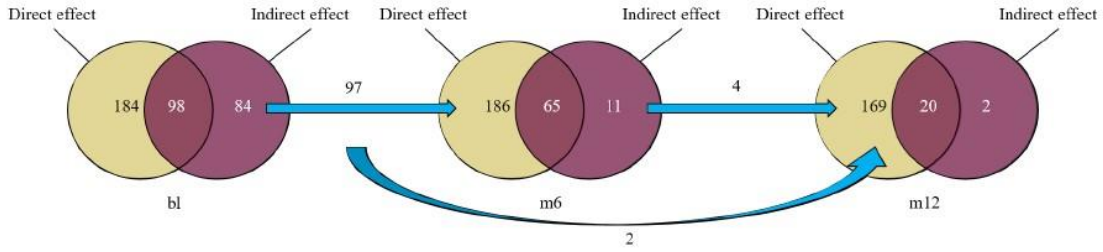


Figure 3: Venn graphs of the SNPs of significant direct effects and indirect effects at different time periods

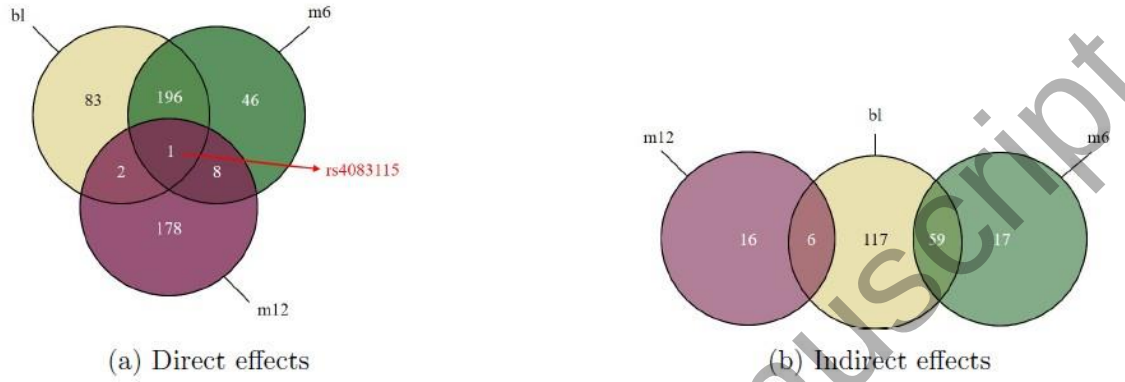


Figure 4: Venn graphs of the significant SNPs at different time periods for direct effects and indirect effects

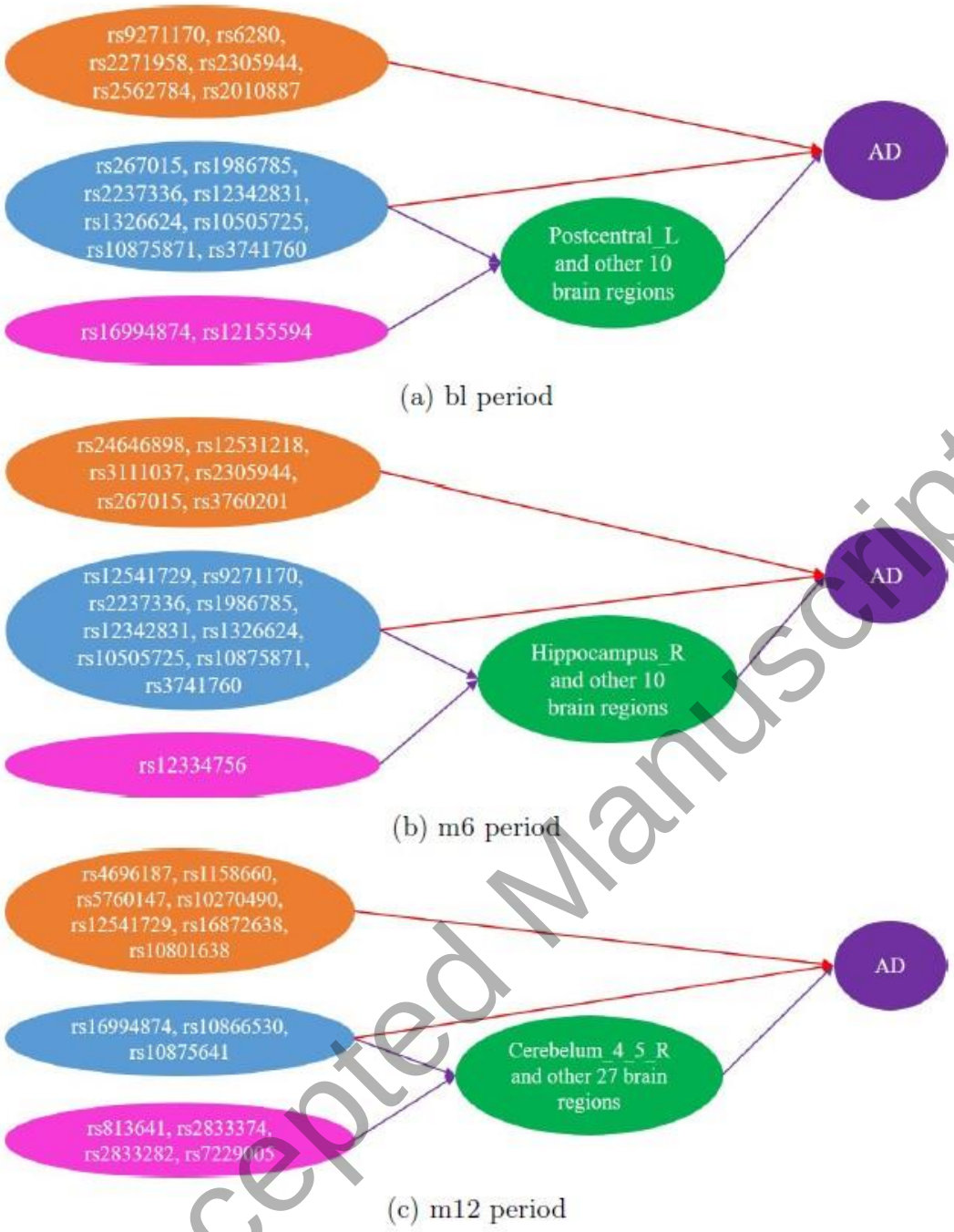


Figure 5: This figure demonstrates the effect pathways of genetic variants on AD at different period. The orange ellipse represents the SNPs which only has significant direct effect, the blue ellipse represents the SNPs which both possess significant direct and indirect effects, and the pink ellipse represents the SNPs which possess only significant indirect effects, while the green ellipse represents the brain region which serves as mediators. For arrows in different colors, the red ones represent the direct effect pathways and the purple ones are the indirect effect pathways.

Table 1: The MSE of direct effects in Example 1 ($\times 1000$)

Signals ＼ Methods	PRIME	Oracle	NoFactor	NVSCAD	HILAMA
-1	1.103	0.975	1.252	1.183	1.722
-0.8	1.078	0.937	1.037	1.193	1.746
-0.6	1.138	1.009	1.121	1.340	1.866
-0.4	1.045	0.968	2.045	3.757	1.789
-0.2	1.095	0.986	1.173	11.033	1.686
0	1.051	0.943	1.390	0.002	0.703
0.2	0.986	0.906	1.185	11.428	1.722
0.4	0.965	0.874	1.297	2.680	1.575
0.6	0.956	0.860	1.395	1.260	1.898
0.8	1.023	0.924	0.963	1.177	1.930
1	1.053	0.964	1.468	1.239	1.855

Table 2: The MSE of indirect effects in Example 1 ($\times 1000$)

Signals ＼ Methods	PRIME	Oracle	NoFactor	NVSCAD
-1	4.861	2.928	7.867	4.255
-0.8	5.141	3.234	8.335	5.287
-0.6	4.126	2.711	17.862	11.478
-0.4	4.080	2.930	34.605	26.694
-0.2	3.868	2.464	25.151	26.085
0	2.581	2.137	2.377	0.217
0.2	4.199	3.220	26.914	26.083
0.4	3.669	2.716	34.799	24.439
0.6	4.509	3.437	21.754	12.482
0.8	3.501	2.805	5.103	5.497
1	5.023	3.424	10.122	4.457

Table 3: FDR of different methods in Example 1

	PRIME	Oracle	NoFactor	NVSCAD	HILAMA
Direct effects	0.038	0.018	0.163	0.056	0.081
Indirect effects	0.077	0.015	0.062	0.443	–

Table 4: The number of signifcant SNPs at each time point

Time Point	Significant SNPs	
	Direct Effect	Indirect Effect
bl	282	182
m6	251	76
m12	189	22

Table 5: Significant SNPs at bl

SNP	DE	p (DE)	IE	p (IE)
rs267015	18.71	0	-3.08	6.24×10^{-5}
rs1986785	18.71	0	-3.08	6.24×10^{-5}
rs2237336	18.71	0	-3.08	6.24×10^{-5}
rs12342831	18.71	0	-3.08	6.24×10^{-5}
rs1326624	18.71	0	-3.08	6.24×10^{-5}
rs10505725	18.71	0	-3.08	6.24×10^{-5}
rs10875871	18.71	0	-3.08	6.24×10^{-5}
rs3741760	18.71	0	-3.08	6.24×10^{-5}
rs9271170	8.59	2.05×10^{-10}	–	–
rs6280	47.38	0	–	–
rs2271958	16.38	1.11×10^{-16}	–	–
rs2305944	47.38	0	–	–
rs2562784	-15.98	2.33×10^{-15}	–	–
rs2010887	7.65	2.42×10^{-8}	–	–
rs16994874	–	–	-0.50	3.35×10^{-7}
rs12155594	–	–	10.28	1.97×10^{-8}

Table 6: Significant SNPs at m6

SNP	DE	p (DE)	IE	p (IE)
rs12541729	2.21	2.07×10^{-7}	-0.07	0
rs9271170	9.04	0	-2.59	9.58×10^{-8}
rs2237336	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs1986785	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs12342831	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs1326624	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs10505725	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs10875871	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs3741760	11.38	1.11×10^{-16}	-6.23	1.45×10^{-6}
rs2464698	0.83	2.10×10^{-8}	–	–
rs12531218	-1.01	1.73×10^{-7}	–	–
rs3111037	1.82	6.06×10^{-7}	–	–
rs2305944	0.93	1.04×10^{-6}	–	–
rs267015	6.71	7.13×10^{-6}	–	–
rs3760201	2.27	1.19×10^{-5}	–	–
rs12334756	–	–	-6.23	1.45×10^{-6}

Table 7: Significant SNPs at m12

SNP	DE	p (DE)	IE	p (IE)
rs16994874	0.60	3.52×10^{-7}	-0.42	1.18×10^{-6}
rs10866530	-0.61	5.18×10^{-11}	0.29	1.72×10^{-7}
rs10875641	-0.43	1.08×10^{-7}	0.12	1.79×10^{-4}
rs4696187	0.35	1.22×10^{-6}	–	–
rs1158660	0.35	3.22×10^{-6}	–	–
rs5760147	0.38	8.51×10^{-7}	–	–
rs10270490	1.93	3.07×10^{-8}	–	–
rs12541729	1.04	9.23×10^{-7}	–	–
rs16872638	2.38	5.88×10^{-14}	–	–
rs10801638	-1.49	5.09×10^{-5}	–	–
rs813641	–	–	-16.22	0
rs2833374	–	–	0.65	1.39×10^{-10}
rs2833282	–	–	-0.84	2.30×10^{-13}
rs7229005	–	–	0.22	9.37×10^{-9}