HIGH-DIMENSIONAL COVARIATE-DEPENDENT GAUSSIAN GRAPHICAL MODELS

JIACHENG WANG

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

November 2025

Abstract

Over the last few decades, network analysis has become a prominent focus of research. One of the most widely used methodologies is known as the Gaussian graphical models (GGMs), which aim to investigate interactions through conditional dependencies. They have been widely applied in various domains such as computational biology, finance, and social network analysis. In recent years, advances in regularized estimation—such as the graphical lasso—have enabled scalable inference of sparse precision matrices. Despite their strengths, standard GGMs assume that the underlying network is static and homogeneous across all observations. However, real-world data often exhibit heterogeneity or dynamics, motivating extensions such as covariate-dependent or structured GGMs that incorporate external information and accommodate more complex dependence structures.

Therefore, in this dissertation, we propose a covariate-dependent GGM (cdexGGM) for capturing network structure that varies with covariates through a novel parameterization. Utilizing a likelihood framework, our methodology jointly estimates all edge and vertex parameters. We further develop statistical inference procedures to test the dynamic nature of the underlying network. Concerning large-scale networks, we perform composite likelihood estimation with an ℓ_1 penalty to discover sparse covariate-dependent graph structures. We establish the estimation error bound in ℓ_2 norm and validate the sign consistency in the high-dimensional context. We apply our method to an influenza vaccine data set to model the gene network that evolves with time. We also investigate a Down syndrome data set to model the protein network, which varies with several covariates under a factorial experimental design.

Moreover, to further address the limitations of GGMs in capturing heterogeneous networks with known structural constraints, we introduce a covariate-dependent colored GGM (CD-CGGM). This model incorporates covariate effects and structured sparsity to model dynamic conditional dependencies. We perform model estimation using penalized composite likelihood, employing coordinate descent and Broyden's method for optimization under different scenarios. We provide theoretical results ensuring both parameter and sign consistency of the proposed estimator. The method is applied to the same influenza vaccine dataset, where it effectively models the time-evolving gene regulatory network under symmetry constraints, thereby demonstrating its empirical performance and interpretability.

Keywords: Covariate-dependent Gaussian Graphical Models, Maximum Likelihood

Estimation, Composite Likelihood Estimation, ℓ_1 Penalization, Coordinate Descent, Symmetry-constrained Gaussian Graphical Models.

Acknowledgements

Though it is impossible to express my gratitude to all individuals who supported me, I wish to extend my deepest appreciation to many individuals whose support, encouragement, and guidance were instrumental in the completion of this dissertation.

I would like to begin by expressing my deepest gratitude towards my supervisor, Professor Xin Gao. Her mentorship and encouragement have been essential to the development and refinement of my research projects, and I am truly thankful for her generous commitment of time and expertise.

I would also like to express my sincere thanks to Professor Cindy Fu, Professor Steven Wang, Professor Rafal Kustra, and Professor Gene Cheung, who served as members of my supervisory committee and dissertation examiners. Their valuable feedback and academic insight have significantly contributed to the quality and direction of this work. I also wish to thank the faculty and staff of the Department of Mathematics and Statistics at York University for providing a supportive academic environment.

My special thanks go to Dr. Yuan Zhong. His technical expertise and willingness to help played an important role in advancing my research. I am also thankful to the many friends and colleagues I have met at York University, whose companionship and encouragement made this journey a meaningful and memorable one.

Finally, I would like to express my deepest gratitude to my parents, whose unconditional love, unwavering support, and constant encouragement have sustained me through every stage of this process. This accomplishment would not have been possible without them.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Gaussian Graphical Models

The study of complex network analysis has become a major research area in recent years, with Gaussian graphical models (GGMs) playing a key role in modeling conditional dependencies among variables. In standard GGMs, each variable corresponds to a node in an undirected graph, where the presence of an edge indicates a non-zero entry in the precision matrix, implying conditional dependence. This property makes GGMs a powerful tool for modeling and visualizing complex relationships in high-dimensional data, such as gene expression, brain connectivity, or social interactions. Key developments have been contributed by a number of authors, such as Lauritzen (1996), Meinshausen and Bühlmann (2006),Højsgaard and Lauritzen (2007), Højsgaard et al. (2012), and many others.

Suppose that a random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_p)^T$ follows a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Without loss of generality, we can assume that the mean of

$\boldsymbol{Y}$ is $\boldsymbol{\mu} = \mathbf{0}$. The inverse of the covariance matrix $\boldsymbol{\Sigma}^{-1}$ is known as the precision matrix (or concentration matrix). In matrix form, we write

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{K} = \begin{pmatrix} k_{11} & \cdots & k_{1p} \\ \vdots & \ddots & \vdots \\ k_{p1} & \cdots & k_{pp} \end{pmatrix}_{p \times p} . \tag{1.1}$$

Define a set $V = \{1, ..., p\}$ that includes all the indices of random variables. The partial correlation between Y_i and Y_j given all the other variables is defined as

$$\rho_{ij|V\backslash\{i,j\}} = \begin{cases} -\frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}} & i \neq j \\ -1 & i = j \end{cases} \tag{1.2}$$

where $1 \leqslant i, j \leqslant p$ and $\rho_{ij|V\backslash\{i,j\}} = 0$ if and only if Y_i and Y_j are conditionally independent given all the other variables.

A Gaussian graphical model is specified by an undirected graph $\mathcal{G} = (V, E)$ generated by the dataset $\left\{\boldsymbol{Y} \sim N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \boldsymbol{K} = \boldsymbol{\Sigma}^{-1} \in \mathcal{S}^+(\mathcal{G})\right\}$, where $V = \{1, ..., p\}$ is the set of index of vertices, E stands for the set of undirected edges in the graph and $\mathcal{S}^+(\mathcal{G})$ is the set of symmetric positive definite matrices. In a graph $\mathcal{G}$, each variable represents a vertex and the edges between vertices are drawn based on partial correlation. In particular, there is an edge between vertex i and j $(i, j \in V)$ if and only if the partial correlation $\rho_{ij|V\backslash\{i,j\}} \neq 0$. On the contrary, no edges connect them if and only if their partial correlation is zero.

Let $a \subset V$, $b \subset V$ and $a \cup b = V$, $a \cap b = \varnothing$, which means that a, b form a partition of the vertex set V. Then, the precision matrix and covariance matrix can be written as

$$\boldsymbol{K} = \begin{pmatrix} k_{aa} & k_{ab} \\ k_{ba} & k_{bb} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{1.3}$$

respectively. We also have

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_a \\ \boldsymbol{Y}_b \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right), \tag{1.4}$$

the conditional distribution $\boldsymbol{Y}_a | \boldsymbol{Y}_b = \boldsymbol{y}_b \sim N_{|a|}(\boldsymbol{\mu}_{a|b}, \Sigma_{a|b})$, where $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\boldsymbol{y}_b - \boldsymbol{\mu}_b) = \Sigma_{ab}\Sigma_{bb}^{-1}\boldsymbol{y}_b$ and $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$. Since

$$\boldsymbol{K} \cdot \boldsymbol{\Sigma} = \begin{pmatrix} k_{aa} & k_{ab} \\ k_{ba} & k_{bb} \end{pmatrix} \cdot \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{|a|} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{|b|} \end{pmatrix} = \boldsymbol{I}, \tag{1.5}$$

we have the following equations

$$k_{aa}\Sigma_{aa} + k_{ab}\Sigma_{ba} = \boldsymbol{I}_{|a|} \tag{1.6}$$

$$k_{aa}\Sigma_{ab} + k_{ab}\Sigma_{bb} = \mathbf{0}. \tag{1.7}$$

From Equation (1.7), $k_{ab} = -k_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}$ and Equation (1.6) becomes

$$k_{aa}\left(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right) = \boldsymbol{I}_{|a|} = k_{aa}k_{aa}^{-1}. \tag{1.8}$$

Therefore, $k_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} = \Sigma_{a|b}$ and

$$k_{aa}^{-1}k_{ab} = k_{aa}^{-1}(-k_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}) = -\Sigma_{ab}\Sigma_{bb}^{-1}. \tag{1.9}$$

As a result, the conditional distribution $\boldsymbol{Y}_a|\boldsymbol{Y}_b = \boldsymbol{y}_b$ can be given as

$$\boldsymbol{Y}_a|\boldsymbol{Y}_b = \boldsymbol{y}_b \sim N_{|a|}\Big(-k_{aa}^{-1}k_{ab}\boldsymbol{y}_b, k_{aa}^{-1}\Big). \tag{1.10}$$

1.2 Symmetry-Constrained Gaussian Graphical Models

High-dimensional data presents a challenge for GGMs due to the excessive number of parameters that must be estimated with limited observations. Moreover, many real-world systems exhibit inherent symmetries or group structures, such as gene families, bilateral brain regions, or social roles, that are not naturally accommodated by traditional GGMs, which may result in poor model fit or misinterpretation. To overcome these limitations, Højsgaard and Lauritzen (2008) proposed a family of symmetry-constrained GGMs (RCON, RCOR, and RCOP models) that integrate symmetries directly into the graphical model structure. The symmetry structure can be visualized by colored graphs, where nodes (variables) and edges (partial correlations) are grouped into color classes that enforce parameter equality across equivalent elements. Therefore, the symmetry-constrained GGMs are commonly referred to as colored GGMs.

1.2.1 RCON Model

One of the colored graphical models is called an RCON model, which refers to equality among specified elements of the precision matrix. Here is a simple example illustrating the basic idea. Suppose that we have 5 vertices A,B,C,D,E, representing 5 variables, and their corresponding precision matrix is given by:

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{K} = \begin{pmatrix} 4.9250 & -0.0186 & 0 & -1.2960 & -0.0027 \\ -0.0186 & 9.9056 & -0.0186 & 0 & -0.4590 \\ 0 & -0.0186 & 4.9250 & -1.2960 & -0.0027 \\ -1.2960 & 0 & -1.2960 & 9.9056 & -0.4590 \\ -0.0027 & -0.4590 & -0.0027 & -0.4590 & 0.1272 \end{pmatrix}_{5\times 5} \tag{1.11}$$

Identical elements in the precision matrix correspond to the same colors; for example, the diagonal elements of A and C are equal, as indicated in orange, and those of B and D are equal in pink. As for off-diagonal elements, for example, the entry of (A, B) is the same as that of (B, C), denoted in saffron yellow. Zero off-diagonal entries imply the absence of edges between vertices, which indicates that they are conditionally independent given all the other variables. The dependency graph is displayed in Figure 1.1.

Mathematically, let $\mathcal{G} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ denote an undirected graph with the associated precision matrix $\boldsymbol{\Sigma}^{-1} = \boldsymbol{K}$, where $\boldsymbol{\mathcal{V}} = \{V_1, ..., V_R\}$ is a partition of the vertex

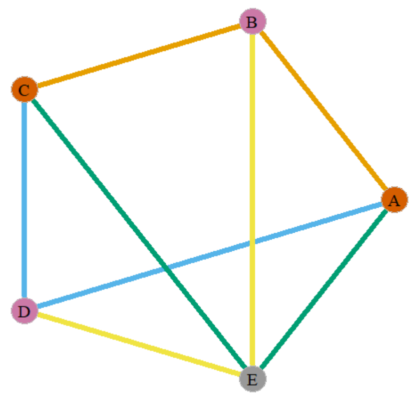


Figure 1.1: Colored graph of an RCON model

set $\boldsymbol{V}$ and each element in $V_r, r = 1, ..., R$ has the same color. These disjoint sets $V_1, \cdots, V_R$ are called vertex color classes. Likewise, $\boldsymbol{\mathcal{E}} = \{E_1, ..., E_S\}$ is a partition of the edge set $\boldsymbol{E}$ and each edge in $E_s, s = 1, \cdots, S$ has the same color. The disjoint sets $E_1, \cdots, E_S$ are called edge color classes. The sets $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{E}}$ are called vertex coloring and edge coloring. In particular, we denote each $(i, j)th$ entry of $\boldsymbol{K}$ as k_{ij}. Thus, an RCON($\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}$) model satisfy the following restrictions:

1. Diagonal entries of $\boldsymbol{K}$: $\forall\ V_r \in \boldsymbol{\mathcal{V}}, r = 1, ..., R$, for all $j \in V_r$, k_{jj} are identical.

2. Off–diagonal entries of $\boldsymbol{K}$: $\forall\ E_s \in \boldsymbol{\mathcal{E}}, s = 1, ..., S$, for all $(i, j) \in E_s$, k_{ij} are identical.

Identical elements of the precision matrix correspond to the same colors. Therefore, the unknown parameters can be written as a vector of length $R+S$, namely $(k_{V_1}, ..., k_{V_R}, k_{E_1}, ..., k_{E_S})$. The set of positive definite matrices satisfying the above restrictions is denoted as $\mathcal{S}^+(\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$.

1.3 Limitations and Motivations

One of the limitations of traditional GGMs is their inapplicability to dynamic settings where conditional dependencies change with covariates, which are commonly observed in biological applications. For example, in the context of genetics, it has been demonstrated that the genetic networks can be impacted by common genetic variants (Franco et al., 2013; Nica and Dermitzakis, 2013). Moreover, the genetic networks can also be affected by conditions such as cell type, cell cycle, stress response, and DNA damage (Zhang et al., 2006; Weighill et al., 2022). As another example, for those highly infectious vaccine-preventable diseases, getting immunized through vaccination serves as an effective strategy to prevent the spread of viruses. For a more in-depth study of the genetic factors involved in the diverse immune responses of healthy adults to a vaccine, researchers may desire to model biological networks across time (i.e., before vaccination and at different time points after vaccination). This may enable us to potentially recognize a group of genes that could be a critical determinant of the vaccine's effect on the immune response, which in turn might unlock innovative research opportunities in vaccine development (Franco et al., 2013). These have prompted the development of covariate-dependent GGMs, which extend classical frameworks by accounting for covariate-driven variations in network structure.

Moreover, in high-dimensional settings where additional structure is known or expected, standard GGMs may face interpretability, estimation, and statistical efficiency challenges. Traditional symmetry-constrained GGMs offer one solution but often lack the flexibility needed to handle the heterogeneous patterns found in real-world data. For example, in biology, genes in the Hox gene family often display similar patterns of interaction (Giampaolo et al., 1989; Mallo et al., 2010), which supports the use of symmetry constraints such as equal partial correlations or identical entries in the precision matrix. However, biological systems are covariate-dependent, with gene networks often exhibiting tissue-specific or species-specific variation (Sonawane et al., 2017; Tu et al., 2023), suggesting that regulatory structures may vary across conditions. This motivates the use of covariate-dependent GGMs with symmetry constraints, which allow the model to adapt structure dynamically, reflecting differences across conditions, developmental stages, or treatments. Similarly, in social networks, members of the same role or group (e.g., employees in the same department) may have interchangeable positions with similar connection patterns (Zhong et al., 2017), a concept captured by structural and role equivalence (Scott, 2019). Although groups may exhibit stable patterns, their behaviors are not static and can vary with time, location, or platform. As a result, modeling covariate-dependent network structures with known symmetry constraints enhances both adaptability and relevance in the analysis of structured and heterogeneous data.

This dissertation develops covariate-dependent Gaussian graphical models that capture how conditional dependencies evolve with observed factors in high-dimensional data. The goal is to construct flexible and scalable models that can accommodate both dynamic changes in network structure and known structural constraints derived from domain knowledge. In chapter 2, we develop the covariate-dependent Gaussian graphical model (cdexGGM), allowing both edge and vertex parameters to vary smoothly with covariates, while maintaining the positive definiteness of the precision matrix under modest assumptions. Building on this, Chapter 3 proposes covariate-dependent RCON models—a class of colored Gaussian graphical models that integrate covariate-driven variation with structural symmetries to model complex network heterogeneity.

1.4 Notations and Collorary

Definition 1.1 A random variable Y with mean $\mu = E\{Y\}$ is sub-exponential if there are non-negative parameters (v^2, c) such that

$$E\Big\{\exp\big\{\lambda(Y-\mu)\big\}\Big\} \leqslant \exp\Big\{\frac{v^2\lambda^2}{2}\Big\} \quad \text{for all } |\lambda| < \frac{1}{c}.$$

Collorary 1.1: If the random vector $\boldsymbol{Y}_m = (Y_{m1}, \ldots, Y_{mp_n})^T$, $m = 1, \ldots, n$, follows multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_m)$, then we have

$$Y_{mj}^2 \sim \text{ Sub-Exponential } (v_{m,jj}^2, c_{m,jj}), \quad Y_{mi}Y_{mj} \sim \text{ Sub-Exponential } (v_{m,ij}^2, c_{m,ij}),$$

where $m = 1, ..., n, i = 1, ..., p_n, j = 1, ..., p_n$ and $i \neq j$.

Notation 1.1: We define a vector $\boldsymbol{v} = (v_1, \cdots, v_d) \in \mathbb{R}^d$ and write its ℓ_q norm as $\|\boldsymbol{v}\|_q = \big(\sum_{i=1}^d |v_i|^q\big)^{\frac{1}{q}}$. In particular, the ℓ_∞ norm of a vector is defined as $\|\boldsymbol{v}\|_\infty = sup_{1\leqslant i\leqslant d}|v_i|$.

Notation 1.2: For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$, we write its spectral norm, infinity norm, max norm, and Frobenius norm as $|||\boldsymbol{A}|||_2 = \Lambda_{\max}(\boldsymbol{A})$, $|||\boldsymbol{A}|||_\infty = \max_{1\leqslant i\leqslant d_1}\sum_{j=1}^{d_2}|a_{ij}|$, $|||\boldsymbol{A}|||_{\max} = \max_{ij}|a_{ij}|$, $|||\boldsymbol{A}|||_F = \sqrt{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}|a_{ij}|^2}$, where $\Lambda_{\max}(\cdot)$ denotes the largest singular value of a matrix and a_{ij} is the (i,j)th element of $\boldsymbol{A}$. Moreover, we let $\Lambda_{\min}(\boldsymbol{A}), \lambda_{\min}(\boldsymbol{A}), \lambda_{\max}(\boldsymbol{A})$ represent the smallest singular value, and smallest and largest eigenvalues of $\boldsymbol{A}$ respectively.

2 High-Dimensional Covariate-Dependent Gaussian Graphical Models

2.1 Introduction

One of the earliest approaches in this direction is the time-varying graphical models, with various methodologies proposed, such as Zhou et al. (2010), Kolar et al. (2010b), Kolar and Xing (2012), Wang and Kolar (2014), Monti et al. (2014), Gibberd and Nelson (2017), and Yang and Peng (2020). Other researchers proposed the covariate-adjusted GGMs, allowing the mean to be dependent on covariates while preserving a constant variance structure. Guo et al. (2011) conducted simultaneous estimation of multiple GGMs using a hierarchical penalty to characterize a series of graphs sharing a certain amount of common structure. Yin et al. (2010) introduced a nonparametric approach for modeling conditional dependencies, enabling the covariance matrix to vary with covariates. Kolar et al. (2010a) focused on the high-dimensional setting and developed a locally weighted kernel estimator for modeling sparse conditional dependencies. Ni et al. (2019) further explored the estimation procedure of covariate-dependent acyclic graphs from the perspective of the Bayesian regression framework. Ni et al. (2022) addressed the general problem of estimating undirected Gaussian graphical models (GGMs) conditioned on covariates (GGMx) using a Bayesian approach, which allows for both the edge strength and the sparsity pattern of the underlying graph structure to change as functions of covariates. By constructing an MCMC algorithm for posterior inference, the estimates of the precision matrices can be obtained, and their positive definiteness can be guaranteed regardless of the covariates (Ni et al., 2022). Zhang and Li (2023) introduced a new type of Gaussian graphical regression model to describe covariate-dependent network structure, allowing both the mean structure and precision matrix to linearly vary with covariates. The model assumes all off-diagonal parameters of the precision matrix to vary with covariates, whereas the diagonal parameters remain constant. They proposed a two-stage procedure that first estimates the ratios of off-diagonal and diagonal parameters by regressing each vertex against all the other vertices and then estimates the diagonal parameters via a moment method. In contrast to existing methods that directly model the underlying graph structure as functions of covariates, Niu et al. (2023) presented a novel Bayesian covariate-dependent graphical model and introduced an intermediate latent variable layer acting as a hidden bridge between the graphs and covariates, which offered a simpler alternative to parameterizing graph structure in terms of covariates.

In this chapter, we introduce a covariate-dependent Gaussian graphical model (cdexGGM), which allows both edge and vertex parameters to vary smoothly with covariates. The proposed model guarantees the positive definiteness of the precision matrix under modest assumptions. We further relax the restriction on the diagonal parameters and allow both off-diagonal and diagonal parameters of the precision matrix to change with covariates. Built within a likelihood framework, the model uses maximum likelihood estimation for small-scale networks, along with formal inference procedures that are developed to test the dynamic nature of the underlying network. For large-scale networks, this model employs composite likelihood and an ℓ_1 penalty to identify sparse, covariate-dependent structures efficiently. Estimation consistency and sign consistency are theoretically justified under high-dimensional settings. To empirically evaluate the proposed model, extensive simulation studies are conducted, including comparisons with several competing methods, displaying satisfactory performance of the proposed likelihood procedure. The method is further applied to biological datasets capturing time-varying gene and protein networks to demonstrate its practical relevance and effectiveness.

2.2 Methodology

2.2.1 Model Setup

Let $\boldsymbol{Y} = \big(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n\big)^T$ be a collection of independent random vectors, with each $\boldsymbol{Y}_m \sim N\big(\boldsymbol{\mu}_m(x_m), \boldsymbol{\Sigma}_m(x_m)\big), m = 1, \cdots, n$, where x_m denotes the observed covariate and $Y_m = (Y_{m1}, Y_{m2}, ..., Y_{mp})^T$. For the sake of simplicity, we assume that the observations are centered, namely $\boldsymbol{\mu}_m(x_m) = \mathbf{0}$. First consider the simple case that x_m is a univariate covariate, which is min-max scaled so that $0 \leqslant x_m \leqslant 1$. We assume that $\boldsymbol{\Sigma}_m^{-1}(x_m)$ can be formulated as

$$\begin{aligned}\boldsymbol{\Sigma}_m^{-1}(x_m) = \boldsymbol{K}_m &= x_m\boldsymbol{Q}_1 + (1 - x_m)\boldsymbol{Q}_0 \\ &= \boldsymbol{Q}_0 + x_m(\boldsymbol{Q}_1 - \boldsymbol{Q}_0) = \boldsymbol{Q}_0 + x_m\boldsymbol{P}_1.\end{aligned} \tag{2.1}$$

The matrix $\boldsymbol{Q}_0$ is the precision matrix when $x_m = 0$, and $\boldsymbol{Q}_1$ is the precision matrix when $x_m = 1$. For any arbitrary value of $x_m \in [0, 1]$, the precision matrix is a weighted average of these two matrices. If $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ are assumed to be positive definite, $\boldsymbol{K}_m$ is guaranteed to be positive definite. After the reformulation, we have $\boldsymbol{K}_m$ expressed in a regression format where $\boldsymbol{Q}_0$ is the intercept matrix, and $\boldsymbol{P}_1$ is the slope matrix associated with the observed covariate.

We can further extend the model (2.1) to the general setting of multiple covariates

$$\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x_m}) = \boldsymbol{K}_m = \sum_{h=1}^{H} x_m^{(h)}\boldsymbol{Q}_h + (1 - \sum_{h=1}^{H}\frac{x_m^{(h)}}{H})\boldsymbol{Q}_0 \tag{2.2}$$

$$= \boldsymbol{Q}_0 + \sum_{h=1}^{H} x_m^{(h)}(\boldsymbol{Q}_h - \boldsymbol{Q}_0/H) = \boldsymbol{Q}_0 + \sum_{h=1}^{H} x_m^{(h)}\boldsymbol{P}_h, \tag{2.3}$$

where $\boldsymbol{P}_h = \boldsymbol{Q}_h - \boldsymbol{Q}_0/H$, H is the number of covariates, $\boldsymbol{x_m} = (x_m^{(1)}, \ldots, x_m^{(H)})^T$, and $x_m^{(h)} \in [0, 1]$ represents the hth covariate of the mth observation. In (2.2), $\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x_m})$ is a linear combination of $\boldsymbol{Q}_0, \ldots, \boldsymbol{Q}_H$ with non-negative coefficients. Under the assumption that all these matrices are positive definite, $\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x_m})$ is guaranteed to be positive definite for any combination of possible values of covariates. The $\boldsymbol{Q}_0$ is the baseline precision matrix when all covariates are set to zero. The matrices $\boldsymbol{P}_1, \cdots, \boldsymbol{P}_h$ can be treated as the slopes of the covariates to describe how the current network structure deviates from its baseline state, resulting from $x_m^{(h)}$. This general model of covariate-dependent GGMs (cdexGGM) encompasses standard GGMs, group-specific GGMs, and time-varying GGMs as special cases, as summarized in Table 2.1. The setting in (2.2) is flexible enough to accommodate scenarios with a known baseline $\boldsymbol{Q}_0$, where the primary interest lies in assessing the covariate effects on the precision matrix, which is characterized by the slope matrices.

We use the following notations $\left(\boldsymbol{Q}_0\right)_{ij} = \alpha_{ij}, \left(\boldsymbol{P}_h\right)_{ij} = [\theta_{ij}]_h$ representing the (i, j)th element of $\boldsymbol{Q}_0$ and $\boldsymbol{P}_h$ respectively. The set of unknown parameters can be formulated as $\boldsymbol{\beta} = (\boldsymbol{\alpha}, \boldsymbol{\theta})^T$, where $\boldsymbol{\alpha}$ is the vector of all $\alpha_{ij}, i \leqslant j$ and $\boldsymbol{\theta}$ is the vector of all

Table 2.1: Special Cases of Covariate-Dependent GGMs

Special Cases	Conditions
Standard GGMs	$\forall\ h = 1, \cdots, H, x_m^{(h)} = 0$
Group-specific GGMs	$\forall\ h = 1, \cdots, H, x_m^{(h)}$ is categorical
Time-varying GGMs	$h = 1, x_m^{(h)}$ is continuous
General covariate-dependent GGMs	$\forall\ h = 1, \cdots, H, x_m^{(h)}$ is continuous or categorical

$[\theta_{ij}]_h, i \leqslant j$. Furthermore, we define an index set $\mathbb{E} = \Big\{(i,j) : \ i \leqslant j;\ i,j = 1, ..., p\Big\}$.

For $\forall\ u \in \mathbb{E}$, we define an indicator matrix $\boldsymbol{T}^u$ as

$$(\boldsymbol{T}^u)_{ij} = \begin{cases} 1 & \text{when } (i,j) = u, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

and let $(\boldsymbol{T}^u)_{ij} = (\boldsymbol{T}^u)_{ji}$. Therefore, each matrix can be written as

$$\boldsymbol{Q}_0 = \sum_{s:s\in\mathbb{E}} \alpha_s \boldsymbol{T}^s, \quad \boldsymbol{P}_h = \sum_{u:u\in\mathbb{E}} [\theta_u]_h \boldsymbol{T}^u. \tag{2.5}$$

Using the notations in (2.4) and (2.5), the precision matrix can be rewritten as

$$\boldsymbol{K}_m = \sum_{s:s\in\mathbb{E}} \alpha_s \boldsymbol{T}^s + \sum_{h=1}^{H} x_m^{(h)} \Big\{ \sum_{u:u\in\mathbb{E}} [\theta_u]_h \boldsymbol{T}^u \Big\}. \tag{2.6}$$

2.2.2 Maximum Likelihood Estimation for Small Networks

In this section, we outline the maximum likelihood estimation procedure for small-scale covariate-dependent networks. Let $\mathcal{L}(\boldsymbol{\beta})$ represent the joint log-likelihood function,

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{m=1}^{n} \mathcal{L}_m(\boldsymbol{\beta}|\boldsymbol{Y}_m) = -\frac{np}{2}\log 2\pi - \frac{1}{2}\sum_{m=1}^{n}\log|\boldsymbol{\Sigma}_m| - \frac{1}{2}\sum_{m=1}^{n} tr(\boldsymbol{Y}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{Y}_m). \tag{2.7}$$

Note that (2.7) belongs to exponential family with the canonical parameters $\boldsymbol{Q}_0, \boldsymbol{P}_1, \cdots, \boldsymbol{P}_H$ and sufficient statistics $T_{Q_0} = \sum_{m=1}^{n} \boldsymbol{Y}_m\boldsymbol{Y}_m^T, T_{P_h} = \sum_{m=1}^{n} x_m^{(h)}\boldsymbol{Y}_m\boldsymbol{Y}_m^T, h = 1, \cdots, H$. For $\forall\ u, v, s, t \in \mathbb{E}$, and $\forall\ h_1, h_2 = 1, \cdots, H$, the first derivatives and the second derivatives of loglikelihood are given by

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial[\theta_u]_{h_1}} &= \frac{1}{2}\sum_{m=1}^{n}\Big\{x_m^{(h_1)} tr(\boldsymbol{\Sigma}_m\boldsymbol{T}^u) - x_m^{(h_1)} tr(\boldsymbol{W}_m\boldsymbol{T}^u)\Big\},\\
\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial\alpha_s} &= \frac{1}{2}\sum_{m=1}^{n}\Big\{tr(\boldsymbol{\Sigma}_m\boldsymbol{T}^s) - tr(\boldsymbol{W}_m\boldsymbol{T}^s)\Big\},\\
\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial[\theta_u]_{h_1}\partial[\theta_v]_{h_2}} &= -\frac{1}{2}\sum_{m=1}^{n} x_m^{(h_1)}x_m^{(h_2)} tr(\boldsymbol{T}^u\boldsymbol{\Sigma}_m\boldsymbol{T}^v\boldsymbol{\Sigma}_m),\\
\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial[\theta_u]_{h_1}\partial\alpha_s} &= -\frac{1}{2}\sum_{m=1}^{n} x_m^{(h_1)} tr(\boldsymbol{T}^u\boldsymbol{\Sigma}_m\boldsymbol{T}^s\boldsymbol{\Sigma}_m),\\
\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial\alpha_s\partial\alpha_t} &= -\frac{1}{2}\sum_{m=1}^{n} tr(\boldsymbol{T}^s\boldsymbol{\Sigma}_m\boldsymbol{T}^t\boldsymbol{\Sigma}_m),
\end{aligned}$$

where $\boldsymbol{W}_m = \boldsymbol{Y}_m\boldsymbol{Y}_m^T$. For simplicity, we use the notation $\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}|\boldsymbol{Y}_m), \boldsymbol{\mathcal{L}}_m^{(2)}(\boldsymbol{\beta}|\boldsymbol{Y}_m)$ representing the score vector and Hessian matrix for each observation respectively.

Define the expected individual negative Hessian matrix as

$$\boldsymbol{V}_m(\boldsymbol{\beta}) = E\Big\{\big\{\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}|\boldsymbol{Y}_m)\big\}\big\{\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}|\boldsymbol{Y}_m)\big\}^T\Big\} = -E\Big\{\boldsymbol{\mathcal{L}}_m^{(2)}(\boldsymbol{\beta}|\boldsymbol{Y}_m)\Big\}. \tag{2.8}$$

It can be shown that the information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ takes the following form

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \sum_{m=1}^{n} \boldsymbol{V}_m(\boldsymbol{\beta}) = \sum_{m=1}^{n} \left\{ \left(\boldsymbol{X}_m \boldsymbol{X}_m^T\right) \otimes \left(\boldsymbol{\Sigma}_m \otimes \boldsymbol{\Sigma}_m\right) \right\},$$

where $\otimes$ represents the Kronecker product and $\boldsymbol{X}_m = \left(1, x_m^{(1)}, x_m^{(2)}, \cdots, x_m^{(H)}\right)^T$. We adopt the iterative coordinate updating procedure proposed by Jensen et al. (1991) and Lauritzen (1996) to find the maximum likelihood estimate (MLE). The detailed updating equations at each step are outlined as follows,

$$\begin{aligned}
[\hat{\theta}_u^{(t+1)}]_{h_1} &= [\hat{\theta}_u^{(t)}]_{h_1} + \frac{\Delta_u}{\sum_{m=1}^n \left\{x_m{}^{(h_1)}\right\}^2 tr(\boldsymbol{T}^u \hat{\boldsymbol{\Sigma}}_m^{(t)} \boldsymbol{T}^u \hat{\boldsymbol{\Sigma}}_m^{(t)}) + \frac{\Delta_u{}^2}{2}}, \\
\hat{\alpha}_s^{(t+1)} &= \hat{\alpha}_s^{(t)} + \frac{\Delta_s}{\sum_{m=1}^n tr(\boldsymbol{T}^s \hat{\boldsymbol{\Sigma}}_m^{(t)} \boldsymbol{T}^s \hat{\boldsymbol{\Sigma}}_m^{(t)}) + \frac{\Delta_s{}^2}{2}}, \\
\Delta_u &= \sum_{m=1}^n x_m^{(h_1)} \Big\{ tr(\hat{\boldsymbol{\Sigma}}_m^{(t)} \boldsymbol{T}^u) - tr(\boldsymbol{W}_m \boldsymbol{T}^u) \Big\}, \\
\Delta_s &= \sum_{m=1}^n \Big\{ tr(\hat{\boldsymbol{\Sigma}}_m^{(t)} \boldsymbol{T}^s) - tr(\boldsymbol{W}_m \boldsymbol{T}^s) \Big\},
\end{aligned} \tag{2.9}$$

where Δ_u, Δ_s are the first derivatives of $\mathcal{L}(\boldsymbol{\beta})$ evaluated at the (t)th iteration.

2.2.3 Theoretical Properties of MLE

In this section, we investigate the asymptotic behavior of the proposed maximum likelihood estimate when p is fixed.

Assumption 1 Let Ω represent the parameter space for $\boldsymbol{\beta}$. We assume there exists an open subset $\boldsymbol{\omega} \subset \boldsymbol{\Omega}$ that includes the true parameter $\boldsymbol{\beta}^0$ such that for $\boldsymbol{\beta} \in \boldsymbol{\omega}$, the eigenvalues of $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_h, h = 1, \cdots, H$, are bounded below by $L_1 > 0$ and bounded above by $L_2 > 0$.

Assumption 2 Assume that the average Hessian matrix $\frac{1}{n}\sum_{m=1}^{n} \boldsymbol{V}_m(\boldsymbol{\beta}^0) \rightarrow \boldsymbol{V}$ as n goes to infinity, where $\boldsymbol{\beta}^0$ represents the true value of $\boldsymbol{\beta}$ and $\boldsymbol{V}$ is positive definite.

Theorem 2.1 Under **Assumptions** 1-2, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0||_2 = \mathcal{O}_p(n^{-\frac{1}{2}})$. Furthermore, the multivariate version of Lindeberg condition (2.10) holds: $\forall\ \epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E\Big\{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2^2 \cdot I\Big(||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2 \geqslant \epsilon\sqrt{n}\Big)\Big\} = 0, \tag{2.10}$$

and thus, the local maximizer $\hat{\boldsymbol{\beta}}$ satisfies the asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N\Big(\mathbf{0}, \boldsymbol{V}^{-1}\Big) \text{ as } n \longrightarrow \infty.$$

Using the asymptotic distribution of the maximum likelihood estimate, we can construct test statistics and conduct statistical inference on the underlying covariate-dependent graph structure. For example, to test if the network is static versus dynamic, we could conduct inference on the slope parameters and test the hypothesis $H_0 : \boldsymbol{\theta} = 0$ versus $H_1 : \boldsymbol{\theta} \neq 0$. If the covariate is categorical, the hypotheses above can be used to test the equality of graphical models among different groups. We can also focus on a single parameter of the slope matrices and construct Wald-type statistic to test hypotheses $H_0 : [\theta_{ij}]_h = 0$ versus $H_1 : [\theta_{ij}]_h \neq 0$. Such test can be used to assess the hth covariate's effect on the (i, j)th edge in the graphical model.

2.2.4 Penalized Composite Likelihood Estimation for Large Networks

In this section, we use the notation p_n to represent the dimension of random vectors because we allow p_n to increase with sample size n. However, all random vectors $\boldsymbol{Y}_m$ are assumed to have the same dimension p_n. Note that the estimation procedure proposed earlier involves matrix inversion in each iteration, leading to computational challenges as p_n grows. Moreover, large-scale networks typically exhibit sparse network structures. Motivated by these insights, we introduce a penalized composite likelihood approach (Lindsay, 1988; Fearnhead and Donnelly, 2002; Cox and Reid, 2004; Varin, 2008; Larribe and Fearnhead, 2011; Ribatet et al., 2012; Gao and Massam, 2015) for large p_n scenarios. The conditional distribution of a random variable Y_{mj} given $\boldsymbol{Y}_{m,V\backslash\{j\}}$ is $N\big(\mu_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}}, \sigma^2_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}}\big)$ with

$$\begin{aligned}
\mu_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}} &= -\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1} \sum_{i=1, i\neq j}^{p_n} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big] y_{mi},\\
\sigma^2_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}} &= \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1},
\end{aligned}$$

where $V = \big\{1, \cdots, p_n\big\}$ denotes the vertex set, and $V\backslash\{j\}$ denotes all the vertices except j. The individual conditional log-likelihood function of Y_{mj} given $\boldsymbol{Y}_{m,V\backslash\{j\}}$

is denoted by $l_c(Y_{mj}, \boldsymbol{\beta})$. We construct the composite log-likelihood function by incorporating all the conditional distributions. Therefore, the overall negative joint composite log-likelihood function is given by

$$
\begin{aligned}
&-l_c(\boldsymbol{Y}, \boldsymbol{\beta}) = -\sum_{m=1}^{n} l_c(\boldsymbol{Y}_m, \boldsymbol{\beta}) = -\sum_{m=1}^{n}\sum_{j=1}^{p_n} l_c(Y_{mj}, \boldsymbol{\beta}) \\
&= -\frac{1}{2}\sum_{m=1}^{n}\sum_{j=1}^{p_n} \log\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big] \\
&+ \frac{1}{2}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]\Big\{y_{mj} + \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\sum_{i=1, i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big\}^2 \\
&+ const.
\end{aligned}
$$

The penalized composite log-likelihood estimator $\hat{\boldsymbol{\beta}}$ is determined by minimizing the following objective function

$$
\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -l_c(\boldsymbol{Y}, \boldsymbol{\beta}) + n\lambda\Big\{\sum_{w\in\mathcal{E}}|\alpha_w| + \sum_{s\in\mathcal{E}}\sum_{h=1}^{H}|[\theta_s]_h|\Big\}, \tag{2.11}
$$

where $\mathcal{E} = \{(i, j) : i < j; \ i, j = 1, \cdots, p_n\}$ is the off-diagonal index set. The ℓ_1 penalty is included in the objective function to enforce sparsity on the network structure. It should be emphasized that the sparsity patterns of $\boldsymbol{Q}_0, \boldsymbol{P}_1, \cdots, \boldsymbol{P}_H$ can differ, leading to graph structures that change with covariates.

To obtain the penalized maximum composite likelihood estimate, we propose to perform the coordinate descent algorithm (Friedman et al., 2007; Tseng, 2001). A thorough exploration including the convergence properties of the estimators using the coordinate descent algorithm with the ℓ_1 penalization can be found in Breheny and Huang (2011) and Mazumder et al. (2011). Specifically, the closed-form updating expressions for off-diagonal parameters can be derived by differentiating $Q(\boldsymbol{\beta})$ with respect to $[\theta_s]_{h_1}, \alpha_w$ for all $s, w = (a, b) \in \mathcal{E}$ and $h_1 = 1, \cdots, H$. The updating equations for each off-diagonal parameter given all the other parameters are as follows,

$$\hat{\alpha}_w = \frac{S\Big(-\Big\{\frac{2}{n}\sum_{m=1}^{n} y_{ma}y_{mb} + \mathcal{I}_{1.1} + \mathcal{I}_{1.2}\Big\}, \lambda\Big)}{\mathcal{I}_{1.3}},$$

$$[\hat{\theta}_s]_{h_1} = \frac{S\Big(-\Big\{\frac{2}{n}\sum_{m=1}^{n} x_m^{(h_1)} y_{ma}y_{mb} + \mathcal{I}_{2.1} + \mathcal{I}_{2.2}\Big\}, \lambda\Big)}{\mathcal{I}_{2.3}},$$

where $S(z, \lambda) = sign(z)(|z| - \lambda)_+$ represents the soft-thresholding operator. The notation $(|z| - \lambda)_+$ is defined as

$$(|z| - \lambda)_+ = \begin{cases} |z| - \lambda, & \text{if } |z| - \lambda > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The terms $\mathcal{I}_{1.1}, \mathcal{I}_{1.2}, \cdots, \mathcal{I}_{2.3}$ are defined as

$$
\begin{aligned}
\mathcal{I}_{1.1} &= \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Big(\sum_{i=1,i\neq a}^{p_n}\sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h y_{mi}\Big) y_{mb} \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}\Big(\sum_{i=1,i\neq b}^{p_n}\sum_{h=1}^{H} x_m^{(h)}[\theta_{bi}]_h y_{mi}\Big) y_{ma}, \\
\mathcal{I}_{1.2} &= \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Big(\sum_{\substack{i=1,i\neq a,\\ i\neq b}}^{p_n} \alpha_{ai} y_{mi}\Big) y_{mb} \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}\Big(\sum_{\substack{i=1,i\neq b,\\ i\neq a}}^{p_n} \alpha_{bi} y_{mi}\Big) y_{ma}, \\
\mathcal{I}_{1.3} &= \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1} y_{mb}^2 + \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1} y_{ma}^2, \\
\mathcal{I}_{2.1} &= \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Big(\sum_{i=1,i\neq a}^{p_n} \alpha_{ai} y_{mi}\Big) x_m^{(h_1)} y_{mb} \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}\Big(\sum_{i=1,i\neq b}^{p_n} \alpha_{bi} y_{mi}\Big) x_m^{(h_1)} y_{ma}, \\
\mathcal{I}_{2.2} &= \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Big(\sum_{\substack{i=1,\\ i\neq a,}}^{p_n}\sum_{\substack{h=1,\\ (i,h)\neq(b,h_1)}}^{H} x_m^{(h)}[\theta_{ai}]_h y_{mi}\Big) x_m^{(h_1)} y_{mb} \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}\Big(\sum_{\substack{i=1,\\ i\neq b,}}^{p_n}\sum_{\substack{h=1,\\ (i,h)\neq(a,h_1)}}^{H} x_m^{(h)}[\theta_{bi}]_h y_{mi}\Big) x_m^{(h_1)} y_{ma}, \\
\mathcal{I}_{2.3} &= \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}[x_m^{(h_1)}]^2 y_{mb}^2 + \frac{1}{n}\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}[x_m^{(h_1)}]^2 y_{ma}^2.
\end{aligned}
$$

For the diagonal parameters, the partial derivatives are given as follows

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\beta})}{\partial \alpha_{jj}} =& -\frac{1}{2}\sum_{m=1}^{n}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big)^2 \\
& -\frac{1}{2}\sum_{m=1}^{n}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}+\frac{1}{2}\sum_{m=1}^{n}y_{mj}^2, \\
\frac{\partial Q(\boldsymbol{\beta})}{\partial [\theta_{jj}]_{h_1}} =& -\frac{1}{2}\sum_{m=1}^{n}x_m^{(h_1)}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big)^2 \\
& -\frac{1}{2}\sum_{m=1}^{n}x_m^{(h_1)}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}+\frac{1}{2}\sum_{m=1}^{n}x_m^{(h_1)}y_{mj}^2,
\end{aligned}
\tag{2.12}
$$

which do not have closed-form solutions. We employ Broyden's method (Nocedal and Wright, 2006) to solve this system of nonlinear equations.

Similar to Zhang and Li (2023), consider a special case where the diagonal elements $(\boldsymbol{\Sigma}_m^{-1})_{jj}$ remain constant, i.e. $[\theta_{jj}]_h = 0, \alpha_{jj} > 0$ for all $j = 1, \cdots, p_n$ and $h = 1, \cdots, H$. This suggests that the conditional variances are not influenced by covariates. Under this scenario, the closed-form updating expressions for diagonal parameters can be directly obtained, which simplifies the estimation procedure. The joint negative composite log-likelihood function is simplified as

$$
-l_c(\boldsymbol{\beta}) = -\sum_{j=1}^{p_n}\sum_{m=1}^{n}\log\alpha_{jj}+\sum_{j=1}^{p_n}\sum_{m=1}^{n}\alpha_{jj}\Big\{y_{mj}+\alpha_{jj}^{-1}\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big\}^2.
$$

The updating equations can be simplified in matrix form, as provided below

$$[\hat{\theta}_s]_h = \frac{S\Big(-tr\Big(\boldsymbol{T}^s\boldsymbol{C}_\theta^{(h)}\Big) + tr\Big(\boldsymbol{T}^s\big(\boldsymbol{T}^{s^c}\odot\big[\boldsymbol{B}_\theta^{(h)}\big]^T\big)\boldsymbol{D}_\theta^{(h)}\Big) + tr\Big(\boldsymbol{T}^s\big(\boldsymbol{T}^{s^1}\odot\boldsymbol{B}_\alpha^T\big)\boldsymbol{C}_\theta^{(h)}\Big) + \Delta, \lambda\Big)}{tr\Big(\boldsymbol{T}^s\boldsymbol{\Sigma}\boldsymbol{T}^s\boldsymbol{D}_\theta^{(h)}\Big)},$$

$$\hat{\alpha}_w = \frac{S\Big(-tr\Big(\boldsymbol{T}^w\boldsymbol{C}_\alpha\Big) + tr\Big(\boldsymbol{T}^w\big(\boldsymbol{T}^{w^c}\odot\boldsymbol{B}_\alpha^T\big)\boldsymbol{C}_\alpha\Big) + \sum_{h=1}^H tr\Big(\boldsymbol{T}^w\big(\boldsymbol{T}^{w^1}\odot\big[\boldsymbol{B}_\theta^{(h)}\big]^T\big)\boldsymbol{C}_\theta^{(h)}\Big), \lambda\Big)}{tr\Big(\boldsymbol{T}^w\boldsymbol{\Sigma}\boldsymbol{T}^w\boldsymbol{C}_\alpha\Big)},$$

$$\hat{\alpha}_{jj} = \frac{2\sum_{m=1}^n\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^H x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big)^2}{-n + \sqrt{n^2 + 4\Big(\sum_{m=1}^n y_{mj}^2\Big)\Big\{\sum_{m=1}^n\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^H x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big)^2\Big\}}},$$

where s^c and w^c denote the complement of s and w respectively, and

$$\begin{aligned}
\boldsymbol{C}_\theta^{(h)} &= \frac{1}{n}\Big(\boldsymbol{X}^{(h)}\odot\boldsymbol{Y}\Big)^T\boldsymbol{Y}, & \boldsymbol{D}_\theta^{(h)} &= \frac{1}{n}\Big(\boldsymbol{X}^{(h)}\odot\boldsymbol{Y}\Big)^T\Big(\boldsymbol{X}^{(h)}\odot\boldsymbol{Y}\Big),\\
\boldsymbol{G}_\theta^{(h,l)} &= \frac{1}{n}\Big(\boldsymbol{X}^{(l)}\odot\boldsymbol{Y}\Big)^T\Big(\boldsymbol{X}^{(h)}\odot\boldsymbol{Y}\Big), & \boldsymbol{C}_\alpha &= \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Y},\\
\Delta &= \sum_{l=1,l\neq h}^H tr\Big(\boldsymbol{T}^s\big(\boldsymbol{T}^{s^1}\odot\big[\boldsymbol{B}_\theta^{(l)}\big]^T\big)\boldsymbol{G}_\theta^{(h,l)}\Big).
\end{aligned}$$

The $n\times p_n$ matrix $\boldsymbol{X}^{(h)}$ is defined as

$$\boldsymbol{X}^{(h)} = \begin{bmatrix} x_1^{(h)} & x_1^{(h)} & \dots & x_1^{(h)} \\ x_2^{(h)} & x_2^{(h)} & \dots & x_2^{(h)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(h)} & x_n^{(h)} & \cdots & x_n^{(h)} \end{bmatrix}_{n\times p_n}. \tag{2.13}$$

Let $\boldsymbol{\Sigma}$ denote a $p_n\times p_n$ diagonal matrix $\text{diag}(\alpha_{11}^{-1},...,\alpha_{p_np_n}^{-1})$, and let $\odot$ denote the component-wise product. Each entry of $\boldsymbol{B}_\theta^{(h)}$ and $\boldsymbol{B}_\alpha$ takes the following form

$$\left(\boldsymbol{B}_\theta^{(h)}\right)_{ij} = \begin{cases} -[\theta_{ij}]_h \alpha_{jj}^{-1} & i \neq j; \quad i,j = 1, ..., p_n, \\ 0 & i = j, \end{cases} \tag{2.14}$$

$$\left(\boldsymbol{B}_\alpha\right)_{ij} = \begin{cases} -\alpha_{ij} \alpha_{jj}^{-1} & i \neq j; \quad i,j = 1, ..., p_n, \\ 0 & i = j, \end{cases} \tag{2.15}$$

Let $\boldsymbol{T}^s$ and $\boldsymbol{T}^w$ denote the edge adjacency matrix, e.g. $(\boldsymbol{T}^s)_{ij} = 1$ if $(i,j) = s$ or $(j,i) = s$, and $(\boldsymbol{T}^s)_{ij} = 0$ otherwise. Let $\boldsymbol{T}^{s^1}$ be defined as: if $(i,j) = s$, all of the elements in the ith and jth row of the matrix $\boldsymbol{T}^{s^1}$ are equal to 1 except the diagonal entries, and all other entries are equal to zero. The matrix $\boldsymbol{T}^{w^1}$ is defined in the same way.

2.2.5 Theoretical Properties of Penalized Composite Likelihood Estimator

In this section, we provide the estimation error bound for the penalized composite likelihood estimator and demonstrate its model selection consistency under some regularity conditions. Let the score function and the Hessian matrix of $l_c(\boldsymbol{\beta})$ be denoted as $l_c^{(1)}(\boldsymbol{\beta}), l_c^{(2)}(\boldsymbol{\beta})$ respectively. We use the notation $\mathcal{S}_1$ to represent the set of true nonzero off-diagonal parameters and $\mathcal{S}_2$ to denote the set of all diagonal parameters. Then the set of true nonzero parameters can be denoted as $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \cup \mathcal{S}_2$.

Define $\boldsymbol{\mathcal{S}}^c$ as the set of all zero parameters. Let s_n denote the maximum number of nonzero parameters within each row across all rows and all matrices $\boldsymbol{Q}_0, \boldsymbol{P}_1, \cdots, \boldsymbol{P}_H$. Let q_n denote the total number of true nonzero off-diagonal parameters, where $q_n = \mathcal{O}(p_n s_n)$.

Assumption 3 Define $H(\boldsymbol{\beta}) = E\big\{ - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta})\big\}$ and there exists a neighborhood $||\boldsymbol{\beta} - \boldsymbol{\beta}^0||_2 < \eta$ for some constant $\eta > 0$ such that $H(\boldsymbol{\beta})$ has eigenvalues bounded away from zero and infinity. Furthermore, the absolute values of eigenvalues of $E\big\{\frac{\partial^3 - l_c(\mathbf{Y}_m, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T \partial \beta_u}\big\}$ are bounded by some constant, where β_u denotes any off-diagonal parameter.

Assumption 4 There exists a constant $d > 0$ such that $s_n^4 = \mathcal{O}(p_n^d)$ and $p_n^{1+d} \log p_n = o(n)$.

Theorem 2.2 Under **Assumptions** 3-4, if the tuning parameter λ satisfies $\delta_1 \Big(\frac{s_n^2 \log p_n}{n}\Big)^{\frac{1}{2}} \leqslant \lambda \leqslant \delta_2 \Big(\frac{s_n^3 \log p_n}{n}\Big)^{\frac{1}{2}}$ for some constants $\delta_1, \delta_2 > 0$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of the objective function $Q(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}||_2 = \mathcal{O}_p\Big\{\Big(\frac{p_n^{1+d} \log p_n}{n}\Big)^{\frac{1}{2}}\Big\}$.

Note that if we further assume that s_n is bounded, the estimation error bound would be

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}||_2 = \mathcal{O}_p\Big\{\Big(\frac{p_n \log p_n}{n}\Big)^{\frac{1}{2}}\Big\}.$$

Assumption 5 Let $H^0_{\mathcal{S}_1\mathcal{S}_1} = E\Big\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big\}$ and $H^0_{\mathcal{S}^c\mathcal{S}_1} = E\Big\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\Big\}$. There exists some positive constant $\xi \in (0,1)$ such that

$$\left|\!\left|\!\left| H^0_{\mathcal{S}^c\mathcal{S}_1}\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1} \right|\!\right|\!\right|_\infty \leqslant 1-\xi.$$

This assumption states that the expected Hessian matrix satisfies the incoherence condition, which assumes that any two variables without an edge between them should not impose a large effect on variable pairs connected by edges in a Gaussian graphical model. This condition is widely used in high-dimensional statistics, including graphical models and regularized regression models (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Raskutti et al., 2008; Ravikumar et al., 2011). In the theorem above, we establish the consistency results for all vertex and edge parameters. As a next step, we investigate the problem of model selection, focusing on the edge parameters only, while the vertex parameters are treated as nuisance parameters and are assumed to be known.

Theorem 2.3 Under **Assumptions** 3-5 and **Theorem** 2.2, suppose that $q_n^2 p_n^{2d}\log p_n = \mathcal{O}(n)$ and the minimum non-zero parameter satisfies

$$\min_{u\in\mathcal{S}_1}\big|\beta_u^0\big| := \min_{\substack{(i,j)\in\mathcal{S}_1\\ h=1,\cdots,H}}\Big\{\big|[\theta^0_{ij}]_h\big|, \big|\alpha^0_{ij}\big|\Big\} \geqslant c^*\sqrt{q_n}\lambda$$

for some constant $c^* > 0$. Then the penalized composite likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^0)$ with probability tending to one.

2.3 Numerical Simulations

In this section, we provide several numerical studies to evaluate the performance of our proposed methods. Simulation results are summarized and reported over 100 data sets.

2.3.1 Covariate-dependent Gaussian Graphical Models with Fixed p

For the first simulation, we consider a small-sized cdexGGM with $p = 10$. Two different graphical structures are simulated for $\boldsymbol{Q}_0$, and $\boldsymbol{Q}_1$ matrices, including a general network and a chain network with the sample sizes $n = 500$ or 3000. One covariate $x_m^{(1)}$ is simulated at 5 different values, and for each specific value of the covariate, there are $n/5$ observations. The precision matrices are established by $\boldsymbol{\Sigma}_m^{-1} = x_m^{(1)}\boldsymbol{Q}_1 + (1 - x_m^{(1)})\boldsymbol{Q}_0$. In the first case, the matrices $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ are generated as two random positive definite matrices by the format of $\boldsymbol{A}\boldsymbol{A}^T + c \cdot \boldsymbol{I}$, where $\boldsymbol{A}$ is a random square matrix, $\boldsymbol{I}$ is the identity matrix, and c is an arbitrary positive constant. In the second case, we generate a tridiagonal precision matrix corresponding to a chain network as outlined in Fan et al. (2009). Specifically, each off-diagonal element of the inverse matrices of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_0$ is generated as $(\boldsymbol{Q}_1^{-1})_{ij} = \exp\big\{-1/2\big|s_i - s_j\big|\big\}$, and $(\boldsymbol{Q}_0^{-1})_{ij} = \exp\Big\{-1/2\Big|s'_i - s'_j\Big|\Big\}$, where s_1 and s'_1 are randomly initialized as

Table 2.2: Simulation results of MLE for small-sized cdexGGM

Sample Size	Case	Matrix	Mean $\|\|\boldsymbol{e}\|\|$	Mean $\|\|\boldsymbol{e}\|\|$ /# of Para.
500	General	$\boldsymbol{Q}_0$	2.6645(1.0731)	0.0484
		$\boldsymbol{Q}_1$	2.6304(0.9897)	0.0478
	Chain	$\boldsymbol{Q}_0$	2.8332(0.6690)	0.0515
		$\boldsymbol{Q}_1$	2.8655(0.6518)	0.0521
3000	General	$\boldsymbol{Q}_0$	0.9573(0.3243)	0.0174
		$\boldsymbol{Q}_1$	0.8403(0.3017)	0.0153
	Chain	$\boldsymbol{Q}_0$	0.9705(0.1571)	0.0176
		$\boldsymbol{Q}_1$	0.9835(0.1603)	0.0179

The number of vertices p is 10 and the total # of parameters is 55 for each matrix. The standard errors are listed in parentheses.

positive numbers, $s_i - s_{i-1} \sim Unif(0.5, 1)$, and $s'_i - s'_{i-1} \sim Unif(0.5, 1)$. We evaluate the average ℓ_2 norm of the error vector $\boldsymbol{e} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ over 100 simulations and further divide it by the number of parameters as a measure of error for the proposed estimates. As shown by Table 2.2, the measure of error is around 0.05 for $n = 500$, and around 0.01 for $n = 3000$.

2.3.2 Covariate-dependent Gaussian Graphical Models with Large p_n

In this section, we consider high-dimensional covariate-dependent GGMs with $p_n = 30$ and $p_n = 50$, along with the same sample size $n = 3000$ and one covariate $H = 1$. We generate random matrices $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ with a sparse network structure, using the construction of $\boldsymbol{A}\boldsymbol{A}^T + c \cdot \boldsymbol{I}$, where $\boldsymbol{A}$ is a random square matrix. The diagonal elements of the precision matrix are set to be linearly varied with the covariate in this study.

The tuning parameter λ is determined by composite loglikelihood EBIC (Foygel and Drton, 2010; Gao and Song, 2010), which is given below

$$EBIC_\gamma = -2l_c(\hat{\boldsymbol{\beta}}_c) + df \log n + 4df\gamma \log p_n, \tag{2.16}$$

where df denotes the total number of nonzero off-diagonal entries and $\gamma \in [0, 1]$ is a user-specified constant. We perform a sensitivity analysis on the influenza vaccination dataset to investigate the impact of γ on the estimation results. Please refer to **Section** 4.2.1 for details. For our simulation studies, we set $\gamma = 1$. We use different model evaluation metrics, including specificity, sensitivity, and Matthews Correlation

Coefficient (MCC), which are defined as

$$sensitivity = \frac{TP}{TP + FN}, \qquad specificity = \frac{TN}{TN + FP},$$
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

where TP, TF, FP, and FN represent the total number of true positives, true negatives, false positives, and false negatives, respectively. In our setting, any nonzero off-diagonal entry is treated as a positive instance (indicating an edge), while a zero entry is considered a negative instance. We avoid relying on MSE for penalized estimators because the primary interest lies in whether the model accurately identifies the presence or absence of edges, rather than how precisely it estimates the numerical values. Nevertheless, we also compute estimation errors across multiple random seeds for verification, and these remain consistently low.

The proposed method (cdexGGM) is compared with four other methods: graphical lasso (Guo et al., 2011), fused graphical lasso, group graphical lasso (Danaher et al., 2014), and the Gaussian graphical regression model (Zhang and Li, 2023). Graphical lasso (glasso) is employed to jointly estimate multiple graphical models for different categories, and the tuning parameter λ is selected by the Bayesian information criterion (BIC) (Guo et al., 2011). The fused graphical lasso (FGL) and the group graphical lasso (GGL) are similar approaches with different penalties, both utilized for modeling Gaussian graphical models across different categories. The GGL imposes a weaker similarity constraint (only encourages similar sparsity patterns) on the multiple precision matrices compared to the FGL, which not only emphasizes similar edge values across multiple categories but also enforces similar sparsity patterns. The tuning parameters λ_1, λ_2 are selected based on an approximation of the Akaike Information Criterion (AIC) (Danaher et al., 2014). The Gaussian graphical regression model (RegGMM) employs a regression-based approach to estimate graphical models that vary with covariates. The tuning parameters λ and λ_g are jointly selected by 5-fold cross-validation, as suggested in Zhang and Li (2023). To ensure a fair comparison, the covariate is uniformly selected from 0 to 1 with 5 different values.

Table 2.3 highlights the average of these metrics over 100 simulation repetitions. Standard errors for each metric are provided in parentheses. For $p_n = 30$, our proposed method demonstrates superior performance in sensitivity and MCC, while maintaining comparable specificity to other approaches. Both the GGL and FGL demonstrate acceptable performance, despite not being intended for modeling covariate-dependent network structure. The RegGMM assumes constant diagonal parameters across all levels of covariates and therefore does not perform well under the varying diagonal parameters setting. Our proposed method outperforms competing approaches across all measured criteria when $p_n = 50$. Other methods perform well in sensitivity, but at the expense of specificity, leading to reduced MCC scores.

Finally, we examine the computational cost associated with the tuning process. Table 2.4 presents the EBIC-based computation time for λ selection across various settings. Simulations were executed on a Windows 11 Home laptop with a 3.4 GHz Intel Core i7-1139 CPU. With $O((H+1)p_n^2)$ parameters, the total computation cost grows roughly quadratically with p_n, as shown in Table 2.4. Our approach enables parallel computation across the 100 simulated datasets, with a runtime of approximately 7.95 minutes per seed when $p_n = 30$ and around 43.53 minutes per seed when $p_n = 50$.

2.3.3 High-Dimensional Covariate-dependent Gaussian Graphical Models with Multiple Covariates

In the third simulation study, we limit our comparison to our proposed method and RegGMM, as both are designed for modeling covariate-dependent network structures. We examine two covariates, generated uniformly from 0 to 1. We set $p_n = 20$ and $p_n = 30$, with the sample size $n = 3000$. For the purpose of a fair comparison, we consider the diagonal entry of the precision matrix to be constant across all levels of covariates. The generation process for the true $\boldsymbol{Q}_0, \boldsymbol{P}_1, \boldsymbol{P}_2$ is as follows:

1. Generate $\boldsymbol{Q}_0, \boldsymbol{Q}_1, \boldsymbol{Q}_2$ as random sparse positive definite matrices as outlined

Table 2.3: Simulation results of cdexGGM and the other four competing methods.

p_n	# of Para.	Method	Matrix	Sensitivity	Specificity	MCC
30	465	cdexGGM	$\boldsymbol{Q}_0$	0.8747(0.0307)	0.6576(0.0526)	0.4401(0.0424)
			$\boldsymbol{Q}_1$	0.9251(0.0232)	0.6275(0.0573)	0.4388(0.0450)
		GGL	$\boldsymbol{Q}_0$	0.7904(0.0207)	0.6183(0.0224)	0.3361(0.0259)
			$\boldsymbol{Q}_1$	0.7876(0.0179)	0.7099(0.0169)	0.4036(0.0217)
		FGL	$\boldsymbol{Q}_0$	0.7902(0.0208)	0.6147(0.0225)	0.3328(0.0259)
			$\boldsymbol{Q}_1$	0.7880(0.0178)	0.7090(0.0168)	0.4030(0.0221)
		glasso	$\boldsymbol{Q}_0$	0.8003(0.0194)	0.5848(0.0226)	0.3160(0.0243)
			$\boldsymbol{Q}_1$	0.7926(0.0162)	0.6918(0.0166)	0.3903(0.0198)
		RegGMM	$\boldsymbol{Q}_0$	0.6035(0.1476)	0.6333(0.1083)	0.2018(0.0574)
			$\boldsymbol{Q}_1$	0.7135(0.1190)	0.5850(0.1442)	0.2458(0.0491)
50	1275	cdexGGM	$\boldsymbol{Q}_0$	0.9091(0.0236)	0.8202(0.0239)	0.3737(0.0251)
			$\boldsymbol{Q}_1$	0.7838(0.0282)	0.9011(0.0145)	0.4152(0.0275)
		GGL	$\boldsymbol{Q}_0$	0.7568(0.0203)	0.7603(0.0101)	0.2454(0.0130)
			$\boldsymbol{Q}_1$	0.8098(0.0292)	0.6551(0.0105)	0.1948(0.0131)
		FGL	$\boldsymbol{Q}_0$	0.7570(0.0203)	0.7577(0.0103)	0.2435(0.0130)
			$\boldsymbol{Q}_1$	0.8104(0.0285)	0.6400(0.0101)	0.1872(0.0127)
		glasso	$\boldsymbol{Q}_0$	0.7573(0.0203)	0.7528(0.0108)	0.2400(0.0129)
			$\boldsymbol{Q}_1$	0.8113(0.0250)	0.6249(0.0098)	0.1801(0.0113)
		RegGMM	$\boldsymbol{Q}_0$	0.8025(0.0739)	0.5431(0.0909)	0.1470(0.0215)
			$\boldsymbol{Q}_1$	0.7904(0.0917)	0.5039(0.1077)	0.1211(0.0251)

The standard errors for each evaluation metric are shown in parentheses. The # of true nonzero parameters of $\boldsymbol{Q}_0, \boldsymbol{Q}_1$ are 123 and 114 when $p_n = 30$, and 106 and 102 when $p_n = 50$.

Table 2.4: Computational cost of cdexGGM

p_n	# of Para.	Computational time
30	930	7.95 mins
50	2550	43.53 mins

in the previous simulation. Let $(\boldsymbol{Q}_0)_{ij} = \alpha_{ij}$, $(\boldsymbol{Q}_1)_{ij} = \eta_{ij}$, and $(\boldsymbol{Q}_2)_{ij} = \pi_{ij}$ represent the $(i,j)th$ element of $\boldsymbol{Q}_0$, $\boldsymbol{Q}_1$, and $\boldsymbol{Q}_2$ respectively.

2. Define $\boldsymbol{N}_1^* = \Lambda_1 \boldsymbol{Q}_1 \Lambda_1$, $\boldsymbol{N}_1^{**} = \Lambda_0 \boldsymbol{N}_1^* \Lambda_0$, and $\boldsymbol{N}_2^* = \Lambda_2 \boldsymbol{Q}_2 \Lambda_2$, $\boldsymbol{N}_2^{**} = \Lambda_0 \boldsymbol{N}_2^* \Lambda_0$, where $\Lambda_0 = diag\Big(\sqrt{\tfrac{1}{2}\alpha_{11}}, \cdots, \sqrt{\tfrac{1}{2}\alpha_{p_np_n}}\Big)$, $\Lambda_1 = diag\Big(\sqrt{\eta_{11}^{-1}}, \cdots, \sqrt{\eta_{p_np_n}^{-1}}\Big)$, $\Lambda_2 = diag\Big(\sqrt{\pi_{11}^{-1}}, \cdots, \sqrt{\pi_{p_np_n}^{-1}}\Big)$.

3. Generate $\boldsymbol{P}_1, \boldsymbol{P}_2$ by $\boldsymbol{P}_1 = \boldsymbol{N}_1^{**} - \frac{1}{2}\boldsymbol{Q}_0$, $\boldsymbol{P}_2 = \boldsymbol{N}_2^{**} - \frac{1}{2}\boldsymbol{Q}_0$.

As shown in Table 2.5, both methods demonstrate strong performance in estimating the sparsity patterns of the slope matrices $\boldsymbol{P}_1, \boldsymbol{P}_2$. In terms of the baseline precision matrix $\boldsymbol{Q}_0$, our method strikes a balance between high sensitivity and relatively good specificity, while RegGMM offers slightly higher sensitivity but at the cost of specificity. Overall, the proposed method exhibits consistent and reliable performance across all settings and evaluation metrics.

Table 2.5: Simulation results of our proposed method and RegGMM on multiple covariates

p_n	Method	Matrix	# of Para.	# of Nonzero Para.	Sensitivity	Specificity	MCC
20	cdexGGM	$\boldsymbol{Q}_0$	210	55	0.8409(0.0409)	0.6870(0.0780)	0.4215(0.0601)
		$\boldsymbol{P}_1$	190	67	0.7576(0.0328)	0.8410(0.0945)	0.5997(0.0986)
		$\boldsymbol{P}_2$	190	70	0.9097(0.0268)	0.8351(0.0820)	0.7286(0.0790)
	RegGMM	$\boldsymbol{Q}_0$	210	55	0.8866(0.0498)	0.3015(0.0393)	0.1652(0.0495)
		$\boldsymbol{P}_1$	190	67	0.6760(0.0433)	0.8913(0.0634)	0.5926(0.0776)
		$\boldsymbol{P}_2$	190	70	0.7943(0.0332)	0.8521(0.0488)	0.6427(0.0589)
30	cdexGGM	$\boldsymbol{Q}_0$	465	116	0.8877(0.0276)	0.7432(0.0505)	0.5202(0.0447)
		$\boldsymbol{P}_1$	435	161	0.8035(0.0296)	0.8954(0.0450)	0.7040(0.0493)
		$\boldsymbol{P}_2$	435	147	0.8301(0.0281)	0.8874(0.0527)	0.7125(0.0598)
	RegGMM	$\boldsymbol{Q}_0$	465	116	0.8888(0.0328)	0.3221(0.0250)	0.1873(0.0303)
		$\boldsymbol{P}_1$	435	161	0.6350(0.0510)	0.8837(0.0602)	0.5481(0.0547)
		$\boldsymbol{P}_2$	435	147	0.6972(0.0300)	0.8345(0.0491)	0.5317(0.0547)

The sample size is 3000, and the standard errors over 100 datasets are listed in parentheses.

2.4 Data Analysis

2.4.1 Influenza Vaccination Study

Nowadays, influenza continues to be a major concern for public health, including the most prevalent strains of influenza A (H1N1), influenza B, and influenza A (H3N2). Insights obtained from animal models have revealed that host genetic factors can have a profound impact on both the immune responses and susceptibility to influenza infection (Trammell and Toth, 2008; Srivastava et al., 2009). The immune response to vaccination, much like in viral infection, is expected to exhibit variability that is influenced by genotype as well. Motivated by this, we apply our proposed model to a global gene expression data set (before and after trivalent influenza vaccination in humans), aiming to characterize the dynamic regulatory gene networks and examine the varying dependency structure at different time points. This analysis may provide insights into pivotal genes that greatly contribute to the immune response to influenza vaccination.

The gene expression data set is downloaded from the National Center for Biotechnology Information website (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48024`). The total sample size is 848, including 119 healthy adult male volunteers (aged 19 - 41 years) and 128 healthy adult female volunteers (aged 19 - 41 years). It records global transcript abundance in peripheral blood RNA specimens before (day 0) and at three time points (days 1, 3, and 14) after vaccination. Raw data is normalized by background adjustment, variance stabilization transformation (Lin et al., 2008) and robust spline normalization using the R package *lumi* (Du et al., 2008).

Similar to Franco et al. (2013), we opt to narrow down our focus to 68 genes that exhibit both a transcriptional response to the vaccine and some evidence of genetic regulation. In summation, the cleaned data set has $p_n = 68$ genes, with the sample size $n = 848$ and four different time points (day 0, days 1, days 3, days 14). Before proceeding with further analysis, we first examine whether the precision matrix shows a roughly linear relationship with time. Specifically, for each time point, we fit a GGM using the graphical lasso (Guo et al., 2011). We then randomly select several estimated entries $(\mathbf{\Sigma}^{-1})_{ij}$ and plot these estimates against time to evaluate whether a linear relationship is approximately satisfied. As demonstrated in Figure 2.1, the randomly selected entries exhibit an approximately linear pattern with time. Therefore, we proceed to analyze the dataset using the proposed cdexGGM. The estimated time-varied regulatory networks and sparsity pattern are given in Figure 2.2 and Figure 2.3, respectively.

As shown by Figure 2.2, several gene interactions experience a rapid decrease after vaccination, including ABCA7 & LRRC37A4, KIAA0391 & RABEP1, C3AR1 & OAS1, and DYNLT1 & OAS1. Some of those even disappear at the end of the 14 days, like D4S234E & OAS1, EMR3 & OAS1, and GRINA & RABEP1. In contrast, getting vaccinated substantially enhances the interactions DIP2A & MGC57346 and OAS1 & TAP2. It also produces the newly-formed edges DIP2A & NT5DC3 and OAS1 & TIMM10. These findings demonstrate some evidence that DIP2A and OAS1 are crucial to people's immune response. This aligns with the results reported in Franco et al. (2013), which recognizes the importance of DIP2A and OAS1 in the humoral immune response to influenza vaccination. In general, our research validates the pivotal role of a specific set of genes in the human immune response to vaccination. Further biological investigations are necessary to fully understand the dynamic relationships among these genes with respect to vaccination time and their biological significance.

2.4.2 Mice Protein Expression

Down syndrome (DS), also known as trisomy 21, is a genetic disorder caused by the presence of an extra copy of chromosome 21, which pertains to intellectual disability, influencing approximately one in a thousand newborns globally. It is generally considered that the excessive expression of genes encoded by the extra

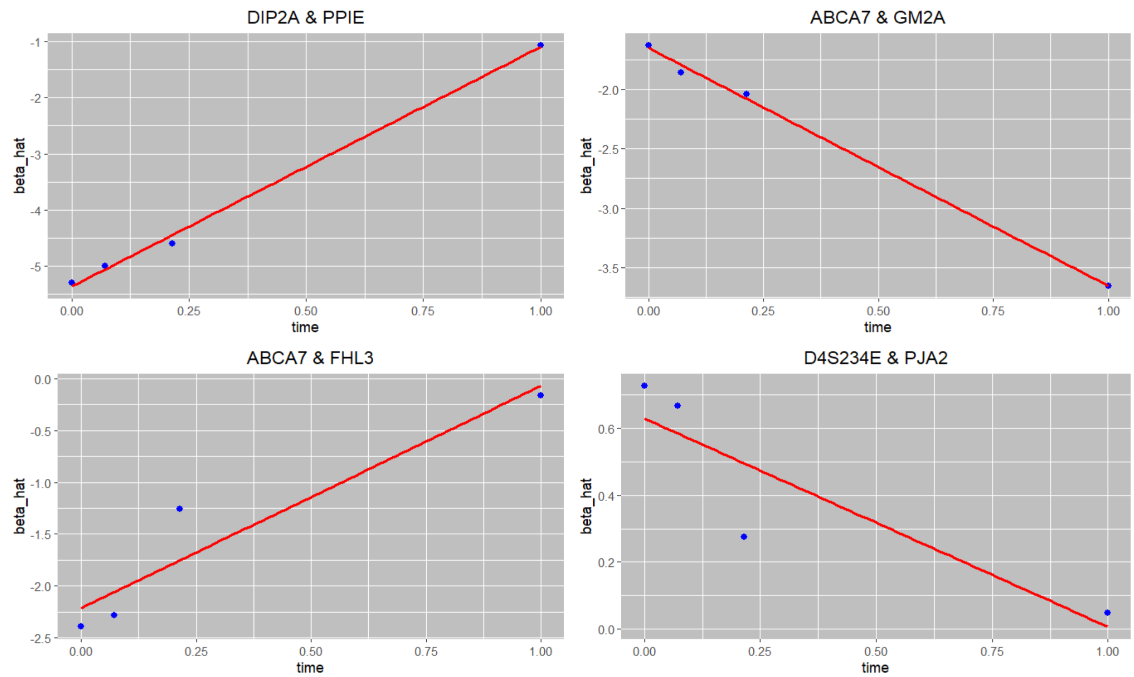


Figure 2.1: Scatter plots of randomly selected gene pairs.

copy of the chromosome gives rise to the disruption of regular pathways and typical reactions towards stimulation, and further leads to learning and memory deficiency. In the following analysis, we utilize the proposed covariate-dependent GGM method to analyze a mice protein expression data set to discover differences and similarities among protein networks influenced by three different covariates. The covariates include the treatment effect of injecting the drug memantine (m) or not (s), the factor of having the stimulation to learn (CS) or not (SC), and the factor of normal

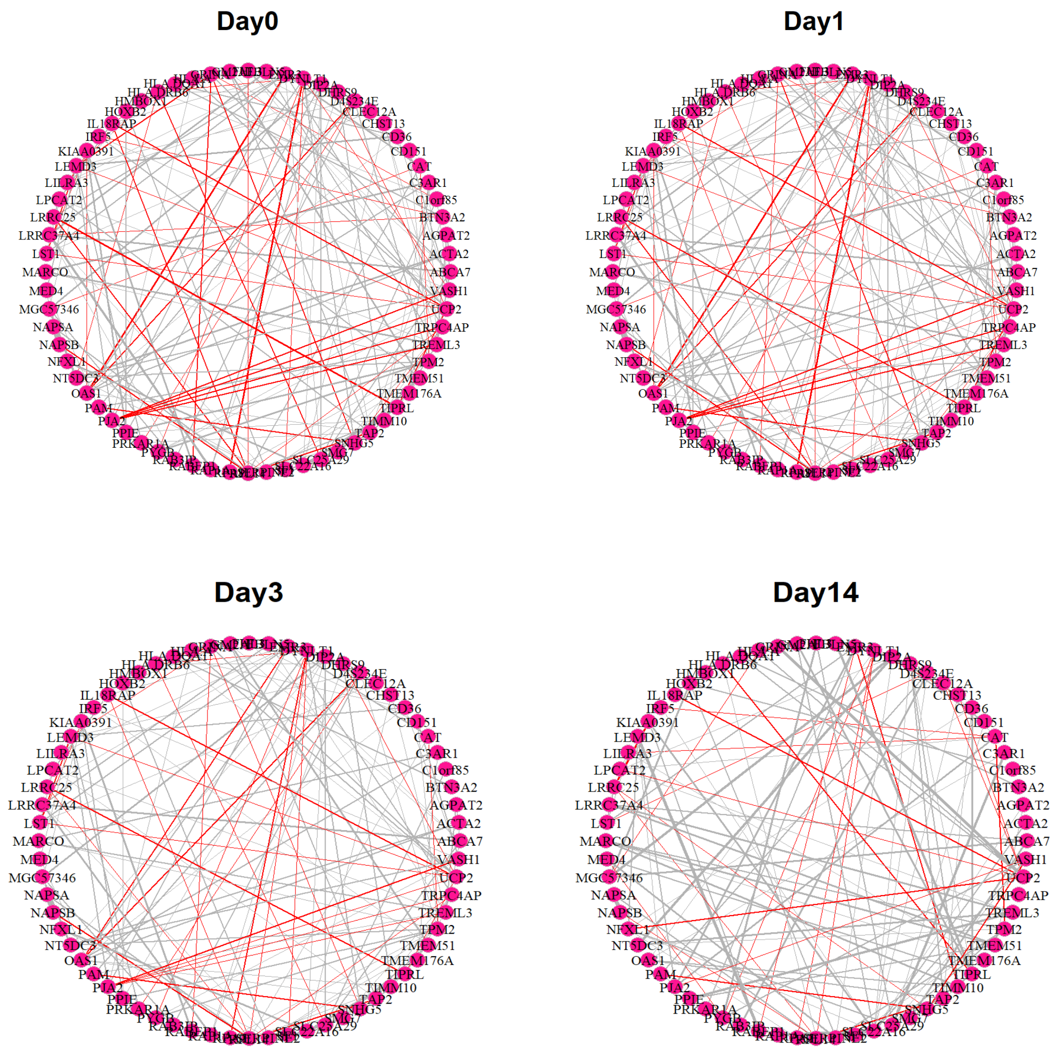


Figure 2.2: The estimated gene regulatory networks for the influenza vaccine data set. The time points "Day0", "Day1", "Day3", and "Day14" represent the day before getting vaccinated, the first day after vaccination, the third day after vaccination, and two weeks after vaccination. Edges in red represent negative partial correlations between genes, while edges in gray stand for positive partial correlations. The strength of interactions between two genes is displayed by the thickness of the edges.

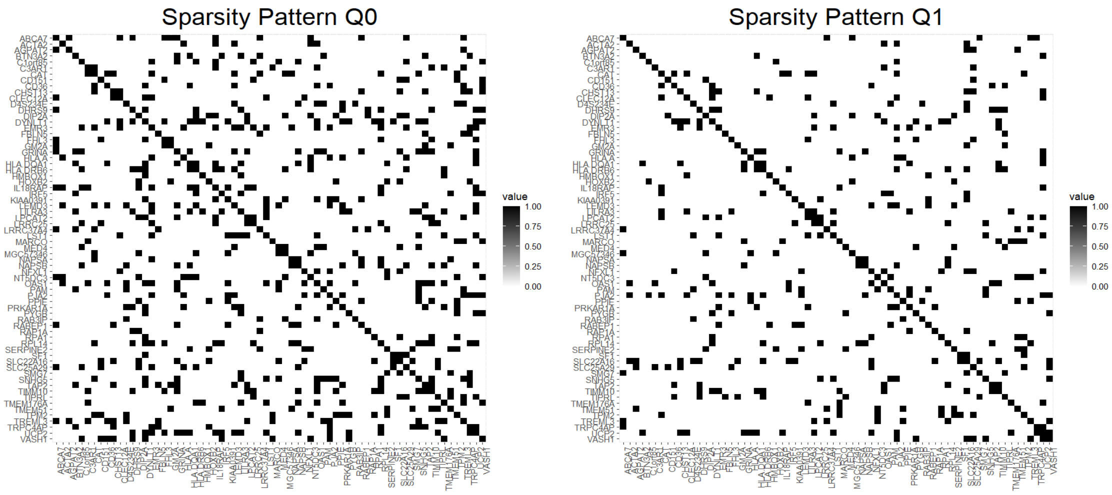


Figure 2.3: The estimated sparsity pattern for the influenza vaccine data set. Each image has 68 rows and 68 columns, representing the elements of $\boldsymbol{Q}_0, \boldsymbol{Q}_1$ respectively. The corresponding cell is marked in black if the estimated entry is nonzero. Otherwise, it is white.

genotype (c) or abnormal trisomy (t). The resulting 8 classes of mice are grouped by the level of these three factors. Each class is examined and labeled by the associated learning outcome, including normal learning, failed learning, and rescued learning (refer to Figure 2.4 for details). We show that our proposed model enables us to investigate the differences in the networks across different levels of three experimental factors. These findings shed light on how the genetic factor, medical treatment, and stimulation factor influence the interactions among the proteins, and how these interactions in turn affect the learning outcomes.

This mice protein expression data set (Higuera et al., 2015) is downloaded from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression`). It comprises the expression levels of 77 proteins that produce detectable signals in the nuclear fraction of the cortex. In total, there are 72 mice involved in the study, with 38 of them serving as control mice and 34 as trisomic mice with Down syndrome. In the experiment, 15 measurements of each protein are recorded per mouse. Accordingly, there are 570 measurements for control mice and 510 measurements for trisomic mice, with a total of 1080 measurements per protein. In particular, based on the information provided by the dataset, the repository treats each measurement as a separate observation. In our analysis, we adopt this simplification but acknowledge that ignoring within-mouse variability may underestimate the true dependence structure. A more refined approach would require a mixed-effects or hierarchical modelling framework, which we list as a potential direction for future work. The mice are evenly distributed across eight classes, with each class comprising around seven to ten mice. Excluding those proteins with missing values and further conducting feature selection through a random forest algorithm (using the eight classes as the response and all protein expression levels as predictors), we finally narrow down our focus to the top 35 proteins with the highest feature importance scores. We proceed to center each observation for the subsequent analysis.

To investigate how covariates influence the underlying dependence structure, we employ the covariate-dependent GGMs through the proposed estimation approach. The following Figure 2.5 illustrates the estimated protein network structures for the 35 selected proteins across eight classes. We further examine the estimated $[\theta_{ij}]_h$ in slope matrices $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, an $\boldsymbol{P}_3$. We select the top-ranked $[\hat{\theta}_{ij}]_h$ in terms of magnitude and perform hypothesis testing $H_0 : [\theta_{ij}]_h = 0$ versus $H_1 : [\theta_{ij}]_h \neq 0$, which is used to test whether or not the corresponding covariate has a significant effect on the specific protein interactions. We construct Wald-type statistics and the standard errors are obtained through the bootstrap method. Table 2.6 highlights those pairs of protein interactions and provides the associated p-values. For example, we observe a positive effect of the trisomy genotype on the interactions Tau_N & P3525_N, PKCA_N & pNUMB_N, and pRSK_N & RSK_N, while the following protein pairs RAPTOR_N & pGSK3B_N, PKCA_N & CAMKII_N, and pRSK_N & RSK_N exhibit a negative response to the injection of memantine. It is interesting to note that hypothesis testing on the pair pRSK_N & RSK_N is significant for both genotype and treatment. Furthermore, the trisomic geneotype is increasing this interaction, whereas the medical treatment is decreasing this interaction. This suggests that the treatment alleviates the harmful impact of the abnormal trisomy genotype on this specific pair of protein interaction. Additionally, we observe that both the memantine treatment and the learning stimulation assistance impose a positive impact on pJNK_N & pGSK3B_N, and meanwhile, negatively influence pNUMB_N & pMTOR_N. These results may offer valuable insights into the importance of those protein pairs in the mechanism of learning disabilities in Down syndrome.

As shown in Figure 2.4, different mouse classes have different learning outcomes, including normal learning, rescued learning, and failed learning. We are interested in comparing dependence structures between two different classes with different learning outcomes. We highlight two sets of comparisons: 1) the class "c-CS-m" demonstrating normal learning outcome versus the class "t-CS-m" demonstrating rescued learning outcome; 2) the class "c-CS-s" demonstrating normal learning outcome versus the class "t-CS-s" demonstrating failed learning outcome. We perform a two-sample test to compare the equality of partial correlations between any given protein pair across the two classes. The test statistic is constructed based on the composite likelihood estimates of the partial correlations, and the standard errors are obtained through the bootstrap method. Both investigations reveal a notable enhancement in the interactions ITSN1_N & BRAF_N, DYRK1A_N & ITSN1_N, pGSK3B_Tyr216_N

& SHH_N, pP70S6_N & pRSK_N, S6_N & ADARB1_N, AcetylH3K9_N & S6_N, and pCAMKII_N & ADARB1_N in normal mice, compared to those experiencing either rescued or failed learning. The p-values of the two-sample tests are provided in Table 2.7. These findings provide evidence that the interactions between these protein pairs are essential to the learning outcome. Some of our results agree with existing findings in biological literature. For example, Malakooti et al. (2020) demonstrates that ITSN1_N contributes to normal learning and memory in mice. Ahmed et al. (2015) reveals that overexpression of DYRK1A can lead to learning and memory deficits. Both proteins have significant results in our two-sample test.

We apply a random forest for feature selection because it enables effective dimension reduction, which in turn significantly decreases the computational cost. We acknowledge that using Random Forest importance scores to pre-select a small set of proteins may introduce a selection bias ("winner's curse"), since the same data are used both for screening and for subsequent inference. This means that some proteins may appear more discriminative by chance, which can affect the interpretation of the bootstrap hypothesis testing results. In light of this, the bootstrap results should be viewed as exploratory rather than fully confirmatory. In summary, our analysis outlines a preliminary framework for how such an inference procedure could be carried out, but additional empirical validation, ideally guided by biological or genetic expertise, will be necessary to draw firm conclusions.

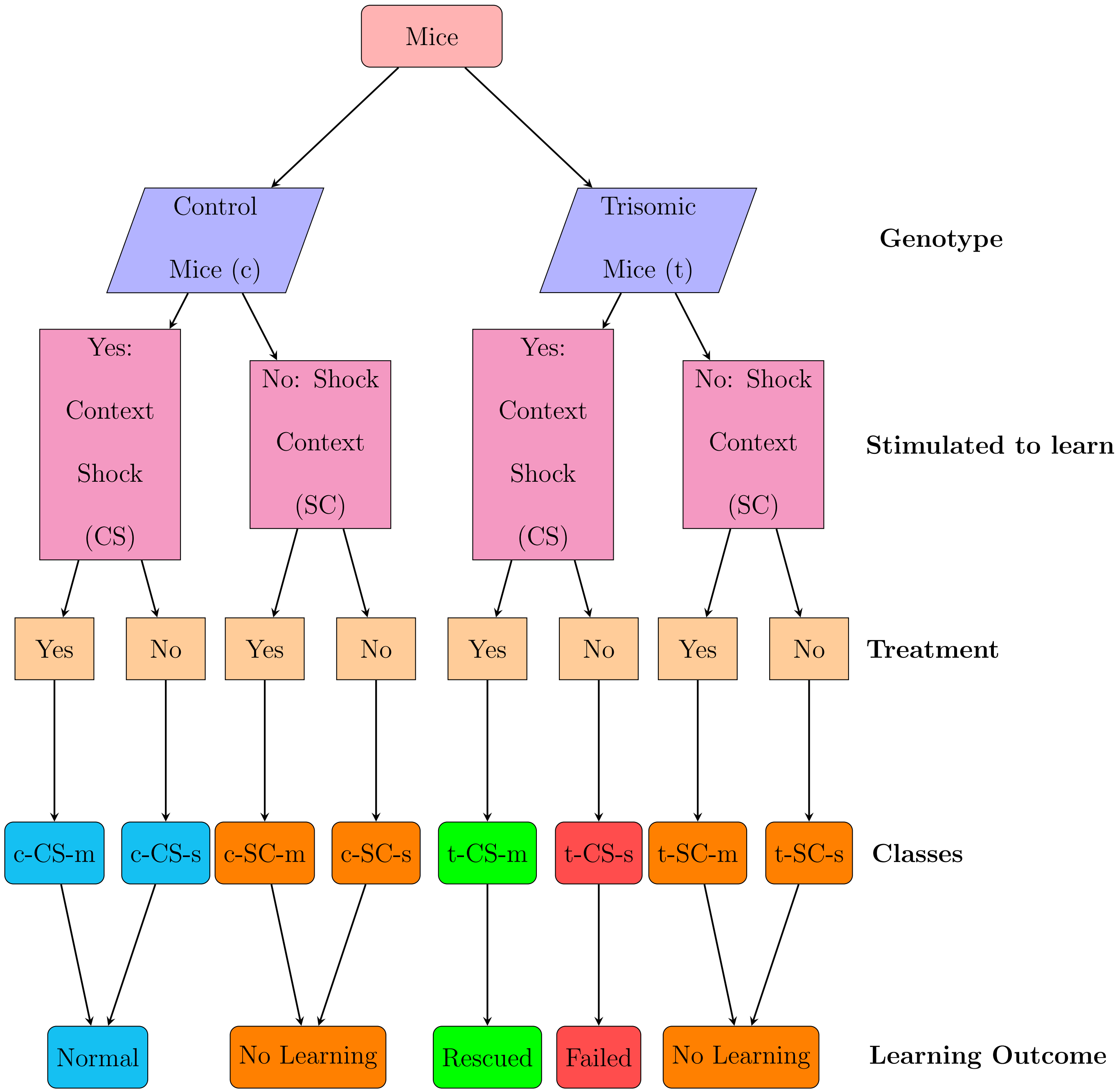


Figure 2.4: Eight Classes of Mice and the associated Learning Outcomes

Table 2.6: Top-ranked gene pairs with significant covariate effects

Covariates	Matrices	Interactions	
		Positive $[\hat{\theta}_{ij}]_h$	Negative $[\hat{\theta}_{ij}]_h$
		Tau_N & P3525_N (9.86E-11)	pNUMB_N & RAPTOR_N (5.19E-3)
		PKCA_N & pNUMB_N (5.83E-3)	pJNK_N & P3525_N (9.12E-3)
Genotype	$\boldsymbol{P}_1$	pRSK_N & RSK_N (8.67E-4)	Tau_N & PKCA_N (1.19E-5)
		RSK_N & P3525_N (4.55E-4)	Tau_N & SHH_N (2.95E-5)
		Tau_N & RSK_N (1.60E-2)	pNUMB_N & pGSK3B_N (1.41E-2)
		pNUMB_N & RSK_N (4.11E-4)	RAPTOR_N & pGSK3B_N (2.86E-2)
		pNUMB_N & pRSK_N (1.75E-7)	pRSK_N & RSK_N (1.54E-3)
Treatment	$\boldsymbol{P}_2$	PKCA_N & P3525_N (1.24E-2)	PKCA_N & CAMKII_N (4.99E-4)
		pJNK_N & pGSK3B_N (7.78E-2)	pNUMB_N & pMTOR_N (3.59E-5)
		pGSK3B_N & P3525_N (8.47E-2)	Tau_N & RSK_N (3.08E-2)
		pJNK_N & pGSK3B_N (2.75E-2)	PKCA_N & pGSK3B_N (1.27E-2)
		GluR3_N & RSK_N (3.50E-2)	pJNK_N & SHH_N (2.74E-7)
Behaviour	$\boldsymbol{P}_3$	pMTOR_N & P3525_N (8.48E-9)	PKCA_N & P3525_N (1.68E-4)
		pNUMB_N & pGSK3B_N (7.12E-2)	RAPTOR_N & RSK_N (8.51E-2)
		CAMKII_N & SHH_N (4.37E-4)	pNUMB_N & pMTOR_N (6.88E-5)

Note: The associated p-values are provided in parentheses.

Table 2.7: Test of equality of partial correlations across two classes with different learning outcomes

Edges	Comparison between two classes			
	Normal learning c-CS-m	Rescued learning t-CS-m	Normal learning c-CS-s	Failed learning t-CS-s
BRAF_N & ITSN1_N	1.75E-11		1.35E-8	
DYRK1A_N & ITSN1_N	5.85E-5		2.86E-4	
pGSK3B_Tyr216_N & SHH_N	<2.2E-16		<2.2E-16	
pP70S6_N & pRSK_N	1.60E-8		3.28E-4	
S6_N & ADARB1_N	<2.2E-16		<2.2E-16	
AcetylH3K9_N & S6_N	<2.2E-16		<2.2E-16	
pCAMKII_N & ADARB1_N	<2.2E-16		<2.2E-16	

Note: The table provides the p-values associated with the two-sample tests to compare the equality of partial correlations between protein pairs across two different classes.

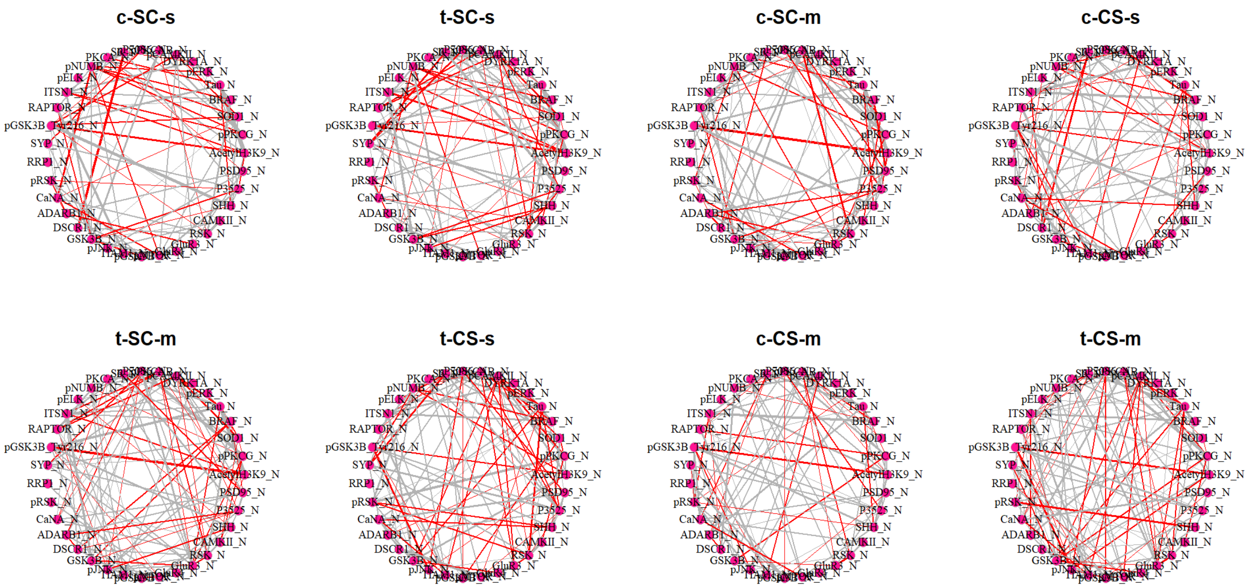


Figure 2.5: Estimated protein dependence structure: c-CS-m (control group, stimulated to learn, memantine); t-CS-m (trisomy group, stimulated to learn, memantine); c-CS-s (control group, stimulated to learn, saline); c-SC-m (control group, not stimulated to learn, memantine); t-CS-s (trisomy group, stimulated to learn, saline); t-SC-m (trisomy group, not stimulated to learn, memantine); c-SC-s (control group, not stimulated to learn, saline); t-SC-s (trisomy group, not stimulated to learn, saline).

2.5 Technical Lemmas and Proofs

Lemma 2.1 If the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_{p_n})^T$ follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, with $\Sigma_{jj} = \sigma_{jj}^2 < \infty$, $j = 1, ..., p_n$, then there exists some constant $0 < \mathcal{M}_1 < \infty$ such that

$$\Big|E\{Y_i^{h_1} Y_j^{h_2}\}\Big| < \mathcal{M}_1, \quad \Big|E\{Y_i^{h_1} Y_j^{h_2} Y_l^{h_3}\}\Big| < \mathcal{M}_1, \quad \Big|E\{Y_i^{h_1} Y_j^{h_2} Y_k^{h_3} Y_l^{h_4}\}\Big| < \mathcal{M}_1,$$

where h_1, h_2, h_3, h_4 are finite positive integers and $i, j, k, l = 1, ...p_n$.

Proof. It is a well-established fact that the central absolute moments for all even orders of normal random variables with mean zero are given by $E\{\big|Y_j\big|^r\} = \sigma_{jj}^r (r-1)!!$, where $(r-1)!!$ denotes the double factorial, that is, $(r-1)!! = (r-1) \cdot (r-3) \cdots 1$ and r is some even number. By Jensen's inequality and Hölder's inequality, $\forall\ i, j, k, l = 1, ...p$, we have

$$\begin{aligned}
&|E\{Y_i^{h_1} Y_j^{h_2}\}| \leqslant E\{|Y_i^{h_1} Y_j^{h_2}|\} \leqslant \Big\{E\{|Y_i^{h_1}|^2\}\Big\}^{\frac{1}{2}} \Big\{E\{|Y_j^{h_2}|^2\}\Big\}^{\frac{1}{2}} \leqslant \mathcal{M}_1, \\
&\Big|E\{Y_i^{h_1} Y_j^{h_2} Y_l^{h_3}\}\Big| \leqslant E\Big\{|Y_i^{h_1} Y_j^{h_2} Y_l^{h_3}|\Big\} \leqslant \Big\{E\{|Y_i^{h_1} Y_j^{h_2}|^2\}\Big\}^{\frac{1}{2}} \Big\{E\{|Y_l^{h_3}|^2\}\Big\}^{\frac{1}{2}} \leqslant \mathcal{M}_1, \\
&\Big|E\{Y_i^{h_1} Y_j^{h_2} Y_k^{h_3} Y_l^{h_4}\}\Big| \leqslant E\Big\{|Y_i^{h_1} Y_j^{h_2} Y_k^{h_3} Y_l^{h_4}|\Big\} \leqslant \Big\{E\{|Y_i^{h_1} Y_j^{h_2}|^2\}\Big\}^{\frac{1}{2}} \Big\{E\{|Y_k^{h_3} Y_l^{h_4}|^2\}\Big\}^{\frac{1}{2}} < \mathcal{M}_1.
\end{aligned}$$

□

Lemma 2.2 For $\forall\ s, w = \{(a,b)\} \in \mathcal{E}$ and $j = 1, \cdots, p_n, h_1 = 1, \cdots, H$, we have

$$\max_{s}\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}}\right| = \mathcal{O}_p\left\{(ns_n^2 \log p_n)^{\frac{1}{2}}\right\}, \quad \max_{w}\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial \alpha_w}\right| = \mathcal{O}_p\left\{(ns_n^2 \log p_n)^{\frac{1}{2}}\right\},$$
$$\max_{1\leqslant j\leqslant p_n}\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}}\right| = \mathcal{O}_p\left\{(ns_n^4 \log p_n)^{\frac{1}{2}}\right\}, \quad \max_{1\leqslant j\leqslant p_n}\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial \alpha_{jj}}\right| = \mathcal{O}_p\left\{(ns_n^4 \log p_n)^{\frac{1}{2}}\right\}.$$

Furthermore, for $\forall\ \epsilon > 0$, there exist some positive constants $\mathcal{K}_1, C_1, c_{z,max}$ and $c_{\mathcal{I}_2,max}$ such that

$$\begin{aligned}
P\left\{\left\|\frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c}\right\|_\infty \geqslant \epsilon\right\} \leqslant\ & 4\exp\left\{-\min\left\{\frac{C_1 n\epsilon^2}{18 s_n^2 \mathcal{K}_1}, \frac{n\epsilon}{6 s_n c_{z,max}}\right\} + \log s_n + \log\left(p_n^2 - p_n - q_n\right)\right\} \\
& + 2\exp\left\{-\min\left\{\frac{C_1 n\epsilon^2}{18 \mathcal{K}_1}, \frac{n\epsilon}{6 c_{\mathcal{I}_2,max}}\right\} + \log\left(p_n^2 - p_n - q_n\right)\right\}, \\
P\left\{\left\|\frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right\|_\infty \geqslant \epsilon\right\} \leqslant\ & 4\exp\left\{-\min\left\{\frac{C_1 n\epsilon^2}{18 s_n^2 \mathcal{K}_1}, \frac{n\epsilon}{6 s_n c_{z,max}}\right\} + \log s_n + \log q_n\right\} \\
& + 2\exp\left\{-\min\left\{\frac{C_1 n\epsilon^2}{18 \mathcal{K}_1}, \frac{n\epsilon}{6 c_{\mathcal{I}_2,max}}\right\} + \log q_n\right\}.
\end{aligned}$$

Proof. We start with $\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}}$ as follows. For $\forall\ s = \{(a,b)\} \in \mathcal{E}$ and $h_1 = 1, \cdots, H$, we show that

$$\frac{\partial - l_c(\boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_s]_{h_1}} = \underbrace{\sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Bigg(\sum_{i=1,i\neq a}^{p_n}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]y_{mi}\Bigg)x_m^{(h_1)}y_{mb}}_{\mathcal{I}_{1.1}}$$

$$+ \underbrace{\sum_{m=1}^{n}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}\Bigg(\sum_{i=1,i\neq b}^{p_n}\Big[\alpha_{bi} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bi}]_h\Big]y_{mi}\Bigg)x_m^{(h_1)}y_{ma}}_{\mathcal{I}_{1.2}}$$

$$+ \underbrace{2\sum_{m=1}^{n} x_m^{(h_1)}y_{ma}y_{mb}}_{\mathcal{I}_2}.$$

The first term $\mathcal{I}_{1.1}$ can be rewritten as

$$\begin{aligned}
\mathcal{I}_{1.1} &= \sum_{m=1}^{n}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Bigg(\sum_{i=1,i\neq a}^{p_n}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]y_{mi}\Bigg)x_m^{(h_1)}y_{mb} \\
&= \sum_{i=1,i\neq a}^{p_n}\underbrace{\Bigg(\sum_{m=1}^{n} x_m^{(h_1)}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]y_{mi}y_{mb}\Bigg)}_{Z_{i,ab}^{(h_1)}} = \sum_{i=1,i\neq a}^{p_n} Z_{i,ab}^{(h_1)},
\end{aligned}$$

where each $Z_{i,ab}^{(h_1)}$ follows a sub-exponential distribution with parameters $\Big(\big[v_{z_{i,ab}}^{(h_1)}\big]^2, c_{z_{i,ab}}^{(h_1)}\Big)$.

The parameters take the following form

$$\begin{aligned}
\big[v_{z_{i,ab}}^{(h_1)}\big]^2 &= \sum_{m=1}^{n}\big[x_m^{(h_1)}\big]^2\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-2}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]^2 v_{m,ib}^2, \\
c_{z_{i,ab}}^{(1)} &= \max_{1\leqslant m\leqslant n}\Big\{x_m^{(h_1)}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-1}\Big|\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]\Big|c_{m,ib}\Big\}.
\end{aligned}$$

According to the property of sub-exponential distribution, we have

$$P\Big\{\Big|Z_{i,ab}^{(h_1)} - E\big\{Z_{i,ab}^{(h_1)}\big\}\Big| \geqslant t\Big\} \leqslant \begin{cases} 2\exp\Big\{-\frac{t^2}{2\big[v_{z_{i,ab}}^{(h_1)}\big]^2}\Big\} & \text{if } 0 \leqslant t \leqslant \frac{\big[v_{z_{i,ab}}^{(h_1)}\big]^2}{c_{z_{i,ab}}^{(h_1)}}, \\ 2\exp\Big\{-\frac{t}{2c_{z_{i,ab}}^{(h_1)}}\Big\} & \text{if } t > \frac{\big[v_{z_{i,ab}}^{(h_1)}\big]^2}{c_{z_{i,ab}}^{(h_1)}}. \end{cases}$$

Using Bonferroni inequalities, we show that

$$\begin{aligned} &P\Big\{\Big|\mathcal{I}_{1.1} - E\big\{\mathcal{I}_{1.1}\big\}\Big| \geqslant t\Big\} = P\left\{\left|\sum_{i=1, i\neq a}^{p_n}\Big[Z_{i,ab}^{(h_1)} - E\big\{Z_{i,ab}^{(h_1)}\big\}\Big]\right| \geqslant t\right\} \\ &\leqslant \sum_{i=1, i\neq a}^{p_n} P\Big\{\Big|Z_{i,ab}^{(h_1)} - E\big\{Z_{i,ab}^{(h_1)}\big\}\Big| \geqslant \frac{t}{s_n}\Big\} \\ &\leqslant 2\exp\left\{-\min\left\{\frac{t^2}{2s_n^2 v_{z,max}^2}, \frac{t}{2s_n c_{z,max}}\right\} + \log s_n\right\}, \end{aligned}$$

where

$$v_{z,max}^2 = \max_{\substack{1\leqslant i,j,k\leqslant p_n,\\ i\neq j, k\neq j,\\ h_1 = 1,\cdots,H}} \left\{\sum_{m=1}^{n}\big[x_m^{(h_1)}\big]^2\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big]^2 v_{m,ik}^2\right\},$$

$$c_{z,max} = \max_{\substack{1\leqslant i,j,k\leqslant p_n,\\ i\neq j, k\neq j\\ h_1 = 1,\cdots,H}} \left\{x_m^{(h_1)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\Big|\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big]\Big| c_{m,ik}\right\}.$$

Likewise, for the second term $\mathcal{I}_{1.2}$, we can show

$$\mathcal{I}_{1.2} = \sum_{i=1, i\neq b}^{p_n}\left(\sum_{m=1}^{n} x_m^{(h_1)}\Big[\alpha_{bb} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bb}]_h\Big]^{-1}\Big[\alpha_{bi} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{bi}]_h\Big] y_{mi} y_{ma}\right) = \sum_{i=1, i\neq b}^{p_n} Z_{i,ba}^{(h_1)},$$

where each $Z_{i,ba}^{(h_1)}$ follows a sub-exponential distribution with parameters $\Big(\big[v_{z_{i,ba}}^{(h_1)}\big]^2, c_{z_{i,ba}}^{(h_1)}\Big)$.

The parameters take the following form

$$\big[v^{(h_1)}_{z_{i,ba}}\big]^2 = \sum_{m=1}^{n} \big[x^{(h_1)}_m\big]^2 \Big[\alpha_{bb} + \sum_{h=1}^{H} x^{(h)}_m [\theta_{bb}]_h\Big]^{-2} \Big[\alpha_{bi} + \sum_{h=1}^{H} x^{(h)}_m [\theta_{bi}]_h\Big]^2 v^2_{m,ia},$$

$$c^{(h_1)}_{z_{i,ba}} = \max_{1\leqslant m\leqslant n} \Big\{ x^{(h_1)}_m \Big[\alpha_{bb} + \sum_{h=1}^{H} x^{(h)}_m [\theta_{bb}]_h\Big]^{-1} \Big|\Big[\alpha_{bi} + \sum_{h=1}^{H} x^{(h)}_m [\theta_{bi}]_h\Big]\Big| c_{m,ia} \Big\}.$$

Therefore, it leads to a similar result

$$P\Big\{\Big|\mathcal{I}_{1.2} - E\{\mathcal{I}_{1.2}\}\Big| \geqslant t\Big\} = P\left\{\left|\sum_{i=1,i\neq b}^{p_n} \Big[Z^{(h_1)}_{i,ba} - E\big\{Z^{(h_1)}_{i,ba}\big\}\Big]\right| \geqslant t\right\}$$

$$\leqslant 2\exp\left\{ -\min\left\{\frac{t^2}{2s_n^2 v^2_{z,max}}, \frac{t}{2s_n c_{z,max}}\right\} + \log s_n \right\}.$$

Lastly, the third term $\mathcal{I}_2 = 2\sum_{m=1}^{n} x^{(h_1)}_m y_{ma} y_{mb}$ follows a sub-exponential distribution with parameters

$$v^2_{\mathcal{I}_{2,ab}} = 4\sum_{m=1}^{n} \big[x^{(h_1)}_m\big]^2 v^2_{m,ab}, \qquad c_{\mathcal{I}_{2,ab}} = \max_{1\leqslant m\leqslant n} \big\{2x^{(h_1)}_m c_{m,ab}\big\}.$$

Hence, for any $t > 0$,

$$P\Big\{\Big|\mathcal{I}_2 - E\{\mathcal{I}_2\}\Big| \geqslant t\Big\} \leqslant 2\exp\left\{ -\min\left\{\frac{t^2}{2v^2_{\mathcal{I}_{2,max}}}, \frac{t}{2c_{\mathcal{I}_{2,max}}}\right\}\right\},$$

where $v^2_{\mathcal{I}_{2,max}} = \max_{\substack{1\leqslant i,j\leqslant p_n,\\ h_1=1,\cdots,H,\\ i\neq j}} \Big\{4\sum_{m=1}^{n} \big[x^{(h_1)}_m\big]^2 v^2_{m,ij}\Big\}$, $\quad c_{\mathcal{I}_{2,max}} = \max_{\substack{1\leqslant i,j\leqslant p_n\\ m=1,\cdots,n\\ h_1=1,\cdots,H\\ i\neq j}} \big\{2x^{(h_1)}_m c_{m,ij}\big\}$.

As a result,

$$
\begin{aligned}
&P\left\{\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}} - E\left\{\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}}\right\}\right| \geqslant t\right\}\\
&= P\Big\{|\mathcal{I}_{1.1} - E\{\mathcal{I}_{1.1}\} + \mathcal{I}_{1.2} - E\{\mathcal{I}_{1.2}\} + \mathcal{I}_2 - E\{\mathcal{I}_2\}| \geqslant t\Big\}\\
&\leqslant P\Big\{|\mathcal{I}_{1.1} - E\{\mathcal{I}_{1.1}\}| \geqslant \frac{t}{3}\Big\} + P\Big\{|\mathcal{I}_{1.2} - E\{\mathcal{I}_{1.2}\}| \geqslant \frac{t}{3}\Big\} + P\Big\{|\mathcal{I}_2 - E\{\mathcal{I}_2\}| \geqslant \frac{t}{3}\Big\}\\
&\leqslant 4\exp\left\{-\min\left\{\frac{t^2}{18 s_n^2 v_{z,max}^2}, \frac{t}{6 s_n c_{z,max}}\right\} + \log s_n\right\} + 2\exp\left\{-\min\left\{\frac{t^2}{18 v_{\mathcal{I}_{2,max}}^2}, \frac{t}{6 c_{\mathcal{I}_{2,max}}}\right\}\right\}.
\end{aligned}
$$

Define the constants $\mathcal{W}_1, \mathcal{K}_1$ as

$$
\mathcal{W}_1 = \max_{m,h_1,i,j,k}\left\{\big[x_m^{(h_1)}\big]^2\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big]^2 v_{m,ik}^2, 4\big[x_m^{(h_1)}\big]^2 v_{m,ij}^2\right\},
$$

$$
\mathcal{K}_1 = C_1\mathcal{W}_1 \text{ for any constant } C_1 > 0,
$$

and let $t = (\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}$, the probability becomes

$$
\begin{aligned}
&P\left\{\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}} - E\left\{\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}}\right\}\right| \geqslant (\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}\right\}\\
&\leqslant 4\exp\left\{-\min\left\{\frac{\mathcal{K}_1 n s_n^2 \log p_n}{18 s_n^2 v_{z,max}^2}, \frac{(\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}}{6 s_n c_{z,max}}\right\} + \log s_n\right\}\\
&+ 2\exp\left\{-\min\left\{\frac{\mathcal{K}_1 n s_n^2 \log p_n}{18 v_{\mathcal{I}_{2,max}}^2}, \frac{(\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}}{6 c_{\mathcal{I}_{2,max}}}\right\}\right\},\\
&\leqslant 4\exp\left\{-\min\left\{\frac{C_1 \log p_n}{18}, \frac{(\mathcal{K}_1 n \log p_n)^{\frac{1}{2}}}{6 c_{z,max}}\right\} + \log s_n\right\}\\
&+ 2\exp\left\{-\min\left\{\frac{C_1 s_n^2 \log p_n}{18}, \frac{(\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}}{6 c_{\mathcal{I}_{2,max}}}\right\}\right\}.
\end{aligned}
$$

By the Bonferroni inequality,

$$P\left\{\max_s\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}}\right| \geqslant (\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}\right\} \leqslant \frac{p_n(p_n-1)}{2} P\left\{\left|\frac{\partial - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}}\right| \geqslant (\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}\right\}$$
$$\leqslant 2\exp\left\{-\min\left\{\frac{C_1 \log p_n}{18}, \frac{(\mathcal{K}_1 n \log p_n)^{\frac{1}{2}}}{6c_{z,max}}\right\} + \log s_n + \log p_n + \log(p_n - 1)\right\}$$
$$+ \exp\left\{-\min\left\{\frac{C_1 s_n^2 \log p_n}{18}, \frac{(\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}}{6c_{\mathcal{I}_{2,max}}}\right\} + \log p_n + \log(p_n - 1)\right\}.$$

By choosing the constant $\mathcal{C}_1 > 18$, the upper bound probability approaches zero as n increases. The proofs of the remaining results are analogous to the one provided above and thus are omitted. $\square$

Lemma 2.3 For $\forall\ s, w \in \mathcal{E}$ and $\forall\ j = 1, ..., p_n, h_1, h_2 = 1, \cdots, H$, regarding all the second order partial derivatives of $-l_c(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, we have

$$\max_{s,w}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}}\right\}\right| = \mathcal{O}_p\big\{(n\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{s,w}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_w]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_w]_{h_2}}\right\}\right| = \mathcal{O}_p\big\{(n\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{s,w}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_w} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_w}\right\}\right| = \mathcal{O}_p\big\{(n\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{s,j}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}}\right\}\right| = \mathcal{O}_p\big\{(ns_n^2\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{s,j}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_{jj}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_{jj}}\right\}\right| = \mathcal{O}_p\big\{(ns_n^2\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{s,j}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_{jj}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_{jj}}\right\}\right| = \mathcal{O}_p\big\{(ns_n^2\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{s,j}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_{jj}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_{jj}]_{h_2}}\right\}\right| = \mathcal{O}_p\big\{(ns_n^2\log p_n)^{\frac{1}{2}}\big\},$$
$$\max_{j}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}}\right\}\right| = \mathcal{O}_p\big\{(ns_n^4\log p_n)^{\frac{1}{2}}\big\}.$$

Proof. Following the approach in **Lemma** 2.2, for $\forall\ \epsilon > 0$, we are able to provide the probabilities outlined below,

$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}} - E\big\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}}\big\}\right| \geqslant \epsilon\right\} \leqslant 2\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2 n}, \frac{\epsilon}{2c_*}\right\}\right\},\ \text{if } s \neq w,$$
$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_s]_{h_2}} - E\big\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_s]_{h_2}}\big\}\right| \geqslant \epsilon\right\} \leqslant 4\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{8\mathcal{K}_2 n}, \frac{\epsilon}{4c_*}\right\}\right\},$$
$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}}\right\}\right| \geqslant \epsilon\right\} \leqslant 2\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2 n s_n^2}, \frac{\epsilon}{2s_n c_{\mathcal{D},max}}\right\} + \log s_n\right\}.$$
$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}}\right\}\right| \geqslant \epsilon\right\} \leqslant 2\exp\left\{-\min\left\{\frac{C_3\epsilon^2}{2\mathcal{K}_3 n s_n^4}, \frac{\epsilon}{2s_n^2 c_{\mathcal{Z},max}}\right\} + 2\log s_n\right\}.$$

for some positive constants $C_2, C_3, \mathcal{K}_2, \mathcal{K}_3, c_{\mathcal{D},max}, c_{\mathcal{Z},max} > 0$. The rest of the probabilities are analogous and are therefore not displayed for brevity. Since

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}} = \underbrace{\frac{1}{2}\sum_{m=1}^{n} x_m^{(h_1)} x_m^{(h_2)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}}_{\mathcal{A}_{2.1}}$$
$$+ \underbrace{\sum_{m=1}^{n} x_m^{(h_1)} x_m^{(h_2)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-3}\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big] y_{mi}\Big)^2}_{\mathcal{A}_{2.2}},$$

the first term $\mathcal{A}_{2.1}$ is some constant and thus can be ignored. The second term $\mathcal{A}_{2.2}$ can be written as

$$\mathcal{A}_{2.2} = \sum_{m=1}^{n} x_m^{(h_1)} x_m^{(h_2)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-3}\Big(\sum_{\substack{r=1\\ r\neq j}}^{p_n}\sum_{\substack{t=1\\ t\neq j}}^{p_n}\Big[\alpha_{jr} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jr}]_h\Big]\Big[\alpha_{jt} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jt}]_h\Big] y_{mr} y_{mt}\Big)$$
$$= \sum_{\substack{r=1\\ r\neq j}}^{p_n}\sum_{\substack{t=1\\ t\neq j}}^{p_n}\underbrace{\Big(\sum_{m=1}^{n} x_m^{(h_1)} x_m^{(h_2)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-3}\Big[\alpha_{jr} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jr}]_h\Big]\Big[\alpha_{jt} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jt}]_h\Big] y_{mr} y_{mt}\Big)}_{\mathcal{Z}_{jrt}^{(h_1,h_2)}}.$$

It can be shown that each $\mathcal{Z}_{jrt}^{(h_1,h_2)}$ follows a sub-exponential distribution with parameters

$$v^2_{\mathcal{Z}_{jrt}^{(h1,h2)}} = \sum_{m=1}^{n}\big[x_m^{(h_1)} x_m^{(h_2)}\big]^2\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-6}\Big[\alpha_{jr} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jr}]_h\Big]^2\Big[\alpha_{jt} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jt}]_h\Big]^2 v^2_{m,rt},$$
$$c_{\mathcal{Z}_{jrt}^{(h_1,h_2)}} = \max_{m,j,r,t}\Big\{x_m^{(h_1)} x_m^{(h_2)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-3}\Big|\Big[\alpha_{jr} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jr}]_h\Big]\Big|\Big|\Big[\alpha_{jt} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jt}]_h\Big]\Big| c_{m,rt}\Big\}.$$

Therefore, we have

$$P\Big\{\Big|\mathcal{Z}_{jrt}^{(h_1,h_2)}-E\big\{\mathcal{Z}_{jrt}^{(h_1,h_2)}\big\}\Big|\geqslant t\Big\}\leqslant\begin{cases}2\exp\Big\{-\frac{t^2}{2\big[v_{\mathcal{Z}_{jrt}}^{(h_1,h_2)}\big]^2}\Big\} & \text{if } 0\leqslant t\leqslant\frac{\big[v_{\mathcal{Z}_{jrt}}^{(h_1,h_2)}\big]^2}{c_{\mathcal{Z}_{jrt}}^{(h_1,h_2)}},\\ 2\exp\Big\{-\frac{t}{2c_{\mathcal{Z}_{jrt}}^{(h_1,h_2)}}\Big\} & \text{if } t>\frac{\big[v_{\mathcal{Z}_{jrt}}^{(h_1,h_2)}\big]^2}{c_{\mathcal{Z}_{jrt}}^{(h_1,h_2)}}.\end{cases}$$

The probability becomes

$$P\left\{\left|\frac{\partial^2-l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}}-E\left\{\frac{\partial^2-l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}}\right\}\right|\geqslant\epsilon\right\}\leqslant P\left\{\left|\sum_{\substack{r=1\\r\neq j}}^{p_n}\sum_{\substack{t=1\\t\neq j}}^{p_n}\Big[\mathcal{Z}_{jrt}^{(h_1,h_2)}-E\Big\{\mathcal{Z}_{jrt}^{(h_1,h_2)}\Big\}\Big]\right|\geqslant\epsilon\right\}$$

$$\leqslant\sum_{\substack{r=1\\r\neq j}}^{p_n}\sum_{\substack{t=1\\t\neq j}}^{p_n}P\left\{\left|\mathcal{Z}_{jrt}^{(h_1,h_2)}-E\Big\{\mathcal{Z}_{jrt}^{(h_1,h_2)}\Big\}\right|\geqslant\frac{\epsilon}{s_n^2}\right\}$$

$$\leqslant 2\exp\left\{-\min\left\{\frac{\epsilon^2}{2s_n^4v_{\mathcal{Z},max}^2},\frac{\epsilon}{2s_n^2c_{\mathcal{Z},max}}\right\}+2\log s_n\right\},$$

where

$$v_{\mathcal{Z},max}^2=\max_{\substack{j,r,t\\h_1,h_2}}\left\{\sum_{m=1}^{n}\big[x_m^{(h_1)}x_m^{(h_2)}\big]^2\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-6}\Big[\alpha_{jr}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jr}]_h\Big]^2\Big[\alpha_{jt}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jt}]_h\Big]^2v_{m,rt}^2\right\},$$

$$c_{\mathcal{Z},max}=\max_{\substack{j,r,t\\m,h_1,h_2}}\left\{x_m^{(h_1)}x_m^{(h_2)}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-3}\Big|\Big[\alpha_{jr}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jr}]_h\Big]\Big|\Big|\Big[\alpha_{jt}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jt}]_h\Big|c_{m,rt}\right\}.$$

Define the constants

$$\mathcal{W}_2=\max_{\substack{m,h_1,h_2\\j,r,t}}\left\{\big[x_m^{(h_1)}x_m^{(h_2)}\big]^2\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-6}\Big[\alpha_{jr}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jr}]_h\Big]^2\Big[\alpha_{jt}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jt}]_h\Big]^2v_{m,rt}^2\right\},$$

$\mathcal{K}_2=C_2\mathcal{W}_2$ for some constant $C_2>0$, leading to

$$P\left\{\left|\frac{\partial^2-l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}}-E\left\{\frac{\partial^2-l_c(\boldsymbol{\beta^0})}{\partial[\theta_{jj}]_{h_1}\partial[\theta_{jj}]_{h_2}}\right\}\right|\geqslant\epsilon\right\}$$

$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2ns_n^4},\frac{\epsilon}{2s_n^2c_{\mathcal{Z},max}}\right\}+2\log s_n\right\}.$$

Likewise, without loss of generality, we consider $\forall\ s = \{(a,b)\} \in \mathcal{E}$ and $j = a$, we have

$$\frac{\partial^2 - l_c(\boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}} = -\sum_{m=1}^{n} x_m^{(h_2)}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-2}\Big(\sum_{i=1, i\neq a}^{p_n}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]y_{mi}\Big)x_m^{(h_1)}y_{mb}$$
$$= \sum_{i=1, i\neq a}^{p_n}\underbrace{\Big(-\sum_{m=1}^{n} x_m^{(h_1)}x_m^{(h_2)}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-2}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]y_{mi}y_{mb}\Big)}_{\mathcal{D}_{iab}^{(h_1,h_2)}}.$$

Each $\mathcal{D}_{iab}^{(h_1,h_2)}$ follows a sub-exponential distribution with parameters

$$v^2_{\mathcal{D}_{iab}^{(h_1,h_2)}} = \sum_{m=1}^{n}\Big[x_m^{(h_1)}x_m^{(h_2)}\Big]^2\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-4}\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]^2 v^2_{m,bi},$$
$$c^2_{\mathcal{D}_{iab}^{(h_1,h_2)}} = \max_m\Big\{x_m^{(h_1)}x_m^{(h_2)}\Big[\alpha_{aa} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{aa}]_h\Big]^{-2}\Big|\Big[\alpha_{ai} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ai}]_h\Big]\Big|c_{m,bi}\Big\}.$$

As a result, we can show

$$P\Bigg\{\Bigg|\frac{\partial^2 - l_c(\boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}} - E\Big\{\frac{\partial^2 - l_c(\boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_s]_{h_1}\partial[\theta_{jj}]_{h_2}}\Big\}\Bigg| \geqslant \epsilon\Bigg\} = P\Bigg\{\Bigg|\sum_{i=1, i\neq a}^{p_n}\Big[\mathcal{D}_{iab}^{(h_1,h_2)} - E\Big\{\mathcal{D}_{iab}^{(h_1,h_2)}\Big\}\Big]\Bigg| \geqslant \epsilon\Bigg\}$$
$$\leqslant 2\exp\Bigg\{-\min\Bigg\{\frac{\epsilon^2}{2s_n^2 v^2_{\mathcal{D},max}}, \frac{\epsilon}{2s_n c_{\mathcal{D},max}}\Bigg\} + \log s_n\Bigg\},$$

where

$$v^2_{\mathcal{D},max} = \max_{i,j,k,h_1,h_2}\Bigg\{\sum_{m=1}^{n}\Big[x_m^{(h_1)}x_m^{(h_2)}\Big]^2\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-4}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big]^2 v^2_{m,ki}\Bigg\},$$
$$c_{\mathcal{D},max} = \max_{i,j,k,h_1,h_2}\Big\{x_m^{(h_1)}x_m^{(h_2)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big|\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big]\Big|c_{m,ki}\Big\}.$$

Define the constants

$$c_* = \max_{\substack{m,h_1,h_2,\\ i,j,k}} \left\{ x_m^{(h_1)} x_m^{(h_2)} \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{jj}]_h \Big]^{-1} c_{m,ki} \right\},$$

$$\mathcal{W}_2 = \max_{\substack{m,h_1,h_2\\ i,j,k}} \left\{ \Big[x_m^{(h_1)} x_m^{(h_2)}\Big]^2 \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{jj}]_h \Big]^{-4} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{ji}]_h \Big]^{2} v_{m,ki}^2, \right.$$

$$\left. \Big[x_m^{(h_1)} x_m^{(h_2)}\Big]^2 \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{jj}]_h \Big]^{-2} v_{m,ki}^2 \right\},$$

$\mathcal{K}_2 = C_2 \mathcal{W}_2$ for some constant $C_2 > 0$, leading to

$$P\left\{ \left| \frac{\partial^2 - l_c(\boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_s]_{h_1} \partial[\theta_{jj}]_{h_2}} - E\left\{ \frac{\partial^2 - l_c(\boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_s]_{h_1} \partial[\theta_{jj}]_{h_2}} \right\} \right| \geqslant \epsilon \right\}$$

$$\leqslant 2 \exp\left\{ -\min\left\{ \frac{C_2 \epsilon^2}{2\mathcal{K}_2 n s_n^2}, \frac{\epsilon}{2 s_n c_{\mathcal{D},max}} \right\} + \log s_n \right\}.$$

The same procedure applies to the remaining terms and is therefore omitted for brevity. □

Lemma 2.4 For $\forall\ \epsilon > 0$, we show that

$$
\begin{aligned}
&P\bigg\{\bigg|\bigg|\bigg| - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\bigg\{ - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\} \\
&\leqslant 8\exp\bigg\{ - \min\bigg\{\frac{C_2 n\epsilon^2}{128\mathcal{K}_2}, \frac{n\epsilon}{16c_*}\bigg\} + \log q_n\bigg\} \\
&+ 4(H+1)\exp\bigg\{ - \min\bigg\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 s_n^2}, \frac{n\epsilon}{8(H+1)s_n c_*}\bigg\} + \log s_n + \log q_n\bigg\}, \\
&P\bigg\{\bigg|\bigg|\bigg| - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - E\bigg\{ - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\bigg\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\} \\
&\leqslant 8\exp\bigg\{ - \min\bigg\{\frac{C_2 n\epsilon^2}{128\mathcal{K}_2}, \frac{n\epsilon}{16c_*}\bigg\} + \log\left(p_n^2 - p_n - q_n\right)\bigg\} \\
&+ 4(H+1)\exp\bigg\{ - \min\bigg\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 s_n^2}, \frac{n\epsilon}{8(H+1)s_n c_*}\bigg\} + \log s_n + \log\left(p_n^2 - p_n - q_n\right)\bigg\}.
\end{aligned}
$$

Proof. For any matrix $A \in \mathbb{R}^{p_n \times p_n}$, the ∞ norm is defined as

$$
|||A|||_\infty = \max_{1 \leqslant i \leqslant p_n} \sum_{j=1}^{p_n} |A_{ij}|.
$$

Let $l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$ denote the $(k,l)th$ entry of $l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$. The probability can be formulated as

$$
\begin{aligned}
&P\bigg\{\bigg|\bigg|\bigg| - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\bigg\{ - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\} \\
&= P\bigg\{\max_k \frac{1}{n}\sum_l \Big| - l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\big\{ - l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\big\}\Big| > \epsilon\bigg\}.
\end{aligned}
$$

In particular, for $\forall\ s, w \in \mathcal{S}_1, s \neq w$, the row summation $\sum_l$ may include

$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}^2} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}^2}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_s} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_s}\right\}\right|,$$
$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_w} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_w}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}}\right\}\right|,$$

or

$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s^2} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s^2}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_s]_{h_1}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_s]_{h_1}}\right\}\right|,$$
$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_w} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_w}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_w]_{h_1}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial[\theta_w]_{h_1}}\right\}\right|.$$

Without loss of generality, we consider the first case. For $\forall\ s = \{(a, b)\} \in \mathcal{S}_1$, we have

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h2}} \begin{cases} = \sum_{m=1}^n \left[x_m^{(h_1)}x_m^{(h_2)}\right]^2\left[\alpha_{aa} + \sum_{h=1}^H x_m^{(h)}[\theta_{aa}]_h\right]^{-1} y_{mb}y_{mc}, & \text{if } w = \{(a, c)\} \in \mathcal{E}, \\ = \sum_{m=1}^n \left[x_m^{(h_1)}x_m^{(h_2)}\right]^2\left[\alpha_{bb} + \sum_{h=1}^H x_m^{(h)}[\theta_{bb}]_h\right]^{-1} y_{ma}y_{mc}, & \text{if } w = \{(b, c)\} \in \mathcal{E}, \\ = 0, & \text{otherwise.} \end{cases}$$

For any fixed $s = \{(a, b)\}$, the total number of nonzero terms $\frac{\partial^2 l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}}$ is equal to the number of possible choices for w, which is less than $2Hs_n$ in the subset $\mathcal{S}_1$. Likewise, for each s, the total number of nonzero terms $\frac{\partial^2 l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_w}$ is less than $2s_n$. As a result, for $\forall\ k$th row of $-l_{kl}^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}$, by **Lemma** 2.3, we have

$$P\left\{\frac{1}{n}\sum_l\left|-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right| \geqslant \epsilon\right\}$$
$$\leqslant 8\exp\left\{-\min\left\{\frac{C_2n\epsilon^2}{128\mathcal{K}_2}, \frac{n\epsilon}{16c_*}\right\}\right\} + 4(H+1)\exp\left\{-\min\left\{\frac{C_2n\epsilon^2}{32(H+1)^2\mathcal{K}_2 s_n^2}, \frac{n\epsilon}{8(H+1)s_nc_*}\right\} + \log s_n\right\}.$$

Through the Bonferroni inequality, we show that

$$P\bigg\{\bigg|\bigg|\bigg| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\bigg\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\}$$
$$\leqslant |\mathcal{S}_1|\, P\bigg\{\frac{1}{n}\sum_l\bigg| -l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\bigg\{-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg\}\bigg| \geqslant \epsilon\bigg\}$$
$$\leqslant 8\exp\bigg\{-\min\bigg\{\frac{C_2 n\epsilon^2}{128\mathcal{K}_2}, \frac{n\epsilon}{16c_*}\bigg\} + \log q_n\bigg\}$$
$$+ 4(H+1)\exp\bigg\{-\min\bigg\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 s_n^2}, \frac{n\epsilon}{8(H+1)s_n c_*}\bigg\} + \log s_n + \log q_n\bigg\}.$$

Likewise, we can show similar results for another subset $\mathcal{S}^c$ as follows,

$$P\bigg\{\bigg|\bigg|\bigg| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - E\bigg\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\bigg\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\}$$
$$\leqslant 8\exp\bigg\{-\min\bigg\{\frac{C_2 n\epsilon^2}{128\mathcal{K}_2}, \frac{n\epsilon}{16c_*}\bigg\} + \log(p_n^2 - p_n - q_n)\bigg\}$$
$$+ 4(H+1)\exp\bigg\{-\min\bigg\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 s_n^2}, \frac{n\epsilon}{8(H+1)s_n c_*}\bigg\} + \log s_n + \log(p_n^2 - p_n - q_n)\bigg\}.$$

□

Lemma 2.5 For $\forall\ \epsilon > 0$, we show that

$$P\bigg\{\bigg|\bigg|\bigg| E\bigg\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg\} - \bigg(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)\bigg|\bigg|\bigg|_2 \geqslant \epsilon\bigg\}$$
$$\leqslant 4(H+1)\exp\bigg\{-\min\bigg\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 q_n^2}, \frac{n\epsilon}{8(H+1)q_n c_*}\bigg\} + \log q_n\bigg\}$$
$$+ 4(H+1)\exp\bigg\{-\min\bigg\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 q_n^2 s_n^2}, \frac{n\epsilon}{8(H+1)q_n s_n c_*}\bigg\} + \log q_n + \log s_n\bigg\}.$$

Proof. For any matrix $A \in \mathbb{R}^{p_n \times p_n}$, we have $|||A|||_2 \leqslant |||A|||_F$, which leads to $P\{|||A|||_2 \geqslant \epsilon\} \leqslant P\{|||A|||_F \geqslant \epsilon\}$. Using the same notation introduced in **Lemma** 2.4, we can show

$$\left|\left|\left|E\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\} - \left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\right|\right|_F = \sqrt{\sum_{k,l}\left|E\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right\} - \left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|^2}$$
$$\leqslant \sum_{k,l}\left|-\frac{1}{n}l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{-\frac{1}{n}l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right|.$$

Therefore, the upper bound of the probability becomes

$$P\left\{\left|\left|\left|E\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\} - \left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\right|\right|_F \geqslant \epsilon\right\}$$
$$\leqslant P\left\{\sum_{k,l}\left|-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right| \geqslant n\epsilon\right\}.$$

The possible terms included in the summation $\sum_{k,l}$ are as follows.

Case 1: For $\forall\ s \in \mathcal{S}_1$, the total number of terms for the following three types

$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_s]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_s]_{h_2}}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s^2} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s^2}\right\}\right|,$$
$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_s} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_s}\right\}\right|,$$

are $q_n(H+1)$.

Case 2: For $\forall\ s \in \mathcal{S}_1, \forall\ w \in \mathcal{S}_1$ and $s \neq w$, it can be shown that the total number of nonzero terms for the following four types is less than $2(H+1)q_n s_n$,

$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h_2}}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_w} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_s\partial\alpha_w}\right\}\right|,$$
$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_w} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial\alpha_w}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_w\partial[\theta_s]_{h_1}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_w\partial[\theta_s]_{h_1}}\right\}\right|.$$

This is because for $\forall\ s = \{(a,b)\} \in \mathcal{S}_1$, we know that

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_s]_{h_1}\partial[\theta_w]_{h2}} \begin{cases} = \sum_{m=1}^n \left[x_m^{(h_1)}x_m^{(h_2)}\right]^2\left[\alpha_{aa} + \sum_{h=1}^H x_m^{(h)}[\theta_{aa}]_h\right]^{-1} y_{mb}y_{mc}, & \text{if } w = \{(a,c)\} \in \mathcal{S}_1, \\ = \sum_{m=1}^n \left[x_m^{(h_1)}x_m^{(h_2)}\right]^2\left[\alpha_{bb} + \sum_{h=1}^H x_m^{(h)}[\theta_{bb}]_h\right]^{-1} y_{ma}y_{mc}, & \text{if } w = \{(b,c)\} \in \mathcal{S}_1, \\ = 0, & \text{otherwise.} \end{cases}$$

The number of nonzero terms for each row would be less than $2(H+1)s_n$. Therefore, by **Lemma** 2.3, for $\forall\ \epsilon > 0$, the probability can be written as

$$\begin{aligned} & P\left\{\sum_{k,l}\left|-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right| \geqslant n\epsilon\right\} \\ & \leqslant \sum_{(H+1)q_n} P\left\{|\text{Any term in Case 1}| \geqslant \frac{n\epsilon}{2(H+1)q_n}\right\} \\ & + \sum_{2(H+1)q_n s_n} P\left\{|\text{Any term in Case 2}| \geqslant \frac{n\epsilon}{4(H+1)q_n s_n}\right\} \\ & \leqslant 4(H+1)\exp\left\{-\min\left\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 q_n^2}, \frac{n\epsilon}{8(H+1)q_n c_*}\right\} + \log q_n\right\} \\ & + 4(H+1)\exp\left\{-\min\left\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 q_n^2 s_n^2}, \frac{n\epsilon}{8(H+1)q_n s_n c_*}\right\} + \log q_n + \log s_n\right\}, \end{aligned}$$

which completes the proof. □

Lemma 2.6 Under **Assumptions** 3-5 and $q_n^2 s_n^2 \log q_n = o(n)$, there exists some constant $\tau_- > 0$ such that $-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$ has minimum eigenvalue lower bounded by τ_- and is invertible with probability tending to one.

Proof. According to the **Assumption** 3, we know that $H(\boldsymbol{\beta^0}) = E\big\{-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})\big\}$ has eigenvalues bounded away from zero and infinity. Hence, for the sub-matrix $H^0_{\mathcal{S}_1\mathcal{S}_1} = E\big\{-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\big\}$, there exists some constant $\tau > 0$ such that $\lambda_{\min}(H^0_{\mathcal{S}_1\mathcal{S}_1}) \geqslant \tau$. By the Courant-Fischer variational representation, we have

$$\begin{aligned}
\lambda_{\min}(H^0_{\mathcal{S}_1\mathcal{S}_1}) &= \lambda_{\min}\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} + H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) \\
&= \min_{||v||_2=1}\bigg\{v^T\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} + H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)v\bigg\} \\
&= \min_{||v||_2=1}\bigg\{v^T\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)v + v^T\bigg(H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)v\bigg\} \\
&\leqslant y^T\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)y + y^T\bigg(H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)y,
\end{aligned}$$

where y is a unit-norm eigenvector of $-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$. Therefore, we show that

$$y^T\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)y \geqslant \lambda_{\min}\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) \geqslant \lambda_{\min}(H^0_{\mathcal{S}_1\mathcal{S}_1}) - y^T\bigg(H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg)y,$$

$$\lambda_{\min}\bigg(-\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) \geqslant \tau - \bigg|\bigg|\bigg|H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg|\bigg|\bigg|_2.$$

By **Lemma** 2.5, we can choose $\epsilon \in (0, \tau)$ such that the following probability approaches zero,

$$
\begin{aligned}
&P\left\{\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}+\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right|\!\right|\!\right|_2 \geqslant \epsilon\right\}\\
&\leqslant 4(H+1)\exp\left\{-\min\left\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 q_n^2},\frac{n\epsilon}{8(H+1)q_n c_*}\right\}+\log q_n\right\}\\
&+4(H+1)\exp\left\{-\min\left\{\frac{C_2 n\epsilon^2}{32(H+1)^2\mathcal{K}_2 q_n^2 s_n^2},\frac{n\epsilon}{8(H+1)q_n s_n c_*}\right\}+\log q_n+\log s_n\right\}.
\end{aligned}
$$

As a result, $\lambda_{\min}\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big) \geqslant \tau-\epsilon=\tau_-$ holds with probability tending to 1. $\square$

Lemma 2.7 (**Incoherence Condition**) Under **Assumptions** 3-5 and $q_n^2 s_n^2\log q_n = o(n)$, the following condition

$$\left|\!\left|\!\left|\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty \leqslant 1-\frac{1}{3}\xi^2 \tag{2.17}$$

holds with probability tending to one.

Proof. To show (2.17), we need to apply several facts derived from **Lemma** 2.4, 2.5 and 2.6, which are summarized as follows

1. By **Lemma** 2.4, for $\forall\ \epsilon>0,\forall\ \epsilon'\in(0,\tau)$, and some constant $\tau_-=\tau-\epsilon'>0$,

we have

$$
\begin{aligned}
&P\left\{\left|\left|\left| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - H^0_{\mathcal{S}_1\mathcal{S}_1}\right|\right|\right|_\infty \geqslant \frac{\tau_-\epsilon}{\sqrt{|\mathcal{S}_1|}}\right\} \\
&\leqslant 8\exp\left\{-\min\left\{\frac{C_2n\tau_-^2\epsilon^2}{128|\mathcal{S}_1|\mathcal{K}_2}, \frac{n\tau_-\epsilon}{16\sqrt{|\mathcal{S}_1|}c_*}\right\} + \log q_n\right\} \\
&+ 4(H+1)\exp\left\{-\min\left\{\frac{C_2n\tau_-^2\epsilon^2}{32(H+1)^2|\mathcal{S}_1|\mathcal{K}_2s_n^2}, \frac{n\tau_-\epsilon}{8(H+1)\sqrt{|\mathcal{S}_1|}s_nc_*}\right\} + \log s_n + \log q_n\right\}, \\
&P\left\{\left|\left|\left| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\right|\right|_\infty \geqslant \frac{\tau\epsilon}{\sqrt{|\mathcal{S}_1|}}\right\} \\
&\leqslant 8\exp\left\{-\min\left\{\frac{C_2n\tau^2\epsilon^2}{128|\mathcal{S}_1|\mathcal{K}_2}, \frac{n\tau\epsilon}{16\sqrt{|\mathcal{S}_1|}c_*}\right\} + \log(p_n^2 - p_n - q_n)\right\} \\
&+ 4(H+1)\exp\left\{-\min\left\{\frac{C_2n\tau^2\epsilon^2}{32(H+1)^2|\mathcal{S}_1|\mathcal{K}_2s_n^2}, \frac{n\tau\epsilon}{8(H+1)\sqrt{|\mathcal{S}_1|}s_nc_*}\right\} + \log s_n + \log(p_n^2 - p_n - q_n)\right\}.
\end{aligned}
$$

2. By **Lemma** 2.5, for $\forall\ \epsilon' > 0$, we have

$$
\begin{aligned}
&P\left\{\left|\left|\left| H^0_{\mathcal{S}_1\mathcal{S}_1} - \left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\right|\right|_2 \geqslant \epsilon'\right\} \\
&\leqslant 4(H+1)\exp\left\{-\min\left\{\frac{C_2n(\epsilon')^2}{32(H+1)^2\mathcal{K}_2q_n^2}, \frac{n\epsilon'}{8(H+1)q_nc_*}\right\} + \log q_n\right\} \\
&+ 4(H+1)\exp\left\{-\min\left\{\frac{C_2n(\epsilon')^2}{32(H+1)^2\mathcal{K}_2q_n^2s_n^2}, \frac{n\epsilon'}{8(H+1)q_ns_nc_*}\right\} + \log q_n + \log s_n\right\}.
\end{aligned}
$$

3. By **Lemma** 2.6, the following inequality holds

$$
\left|\left|\left|\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_2 \leqslant \frac{1}{\tau_-}
$$

with the probability tending to one, where $\tau_- = \tau - \epsilon' > 0$.

We decompose the term $\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\big(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}$ as

$$
\begin{aligned}
&\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\\
&=\left[H^0_{\mathcal{S}^c\mathcal{S}_1}-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right]\cdot\left[\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}-\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right]\\
&=\underbrace{H^0_{\mathcal{S}^c\mathcal{S}_1}\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}}_{\mathcal{I}_1}+\underbrace{H^0_{\mathcal{S}^c\mathcal{S}_1}\left[\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right]}_{\mathcal{I}_2}\\
&+\underbrace{\left[-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right]\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}}_{\mathcal{I}_3}\\
&+\underbrace{\left[-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right]\left[\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right]}_{\mathcal{I}_4}.
\end{aligned}
$$

Therefore, $|||\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\big(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}|||_\infty \leqslant |||\mathcal{I}_1|||_\infty + |||\mathcal{I}_2|||_\infty + |||\mathcal{I}_3|||_\infty + |||\mathcal{I}_4|||_\infty$. By **Assumption** 5, $|||\mathcal{I}_1|||_\infty \leqslant 1-\xi$. The second term can be formulated as

$$
\begin{aligned}
\mathcal{I}_2&=H^0_{\mathcal{S}^c\mathcal{S}_1}\left[\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right]\\
&=H^0_{\mathcal{S}^c\mathcal{S}_1}\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right]\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}.
\end{aligned}
$$

By **Facts** 1, 3 and **Assumption** 5, we show that

$$
\begin{aligned}
|||\mathcal{I}_2|||_\infty &\leqslant \left|\left|\left|H^0_{\mathcal{S}^c\mathcal{S}_1}\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right|\right|\right|_\infty \left|\left|\left| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - H^0_{\mathcal{S}_1\mathcal{S}_1}\right|\right|\right|_\infty \left|\left|\left|\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}\right|\right|\right|_\infty \\
&\leqslant (1-\xi)\left|\left|\left| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - H^0_{\mathcal{S}_1\mathcal{S}_1}\right|\right|\right|_\infty \sqrt{|\mathcal{S}_1|}\left|\left|\left|\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}\right|\right|\right|_2 \\
&\leqslant (1-\xi)\frac{\tau_-\epsilon}{\sqrt{|\mathcal{S}_1|}}\frac{\sqrt{|\mathcal{S}_1|}}{\tau_-} = (1-\xi)\epsilon,
\end{aligned}
$$

with a probability of at least $1-8\exp\Big\{-\min\Big\{\frac{C_2 n\tau_-^2\epsilon^2}{128|\mathcal{S}_1|\mathcal{K}_2}, \frac{n\tau_-\epsilon}{16\sqrt{|\mathcal{S}_1|}c_*}\Big\}+\log q_n\Big\}-4(H+1)\exp\Big\{-\min\Big\{\frac{C_2 n\tau_-^2\epsilon^2}{32(H+1)^2|\mathcal{S}_1|\mathcal{K}_2 s_n^2}, \frac{n\tau_-\epsilon}{8(H+1)\sqrt{|\mathcal{S}_1|}s_n c_*}\Big\}+\log s_n+\log q_n\Big\}-4(H+1)\exp\Big\{-\min\Big\{\frac{C_2 n(\epsilon')^2}{32(H+1)^2\mathcal{K}_2 q_n^2}, \frac{n\epsilon'}{8(H+1)q_n c_*}\Big\}+\log q_n\Big\}-4(H+1)\exp\Big\{-\min\Big\{\frac{C_2 n(\epsilon')^2}{32(H+1)^2\mathcal{K}_2 q_n^2 s_n^2}, \frac{n\epsilon'}{8(H+1)q_n s_n c_*}\Big\}+\log q_n+\log s_n\Big\}$.

Likewise, by **Fact** 1, the third term $|||\mathcal{I}_3|||_\infty$ can be bounded by

$$
\begin{aligned}
|||\mathcal{I}_3|||_\infty &\leqslant \left|\left|\left| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\right|\right|_\infty \left|\left|\left|\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right|\right|\right|_\infty \\
&\leqslant \left|\left|\left| -\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\right|\right|_\infty \sqrt{|\mathcal{S}_1|}\left|\left|\left|\big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\right|\right|\right|_2 \leqslant \frac{\tau\epsilon}{\sqrt{|\mathcal{S}_1|}}\frac{\sqrt{|\mathcal{S}_1|}}{\tau} = \epsilon,
\end{aligned}
$$

with a probability of at least $1-8\exp\Big\{-\min\Big\{\frac{C_2 n\tau^2\epsilon^2}{128|\mathcal{S}_1|\mathcal{K}_2}, \frac{n\tau\epsilon}{16\sqrt{|\mathcal{S}_1|}c_*}\Big\}+\log(p_n^2-p_n-q_n)\Big\}-4(H+1)\exp\Big\{-\min\Big\{\frac{C_2 n\tau^2\epsilon^2}{32(H+1)^2|\mathcal{S}_1|\mathcal{K}_2 s_n^2}, \frac{n\tau\epsilon}{8(H+1)\sqrt{|\mathcal{S}_1|}s_n c_*}\Big\}+\log s_n+\log(p_n^2-p_n-q_n)\Big\}$.

For $\mathcal{I}_4$, we first show that

$$
\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1} - \big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1} = \big(H^0_{\mathcal{S}_1\mathcal{S}_1}\big)^{-1}\Big[H^0_{\mathcal{S}_1\mathcal{S}_1} - \Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big)\Big]\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}.
$$

The infinity norm becomes

$$\left|\!\left|\!\left|\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}^{-1}-\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty=\left|\!\left|\!\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1}+\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right]\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}^{-1}\right|\!\right|\!\right|_\infty$$

$$\leqslant\sqrt{|\mathcal{S}_1|}\cdot\left|\!\left|\!\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right]\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}^{-1}\right|\!\right|\!\right|_2$$

$$\leqslant\sqrt{|\mathcal{S}_1|}\cdot\left|\!\left|\!\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_2\cdot\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2\cdot\left|\!\left|\!\left|\left\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}^{-1}\right|\!\right|\!\right|_2$$

$$\leqslant\frac{\sqrt{|\mathcal{S}_1|}}{\tau\cdot\tau_-}\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2.$$

By **Facts** 1, 2, we can calculate the upper bound as

$$\begin{aligned}|||\mathcal{I}_4|||_\infty&\leqslant\left|\!\left|\!\left|-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\!\right|\!\right|_\infty\left|\!\left|\!\left|\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty\\&\leqslant\left|\!\left|\!\left|-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\!\right|\!\right|_\infty\frac{\sqrt{|\mathcal{S}_1|}}{\tau\cdot\tau_-}\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2\\&\leqslant\frac{\tau\epsilon}{\sqrt{|\mathcal{S}_1|}}\frac{\sqrt{|\mathcal{S}_1|}}{\tau\cdot\tau_-}\epsilon'=\frac{\epsilon'}{\tau_-}\epsilon.\end{aligned}$$

We let $\epsilon'<\tau_-=\tau-\epsilon'$, which is equivalent to $\epsilon'<0.5\tau$, and also choose $\epsilon<\frac{\xi}{3}$ for some constant $\xi\in(0,1)$ defined in **Assumption** 5, then we have

$$\begin{aligned}\left|\!\left|\!\left|\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty&\leqslant|||\mathcal{I}_1|||_\infty+|||\mathcal{I}_2|||_\infty+|||\mathcal{I}_3|||_\infty+|||\mathcal{I}_4|||_\infty\\&\leqslant1-\xi+(1-\xi)\epsilon+\epsilon+\frac{\epsilon'}{\tau_-}\epsilon\\&\leqslant1-\xi+\xi-\frac{1}{3}\xi^2=1-\frac{1}{3}\xi^2,\end{aligned}$$

with a probability of at least $1-8\exp\left\{-\min\left\{\frac{C_2n\tau_-^2\epsilon^2}{128|\mathcal{S}_1|\mathcal{K}_2},\frac{n\tau_-\epsilon}{16\sqrt{|\mathcal{S}_1|}c_*}\right\}+\log q_n\right\}-4(H+1)\exp\left\{-\min\left\{\frac{C_2n\tau_-^2\epsilon^2}{32(H+1)^2|\mathcal{S}_1|\mathcal{K}_2s_n^2},\frac{n\tau_-\epsilon}{8(H+1)\sqrt{|\mathcal{S}_1|}s_nc_*}\right\}+\log s_n+\log q_n\right\}-4(H+1)\exp\Big\{-$

$\min\left\{\frac{C_2 n(\epsilon')^2}{32(H+1)^2\mathcal{K}_2 q_n^2}, \frac{n\epsilon'}{8(H+1)q_n c_*}\right\} + \log q_n\Bigg\} - 4(H+1)\exp\Bigg\{-\min\left\{\frac{C_2 n(\epsilon')^2}{32(H+1)^2\mathcal{K}_2 q_n^2 s_n^2}, \frac{n\epsilon'}{8(H+1)q_n s_n c_*}\right\} + \log q_n + \log s_n\Bigg\} - 8\exp\Bigg\{ - \min\left\{\frac{C_2 n\tau^2\epsilon^2}{128|\mathcal{S}_1|\mathcal{K}_2}, \frac{n\tau\epsilon}{16\sqrt{|\mathcal{S}_1|}c_*}\right\} + \log(p_n^2 - p_n - q_n)\Bigg\} - 4(H + 1)\exp\Bigg\{ - \min\left\{\frac{C_2 n\tau^2\epsilon^2}{32(H+1)^2|\mathcal{S}_1|\mathcal{K}_2 s_n^2}, \frac{n\tau\epsilon}{8(H+1)\sqrt{|\mathcal{S}_1|}s_n c_*}\right\} + \log s_n + \log(p_n^2 - p_n - q_n)\Bigg\}$. □

Proof of Theorem 2.1

Proof. The multivariate version of the Lindeberg condition can be bounded by

$$
\begin{aligned}
& n^{-1}\sum_{m=1}^{n} E\Big\{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2^2 \cdot I\Big(||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2 \geqslant \epsilon\sqrt{n}\Big)\Big\} \\
& \leqslant n^{-1}\sum_{m=1}^{n} E\Bigg\{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2^2 \cdot \left\{\frac{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2}{\epsilon\sqrt{n}}\right\}^2\Bigg\} \\
& = \epsilon^{-2} n^{-2}\sum_{m=1}^{n} E\Big\{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2^4\Big\}.
\end{aligned}
$$

Denote each (i,j)th element of $\boldsymbol{\Sigma}_m$ as $\sigma_{ij,m}^2$ and it can be shown that

$$
\begin{aligned}
\frac{\partial \boldsymbol{\mathcal{L}}_m(\boldsymbol{\beta})}{\partial[\theta_u]_{h_1}} &= \begin{cases} \frac{1}{2}x_m^{(h_1)}[\sigma_{jj,m}^2 - y_{mj}^2] & \text{w.r.t } [\theta_{jj}]_h, \\ x_m^{(h_1)}[\sigma_{ji,m}^2 - y_{mj}y_{mi}] & \text{w.r.t } [\theta_{ji}]_h, \end{cases} \\
\frac{\partial \boldsymbol{\mathcal{L}}_m(\boldsymbol{\beta})}{\partial \alpha_s} &= \begin{cases} \frac{1}{2}[\sigma_{jj,m}^2 - y_{mj}^2] & \text{w.r.t } \alpha_{jj}, \\ \sigma_{ji,m}^2 - y_{mj}y_{mi} & \text{w.r.t } \alpha_{ji}, \end{cases}
\end{aligned}
\tag{2.18}
$$

for all $h = 1, \cdots, H, m = 1, \cdots, n$. By **Assumption** 1 and **Lemma** 2.1, we show that

$$E\Big\{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2^4\Big\} = E\Bigg\{\Bigg(\underbrace{\sum_{u,h_1}\Bigg\{\frac{\partial \boldsymbol{\mathcal{L}}_m(\boldsymbol{\beta}^0)}{\partial[\theta_u]_{h_1}}\Bigg\}^2 + \sum_{s}\Bigg\{\frac{\partial \boldsymbol{\mathcal{L}}_m(\boldsymbol{\beta}^0)}{\partial \alpha_s}\Bigg\}^2}_{\text{the sum of } 0.5(H+1)p(p+1) \text{ terms}}\Bigg)^2\Bigg\} = \mathcal{O}(p^4),$$

which leads to $\lim_{n\to\infty}\epsilon^{-1}n^{-2}\sum_{m=1}^n E\Big\{||\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0)||_2^4\Big\} = 0$. Therefore, by the multivariate version of the Lindeberg-Feller central limit theorem, we conclude that

$$n^{-\frac{1}{2}}\boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0) = n^{-\frac{1}{2}}\sum_{m=1}^{n}\boldsymbol{\mathcal{L}}_m^{(1)}(\boldsymbol{\beta}^0) \xrightarrow{d} N\Big(\mathbf{0}, \boldsymbol{V}\Big). \tag{2.19}$$

Our next step is to show the existence of a local maximizer $\hat{\boldsymbol{\beta}}$ of $L(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0||_2 = \mathcal{O}_p(n^{-\frac{1}{2}})$. Define a ball $\mathcal{A} = \Big\{\boldsymbol{\beta} : \boldsymbol{\beta} \leqslant \boldsymbol{\beta}^0 + Cn^{-\frac{1}{2}}\boldsymbol{v}\Big\}$, where $\boldsymbol{v}$ is any unit vector and C is a positive constant. Since $\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^0)$ is a continuous function with $\mathcal{L}(\boldsymbol{\beta}^0) - \mathcal{L}(\boldsymbol{\beta}^0) = 0$, if one can show that it is strictly negative on the boundary of the ball, we can conclude that there exists a local maximizer of $\mathcal{L}(\boldsymbol{\beta})$ inside $\mathcal{A}$. By Taylor expansion at $\boldsymbol{\beta}^0$, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^0) =& \boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0) \\ &+ \frac{1}{6}\sum_{r,t,u}[(\boldsymbol{\beta} - \boldsymbol{\beta}^0)]_{[r]}[(\boldsymbol{\beta} - \boldsymbol{\beta}^0)]_{[t]}[(\boldsymbol{\beta} - \boldsymbol{\beta}^0)]_{[u]}\frac{\partial^3\mathcal{L}(\tilde{\boldsymbol{\beta}})}{\partial\beta_r\partial\beta_t\partial\beta_u},\end{aligned} \tag{2.20}$$

where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^0$. From (2.19), we have $n^{-\frac{1}{2}}\boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0) = \mathcal{O}_p(1)$. The first term on the right-hand side of (2.20) is of the order $\mathcal{O}_p(1)$. As $\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)$ is deterministic, the second term can be written as

$$\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0) = -\frac{1}{2}n(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}^0)}{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^0).$$

From **Assumption** 2, we know that $\frac{\mathcal{I}(\boldsymbol{\beta}^0)}{n} \to \boldsymbol{V}$ as $n \to \infty$ and $\boldsymbol{V}$ is positive definite. Thus, the second term is less than $-\frac{1}{2}C^2\lambda_{\min}(\boldsymbol{V})$ as n goes to infinity. By choosing a sufficiently large C, the second term can dominate the first term. The third order partial derivative of $\mathcal{L}(\boldsymbol{\beta})$, for example, can be given as

$$\frac{\partial^3 \mathcal{L}(\boldsymbol{\beta})}{\partial[\theta_u]_{h_1}\partial[\theta_v]_{h_2}\partial\alpha_s} = \sum_{m=1}^{n} x_m^{(h_1)} x_m^{(h_2)} tr(\boldsymbol{T}^u\boldsymbol{\Sigma}_m\boldsymbol{T}^s\boldsymbol{\Sigma}_m\boldsymbol{T}^v\boldsymbol{\Sigma}_m). \tag{2.21}$$

Other third-order partial derivatives take similar forms to (2.21). From **Assumption** 1, we have $\frac{\partial^3\mathcal{L}(\tilde{\boldsymbol{\beta}})}{\partial\beta_r\partial\beta_t\partial\beta_u} = \mathcal{O}(n)$ and thus, the third term is of the order $\mathcal{O}(n^{-\frac{1}{2}})$. This is dominated by the second term as well. As a result, we have shown that with probability tending to 1, the function $\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^0)$ is strictly negative on the ball, which completes the proof of the consistency of the estimator.

Lastly, we show the asymptotic normality for $\hat{\boldsymbol{\beta}}$. By Taylor expansion, there exists $\tilde{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^0$ such that for $\forall\ r, t, u = 1, \cdots, \frac{p(p+1)(H+1)}{2}$,

$$\begin{aligned}\big[\boldsymbol{\mathcal{L}}^{(1)}(\hat{\boldsymbol{\beta}})\big]_r = 0 = \big[\boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0)\big]_r + \sum_t \big[\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)\big]_{rt} \cdot \big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\big]_t \\ + \sum_{t,u} \frac{\partial^3\boldsymbol{\mathcal{L}}(\tilde{\boldsymbol{\beta}})}{\partial\beta_r\partial\beta_t\partial\beta_u}\big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\big]_t\big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\big]_u,\end{aligned} \tag{2.22}$$

where r, t, u denote the row or column index of the vector and matrices. The third term on the right-hand side is $\mathcal{O}_p(1)$, whereas the second term has an order of $\mathcal{O}_p(n^{\frac{1}{2}})$, which dominates the third term. Then (2.22) can be written as

$$\frac{1}{n}\big[\boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0)\big]_r = \sum_t \big[-\frac{1}{n}\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)\big]_{rt} \cdot \big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\big]_t(1 + o_p(1)).$$

In matrix form, we have

$$\frac{1}{n}\boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0) = -\frac{1}{n}\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big) + \Big\{-\frac{1}{n}\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)\Big\}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big)\boldsymbol{o}_p(1), \tag{2.23}$$

where $\boldsymbol{o}_p(1)$ denotes the residual vector. Since the set of positive definite matrices is an open convex cone and the matrix $\boldsymbol{V}$ is positive definite by **Assumption** 2, we have the positive definiteness of $-\frac{1}{n}\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)$. Taking the inverse transformation, (2.23) becomes

$$n^{\frac{1}{2}}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big) = \Big\{-\frac{1}{n}\boldsymbol{\mathcal{L}}^{(2)}(\boldsymbol{\beta}^0)\Big\}^{-1}\Big\{n^{-\frac{1}{2}}\boldsymbol{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^0)\Big\} + \boldsymbol{o}_p(1).$$

Using (2.19), we show the asymptotic normality for $\hat{\boldsymbol{\beta}}$ as

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N\Big(\mathbf{0}, \boldsymbol{V}^{-1}\Big). \tag{2.24}$$

□

Proof of Theorem 2.2

Proof. The objective function is given by

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -l_c(\boldsymbol{\beta}) + n\sum_{h=1}^{H}\sum_{s\in\mathcal{E}}\rho_\lambda(|[\theta_s]_h|) + n\sum_{w\in\mathcal{E}}\rho_\lambda(|\alpha_w|).$$

Define a function $G(\boldsymbol{\Delta}) = Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}^0)$, where $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$. If we take a closed convex set $\mathcal{A}$ which contains $\mathbf{0}$ and further show that the function $G(\boldsymbol{\Delta})$ is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that $G(\boldsymbol{\Delta})$ has a local minimum inside the convex set $\mathcal{A}$, since $G(\mathbf{\Delta})$ is continuous and $G(\mathbf{0}) = \mathbf{0}$. In particular, we define the closed convex set $\mathcal{A} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{0}}||_2 \leqslant Cr_n\}$ and its boundary $\partial\mathcal{A} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{0}}||_2 = Cr_n\}$, where C is some positive constant and $r_n = \left\{\frac{p_n^{1+d}\log p_n}{n}\right\}^{\frac{1}{2}}$.

By the definition, we show that

$$G(\mathbf{\Delta}) = l_c(\boldsymbol{\beta}^0) - l_c(\boldsymbol{\beta}) + \underbrace{n\sum_{h=1}^{H}\sum_{s\in\mathcal{E}}\left\{\rho_\lambda(|[\theta_s]_h|) - \rho_\lambda(|[\theta_s^0]_h|)\right\}}_{\mathcal{I}_4} + \underbrace{n\sum_{w\in\mathcal{E}}\left\{\rho_\lambda(|\alpha_w|) - \rho_\lambda(|\alpha_w^0|)\right\}}_{\mathcal{I}_5}.$$

We start with the composite log-likelihood function and apply Taylor's expansion to $l_c(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathbf{0}}$:

$$l_c(\boldsymbol{\beta}) = l_c(\boldsymbol{\beta}^0) + l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot l_c^{(2)}(\boldsymbol{\beta}^*) \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0), \tag{2.25}$$

where $\boldsymbol{\beta}^* = a\boldsymbol{\beta} + (1-a)\boldsymbol{\beta}^0$ for some constant $a \in (0, 1)$. The difference between the two composite log-likelihood functions can be written as

$$\begin{aligned} l_c(\boldsymbol{\beta}^0) - l_c(\boldsymbol{\beta}) &= -l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot l_c^{(2)}(\boldsymbol{\beta}^*) \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) \\ &= \underbrace{-l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)}_{\mathcal{I}_1} + \underbrace{\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\} \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)}_{\mathcal{I}_2} \\ &\quad + \underbrace{\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot \left\{ -l_c^{(2)}(\boldsymbol{\beta}^*) - E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\right\} \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)}_{\mathcal{I}_3}. \end{aligned} \tag{2.26}$$

By the assumption that $E\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^*)\}$ has eigenvalues uniformly bounded away from zero and infinity when $\boldsymbol{\beta}^*$ is in the η-neighborhood of $\boldsymbol{\beta}^0$, there exist $\kappa_-, \kappa_+ > 0$ such that $0 < \kappa_- \leqslant \lambda\Big(E\{-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^*)\}\Big) \leqslant \kappa_+ < \infty$. Therefore, we have $\boldsymbol{v}^T E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\boldsymbol{v} = \mathcal{O}_p(n) > 0$, where $\boldsymbol{v}$ is any unit-norm vector. Consider on the boundary $\partial\mathcal{A}$ where $\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathbf{0}} + Cr_n\boldsymbol{v}$ for some constant $C > 0$. We show that

$$\begin{aligned}
\mathcal{I}_2 &= \frac{1}{2}C^2r_n^2 \cdot \boldsymbol{v}^T E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\boldsymbol{v} = \frac{1}{2}C^2r_n^2 \cdot \mathcal{O}_p(n) > 0, \\
\mathcal{I}_3 &= \frac{1}{2}C^2r_n^2 \cdot \boldsymbol{v}^T\Big\{-l_c^{(2)}(\boldsymbol{\beta}^*) - E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\Big\}\boldsymbol{v} \\
&= \frac{1}{2}C^2r_n^2 \cdot \sum_{r,t} v_r o_p\Big\{\Big[E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\Big]_{rt}\Big\}v_t,
\end{aligned}$$

which is dominated by $\mathcal{I}_2$. For $\mathcal{I}_1$, we have

$$\begin{aligned}
|\mathcal{I}_1| &= \Big|-l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)\Big| = \Big|\sum_t \big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t]} \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{[t]}\Big| \\
&\leqslant \underbrace{Cr_n \sum_{t_1:diagonal}\Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_1]}\Big||v_{t_1}|}_{|\mathcal{I}_{1.1}|} + \underbrace{Cr_n \sum_{t_2:\beta_{t_2}^0 \neq 0}\Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_2]}\Big||v_{t_2}|}_{|\mathcal{I}_{1.2}|} \\
&\quad + \underbrace{Cr_n \sum_{t_3:\beta_{t_3}^0 = 0}\Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_3]}\Big||v_{t_3}|}_{|\mathcal{I}_{1.3}|}.
\end{aligned}$$

We start with the second term $|\mathcal{I}_{1.2}|$ as follows. Since the # of nonzero true off-diagonal parameters is q_n and $\max_s\Big|\frac{\partial - l_c(\boldsymbol{\beta}^0)}{\partial[\theta_s]_{h_1}}\Big| = \max_w\Big|\frac{\partial - l_c(\boldsymbol{\beta}^0)}{\partial\alpha_w}\Big| = \mathcal{O}_p\Big\{(ns_n^2\log p_n)^{\frac{1}{2}}\Big\}$.

By Cauchy–Schwarz inequality, $|\mathcal{I}_{1.2}|$ becomes

$$
\begin{aligned}
|\mathcal{I}_{1.2}| = Cr_n \sum_{t_2:\beta^0_{t_2}\neq 0} \Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_2]}\Big| |v_{t_2}| &\leqslant Cr_n \sqrt{\sum_{t_2:\beta^0_{t_2}\neq 0} \big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]^2_{[t_2]}} \sqrt{\sum_{t_2:\beta^0_{t_2}\neq 0} [v_{t_2}]^2} \\
&\leqslant Cr_n \mathcal{O}_p\Big\{(q_n n s_n^2 \log p_n)^{\frac{1}{2}}\Big\}.
\end{aligned}
$$

Divide $|\mathcal{I}_{1.2}|$ by $\mathcal{I}_2$, we show that

$$
\begin{aligned}
\frac{|\mathcal{I}_{1.2}|}{\mathcal{I}_2} \leqslant \frac{Cr_n \mathcal{O}_p\Big\{(q_n n s_n^2 \log p_n)^{\frac{1}{2}}\Big\}}{\frac{1}{2}C^2 r_n^2 n \kappa_-} = \frac{2\mathcal{O}_p\Big\{(q_n s_n^2 \log p_n)^{\frac{1}{2}}\Big\}}{Cr_n n^{\frac{1}{2}} \kappa_-} &= \frac{2\mathcal{O}_p\Big\{(p_n s_n^3 \log p_n)^{\frac{1}{2}}\Big\}}{C\kappa_-(p_n^{1+d}\log p_n)^{\frac{1}{2}}} \\
&= \frac{2}{C\kappa_-}\mathcal{O}_p\Bigg\{\bigg(\frac{s_n^3}{p_n^d}\bigg)^{\frac{1}{2}}\Bigg\}.
\end{aligned}
$$

Since $r_n = \left(\frac{p_n^{1+d}\log p_n}{n}\right)^{\frac{1}{2}}$, $q_n = \mathcal{O}(p_n s_n)$, and $s_n^4 = \mathcal{O}(p_n^d)$, $|\mathcal{I}_{1.2}|$ can be dominated by $\mathcal{I}_2$ with an appropriate $C > 0$.

Third Term $|\mathcal{I}_{1.3}|$: We further decompose the penalty terms $\mathcal{I}_4$ and $\mathcal{I}_5$ as

$$\begin{aligned}
\mathcal{I}_4 &= n\sum_{h=1}^{H}\sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h=0}}\Big\{\rho_\lambda(|[\theta_s]_h|)-\rho_\lambda(|[\theta_s^0]_h|)\Big\} + n\sum_{h=1}^{H}\sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}}\Big\{\rho_\lambda(|[\theta_s]_h|)-\rho_\lambda(|[\theta_s^0]_h|)\Big\}\\
&= \underbrace{n\sum_{h=1}^{H}\sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h=0}}\Big\{\rho_\lambda(|[\theta_s]_h|)\Big\}}_{\mathcal{I}_{4.1}} + \underbrace{n\sum_{h=1}^{H}\sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}}\Big\{\rho_\lambda(|[\theta_s]_h|)-\rho_\lambda(|[\theta_s^0]_h|)\Big\}}_{\mathcal{I}_{4.2}},\\
\mathcal{I}_5 &= n\sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0=0}}\Big\{\rho_\lambda(|\alpha_w|)-\rho_\lambda(|\alpha_w^0|)\Big\} + n\sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0\neq 0}}\Big\{\rho_\lambda(|\alpha_w|)-\rho_\lambda(|\alpha_w^0|)\Big\}\\
&= \underbrace{n\sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0=0}}\Big\{\rho_\lambda(|\alpha_w|)\Big\}}_{\mathcal{I}_{5.1}} + \underbrace{n\sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0\neq 0}}\Big\{\rho_\lambda(|\alpha_w|)-\rho_\lambda(|\alpha_w^0|)\Big\}}_{\mathcal{I}_{5.2}}.
\end{aligned} \tag{2.27}$$

by **Lemma** 2.2, we show that

$$\begin{aligned}
|\mathcal{I}_{1.3}| &= \sum_{t_3:\boldsymbol{\beta}_{t_3}^0=0}\Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_3]}\Big|\cdot\big|(\boldsymbol{\beta}-\boldsymbol{\beta}^0)_{[t_3]}\big| = Cr_n\sum_{t_3:\beta_{t_3}^0=0}\Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_3]}\Big||v_{t_3}|\\
&= Cr_n\sum_{t_3:\beta_{t_3}^0=0}\mathcal{O}_p\Big\{(ns_n^2\log p_n)^{\frac{1}{2}}\Big\}|v_{t_3}|,\\
\mathcal{I}_{4.1}+\mathcal{I}_{5.1} &= n\lambda\sum_{h=1}^{H}\sum_{s:[\theta_s^0]_h=0}|[\theta_s]_h| + n\lambda\sum_{w:\alpha_w^0=0}|\alpha_w|\\
&= n\lambda\sum_{\substack{t_3,h=1,\cdots,H:\\ [\theta_s^0]_h=0,\alpha_w^0=0}}Cr_n|v_{t_3}| = Cr_n\sum_{\substack{t_3,h=1,\cdots,H:\\ [\theta_s^0]_h=0,\alpha_w^0=0}}n\lambda|v_{t_3}|>0.
\end{aligned}$$

We choose $\lambda \geqslant \delta_1\left(\frac{s_n^2\log p_n}{n}\right)^{\frac{1}{2}}$ for some constant $\delta_1>0$ such that $|\mathcal{I}_{1.3}|\leqslant \mathcal{I}_{4.1}+\mathcal{I}_{5.1}$.

First Term $|\mathcal{I}_{1.1}|$: By **Lemma** 2.2 and Cauchy-Schwarz inequality, the upper bound is

$$|\mathcal{I}_{1.1}| = Cr_n \sum_{t_1:diagonal} \Big| \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t_1]} \Big| |v_{t_1}| \leqslant Cr_n \sqrt{\sum_{t_1} \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t_1]}^2} \sqrt{\sum_{t_1} v_{t_1}^2}$$

$$\leqslant Cr_n \mathcal{O}_p \Big\{ (p_n n s_n^4 \log p_n)^{\frac{1}{2}} \Big\}.$$

Divide $|\mathcal{I}_{1.1}|$ by $\mathcal{I}_2$, we have

$$\frac{|\mathcal{I}_{1.1}|}{\mathcal{I}_2} \leqslant \frac{Cr_n \mathcal{O}_p \Big\{ (p_n n s_n^4 \log p_n)^{\frac{1}{2}} \Big\}}{C^2 \kappa_- r_n^2 n} = \frac{\mathcal{O}_p \Big\{ (p_n s_n^4 \log p_n)^{\frac{1}{2}} \Big\}}{C \kappa_- r_n n^{\frac{1}{2}}}.$$

If $r_n = \Big(\frac{p_n^{1+d} \log p_n}{n} \Big)^{\frac{1}{2}}$, then

$$\frac{|\mathcal{I}_{1.1}|}{\mathcal{I}_2} \leqslant \frac{1}{C\kappa_-} \mathcal{O}_p \Big\{ \Big(\frac{s_n^4}{p_n^d} \Big)^{\frac{1}{2}} \Big\}.$$

$\mathcal{I}_2$ can dominate $|\mathcal{I}_{1.1}|$ with an appropriate constant $C > 0$ if $s_n^4 = \mathcal{O}(p_n^d)$.

Next, we move on to the rest of the penalty terms. Since

$$\rho_\lambda(|[\theta_s]_h|) - \rho_\lambda(|[\theta_s^0]_h|) = \begin{cases} \lambda([\theta_s]_h - [\theta_s^0]_h), & \text{if } [\theta_s]_h \geqslant 0, [\theta_s^0]_h > 0; \\ -\lambda([\theta_s]_h - [\theta_s^0]_h), & \text{if } [\theta_s]_h \leqslant 0, [\theta_s^0]_h < 0; \\ -\lambda[\theta_s]_h - \lambda[\theta_s^0]_h, & \text{if } [\theta_s]_h < 0, [\theta_s^0]_h > 0; \\ \lambda[\theta_s]_h + \lambda[\theta_s^0]_h, & \text{if } [\theta_s]_h > 0, [\theta_s^0]_h < 0; \end{cases}$$

and

$$\text{when } [\theta_s]_h < 0, [\theta_s^0]_h > 0, \left| -\lambda[\theta_s]_h - \lambda[\theta_s^0]_h \right| \leqslant \left| \lambda([\theta_s]_h - [\theta_s^0]_h) \right|,$$

$$\text{when } [\theta_s]_h > 0, [\theta_s^0]_h < 0, \left| \lambda[\theta_s]_h + \lambda[\theta_s^0]_h \right| \leqslant \left| \lambda([\theta_s]_h - [\theta_s^0]_h) \right|.$$

Following similar steps, the expression $\rho_\lambda(|\alpha_w|) - \rho_\lambda(|\alpha_w^0|)$ yields equivalent bounds, leading us to conclude that

$$|\mathcal{I}_{4.2}| = \Big|n \sum_{h=1}^{H} \sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} \Big\{\rho_\lambda(|[\theta_s]_h|) - \rho_\lambda(|[\theta_s^0]_h|)\Big\}\Big| \leqslant n \sum_{h=1}^{H} \sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} \big|\rho_\lambda(|[\theta_s]_h|) - \rho_\lambda(|[\theta_s^0]_h|)\big|$$

$$\leqslant n \sum_{h=1}^{H} \sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} \big|\lambda([\theta_s]_h - [\theta_s^0]_h)\big| = n\lambda \sum_{h=1}^{H} \sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} Cr_n|[v_s]_h| \leqslant n\lambda Cr_n \sum_{h=1}^{H} \sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} |[v_s]_h| \cdot 1,$$

$$|\mathcal{I}_{5.2}| \leqslant n \sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0\neq 0}} \big|\rho_\lambda(|\alpha_w|) - \rho_\lambda(|\alpha_w^0|)\big| \leqslant n\lambda Cr_n \sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0\neq 0}} |v_w| \cdot 1.$$

Since the # of terms included in the summation $\big(\sum_{h=1}^{H}\sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} + \sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0\neq 0}}\big)$ is q_n, by Cauchy-Schwarz inequality, the upper bound of $|\mathcal{I}_{4.2}| + |\mathcal{I}_{5.2}|$ can be given as

$$\begin{aligned} |\mathcal{I}_{4.2}| + |\mathcal{I}_{5.2}| &\leqslant n\lambda Cr_n \Bigg(\sum_{h=1}^{H} \sum_{\substack{s\in\mathcal{E}:\\ [\theta_s^0]_h\neq 0}} |[v_s]_h| \cdot 1 + \sum_{\substack{w\in\mathcal{E}:\\ \alpha_w^0\neq 0}} |v_w| \cdot 1 \Bigg) \\ &\leqslant n\lambda Cr_n \sqrt{\sum_{s,w,h} ([v]_h)^2} \sqrt{\sum_{s,w,h} (1)^2} \leqslant n\lambda Cr_n \cdot (q_n)^{\frac{1}{2}}. \end{aligned} \tag{2.28}$$

We divide it by $\mathcal{I}_2$

$$\frac{|\mathcal{I}_{4.2}| + |\mathcal{I}_{5.2}|}{\mathcal{I}_2} \leqslant \frac{2n\lambda Cr_n q_n^{\frac{1}{2}}}{C^2 r_n^2 n\kappa_-} = \frac{2\lambda q_n^{\frac{1}{2}}}{Cr_n\kappa_-} \leqslant \frac{2\delta_2}{C\kappa_-}\left(\frac{s_n^3}{p_n^d}\right)^{\frac{1}{2}}\left(\frac{q_n}{p_n}\right)^{\frac{1}{2}} = \frac{2\delta_2}{C\kappa_-}\mathcal{O}\left\{\left(\frac{s_n^4}{p_n^d}\right)^{\frac{1}{2}}\right\}$$

The numerator can be dominated by $\mathcal{I}_2$ with an appropriate constant $C > 0$ if $\lambda \leqslant \delta_2\left(\frac{s_n^3 \log p_n}{n}\right)^{\frac{1}{2}}$ for some constant $\delta_2 > 0$.

In summary, we have shown that $G(\boldsymbol{\Delta}) > 0$ on the boundary $\partial\mathcal{A}$ by choosing $C > 0$ sufficiently large and $\delta_1\left(\frac{s_n^2 \log p_n}{n}\right)^{\frac{1}{2}} \leqslant \lambda \leqslant \delta_2\left(\frac{s_n^3 \log p_n}{n}\right)^{\frac{1}{2}}$ for some constants $\delta_1, \delta_2 > 0$, which completes the proof. □

Proof of Theorem 2.3

Proof. Since the property of sign consistency is regarding the edge parameters, we treat diagonal parameters as known in this theorem. Recall that the composite likelihood estimator $\hat{\boldsymbol{\beta}}$ is obtained by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \big\{ -l_c(\boldsymbol{\beta}) + n\sum_{s\in\mathcal{E}}\sum_{h=1}^{H}\rho_\lambda(|[\theta_s]_h|) + n\sum_{w\in\mathcal{E}}\rho_\lambda(|\alpha_w|)\big\}. \tag{2.29}$$

Here $\boldsymbol{\beta}$ excludes diagonal parameters. Similar to Ravikumar and Yu (2008), we introduce another estimator

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}=(\boldsymbol{\beta}_{\mathcal{S}_1},\mathbf{0})} \big\{ -l_c(\boldsymbol{\beta}) + n\sum_{s\in\mathcal{E}}\sum_{h=1}^{H}\rho_\lambda(|[\theta_s]_h|) + n\sum_{w\in\mathcal{E}}\rho_\lambda(|\alpha_w|)\big\}, \tag{2.30}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_{\mathcal{S}_1}, \tilde{\boldsymbol{\beta}}_{\mathcal{S}^c})$ with the constraint $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}$. The set of first derivative equations $\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$ can be partitioned into two sets of equations

$$\frac{1}{n} l_c^{(1)}(\tilde{\boldsymbol{\beta}})_{\mathcal{S}_1} = \lambda \tilde{z}_{\mathcal{S}_1}, \tag{2.31}$$

$$\frac{1}{n} l_c^{(1)}(\tilde{\boldsymbol{\beta}})_{\mathcal{S}^c} = \lambda \tilde{z}_{\mathcal{S}^c}, \tag{2.32}$$

where the sub-differential of the absolute value function is defined as

$$\tilde{z} = \frac{\partial |x|}{\partial x} = \begin{cases} 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

If the estimator $\tilde{\boldsymbol{\beta}}$ satisfies conditions (2.31) and (2.32), then $\tilde{\boldsymbol{\beta}}$ is the local minimizer $\hat{\boldsymbol{\beta}}$ of the objective function $Q(\boldsymbol{\beta})$. We show the proof in two steps.

Step 1: Verify $\|\tilde{z}_{\mathcal{S}^c}\|_\infty < 1$.

By Taylor's expansion, we have

$$\frac{1}{n} l_c^{(1)}(\tilde{\boldsymbol{\beta}}) = \frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0) + \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \underbrace{\frac{1}{n}\left\{ l_c^{(2)}(\boldsymbol{\beta}^*) - l_c^{(2)}(\boldsymbol{\beta}^0)\right\}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}_{R}, \quad (2.33)$$

where $\boldsymbol{\beta}^* = a\boldsymbol{\beta}^0 + (1-a)\tilde{\boldsymbol{\beta}}$ for some constant $a \in [0, 1]$. In block format, we can write

$$\begin{pmatrix} \frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1} \\ \frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}^c} \end{pmatrix} + \frac{1}{n} \begin{pmatrix} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1} & l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}^c} \\ l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1} & l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}^c} \end{pmatrix} \begin{pmatrix} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)_{\mathcal{S}_1} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} R_{\mathcal{S}_1} \\ R_{\mathcal{S}^c} \end{pmatrix} = \lambda \begin{pmatrix} \tilde{z}_{\mathcal{S}_1} \\ \tilde{z}_{\mathcal{S}^c} \end{pmatrix},$$

where $R_{\mathcal{S}_1}, R_{\mathcal{S}^c}$ are the partitions of the residual vector. From the block format, we know that

$$\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)_{\mathcal{S}_1} = \left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} \left(\lambda \tilde{z}_{\mathcal{S}_1} - R_{\mathcal{S}_1} - \frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right), \quad (2.34)$$

$$\tilde{z}_{\mathcal{S}^c} = \frac{1}{\lambda}\left(\frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} + \frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)_{\mathcal{S}_1} + R_{\mathcal{S}^c}\right). \tag{2.35}$$

By equations (2.34) and (2.35), we have

$$\begin{aligned}
\tilde{z}_{\mathcal{S}^c} =& \frac{1}{\lambda}\left\{\frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} + \frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left(\lambda\tilde{z}_{\mathcal{S}_1} - R_{\mathcal{S}_1} - \frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right) + R_{\mathcal{S}^c}\right\} \\
=& \underbrace{\frac{1}{\lambda}\left\{\frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} - \frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left(\frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right)\right\}}_{\mathcal{I}_1} \\
&+ \underbrace{\frac{1}{\lambda}\left\{R_{\mathcal{S}^c} - \frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}R_{\mathcal{S}_1}\right\}}_{\mathcal{I}_2} + \underbrace{\frac{1}{\lambda}\left\{\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\lambda\tilde{z}_{\mathcal{S}_1}\right\}}_{\mathcal{I}_3}.
\end{aligned}$$

Since $\|\tilde{z}_{\mathcal{S}^c}\|_\infty = \|\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3\|_\infty \leqslant \|\mathcal{I}_1\|_\infty + \|\mathcal{I}_2\|_\infty + \|\mathcal{I}_3\|_\infty$, we start with the upper bound of $\|\mathcal{I}_2\|_\infty$. By Taylor's expansion, we show that each element of R can be written as

$$(R)_u = \frac{1}{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{\mathbf{0}})^T\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}),$$

where β_u denotes any off-diagonal parameter and $\boldsymbol{\beta}^{**} = b\boldsymbol{\beta}^0 + (1-b)\boldsymbol{\beta}^*$ for some constant $b \in [0,1]$. Also note that $(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{\mathbf{0}})^T = (1-a)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T$, we can rewrite it as

$$
\begin{aligned}
(R)_u &= \frac{1}{n}(1-a)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\mathbf{0}})^T \frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\mathbf{0}}) \\
&= \underbrace{\frac{1}{n}(1-a)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\mathbf{0}})^T\left\{\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u} - E\left\{\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\right\}\right\}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\mathbf{0}})}_{\mathcal{M}_1} \\
&\quad \underbrace{-\frac{1}{n}(1-a)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\mathbf{0}})^T E\left\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\right\}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\mathbf{0}})}_{\mathcal{M}_2}.
\end{aligned}
$$

Following the same methodology as in **Theorem** 2.2, we can show $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2 = \mathcal{O}_p\left\{\left(\frac{p_n^{1+d}\log p_n}{n}\right)^{\frac{1}{2}}\right\}$. Thus, we have $\mathcal{M}_2 \leqslant \frac{1}{n}(1-a)\mathcal{O}_p\left\{\frac{p_n^{1+d}\log p_n}{n}\right\}\boldsymbol{v}^T E\left\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\right\}\boldsymbol{v}$ and

$$
\begin{aligned}
\left\|(R)_u\right\|_\infty &= \left\|\mathcal{M}_1-\mathcal{M}_2\right\|_\infty \leqslant \left\|\mathcal{M}_1\right\|_\infty + \left\|\mathcal{M}_2\right\|_\infty, \\
&\leqslant \left\|\mathcal{M}_2\right\|_\infty\big(1+o_p(1)\big) \leqslant \frac{1}{n}\mathcal{O}_p\left\{\frac{p_n^{1+d}\log p_n}{n}\right\}\left|\left|\left|E\left\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\right\}\right|\right|\right|_2\big(1+o_p(1)\big).
\end{aligned}
$$

By **Assumptions** 3 and 4, we have $\left|\left|\left|E\left\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\right\}\right|\right|\right|_2 \leqslant n\mathcal{W}^*$ for some constant $\mathcal{W}^* > 0$ and show that

$$
\left\|(R)_u\right\|_\infty \leqslant \mathcal{O}_p\left\{\frac{p_n^{1+d}\log p_n}{n}\right\}\frac{1}{n}n\mathcal{W}^*\big(1+o_p(1)\big) \leqslant \mathcal{O}_p\left\{\frac{p_n^{1+d}\log p_n}{n}\right\}\big(1+o_p(1)\big) = o_p(1).
$$

Therefore, by **Lemma** 2.7, the upper bound of $\|\mathcal{I}_2\|_\infty$ can be given as

$$\begin{aligned}
\|\mathcal{I}_2\|_\infty &= \frac{1}{\lambda}\left|\left|\left| R_{\mathcal{S}^c} - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} R_{\mathcal{S}_1}\right|\right|\right|_\infty \\
&\leqslant \frac{1}{\lambda}\|R_{\mathcal{S}^c}\|_\infty + \frac{1}{\lambda}\left|\left|\left|\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \|R_{\mathcal{S}_1}\|_\infty \\
&\leqslant \frac{1}{\delta_1}\left(\frac{n}{s_n^2 \log p_n}\right)^{\frac{1}{2}} \mathcal{O}_p\left\{\frac{p_n^{1+d}\log p_n}{n}\right\}\big(1+o_p(1)\big) \\
&+ \frac{1}{\delta_1}\left(\frac{n}{s_n^2 \log p_n}\right)^{\frac{1}{2}} (1-\frac{1}{3}\xi^2)\mathcal{O}_p\left\{\frac{p_n^{1+d}\log p_n}{n}\right\}\big(1+o_p(1)\big) \\
&\leqslant \mathcal{O}_p\left\{\left(\frac{p_n^{2+2d}\log p_n}{n s_n^2}\right)^{\frac{1}{2}}\right\} \leqslant \mathcal{O}_p\left\{\left(\frac{q_n^2 p_n^{2d}\log p_n}{n}\right)^{\frac{1}{2}}\frac{1}{s_n}\right\} = o_p(1).
\end{aligned}$$

Similarly, by **Lemma** 2.2 and 2.7, the rest two terms can be bounded by

$$\begin{aligned}
\|\mathcal{I}_1\|_\infty &= \frac{1}{\lambda}\left\|\frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} - \frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left(\frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right)\right\|_\infty \\
&\leqslant \frac{1}{\lambda}\left\{\left\|\frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c}\right\|_\infty + \left|\left|\left|\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \left\|\frac{1}{n} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right\|_\infty\right\} \\
&\leqslant \frac{1}{\lambda}\left\{\epsilon + (1-\frac{1}{3}\xi^2)\epsilon\right\} \leqslant \frac{1}{\lambda}(2-\frac{1}{3}\xi^2)\epsilon, \\
\|\mathcal{I}_3\|_\infty &= \frac{1}{\lambda}\left\|\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\lambda\tilde{z}_{\mathcal{S}_1}\right\|_\infty \leqslant \left|\left|\left|\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \\
&\leqslant 1-\frac{1}{3}\xi^2.
\end{aligned}$$

Therefore, we choose $\epsilon = \frac{1}{6}\xi^2\lambda$ and

$$\begin{aligned}
\|\tilde{z}_{\mathcal{S}^c}\|_\infty \leqslant \|\mathcal{I}_1\|_\infty + \|\mathcal{I}_2\|_\infty + \|\mathcal{I}_3\|_\infty &\leqslant \frac{1}{\lambda}(2-\frac{1}{3}\xi^2)\epsilon + o(1) + 1 - \frac{1}{3}\xi^2 \\
&\leqslant \frac{1}{3}\xi^2 - \frac{1}{18}\xi^4 + 1 - \frac{1}{3}\xi^2 + o(1) < 1.
\end{aligned}$$

Once the first step is completed, it demonstrates that the solution $\tilde{\boldsymbol{\beta}}$ to the constrained problem (**??**) is equivalent to the solution $\hat{\boldsymbol{\beta}}$ to the original unrestricted problem (2.31).

Step 2: Verify $sign(\hat{\boldsymbol{\beta}}_{\mathcal{S}_1}) = sign(\boldsymbol{\beta}^0_{\mathcal{S}_1})$. From equation (2.34) and **Lemma** 2.2 and 2.6, we show that for some constant $c^* = \frac{1+\xi^2/6}{\tau_-}\big(1+o_p(1)\big)$,

$$\begin{aligned}
\Big\|\big(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big)_{\mathcal{S}_1}\Big\|_\infty &\leqslant \left|\left|\left|\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}\right|\right|\right|_\infty \left\|\Big(-\lambda\tilde{z}_{\mathcal{S}_1} + R_{\mathcal{S}_1} + \frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\Big)\right\|_\infty \\
&\leqslant \sqrt{|\mathcal{S}_1|}\left|\left|\left|\Big(-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}\right|\right|\right|_2 \Big(\lambda\|\tilde{z}_{\mathcal{S}_1}\|_\infty + \|R_{\mathcal{S}_1}\|_\infty + \|\frac{1}{n}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\|_\infty\Big) \\
&\leqslant \frac{\sqrt{q_n}}{\tau_-}\Big(\lambda + \mathcal{O}_p\Big\{\frac{p_n^{1+d}\log p_n}{n}\Big\}\big(1+o_p(1)\big) + \frac{1}{6}\xi^2\lambda\Big) \\
&\leqslant \frac{\sqrt{q_n}}{\tau_-}(1+\frac{1}{6}\xi^2)\lambda\big(1+o_p(1)\big) = c^*\sqrt{q_n}\lambda \leqslant \min_{u\in\mathcal{S}_1}\big\{\big|\beta_u^0\big|\big\}.
\end{aligned}$$

Based on all the results presented above, we have $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^0)$ with probability tending to one. □

3 High-Dimensional Covariate-Dependent Colored Gaussian Graphical Models

3.1 Introduction

Subsequent research has extended symmetry-constrained GGMs in several directions. Gehrmann (2011) demonstrated that both RCON and RCOR models exhibit complete lattice structures, making them suitable for model selection using the Edwards–Havránek procedure. Gao and Massam (2015) presented a robust and efficient framework for estimating symmetry-constrained GGMs in clustered dense networks. The method combined penalized likelihood with composite likelihood estimation under the colored GGM paradigm, achieving a desirable trade-off between model complexity and computational feasibility. Moreover, Massam et al. (2018) proposed a Bayesian approach to estimate precision and covariance matrices in GGMs subject to symmetry constraints, using the colored G-Wishart distribution. Several other works investigated model selection strategies within the framework of colored GGMs. Li et al. (2020) introduced a Bayesian model selection approach for symmetry-constrained GGMs, combining linear regression techniques with a double reversible jump MCMC algorithm. Additionally, Li et al. (2021) also proposed a penalized composite likelihood approach that simultaneously performs model selection and estimation in colored GGMs. Using the representation theory of the symmetric group, Graczyk et al. (2022) proposed a Bayesian model selection strategy for RCOP models. Following the previous work, the R package **gips**, introduced by Chojecki et al. (2025), enables the identification of hidden permutation symmetries and the corresponding estimation of covariance structures. Recent studies have also explored its applications to paired data. Ranciati et al. (2021) developed the symmetric graphical lasso (SGL) to perform parameter estimation and recover graph structures in pdRCON models. Ranciati and Roverato (2024) further developed pdglasso, which generalizes the symmetric graphical lasso to effectively model paired data through the pdRCON framework.

In this section, we propose a type of covariate-dependent colored Gaussian graphical model (CD–GGM), termed covariate-dependent RCON models, which integrates covariate effects with symmetry-induced colorings to model structured, heterogeneous networks. This setting arises naturally when domain knowledge suggests groups of edges or vertices should be constrained to share the same parameter, such as in gene regulatory networks or brain connectivity studies. We estimate the model parameters through a composite likelihood framework, applying coordinate descent for optimization. This method is fully applicable when the precision matrix has fixed diagonal entries. However, in situations where these diagonal entries vary with covariates, we adopt a hybrid approach that integrates coordinate descent with Broyden's method for estimation. An ℓ_1 penalty is incorporated to address sparsity in large-scale networks. We establish the parameter consistency and sign consistency of the proposed estimator. These results are applicable to models with either fixed or covariate-dependent diagonal components of the precision matrix. Our simulation studies, including both penalized and non-penalized, demonstrate strong empirical performance. A real data application to influenza vaccine responses is also provided, demonstrating the model's practical relevance and effectiveness.

3.2 Methodology

3.2.1 Model Setup

We begin with a brief review of the covariate-dependent Gaussian graphical models. Let $\boldsymbol{Y} = \big(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n\big)^T$ be a collection of independent random vectors of dimension p_n, with each $\boldsymbol{Y}_m \sim N\big(\boldsymbol{\mu}_m(\boldsymbol{x}_m), \boldsymbol{\Sigma}_m(\boldsymbol{x}_m)\big), m = 1, \cdots, n$, where $\boldsymbol{x}_m = (x_m^{(1)}, \ldots, x_m^{(H)})^T$ denotes the vector of observed covariates and $\boldsymbol{Y}_m = (Y_{m1}, Y_{m2}, ..., Y_{mp_n})^T$.

The quantity H is the number of covariates. We assume the observations are centered, namely $\boldsymbol{\mu}_m(\boldsymbol{x}_m) = \mathbf{0}$. The precision matrix varies with observed covariates in a linear form shown below

$$\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_m) = \boldsymbol{K}_m = \sum_{h=1}^{H} x_m^{(h)} \boldsymbol{Q}_h + (1 - \sum_{h=1}^{H} \frac{x_m^{(h)}}{H}) \boldsymbol{Q}_0 \tag{3.1}$$

$$= \boldsymbol{Q}_0 + \sum_{h=1}^{H} x_m^{(h)} (\boldsymbol{Q}_h - \boldsymbol{Q}_0/H) = \boldsymbol{Q}_0 + \sum_{h=1}^{H} x_m^{(h)} \boldsymbol{P}_h, \tag{3.2}$$

where $\boldsymbol{P}_h = \boldsymbol{Q}_h - \boldsymbol{Q}_0/H$. Each value $x_m^{(h)}$ is min-max scaled such that $x_m^{(h)} \in [0, 1]$, representing the mth observation of the hth covariate. We still use the following notations $\left(\boldsymbol{Q}_0\right)_{ij} = \alpha_{ij}$, $\left(\boldsymbol{P}_h\right)_{ij} = [\theta_{ij}]_h$ representing the (i, j)th element of $\boldsymbol{Q}_0$ and $\boldsymbol{P}_h$ respectively. The set of unknown parameters can be formulated as $\boldsymbol{\beta} = (\boldsymbol{\alpha}, \boldsymbol{\theta})^T$, where $\boldsymbol{\alpha}$ is the vector of all $\alpha_{ij}, i \leqslant j$ and $\boldsymbol{\theta}$ is the vector of all $[\theta_{ij}]_h, i \leqslant j, h = 1, \cdots, H$. The true values of parameters are denoted as $\boldsymbol{\beta}^{\mathbf{0}} = (\boldsymbol{\alpha}^{\mathbf{0}}, \boldsymbol{\theta}^{\mathbf{0}})^T$.

Based on the setting in (3.1), we further impose symmetry constraints on the precision matrix to introduce covariate-dependent RCON models. In the presence of known or hypothesized symmetries, the network's vertices and edges are organized into predefined groups (e.g., based on biological pathways in genetics), which are treated as independent of observed covariates. The predefined groups for vertices and edges are known as vertex and edge colorings, respectively. Fixed vertex and edge colorings are essential to maintain the symmetry conditions imposed by the RCON model, since allowing the colorings to vary with covariates would invalidate these structural constraints under the setting (3.1). If this assumption is relaxed and colorings are covariate-dependent, the model would reduce to a general covariate-dependent GGM, which can be handled using existing estimation procedures. As a result, the proposed covariate-dependent RCON($\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}$) models are given as

1. We denote the vertex coloring and the vertex color classes as $\boldsymbol{\mathcal{V}} = \{V_1, \cdots, V_c\}$, where c denotes the number of vertex color classes. For all $j \in V_r$, α_{jj} in $\boldsymbol{Q}_0$ are identical, $[\theta_{jj}]_h$ in $\boldsymbol{P}_h$ are identical.

2. We denote the common edge coloring and edge color classes as $\boldsymbol{\mathcal{E}} = \{E_1, \cdots, E_w\}$, where w denotes the number of edge color classes. For all $(i, j) \in E_s$, α_{ij} in $\boldsymbol{Q}_0$ are identical, $[\theta_{ij}]_h$ in $\boldsymbol{P}_h$ are identical.

Table 3.1 categorizes possible structures of covariate-dependent RCON models. Note that although we allow the dimension p_n to grow with n, the number of vertex and edge color classes is treated as finite, which substantially reduces the parameter space relative to general GGMs. Therefore, the unknown parameter set can be represented as

$$\begin{aligned}\boldsymbol{\beta} = (&\alpha_{V_1}, \cdots, \alpha_{V_c}, [\theta_{V_1}]_1, \cdots, [\theta_{V_c}]_1, \cdots, [\theta_{V_1}]_H, \cdots, [\theta_{V_c}]_H, \\ &\alpha_{E_1}, \cdots, \alpha_{E_w}, [\theta_{E_1}]_1, \cdots, [\theta_{E_w}]_1, \cdots, [\theta_{E_1}]_H, \cdots, [\theta_{E_w}]_H).\end{aligned}$$

Table 3.1: An overview of various covariate-dependent RCON Models

Special Cases	Conditions
Standard RCON models	$\forall\ h = 1, \cdots, H, x_m^{(h)} = 0$
Group-specific RCON models	$\forall\ h = 1, \cdots, H, x_m^{(h)}$ is categorical
Time-varying RCON models	$h = 1, x_m^{(h)}$ is continuous
General covariate-dependent RCON models	$\forall\ h = 1, \cdots, H, x_m^{(h)}$ is continuous or categorical

We define indicator matrices $\boldsymbol{T}_{V_r}, \boldsymbol{T}_{E_s}$ for each vertex and edge color class as

$$(\boldsymbol{T}_{V_r})_{ij} = \begin{cases} 1 & \text{if } i = j, j \in V_r, \\ 0 & \text{otherwise.} \end{cases} \quad (\boldsymbol{T}_{E_s})_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E_s \text{ or } (j,i) \in E_s, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Therefore, each matrix can be written as

$$\boldsymbol{Q}_0 = \sum_{s=1}^{w} \alpha_{E_s} \boldsymbol{T}_{E_s} + \sum_{r=1}^{c} \alpha_{V_r} \boldsymbol{T}_{V_r}, \quad \boldsymbol{P}_h = \sum_{s=1}^{w} [\theta_{E_s}]_h \boldsymbol{T}_{E_s} + \sum_{r=1}^{c} [\theta_{V_r}]_h \boldsymbol{T}_{V_r}. \tag{3.4}$$

Using the notations in (3.3) and (3.4), the precision matrix can be rewritten as

$$\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_m) = \boldsymbol{K}_m = \sum_{s=1}^{w} \alpha_{E_s} \boldsymbol{T}_{E_s} + \sum_{r=1}^{c} \alpha_{V_r} \boldsymbol{T}_{V_r} + \sum_{h=1}^{H} x_m^{(h)} \Big\{ \sum_{s=1}^{w} [\theta_{E_s}]_h \boldsymbol{T}_{E_s} + \sum_{r=1}^{c} [\theta_{V_r}]_h \boldsymbol{T}_{V_r} \Big\}. \tag{3.5}$$

3.2.2 Composite Likelihood Estimation

We introduce a penalized composite likelihood approach (Lindsay, 1988; Fearnhead and Donnelly, 2002; Cox and Reid, 2004; Varin, 2008; Larribe and Fearnhead, 2011; Ribatet et al., 2012; Gao and Massam, 2015) for estimation. For our proposed covariate-dependent RCON models, each random variable Y_{mj} given $\boldsymbol{Y}_{m,V\backslash\{j\}}$ follows a normal distribution $N\big(\mu_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}}, \sigma^2_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}}\big)$ with

$$\mu_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}} = -\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1} \sum_{i=1,i\neq j}^{p_n} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big] y_{mi}, \tag{3.6}$$

$$\sigma^2_{Y_{mj}|\boldsymbol{Y}_{m,V\backslash\{j\}}} = \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}. \tag{3.7}$$

where $V = \big\{1, \cdots, p_n\big\}$ denotes the vertex set, and $V\backslash\{j\}$ denotes all the vertices except j. Define $l_c(Y_{mj}, \boldsymbol{\beta})$ as the log-likelihood function of Y_{mj} given $\boldsymbol{Y}_{m,V\backslash\{j\}}$, and further define $l_c(\boldsymbol{Y}_m, \boldsymbol{\beta})$ as the joint log-likelihood function of $Y_{m1}|\boldsymbol{Y}_{m,V\backslash\{1\}}, Y_{m2}|\boldsymbol{Y}_{m,V\backslash\{2\}}$, $\cdots Y_{mp_n}|\boldsymbol{Y}_{m,V\backslash\{p_n\}}$, which incorporates all the conditional distributions for the mth observation. Lastly, we introduce the notation $l_c(\boldsymbol{Y}, \boldsymbol{\beta})$ for the joint composite log-likelihood function, which is constructed by incorporating all the conditional distributions. These notations lead to the expression: $l_c(\boldsymbol{Y}, \boldsymbol{\beta}) = \sum_{m=1}^{n} l_c(\boldsymbol{Y}_m, \boldsymbol{\beta}) = \sum_{m=1}^{n}\sum_{j=1}^{p_n} l_c(Y_{mj}, \boldsymbol{\beta})$. Therefore, the overall negative joint composite log-likelihood function can be shown as

$$
\begin{aligned}
&-l_c(\boldsymbol{Y},\boldsymbol{\beta}) = -\sum_{m=1}^{n} l_c(\boldsymbol{Y}_m,\boldsymbol{\beta}) = -\sum_{m=1}^{n}\sum_{j=1}^{p_n} l_c(Y_{mj},\boldsymbol{\beta})\\
&= -\frac{1}{2}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\log\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]+\frac{1}{2}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]y_{mj}^2\\
&+\frac{1}{2}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big[\alpha_{jj}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\Big\{\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big\}^2\\
&+\sum_{m=1}^{n}\sum_{j=1}^{p_n}y_{mj}\Big\{\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big\}+const.
\end{aligned}
$$

To enforce sparsity in the covariate-dependent RCON model, an ℓ_1 penalty is introduced, leading to the following optimization problem,

$$
\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -l_c(\boldsymbol{Y},\boldsymbol{\beta}) + n\phi\lambda\Big\{\sum_{s=1}^{w}|\alpha_{E_s}| + \sum_{s=1}^{w}\sum_{h=1}^{H}|[\theta_{E_s}]_h|\Big\}, \tag{3.8}
$$

where $\phi = \max_r\big\{|V_r|\big\}$ denotes the maximum cardinality among all V_r. As the composite log-likelihood incorporates multiple conditional densities, the factor ϕ is introduced to balance the relative scale of the penalty terms and likelihood. To simplify notation, we write $-l_c(\boldsymbol{Y},\boldsymbol{\beta}) = -l_c(\boldsymbol{\beta})$ throughout the remainder of the paper when there is no ambiguity.

We start with the coordinate descent algorithm (Friedman et al., 2007; Tseng, 2001) to solve (3.8) and obtain the penalized composite log-likelihood estimator $\hat{\boldsymbol{\beta}}_c$ for the edge parameters. A thorough exploration including the convergence properties of the estimators using the coordinate descent algorithm with the ℓ_1 penalization can be found in Breheny and Huang (2011) and Mazumder et al. (2011).

For $\forall\ E_s \in \boldsymbol{\mathcal{E}}, h_1 = 1, \cdots, H$, taking the partial derivatives of $Q(\boldsymbol{\beta})$ with respect to $[\theta_{E_s}]_{h_1}, \alpha_{E_s}$, and setting them to zero leads to the following updating equations for each edge color class

$$\hat{\alpha}_{E_s} = \frac{S\Big(-\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big\{ y_{mj}\big(\sum_{k:(k,j)\in E_s} y_{mk}\big) + \mathcal{I}_{1.1} + \mathcal{I}_{1.2}\Big\}, \lambda\Big)}{\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{k:(k,j)\in E_s} y_{mk}\big)\big(\sum_{k:(k,j)\in E_s} y_{mk}\big)}, \tag{3.9}$$

$$[\hat{\theta}_{E_s}]_{h_1} = \frac{S\Big(-\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big\{ y_{mj}\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big) + \mathcal{I}_{2.1} + \mathcal{I}_{2.2} + \mathcal{I}_{2.3}\Big\}, \lambda\Big)}{\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big)\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big)}, \tag{3.10}$$

where $S(z, \lambda) = sign(z)(|z| - \lambda)_+$ represents the soft-thresholding operator. The notation $(|z| - \lambda)_+$ is defined as

$$(|z| - \lambda)_+ = \begin{cases} |z| - \lambda, & \text{if } |z| - \lambda > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The expressions for $\mathcal{I}_{1.1}, \cdots, \mathcal{I}_{2.3}$ are provided below

$$\mathcal{I}_{1.1} = \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{r:(r,j)\in E_s^c} \alpha_{jr} y_{mr}\big)\big(\sum_{k:(k,j)\in E_s} y_{mk}\big),$$

$$\mathcal{I}_{1.2} = \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n} \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h y_{mi}\big)\big(\sum_{k:(k,j)\in E_s} y_{mk}\big),$$

$$\mathcal{I}_{2.1} = \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n} \alpha_{ji} y_{mi}\big)\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big),$$

$$\mathcal{I}_{2.2} = \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{r:(r,j)\in E_s^c} x_m^{(h_1)}[\theta_{jr}]_{h_1} y_{mr}\big)\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big),$$

$$\mathcal{I}_{2.3} = \Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1}\big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n} \sum_{\substack{h=1,\\ h\neq h_1}}^{H} x_m^{(h)}[\theta_{ji}]_h y_{mi}\big)\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big).$$

For the vertex parameters, the corresponding partial derivatives do not yield closed-form solutions when set to zero. Consequently, the set of equations presented below is solved numerically using Broyden's method (Nocedal and Wright, 2006).

$$\begin{aligned}
\frac{\partial - l_c(\boldsymbol{Y}, \boldsymbol{\beta})}{\partial \alpha_{V_r}} &= \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r} y_{mj}^2 - \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1} \\
&\quad - \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big\{\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big] y_{mi}\Big\}^2, \\
\frac{\partial - l_c(\boldsymbol{Y}, \boldsymbol{\beta})}{\partial [\theta_{V_r}]_{h_1}} &= \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r} x_m^{(h_1)} y_{mj}^2 - \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r} x_m^{(h_1)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-1} \\
&\quad - \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r} x_m^{(h_1)}\Big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\Big]^{-2}\Big\{\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big] y_{mi}\Big\}^2.
\end{aligned}$$

To the best of our knowledge, (3.9) and (3.10) cannot be reformulated as trace (matrix) expressions, making them computationally demanding when p_n is large. In contrast, under the assumption that the diagonal components of the precision matrix do not vary with covariates, as considered in Zhang and Li (2023), the updating equations would be reduced to

$$\hat{\alpha}_{E_s} = \frac{S\Big(-\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big\{ y_{mj}\big(\sum_{k:(k,j)\in E_s} y_{mk}\big) + \mathcal{I}_{3.1} + \mathcal{I}_{3.2}\Big\}, \lambda\Big)}{\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\alpha_{jj}^{-1}\big(\sum_{k:(k,j)\in E_s} y_{mk}\big)\big(\sum_{k:(k,j)\in E_s} y_{mk}\big)}, \tag{3.11}$$

$$[\hat{\theta}_{E_s}]_{h_1} = \frac{S\Big(-\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\Big\{ y_{mj}\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big) + \mathcal{I}_{4.1} + \mathcal{I}_{4.2} + \mathcal{I}_{4.3}\Big\}, \lambda\Big)}{\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n}\alpha_{jj}^{-1}\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big)\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big)}, \tag{3.12}$$

where the terms $\mathcal{I}_{3.1}, \mathcal{I}_{3.2}, \cdots, \mathcal{I}_{4.3}$ are defined below.

$$\mathcal{I}_{3.1} = \alpha_{jj}^{-1}\big(\sum_{r:(r,j)\in E_s^c} \alpha_{jr} y_{mr}\big)\big(\sum_{k:(k,j)\in E_s} y_{mk}\big), \qquad \mathcal{I}_{3.2} = \alpha_{jj}^{-1}\big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n}\sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h y_{mi}\big)\big(\sum_{k:(k,j)\in E_s} y_{mk}\big),$$

$$\mathcal{I}_{4.1} = \alpha_{jj}^{-1}\big(\sum_{\substack{r:\\ (r,j)\in E_s^c}} x_m^{(h_1)}[\theta_{jr}]_{h_1} y_{mr}\big)\big(\sum_{\substack{k:\\ (k,j)\in E_s}} x_m^{(h_1)} y_{mk}\big), \quad \mathcal{I}_{4.2} = \alpha_{jj}^{-1}\big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n} \alpha_{ji} y_{mi}\big)\big(\sum_{\substack{k:\\ (k,j)\in E_s}} x_m^{(h_1)} y_{mk}\big),$$

$$\mathcal{I}_{4.3} = \alpha_{jj}^{-1}\big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n}\sum_{\substack{h=1,\\ h\neq h_1}}^{H} x_m^{(h)}[\theta_{ji}]_h y_{mi}\big)\big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\big).$$

Through appropriate matrix transformations, the updating equations (3.11) (3.12) can be simplified as

$$\hat{\alpha}_{E_s} = \frac{S\Big(-tr\Big(\boldsymbol{T}_{E_s}\boldsymbol{C}_\alpha\Big) + tr\Big(\boldsymbol{T}_{E_s}\big(\boldsymbol{T}_{E_s^c} \odot \boldsymbol{B}_\alpha^T\big)\boldsymbol{C}_\alpha\Big) + \sum_{h=1}^{H} tr\Big(\boldsymbol{T}_{E_s}\big(\boldsymbol{T}_{E_s}^1 \odot \big[\boldsymbol{B}_\theta^{(h)}\big]^T\big)\boldsymbol{C}_\theta^{(h)}\Big), \lambda\Big)}{tr\Big(\boldsymbol{T}_{E_s}\boldsymbol{\Sigma}\boldsymbol{T}_{E_s}\boldsymbol{C}_\alpha\Big)}, \tag{3.13}$$

$$[\hat{\theta}_{E_s}]_{h_1} = \frac{S\Big(-tr\Big(\boldsymbol{T}_{E_s}\boldsymbol{C}_\theta^{(h_1)}\Big) + tr\Big(\boldsymbol{T}_{E_s}\big(\boldsymbol{T}_{E_s^c} \odot \big[\boldsymbol{B}_\theta^{(h_1)}\big]^T\big)\boldsymbol{D}_\theta^{(h_1)}\Big) + tr\Big(\boldsymbol{T}_{E_s}\big(\boldsymbol{T}_{E_s}^1 \odot \boldsymbol{B}_\alpha^T\big)\boldsymbol{C}_\theta^{(h_1)}\Big) + \Delta, \lambda\Big)}{tr\Big(\boldsymbol{T}_{E_s}\boldsymbol{\Sigma}\boldsymbol{T}_{E_s}\boldsymbol{D}_\theta^{(h_1)}\Big)}, \tag{3.14}$$

where E_s^c denotes the complement of E_s, and the matrix notations in (3.13), (3.14) are given as

$$\boldsymbol{C}_\theta^{(h_1)} = \frac{1}{n}\Big(\boldsymbol{X}^{(h_1)} \odot \boldsymbol{Y}\Big)^T \boldsymbol{Y}, \qquad \boldsymbol{D}_\theta^{(h_1)} = \frac{1}{n}\Big(\boldsymbol{X}^{(h_1)} \odot \boldsymbol{Y}\Big)^T \Big(\boldsymbol{X}^{(h_1)} \odot \boldsymbol{Y}\Big),$$

$$\boldsymbol{G}_\theta^{(h_1,l)} = \frac{1}{n}\Big(\boldsymbol{X}^{(l)} \odot \boldsymbol{Y}\Big)^T \Big(\boldsymbol{X}^{(h_1)} \odot \boldsymbol{Y}\Big), \qquad \boldsymbol{C}_\alpha = \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Y},$$

$$\Delta = \sum_{l=1, l\neq h_1}^{H} tr\Big(\boldsymbol{T}_{E_s}\big(\boldsymbol{T}_{E_s}^1 \odot \big[\boldsymbol{B}_\theta^{(l)}\big]^T\big)\boldsymbol{G}_\theta^{(h_1,l)}\Big).$$

The $n \times p_n$ matrix $\boldsymbol{X}^{(h_1)}$ is defined as

$$\boldsymbol{X}^{(h_1)} = \begin{bmatrix} x_1^{(h_1)} & x_1^{(h_1)} & \dots & x_1^{(h_1)} \\ x_2^{(h_1)} & x_2^{(h_1)} & \dots & x_2^{(h_1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(h_1)} & x_n^{(h_1)} & \cdots & x_n^{(h_1)} \end{bmatrix}_{n \times p_n}. \tag{3.15}$$

Let $\boldsymbol{\Sigma}$ denote a $p_n \times p_n$ diagonal matrix $\text{diag}(\alpha_{11}^{-1}, ..., \alpha_{p_n p_n}^{-1})$, and let $\odot$ denote the component-wise product. Each entry of $\boldsymbol{B}_\theta^{(h)}$ and $\boldsymbol{B}_\alpha$ takes the following form

$$\big(\boldsymbol{B}_\theta^{(h)}\big)_{ij} = \begin{cases} -[\theta_{ij}]_h \alpha_{jj}^{-1} & i \neq j; \quad i,j = 1,...,p_n, \\ 0 & i = j, \end{cases} \tag{3.16}$$

$$\big(\boldsymbol{B}_\alpha\big)_{ij} = \begin{cases} -\alpha_{ij} \alpha_{jj}^{-1} & i \neq j; \quad i,j = 1,...,p_n, \\ 0 & i = j, \end{cases} \tag{3.17}$$

Let $\boldsymbol{T}_{E_s}$ denote the edge adjacency matrix, e.g. $(\boldsymbol{T}_{E_s})_{ij} = 1$ if $(i,j) \in E_s$ or $(j,i) \in E_s$, and $(\boldsymbol{T}_{E_s})_{ij} = 0$ otherwise. Define $\boldsymbol{T}_{E_s}^1$ as: if $(i,j) \in E_s$, all of the elements in the ith and jth row of the matrix $\boldsymbol{T}_{E_s}^1$ are equal to 1 except the diagonal entries, and all other entries are equal to zero.

The matrix-form solutions presented above substantially improve computational efficiency, enabling fast and straightforward estimation in large-scale network settings. Furthermore, since the vertex parameters have closed-form solutions in this scenario, a coordinate descent algorithm can be employed. Specifically, for the diagonal parameter α_{V_r}, the partial derivative is given as follows

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \alpha_{V_r}} = -\frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r}\alpha_{jj}^{-2}\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Big] y_{mi}\Big)^2 - \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r}\alpha_{jj}^{-1} + \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r} y_{mj}^2. \tag{3.18}$$

The solution to $\frac{\partial Q(\boldsymbol{\beta})}{\partial \alpha_{V_r}} = 0$ is

$$\hat{\alpha}_{V_r} = \frac{\frac{2}{n}\sum_{m=1}^{n}\sum_{j\in V_r}\left(\sum_{i=1,i\neq j}^{p_n}\left[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\right]y_{mi}\right)^2}{-|V_r| + \sqrt{|V_r|^2 + \frac{4}{n^2}\left(\sum_{m=1}^{n}\sum_{j\in V_r}y_{mj}^2\right)\left(\sum_{m=1}^{n}\sum_{j\in V_r}\left(\sum_{\substack{i=1,\\ i\neq j}}^{p_n}\left[\alpha_{ji}+\sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\right]y_{mi}\right)^2\right)}},$$

where $|V_r|$ is the cardinality of set V_r.

3.2.3 Theoretical Properties of Penalized Composite Likelihood Estimator

We now analyze the estimation error and model selection properties of the penalized composite likelihood estimator under regularity assumptions. Define s_n as the largest number of nonzero parameters in any row across all matrices $\boldsymbol{Q}_0, \boldsymbol{P}_1, \cdots, \boldsymbol{P}_H$. We denote the score function and the Hessian matrix of $l_c(\boldsymbol{\beta})$ as $l_c^{(1)}(\boldsymbol{\beta}), l_c^{(2)}(\boldsymbol{\beta})$ respectively. Further define $\mathcal{S}_1$ as the set of all true nonzero edge parameters in $\boldsymbol{Q}_0, \boldsymbol{P}_1, \cdots, \boldsymbol{P}_H$, namely $\mathcal{S}_1 = \Big\{\text{all } \alpha_{E_s}^0, [\theta_{E_s}^0]_h \text{ such that: } \alpha_{E_s}^0 \neq 0, [\theta_{E_s}^0]_h \neq 0\Big\}$, and denote u_e as the maximum number of edges in a single edge color class. Similarly, denote $\mathcal{S}_2$ as the set of diagonal parameters. Then we let $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \cup \mathcal{S}_2$ represent the set of true nonzero parameters. We use its complement $\boldsymbol{\mathcal{S}}^c$ representing the set of true zero parameters. Let $|\mathcal{S}_1|$,$|\boldsymbol{\mathcal{S}}^c|$ denote the cardinality of sets $\mathcal{S}_1, \boldsymbol{\mathcal{S}}^c$. Therefore, the total number of true nonzero off-diagonal parameters is $|\mathcal{S}_1|$, which is finite.

Assumption 6 The expected Hessian matrix $E\Big\{-l_c^{(2)}(Y_{mj}, \boldsymbol{\beta})\Big\}$ is assumed to have eigenvalues bounded away from zero and infinity within a neighborhood of the true parameter $\boldsymbol{\beta}^{\mathbf{0}}$: $||\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{0}}||_2 < \eta$ for some constant $\eta > 0$. Furthermore, the spectral norms of $E\left\{\frac{\partial^3 - l_c(Y_{mj}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T \partial \beta_u}\right\}$ are bounded by some constant, where β_u denotes any off-diagonal parameter.

Assumption 7 Assume $s_n^4 = \mathcal{O}(p_n), s_n^2 u_e^2 = \mathcal{O}(p_n)$, and $p_n \log p_n = o(n)$.

Theorem 3.1 Under **Assumptions** 6-7, for any constants $\delta_1, \delta_2 > 0$, let $\delta_1 \left(\frac{\log p_n}{n}\right)^{\frac{1}{2}} \leqslant \lambda \leqslant \delta_2 \left(\frac{s_n^4 \log p_n}{n}\right)^{\frac{1}{2}}$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of the objective function $Q(\boldsymbol{\beta})$ such that

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}||_2 = \mathcal{O}_p \left\{ \left(\frac{p_n \log p_n}{n} \right)^{\frac{1}{2}} \right\}.$$

Assumption 8 Let $H^0_{\mathcal{S}_1 \mathcal{S}_1} = E\left\{-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1 \mathcal{S}_1}\right\}$ and $H^0_{\boldsymbol{\mathcal{S}}^c \mathcal{S}_1} = E\left\{-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\boldsymbol{\mathcal{S}}^c \mathcal{S}_1}\right\}$. There exists some positive constant $\xi \in (0, 1)$ such that

$$\left|\left|\left| H^0_{\boldsymbol{\mathcal{S}}^c \mathcal{S}_1} \left(H^0_{\mathcal{S}_1 \mathcal{S}_1} \right)^{-1} \right|\right|\right|_\infty \leqslant 1 - \xi,$$

where $|||\cdot|||_\infty$ is the maximum absolute row sum of a matrix.

Under the Gaussian graphical model framework, **Assumption** 8 requires that the expected Hessian matrix meets the incoherence condition, indicating that variables without direct edges impose minor influence on connected variable pairs, which has become a standard assumption in high-dimensional statistical theory, especially in the study of graphical models and penalized regression frameworks (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Raskutti et al., 2008; Ravikumar et al., 2011). **Theorem** 3.1 demonstrates consistency for both vertex and edge parameters. We now turn to the model selection problem, concentrating solely on the edge parameters while treating the vertex parameters as known nuisance terms.

Theorem 3.2 Under **Assumptions** 6-8 and **Theorem** 3.1, let $\lambda = \delta_2 \left(\frac{s_n^4 \log p_n}{n} \right)^{\frac{1}{2}}$ for any constant δ_2 defined in **Theorem** 3.1, suppose that $u_e^6 \log u_e = o(n\phi^2)$, $p_n^2 \log p_n = \mathcal{O}(n)$ and the minimum non-zero parameter satisfies

$$\min_{u \in \mathcal{S}_1} \left| \beta_u^0 \right| := \min_{\substack{(i,j) \in \mathcal{S}_1 \\ h=1,\cdots,H}} \left\{ \left| [\theta_{ij}^0]_h \right|, \left| \alpha_{ij}^0 \right| \right\} \geqslant c^* \sqrt{|\mathcal{S}_1|} \lambda$$

for some constant $c^* > 0$. Then the penalized composite likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^{\mathbf{0}})$ with probability tending to one.

Note that the sign consistency applies asymptotically, since parameter consistency holds only in the asymptotic setting.

3.3 Numerical Simulations

The study begins with fixed diagonal elements of the precision matrix as they are more computationally efficient and easier to implement than those involving covariate-dependent diagonal parameters. The complete simulation study consists of both penalized and non-penalized cases.

3.3.1 Fixed Diagonal Entries: Without Penalization

We evaluate the performance of our proposed composite likelihood estimation on large-scale networks. Our simulations explore scenarios where n varies between 2000 and 3000 and p_n takes values of $60, 80$, and 100. The analysis includes 30 edge color classes and 20 vertex color classes. We consider two covariates, each randomly drawn from a uniform distribution between 0 and 1. Therefore, the underlying precision matrix takes the form $\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_m) = \boldsymbol{Q}_0 + x_m^{(1)}\boldsymbol{P}_1 + x_m^{(2)}\boldsymbol{P}_2$. We generate the baseline matrix $\boldsymbol{Q}_0$ with off-diagonal entries uniformly sampled from -0.6 to 0.6, and the diagonal entries drawn uniformly from 2 to 3 or 3 to 4. The off-digaonal entries of slope matrices $\boldsymbol{P}_1, \boldsymbol{P}_2$ are uniformly drawn from -0.3 to 0.3. To assess model performance, we use the average of the root mean squared error and the relative error. The average of root mean squared error is computed as the mean of ℓ_2 norm, while the mean relative error is calculated as the average of $\left|([\hat{\theta}_{E_s}]_h - [\theta^0_{E_s}]_h)/[\theta^0_{E_s}]_h\right|$ for $\boldsymbol{P}_1, \boldsymbol{P}_2$ and $\left|(\hat{\alpha}_{E_s} - \alpha^0_{E_s})/\alpha^0_{E_s}\right|$ for $\boldsymbol{Q}_0$. Note that the mean relative error is only computed over the nonzero subset of parameters. The simulation results over 100 simulated data sets are given in Table 3.2, 3.3, and 3.4. We observe that the average root mean squared error for $\boldsymbol{Q}_0$ appears to be slightly higher than that for the slope matrices $\boldsymbol{P}_1, \boldsymbol{P}_2$ across all scenarios. This could be due to $\boldsymbol{Q}_0$ having a magnitude twice that of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. Regarding the mean relative error, we find that all errors are below 0.1 with minimal standard errors. These findings demonstrate the effectiveness of the proposed estimation method in high-dimensional networks.

3.3.2 Fixed Diagonal Entries: With Penalization

This section presents penalization and investigates the performance of the penalized composite likelihood estimator in the high-dimensional context. We simulate the covariate-dependent RCON models using the same setup as introduced before. Specifically, we conduct simulations across different settings, where n ranges from 2000 to 3500 and $p_n = 60, 80, 100$. The study accounts for 30 edge color classes and 20 vertex color classes, with two covariates randomly sampled from a uniform distribution in the range [0,1]. We use an ℓ_1 penalization to achieve sparse network structures. The composite likelihood Bayesian information criterion is used to determine the tuning parameter, defined as $BIC_{comp} = -2l_c(\hat{\boldsymbol{\beta}}_c) + df \log n$, where df denotes the count of nonzero parameters and $\hat{\boldsymbol{\beta}}_c$ denotes the penalized composite likelihood estimator. Model evaluation is conducted using sensitivity, specificity, F1 score, and Matthews Correlation Coefficient (MCC), which are defined as

$$sensitivity = \frac{TP}{TP + FN}, \quad specificity = \frac{TN}{TN + FP}, \quad F_1 = \frac{2 \times TP}{2 \times TP + FP + FN},$$
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}},$$

where TP, TF, FP, and FN represent the count of true positives, true negatives, false positives, and false negatives, respectively. The results over 100 simulated data sets are reported in Table 3.5, 3.6, and 3.7. We find that the average sensitivity and specificity exceed 0.9 across all simulation settings, indicating that the proposed method effectively minimizes false negatives and false positives on average. Consequently, the proposed method achieves high F1 scores and MCC, highlighting its strong performance in recovering colored sparse network structures.

3.3.3 Covariate-dependent Diagonal Entries

Scenarios with covariate-dependent diagonal components are also included in this section, under both penalized and non-penalized settings. Because of the intensive computational cost, we limit our study to the case where $p_n = 30$ and $n = 2000$. Similar to the previous section, we consider two covariates and the graph structure is characterized by $\boldsymbol{\Sigma}_m^{-1} = \boldsymbol{Q}_0 + x_m^{(1)}\boldsymbol{P}_1 + x_m^{(2)}\boldsymbol{P}_2$. We simulate the matrices $\boldsymbol{Q}_0, \boldsymbol{P}_1, \boldsymbol{P}_2$ with 15 edge color classes and 5 vertex color classes. We uniformly draw diagonal elements of $\boldsymbol{Q}_0$ from 1.5 to 2.5 and those of $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ from 1 to 1.5, with all off-diagonal entries sampled uniformly from -0.4 to 0.4. The average root mean squared error and relative error are reported for the non-penalized setting, while sensitivity, specificity, F1 score, and MCC are presented for the penalized case in Table 3.8 and

Table 3.2: Simulation results for covariate-dependent RCON models without penalization (Fixed diagonal entries)

n	p_n	Matrix	Parameter	mrmse	mre
2000	60	$\boldsymbol{Q}_0$	edge	0.1351 (0.0174)	0.0878 (0.0443)
			vertex	0.8046 (0.1434)	0.0575 (0.0104)
		$\boldsymbol{P}_1$	edge	0.0341 (0.0056)	0.0611 (0.0248)
		$\boldsymbol{P}_2$	edge	0.0281 (0.0056)	0.0375 (0.0167)
2000	80	$\boldsymbol{Q}_0$	edge	0.1113 (0.0149)	0.0915 (0.0422)
			vertex	0.7221 (0.1088)	0.0510 (0.0080)
		$\boldsymbol{P}_1$	edge	0.0203 (0.0037)	0.0387 (0.0147)
		$\boldsymbol{P}_2$	edge	0.0184 (0.0057)	0.0247 (0.0107)
2000	100	$\boldsymbol{Q}_0$	edge	0.1263 (0.0161)	0.0830 (0.0302)
			vertex	0.8946 (0.1348)	0.0454 (0.0072)
		$\boldsymbol{P}_1$	edge	0.0156 (0.0028)	0.0277 (0.0112)
		$\boldsymbol{P}_2$	edge	0.0161 (0.0062)	0.0231 (0.0120)

The "mrmse" stands for the mean of root mean squared error, and "mre" stands for mean relative error. The standard errors are listed in parentheses.

Table 3.3: Simulation results for covariate-dependent RCON models without penalization (Fixed diagonal entries)

n	p_n	Matrix	Parameter	mrmse	mre
2500	60	$\boldsymbol{Q}_0$	edge	0.1207 (0.0161)	0.0790 (0.0376)
			vertex	0.7092 (0.1196)	0.0505 (0.0091)
		$\boldsymbol{P}_1$	edge	0.0310 (0.0049)	0.0576 (0.0227)
		$\boldsymbol{P}_2$	edge	0.0252 (0.0047)	0.0347 (0.0144)
2500	80	$\boldsymbol{Q}_0$	edge	0.0988 (0.0130)	0.0791 (0.0333)
			vertex	0.6312 (0.0959)	0.0448 (0.0071)
		$\boldsymbol{P}_1$	edge	0.0179 (0.0033)	0.0336 (0.0137)
		$\boldsymbol{P}_2$	edge	0.0161 (0.0049)	0.0220 (0.0100)
2500	100	$\boldsymbol{Q}_0$	edge	0.1118 (0.0144)	0.0733 (0.0264)
			vertex	0.7923 (0.1255)	0.0404 (0.0066)
		$\boldsymbol{P}_1$	edge	0.0139 (0.0025)	0.0242 (0.0088)
		$\boldsymbol{P}_2$	edge	0.0138 (0.0045)	0.0198 (0.0108)

The "mrmse" stands for the mean of root mean squared error, and "mre" stands for mean relative error. The standard errors are listed in parentheses.

Table 3.4: Simulation results for covariate-dependent RCON models without penalization (Fixed diagonal entries)

n	p_n	Matrix	Parameter	mrmse	mre
3000	60	$\boldsymbol{Q}_0$	edge	0.1103 (0.0152)	0.0723 (0.0357)
			vertex	0.6442 (0.1016)	0.0459 (0.0077)
		$\boldsymbol{P}_1$	edge	0.0282 (0.0045)	0.0522 (0.0207)
		$\boldsymbol{P}_2$	edge	0.0230 (0.0041)	0.0313 (0.0136)
3000	80	$\boldsymbol{Q}_0$	edge	0.0902 (0.0110)	0.0723 (0.0300)
			vertex	0.5764 (0.0941)	0.0408 (0.0065)
		$\boldsymbol{P}_1$	edge	0.0165 (0.0029)	0.0308 (0.0133)
		$\boldsymbol{P}_2$	edge	0.0147 (0.0046)	0.0201 (0.0090)
3000	100	$\boldsymbol{Q}_0$	edge	0.1016 (0.0131)	0.0661 (0.0267)
			vertex	0.7186 (0.1189)	0.0366 (0.0062)
		$\boldsymbol{P}_1$	edge	0.0126 (0.0024)	0.0224 (0.0089)
		$\boldsymbol{P}_2$	edge	0.0127 (0.0044)	0.0180 (0.0092)

The "mrmse" stands for the mean of root mean squared error, and "mre" stands for mean relative error. The standard errors are listed in parentheses.

Table 3.5: Simulation results for covariate-dependent RCON models with penalization (Fixed diagonal entries)

n	p_n	Matrix	Sensitivity	Specificity	F1 Score	MCC
2000	60	$\boldsymbol{Q}_0$	0.9738	0.9894	0.9623	0.9576
			(0.0749)	(0.0241)	(0.0616)	(0.0684)
		$\boldsymbol{P}_1$	0.9882	0.9760	0.9461	0.9390
			(0.0469)	(0.0461)	(0.0894)	(0.1006)
		$\boldsymbol{P}_2$	0.9939	0.9645	0.9256	0.9149
			(0.0346)	(0.0500)	(0.0917)	(0.1050)
2000	80	$\boldsymbol{Q}_0$	0.9350	0.9737	0.9098	0.8972
			(0.0934)	(0.0405)	(0.0826)	(0.0950)
		$\boldsymbol{P}_1$	0.9981	0.9600	0.9250	0.9158
			(0.0189)	(0.0717)	(0.1131)	(0.1254)
		$\boldsymbol{P}_2$	1.0000	0.9664	0.9325	0.9238
			(0.0000)	(0.0503)	(0.0944)	(0.1051)
2500	60	$\boldsymbol{Q}_0$	0.9918	0.9911	0.9762	0.9728
			(0.0403)	(0.0221)	(0.0497)	(0.0564)
		$\boldsymbol{P}_1$	0.9763	0.9880	0.9626	0.9579
			(0.0648)	(0.0311)	(0.0667)	(0.0751)
		$\boldsymbol{P}_2$	0.9960	0.9790	0.9544	0.9479
			(0.0280)	(0.0364)	(0.0709)	(0.0804)

The standard errors are listed in parentheses.

Table 3.6: Simulation results for covariate-dependent RCON models with penalization (Fixed diagonal entries)

n	p_n	Matrix	Sensitivity	Specificity	F1 Score	MCC
2500	80	$\boldsymbol{Q}_0$	0.9812	0.9793	0.9464	0.9392
			(0.0602)	(0.0373)	(0.0753)	(0.0847)
		$\boldsymbol{P}_1$	0.9982	0.9726	0.9451	0.9381
			(0.0184)	(0.0510)	(0.0921)	(0.1026)
		$\boldsymbol{P}_2$	1.0000	0.9753	0.9478	0.9406
			(0.0000)	(0.0377)	(0.0749)	(0.0844)
2500	100	$\boldsymbol{Q}_0$	0.9012	0.9969	0.9369	0.9306
			(0.1176)	(0.0124)	(0.0709)	(0.0768)
		$\boldsymbol{P}_1$	0.9902	0.9865	0.9652	0.9598
			(0.0432)	(0.0254)	(0.0630)	(0.0732)
		$\boldsymbol{P}_2$	0.9960	0.9811	0.9583	0.9525
			(0.0280)	(0.0351)	(0.0694)	(0.0784)
3000	60	$\boldsymbol{Q}_0$	0.9977	0.9948	0.9871	0.9852
			(0.0227)	(0.0154)	(0.0359)	(0.0411)
		$\boldsymbol{P}_1$	0.9786	0.9923	0.9715	0.9679
			(0.0614)	(0.0222)	(0.0539)	(0.0607)
		$\boldsymbol{P}_2$	0.9979	0.9863	0.9692	0.9648
			(0.0207)	(0.0265)	(0.0572)	(0.0650)

The standard errors are listed in parentheses.

Table 3.7: Simulation results for covariate-dependent RCON models with penalization (Fixed diagonal entries)

n	p_n	Matrix	Sensitivity	Specificity	F1 Score	MCC
3000	80	$\boldsymbol{Q}_0$	0.9935	0.9813	0.9559	0.9497
			(0.0370)	(0.0302)	(0.0639)	(0.0724)
		$\boldsymbol{P}_1$	0.9963	0.9772	0.9522	0.9461
			(0.0262)	(0.0444)	(0.0841)	(0.0940)
		$\boldsymbol{P}_2$	1.0000	0.9880	0.9736	0.9696
			(0.0000)	(0.0236)	(0.0512)	(0.0586)
3000	100	$\boldsymbol{Q}_0$	0.9604	0.9956	0.9683	0.9644
			(0.0798)	(0.0139)	(0.0496)	(0.0556)
		$\boldsymbol{P}_1$	0.9925	0.9908	0.9755	0.9720
			(0.0368)	(0.0205)	(0.0473)	(0.0540)
		$\boldsymbol{P}_2$	0.9980	0.9900	0.9776	0.9744
			(0.0196)	(0.0261)	(0.0533)	(0.0603)
3500	100	$\boldsymbol{Q}_0$	0.9900	0.9920	0.9771	0.9739
			(0.0437)	(0.0215)	(0.0493)	(0.0559)
		$\boldsymbol{P}_1$	0.9923	0.9936	0.9817	0.9791
			(0.0380)	(0.0189)	(0.0434)	(0.0493)
		$\boldsymbol{P}_2$	0.9980	0.9959	0.9900	0.9886
			(0.0196)	(0.0161)	(0.0355)	(0.0404)

The standard errors are listed in parentheses.

Table 3.8: Simulation results for covariate-dependent RCON models without penalization (Covariate-dependent diagonal entries)

n	p_n	Matrix	Parameter	mrmse	mre
2000	30	$\boldsymbol{Q}_0$	edge	0.0822 (0.0174)	0.0697 (0.0257)
			vertex	0.4146 (0.1590)	0.0329 (0.0135)
		$\boldsymbol{P}_1$	edge	0.0728 (0.0152)	0.0771 (0.0278)
			vertex	0.4067 (0.1151)	0.0505 (0.0158)
		$\boldsymbol{P}_2$	edge	0.0574 (0.0131)	0.0593 (0.0242)
			vertex	0.3118 (0.1283)	0.0370 (0.0167)

The standard errors are listed in parentheses.

3.9. We observe that all mean relative errors are under 0.1 with minimal variation. Moreover, the model also exhibits strong performance in recovering sparsity patterns, as indicated by high sensitivity, specificity, F1, and MCC scores. These findings highlight its effectiveness and practical applicability in handling varying diagonal entries of $\boldsymbol{\Sigma}_m^{-1}$.

Finally, we investigate the computational cost for the simulations. The results,

Table 3.9: Simulation results for covariate-dependent RCON models with penalization (Covariate-dependent diagonal entries)

n	p_n	Matrix	Sensitivity	Specificity	F1 Score	MCC
2000	30	$\boldsymbol{Q}_0$	1.0000	1.0000	1.0000	1.0000
			(0.0000)	(0.0000)	(0.0000)	(0.0000)
		$\boldsymbol{P}_1$	0.8911	0.9907	0.9267	0.9084
			(0.1561)	(0.0408)	(0.1008)	(0.1283)
		$\boldsymbol{P}_2$	0.9086	0.9836	0.9331	0.9107
			(0.1089)	(0.0411)	(0.0683)	(0.0879)

The standard errors are listed in parentheses.

summarized in Table 3.10, show the computational expense per seed under various settings. As the number of color classes is fixed, the runtime does not rise sharply with dimension. Nevertheless, the presence of covariate-dependent diagonal parameters significantly increases the computational burden.

3.4 Data Analysis

Among current public health threats, influenza remains prominent, with influenza A (H1N1, H3N2) and influenza B being the most widespread strains. Host genetic variation has been shown to strongly influence both the immune response and sus-

Table 3.10: Computational cost of covariate-dependent colored RCON models

$(\boldsymbol{\Sigma}_m^{-1})_{jj}$	Penalty	(p_n, n)	Computational time
Fixed	No	(60, 2000)	0.68 mins
		(80, 3000)	0.88 mins
		(100, 3000)	1.62 mins
	Yes	(60, 2000)	6.20 mins
		(80, 3000)	9.84 mins
		(100, 3000)	20.38 mins
Covariate-dependent	No	(30, 2000)	21.39 mins
	Yes		27.50 mins

ceptibility to influenza infection (Trammell and Toth, 2008; Srivastava et al., 2009). Moreover, genes that are involved in similar biological processes or pathways often show similar expression patterns, especially in response to stimuli like vaccination. Clustering techniques such as k-means, hierarchical, or spectral methods are widely used in genomics to group genes with similar expression profiles, potentially uncovering co-regulated or functionally related gene sets (Oyelade et al., 2016). Driven by these findings, we implement our proposed covariate-dependent RCON models on a global gene expression data set collected before and after the injection of trivalent influenza vaccine in humans, to uncover dynamic regulatory gene networks and assess how dependency structures evolve across time points. This study may shed light on critical genes with significant regulatory influence on the host immune response to the influenza vaccine.

The data set was obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) under accession number GSE48024 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48024`). The study involved 119 healthy adult male and 128 female volunteers, all aged 19–41, resulting in a total of 848 observations. To assess global transcript abundance, peripheral blood RNA was sampled at baseline (day 0) and at three post-vaccination time points: days 1, 3, and 14. The raw data were pre-processed using background correction, variance-stabilizing transformation (Lin et al., 2008), and robust spline normalization, implemented via the R package lumi (Du et al., 2008).

Following the work of Franco et al. (2013), we restrict our analysis to 68 genes that demonstrate both transcriptional activity in response to the vaccine and evidence of regulatory genetic effects. We use spectral clustering to organize the genes into groups (Ng et al., 2002; Qin and Rohe, 2013; Hennig et al., 2015), determining the optimal number of clusters by examining the largest gap between successive eigenvalues of the Laplacian matrix. As the first five eigenvalues are significantly larger than the rest, we determine the optimal number of groups to be five. Genes grouped within the same cluster tend to exhibit similar expression patterns. The spectral clustering procedure is conducted through the Spectrum package in R (John et al., 2020). The sparsity patterns and time-varying regulatory networks across distinct color classes are given in Figure 3.1, 3.2, and 3.3, respectively. Moreover, the full network structures are also provided on the left panel of Figure 3.2.

Due to the complexity of the full graph, we restrict our focus to the underlying network structures of specific color classes. Among the five classes, some remain stable throughout the two-week period, while others demonstrate an increasing trend or emerge in response to vaccine exposure. The right panel of Figure 3.2 presents one of the two color classes that remain consistent over time, indicating that its internal interactions are likely unaffected by vaccine administration. The left-hand side of Figure 3.3 shows the third color class, which exhibits a modest increase following vaccine injection, ultimately reaching twice its original edge interaction intensity by the end of the two weeks. In contrast, the fifth color class is not detected until the vaccine administration, and it demonstrates a steady increase over the observation period, as depicted in the right panel of Figure 3.3. Our analysis reveals compelling evidence of the essential role played by several gene groups in the immune response to vaccination. To gain a deeper understanding of how these genes interact over time, and to better understand their biological relevance, more in-depth biological

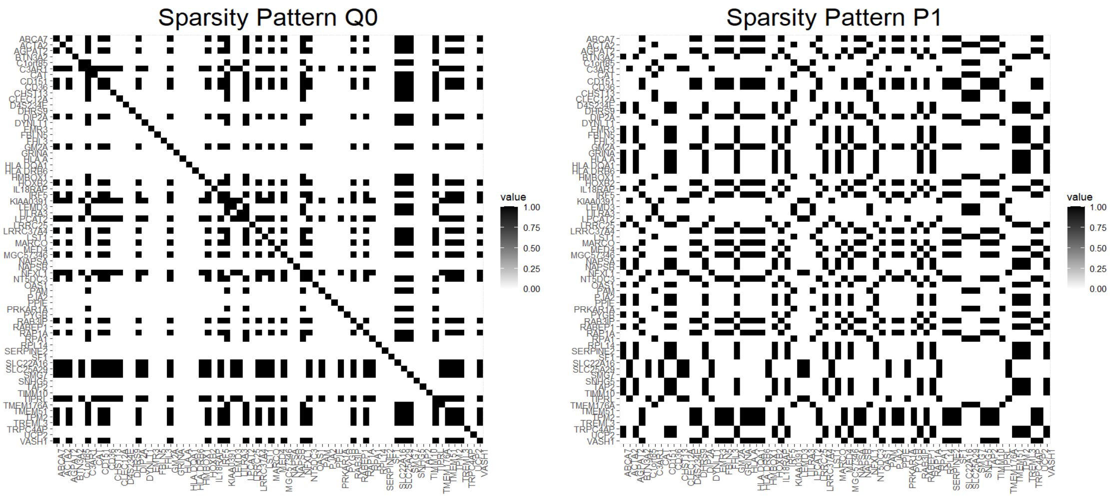


Figure 3.1: The estimated sparsity pattern for the influenza vaccine data set. Each image has 68 rows and 68 columns, representing the elements of $\boldsymbol{Q}_0, \boldsymbol{P}_1$ respectively. The corresponding cell is marked in black if the estimated entry is nonzero. Otherwise, it is white.

research is required.

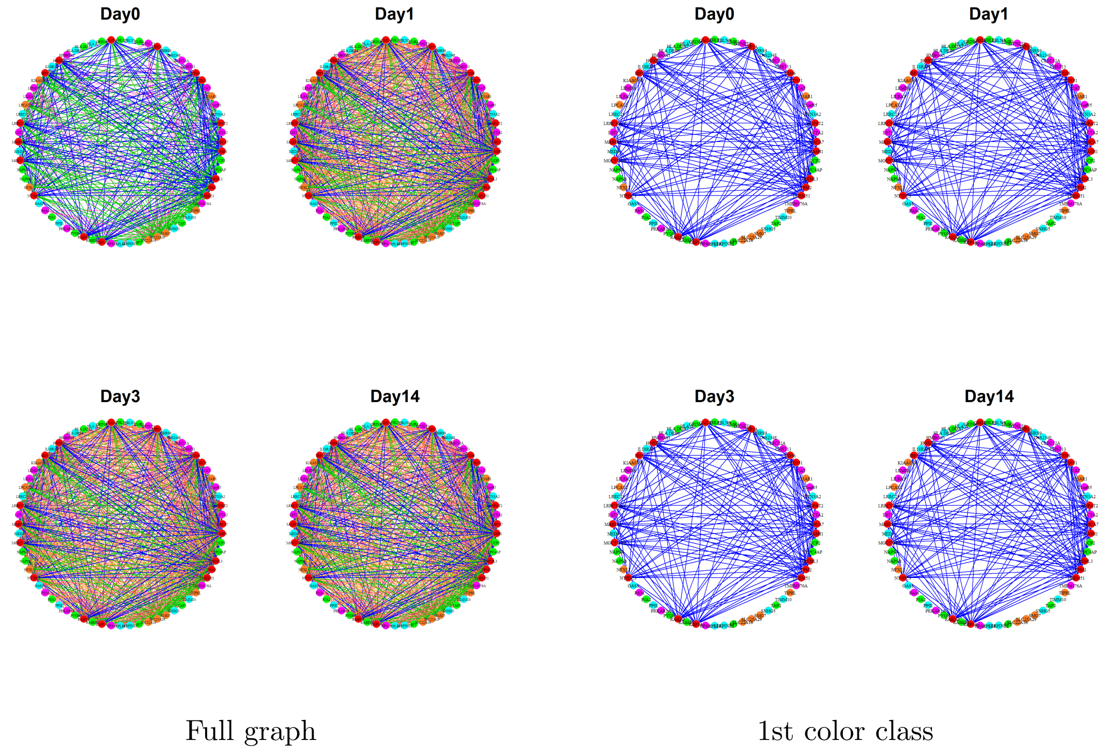
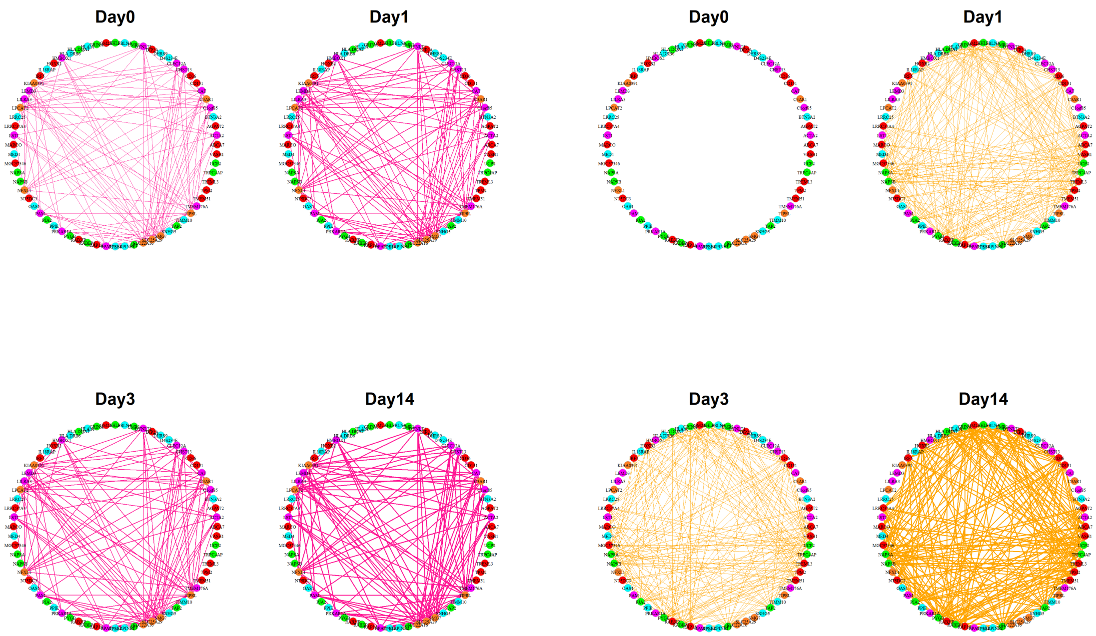


Figure 3.2: The estimated gene regulatory networks for the full graph and the first edge color class, respectively. The time points "Day0", "Day1", "Day3", and "Day14" represent the day before getting vaccinated, the first day after vaccination, the third day after vaccination, and two weeks after vaccination. The thickness of the edges displays the strength of interactions between two genes.



3rd color class 5th color class

Figure 3.3: The estimated gene regulatory networks for the third and fifth edge color classes. The time points "Day0", "Day1", "Day3", and "Day14" represent the day before getting vaccinated, the first day after vaccination, the third day after vaccination, and two weeks after vaccination. The thickness of the edges displays the strength of interactions between two genes.

3.5 Technical Lemmas and Proofs

We present the technical lemmas and their proofs in this section, starting with a few preliminary notations.

Notation 3.1: Consider the summation $\sum_{j=1}^{p_n}$ in equations

$$\sum_{j=1}^{p_n}\sum_{m=1}^{n}\alpha_{jj}^{-1}\big(\sum_{i=1,i\neq j}^{p_n}\big[\alpha_{ji}+\sum_{h=1}^{H}x_m^h[\theta_{ji}]_h\big]y_{mi}\big)\big(\sum_{k:(k,j)\in E_s}x_m^{(h_1)}y_{mk}\big). \tag{3.19}$$

While j is selected from 1 to p_n, it also needs to fulfill the condition $\{j : (k,j) \in E_s \text{ for some } k\}$. As a result, the total number of terms in the summation $\sum_{j=1}^{p_n}$ is less than the maximum number of unique row/column indices that appear in the pair $(k,j) \in E_s$, denoted as q_e, where $q_e \leqslant p_n$. That is, for example,

$$\#\text{ of defined terms in } \sum_{j=1}^{p_n}\alpha_{jj}^{-1}\Big(\sum_{\substack{i=1,\\ i\neq j}}^{p_n}\alpha_{ji}y_{mi}\Big)\Big(\sum_{\substack{k:\\ (k,j)\in E_s}}x_m^{(h_1)}y_{mk}\Big) \leqslant q_e s_n u_e.$$

By the definition, $q_e \leqslant u_e$ holds, and under most circumstances, q_e tends to be much smaller than u_e.

Notation 3.2: Consider the summation $\sum_{k:(k,j)\in E_s}$ in equation (3.19). For any fixed j and fixed $E_s \in \boldsymbol{\mathcal{E}}$, denote the maximum number of indices k satisfying the condition $\{k : (k,j) \in E_s\}$ as ψ. That is, the number of defined terms in $\sum_{k:(k,j)\in E_s}$ is bounded above by the maximum number of row index k associated with the pair $(k,j) \in E_s$. Normally, ψ is a small number.

The following lemmas are established under the assumption of fixed diagonal parameters. Nevertheless, the results remain valid in the case of varying diagonal components, by replacing the conditional variance α_{jj}^{-1} with $\big[\alpha_{jj} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{jj}]_h\big]^{-1}$.

Lemma 3.1 To simplify the notation, we write $-\sum_{m=1}^{n} l_c(Y_{mj}, \boldsymbol{\beta}^{\mathbf{0}}) = -l_c(\mathbf{Y}_j, \boldsymbol{\beta}^0)$, where j denotes a fixed index in $\{1, \cdots, p_n\}$. For $\forall\ E_s \in \boldsymbol{\mathcal{E}}$, $\forall\ V_r \in \boldsymbol{\mathcal{V}}$, and $\forall\ h_1 = 1, \cdots, H$, we have

$$\max_{E_s, h_1} \left| \frac{\partial - l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial [\theta_{E_s}]_{h_1}} \right| = \mathcal{O}_p \left\{ (n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}} \right\}, \quad \max_{E_s} \left| \frac{\partial - l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial \alpha_{E_s}} \right| = \mathcal{O}_p \left\{ (n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}} \right\},$$
$$\max_{V_r} \left| \frac{\partial - l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial \alpha_{V_r}} \right| = \mathcal{O}_p \left\{ (n s_n^4 \log p_n)^{\frac{1}{2}} \right\}.$$

Furthermore, for $\forall\ \epsilon > 0$, there exist some positive constants $\mathcal{K}_1, C_1, c_{\mathcal{I}_1,max}$ and $c_{\mathcal{I}_2,max}$ such that

$$\begin{aligned} P\left\{ \left\| \frac{1}{n\phi} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} \right\|_\infty \geqslant \epsilon \right\} &\leqslant 2q_e s_n \psi \exp\left\{ -\min\left\{ \frac{C_1 n \phi^2 \epsilon^2}{8\mathcal{K}_1 q_e^2 s_n^2 \psi^2}, \frac{n\phi\epsilon}{4q_e s_n \psi c_{\mathcal{I}_1,max}} \right\} + \log|\mathcal{S}^c| \right\} \\ &\quad + 2q_e \psi \exp\left\{ -\min\left\{ \frac{C_1 n \phi^2 \epsilon^2}{8\mathcal{K}_1 q_e^2 \psi^2}, \frac{n\phi\epsilon}{4q_e \psi c_{\mathcal{I}_2,max}} \right\} + \log|\mathcal{S}^c| \right\}, \\ P\left\{ \left\| \frac{1}{n\phi} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1} \right\|_\infty \geqslant \epsilon \right\} &\leqslant 2q_e s_n \psi \exp\left\{ -\min\left\{ \frac{C_1 n \phi^2 \epsilon^2}{8\mathcal{K}_1 q_e^2 s_n^2 \psi^2}, \frac{n\phi\epsilon}{4q_e s_n \psi c_{\mathcal{I}_1,max}} \right\} + \log|\mathcal{S}_1| \right\} \\ &\quad + 2q_e \psi \exp\left\{ -\min\left\{ \frac{C_1 n \phi^2 \epsilon^2}{8\mathcal{K}_1 q_e^2 \psi^2}, \frac{n\phi\epsilon}{4q_e \psi c_{\mathcal{I}_2,max}} \right\} + \log|\mathcal{S}_1| \right\}, \end{aligned}$$

Proof. We start with $\frac{\partial - l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial [\theta_{E_s}]_{h_1}}$ as follows. For any edge color class $\forall\ E_s \in \mathcal{E}$ and any

$h_1 = 1, \cdots, H$, we show that

$$\frac{\partial - l_c(\boldsymbol{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial [\theta_{E_s}]_{h_1}} = \underbrace{\sum_{m=1}^{n} \alpha_{jj}^{-1} \Big(\sum_{i=1, i \neq j}^{p_n} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{ji}]_h \Big] y_{mi} \Big) \Big(\sum_{k:(k,j) \in E_s} x_m^{(h_1)} y_{mk} \Big)}_{\mathcal{I}_1}$$

$$+ \underbrace{\sum_{m=1}^{n} y_{mj} \Big(\sum_{k:(k,j) \in E_s} x_m^{(h_1)} y_{mk} \Big)}_{\mathcal{I}_2}.$$

The first term $\mathcal{I}_1$ can be written as

$$\mathcal{I}_1 = \sum_{i=1, i \neq j}^{p_n} \sum_{k:(k,j) \in E_s} \underbrace{\Big\{ \sum_{m=1}^{n} \alpha_{jj}^{-1} x_m^{(h_1)} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{ji}]_h \Big] y_{mi} y_{mk} \Big\}}_{Z_{ijk}^{(h_1)}}.$$

It can be shown that $Z_{ijk}^{(h_1)}$ follows a sub-exponential distribution with parameters

$$\big[v_{ijk}^{(h_1)}\big]^2 = \sum_{m=1}^{n} \Big\{ \alpha_{jj}^{-1} x_m^{(h_1)} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{ji}]_h \Big] \Big\}^2 v_{m,ik}^2,$$

$$c_{ijk}^{(h_1)} = \max_{1 \leqslant m \leqslant n} \Big\{ \alpha_{jj}^{-1} x_m^{(h_1)} \Big(\big|\alpha_{ji}\big| + \sum_{h=1}^{H} x_m^{(h)} \big|[\theta_{ji}]_h\big| \Big) c_{m,ik} \Big\}.$$

According to the property of sub-exponential distribution, we have

$$P\Big\{ \Big| Z_{ijk}^{(h_1)} - E\big\{Z_{ijk}^{(h_1)}\big\} \Big| \geqslant t \Big\} \leqslant \begin{cases} 2\exp\Big\{ -\frac{t^2}{2\big[v_{ijk}^{(h_1)}\big]^2} \Big\} & \text{if } 0 \leqslant t \leqslant \frac{\big[v_{ijk}^{(h_1)}\big]^2}{c_{ijk}^{(h_1)}}, \\ 2\exp\Big\{ -\frac{t}{2c_{ijk}^{(h_1)}} \Big\} & \text{if } t > \frac{\big[v_{ijk}^{(h_1)}\big]^2}{c_{ijk}^{(h_1)}}. \end{cases}$$

Therefore, for any $t > 0$, we show that

$$
\begin{aligned}
P\Big\{\Big|\mathcal{I}_1 - E\{\mathcal{I}_1\}\Big| \geqslant t\Big\} &= P\left\{\left|\sum_{\substack{i=1\\ i\neq j}}^{p_n} \sum_{k:(k,j)\in E_s} \Big[Z_{ijk}^{(h_1)} - E\big\{Z_{ijk}^{(h_1)}\big\}\Big]\right| \geqslant t\right\} \\
&\leqslant \sum_{\substack{i=1\\ i\neq j}}^{p_n} \sum_{k:(k,j)\in E_s} P\left\{\Big|Z_{ijk}^{(h_1)} - E\big\{Z_{ijk}^{(h_1)}\big\}\Big| \geqslant \frac{t}{\text{total \# of terms}}\right\} \\
&\leqslant 2 s_n u_e \exp\left\{-\min\left\{\frac{t^2}{2 s_n^2 u_e^2 v_{\mathcal{I}_1,max}^2}, \frac{t}{2 s_n u_e c_{\mathcal{I}_1,max}}\right\}\right\},
\end{aligned}
\tag{3.20}
$$

where

$$
\begin{aligned}
v_{\mathcal{I}_1,max}^2 &= \max_{i,j,k,h_1} \left\{ \sum_{m=1}^{n} \left\{\alpha_{jj}^{-1} x_m^{(h_1)} \Big[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)} [\theta_{ji}]_h\Big]\right\}^2 v_{m,ik}^2 \right\}, \\
c_{\mathcal{I}_1,max} &= \max_{\substack{i,j,k,\\ h_1,m}} \left\{ \alpha_{jj}^{-1} x_m^{(h_1)} \Big(\big|\alpha_{ji}\big| + \sum_{h=1}^{H} x_m^{(h)} \big|[\theta_{ji}]_h\big|\Big) c_{m,ik} \right\}.
\end{aligned}
$$

Next, we focus on $\mathcal{I}_2 = \sum_{m=1}^{n} y_{mj}\Big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\Big)$. We show that

$$
\mathcal{I}_2 = \sum_{m=1}^{n} y_{mj}\Big(\sum_{k:(k,j)\in E_s} x_m^{(h_1)} y_{mk}\Big) = \sum_{k:(k,j)\in E_s} \Big(\sum_{m=1}^{n} x_m^{(h_1)} y_{mj} y_{mk}\Big),
$$

where $\sum_{m=1}^{n} x_m^{(h_1)} y_{mj} y_{mk}$ follows a sub-exponential distribution with parameters

$$
v_{\mathcal{I}_{2,jk}}^2 = \sum_{m=1}^{n} [x_m^{(h_1)}]^2 v_{m,jk}^2, \quad c_{\mathcal{I}_{2,jk}} = \max_{1\leqslant m\leqslant n} \big\{x_m^{(h_1)} c_{m,jk}\big\}.
$$

Consider the probability

$$P\Big\{\Big|\mathcal{I}_2 - E\{\mathcal{I}_2\}\Big| \geqslant t\Big\} = P\Bigg\{\Bigg|\sum_{k:(k,j)\in E_s}\Bigg[\sum_{m=1}^{n} x_m^{(h_1)} y_{mj} y_{mk} - E\Big\{\sum_{m=1}^{n} x_m^{(h_1)} y_{mj} y_{mk}\Big\}\Bigg]\Bigg| \geqslant t\Bigg\}$$

$$\leqslant \sum_{k:(k,j)\in E_s} P\Bigg\{\Bigg|\sum_{m=1}^{n} x_m^{(h_1)} y_{mj} y_{mk} - E\Big\{\sum_{m=1}^{n} x_m^{(h_1)} y_{mj} y_{mk}\Big\}\Bigg| \geqslant \frac{t}{\text{total \# of terms}}\Bigg\}$$

$$\leqslant 2u_e \exp\Bigg\{-\min\Bigg\{\frac{t^2}{2u_e^2 v_{\mathcal{I}_2,max}^2}, \frac{t}{2u_e c_{\mathcal{I}_2,max}}\Bigg\}\Bigg\}, \tag{3.21}$$

where

$$v_{\mathcal{I}_2,max}^2 = \max_{j,k}\Bigg\{\sum_{m=1}^{n}[x_m^{(h_1)}]^2 v_{m,jk}^2\Bigg\}, \quad c_{\mathcal{I}_2,max} = \max_{m,j,k,h_1}\big\{x_m^{(h_1)} c_{m,jk}\big\}.$$

Therefore, we have

$$P\Bigg\{\Bigg|\frac{\partial - l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_{E_s}]_{h_1}} - E\Bigg\{\frac{\partial - l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial[\theta_{E_s}]_{h_1}}\Bigg\}\Bigg| \geqslant t\Bigg\} = P\Big\{|\mathcal{I}_1 - E\{\mathcal{I}_1\} + \mathcal{I}_2 - E\{\mathcal{I}_2\}| \geqslant t\Big\}$$

$$\leqslant P\Big\{|\mathcal{I}_1 - E\{\mathcal{I}_1\}| \geqslant \frac{t}{2}\Big\} + P\Big\{|\mathcal{I}_2 - E\{\mathcal{I}_2\}| \geqslant \frac{t}{2}\Big\}$$

$$\leqslant 2s_n u_e \exp\Bigg\{-\min\Bigg\{\frac{t^2}{8s_n^2 u_e^2 v_{\mathcal{I}_1,max}^2}, \frac{t}{4s_n u_e c_{\mathcal{I}_1,max}}\Bigg\}\Bigg\}$$

$$+ 2u_e \exp\Bigg\{-\min\Bigg\{\frac{t^2}{8u_e^2 v_{\mathcal{I}_2,max}^2}, \frac{t}{4u_e c_{\mathcal{I}_2,max}}\Bigg\}\Bigg\}.$$

Define constants $\mathcal{W}_1, \mathcal{K}_1$ as

$$\mathcal{W}_1 = \max_{\substack{m,i,j,\\ k,h_1}}\Bigg\{\Bigg\{\alpha_{jj}^{-1} x_m^{(h_1)}\Bigg[\alpha_{ji} + \sum_{h=1}^{H} x_m^{(h)}[\theta_{ji}]_h\Bigg]\Bigg\}^2 v_{m,ik}^2, [x_m^{(h_1)}]^2 v_{m,jk}^2\Bigg\},$$

$$\mathcal{K}_1 = C_1\mathcal{W}_1 \ \text{ for any constant } C_1 > 0,$$

and let $t = (\mathcal{K}_1 n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}}$. The probability becomes

$$
\begin{aligned}
& P\left\{\left|\frac{\partial - l_c(\boldsymbol{Y}_j, \boldsymbol{\beta^0})}{\partial [\theta_{E_s}]_{h_1}} - E\left\{\frac{\partial - l_c(\boldsymbol{Y}_j, \boldsymbol{\beta^0})}{\partial [\theta_{E_s}]_{h_1}}\right\}\right| \geqslant (\mathcal{K}_1 n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}}\right\} \\
& \leqslant 2 s_n u_e \exp\left\{ - \min\left\{\frac{\mathcal{K}_1 n s_n^2 u_e^2 \log p_n}{8 s_n^2 u_e^2 n \mathcal{W}_1}, \frac{(\mathcal{K}_1 n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}}}{4 s_n u_e c_{\mathcal{I}_1, max}}\right\}\right\} \\
& + 2 u_e \exp\left\{ - \min\left\{\frac{\mathcal{K}_1 n s_n^2 u_e^2 \log p_n}{8 u_e^2 n \mathcal{W}_1}, \frac{(\mathcal{K}_1 n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}}}{4 u_e c_{\mathcal{I}_2, max}}\right\}\right\} \\
& \leqslant 2 \exp\left\{ - \min\left\{\frac{C_1 \log p_n}{8}, \frac{(\mathcal{K}_1 n \log p_n)^{\frac{1}{2}}}{4 c_{\mathcal{I}_1, max}}\right\} + \log s_n + \log u_e\right\} \\
& + 2 \exp\left\{ - \min\left\{\frac{C_1 s_n^2 \log p_n}{8}, \frac{(\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}}{4 c_{\mathcal{I}_2, max}}\right\} + \log u_e\right\}.
\end{aligned}
$$

By the Bonferroni inequality,

$$
\begin{aligned}
& P\left\{\max_{E_s, h_1}\left|\frac{\partial - l_c(\boldsymbol{Y}_j, \boldsymbol{\beta^0})}{\partial [\theta_{E_s}]_{h_1}}\right| \geqslant (\mathcal{K}_1 n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}}\right\} \leqslant H |\boldsymbol{\mathcal{E}}|\, P\left\{\left|\frac{\partial - l_c(\boldsymbol{Y}_j, \boldsymbol{\beta^0})}{\partial [\theta_{E_s}]_{h_1}}\right| \geqslant (\mathcal{K}_1 n s_n^2 u_e^2 \log p_n)^{\frac{1}{2}}\right\} \\
& \leqslant 2 |\boldsymbol{\mathcal{E}}|\, H \exp\left\{ - \min\left\{\frac{C_1 \log p_n}{8}, \frac{(\mathcal{K}_1 n \log p_n)^{\frac{1}{2}}}{4 c_{\mathcal{I}_1, max}}\right\} + \log s_n + \log u_e\right\} \\
& + 2 |\boldsymbol{\mathcal{E}}|\, H \exp\left\{ - \min\left\{\frac{C_1 s_n^2 \log p_n}{8}, \frac{(\mathcal{K}_1 n s_n^2 \log p_n)^{\frac{1}{2}}}{4 c_{\mathcal{I}_2, max}}\right\} + \log u_e\right\},
\end{aligned}
$$

where $|\boldsymbol{\mathcal{E}}|$ denotes the cardinality of the edge color classes. Provided that the constant $C_1 > 8$, the probability above approaches zero. The proofs of the remaining results are analogous to the one provided above and thus are omitted. □

Lemma 3.2 For $\forall\ E_{s_1}, E_{s_2} \in \boldsymbol{\mathcal{E}}$, $\forall\ V_r \in \boldsymbol{\mathcal{V}}$, and $h_1, h_2 = 1, \cdots, H$, we have

$$\max_{\substack{E_{s_1},E_{s_2},\\h_2}}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial[\theta_{E_{s_2}}]_{h_2}}\right| = \mathcal{O}_p\left\{(nq_e^2u_e^4\log p_n)^{\frac{1}{2}}\right\},\quad \max_{\substack{E_{s_1},V_r,\\h_1}}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial\alpha_{V_r}}\right| = \mathcal{O}_p\left\{(n\phi^2s_n^2u_e^2\log p_n)^{\frac{1}{2}}\right\},$$

$$\max_{\substack{E_{s_1},E_{s_2},\\h_1,h_2}}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial[\theta_{E_{s_2}}]_{h_2}}\right| = \mathcal{O}_p\left\{(nq_e^2u_e^4\log p_n)^{\frac{1}{2}}\right\},\quad \max_{V_r}\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial^2\alpha_{V_r}}\right| = \mathcal{O}_p\left\{(n\phi^2s_n^4\log p_n)^{\frac{1}{2}}\right\},$$

Proof. The second derivatives are given as follows

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial[\theta_{E_{s_2}}]_{h_2}} = \sum_{m=1}^{n}\sum_{j=1}^{p_n}\alpha_{jj}^{-1}\Big(\sum_{l:(l,j)\in E_{s_2}} x_m^{(h_2)}y_{ml}\Big)\Big(\sum_{k:(k,j)\in E_{s_1}} x_m^{(h_1)}y_{mk}\Big),$$

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial[\theta_{E_{s_2}}]_{h_2}} = \sum_{m=1}^{n}\sum_{j=1}^{p_n}\alpha_{jj}^{-1}\Big(\sum_{l:(l,j)\in E_{s_2}} x_m^{(h_2)}y_{ml}\Big)\Big(\sum_{k:(k,j)\in E_{s_1}} y_{mk}\Big),$$

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial\alpha_{V_r}} = -\sum_{m=1}^{n}\sum_{j\in V_r}\alpha_{jj}^{-2}\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big)\Big(\sum_{k:(k,j)\in E_{s_1}} x_m^{(h_1)}y_{mk}\Big),$$

$$\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{V_r}^2} = \frac{1}{2}\sum_{m=1}^{n}\sum_{j\in V_r}\alpha_{jj}^{-2} + \sum_{m=1}^{n}\sum_{j\in V_r}\alpha_{jj}^{-3}\Big(\sum_{i=1,i\neq j}^{p_n}\Big[\alpha_{ji} + \sum_{h=1}^{H}x_m^{(h)}[\theta_{ji}]_h\Big]y_{mi}\Big)^2.$$

Following the approach in **Lemma** 3.1, for $\forall\ \epsilon > 0$, we provide the probabilities outlined below,

$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial[\theta_{E_{s_2}}]_{h_1}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial[\theta_{E_{s_2}}]_{h_1}}\right\}\right| \geqslant \epsilon\right\}$$

$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2 n q_e^2 u_e^4}, \frac{\epsilon}{2q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e\right\},$$

$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial\alpha_{V_r}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial\alpha_{V_r}}\right\}\right| \geqslant \epsilon\right\}$$

$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2 n\phi^2 s_n^2 u_e^2}, \frac{\epsilon}{2\phi s_n u_e c_*}\right\} + \log\phi + \log s_n + \log u_e\right\},$$

$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial[\theta_{E_{s_1}}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial[\theta_{E_{s_1}}]_{h_2}}\right\}\right| \geqslant \epsilon\right\}$$

$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2 n q_e^2 u_e^4}, \frac{\epsilon}{2q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e\right\},$$

$$P\left\{\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{V_r}^2} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{V_r}^2}\right\}\right| \geqslant \epsilon\right\} \leqslant 2\phi\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{8\mathcal{K}_2 n\phi^2 s_n^2}, \frac{\epsilon}{4\phi s_n c_*}\right\} + \log s_n\right\}$$

$$+\exp\left\{-\min\left\{\frac{C_2\epsilon^2}{2\mathcal{K}_2 n\phi^2 s_n^4}, \frac{\epsilon}{2\phi s_n^2 c_*}\right\} + \log s_n + \log(s_n - 1) + \log\phi\right\},$$

for some positive constants $C_2, \mathcal{K}_2, c_* > 0$. The detailed proof is similar to **Lemma 3.1** and therefore is omitted. $\square$

Lemma 3.3 For $\forall\ \epsilon > 0$, we show that

$$P\left\{\left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{ -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right|\right|\right|_\infty \geqslant \epsilon \right\}$$

$$\leqslant 2\exp\left\{ -\min\left\{\frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^2\,\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|\, q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\right\},$$

$$P\left\{\left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - E\left\{ -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\right\}\right|\right|\right|_\infty \geqslant \epsilon \right\}$$

$$\leqslant 2\exp\left\{ -\min\left\{\frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^2\,\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|\, q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + \log|\mathcal{S}_1| + \log|\mathcal{S}^c|\right\}.$$

Proof. For any matrix $\boldsymbol{A} \in \mathbb{R}^{p_n\times p_n}$, the ∞ norm is defined as

$$|||\boldsymbol{A}|||_\infty = \max_{1\leqslant i\leqslant p_n}\sum_{j=1}^{p_n}|a_{ij}|.$$

Note that $-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$ is a $|\mathcal{S}_1|\times|\mathcal{S}_1|$ matrix. For simplicity, we use the notation $-\frac{1}{n\phi}l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$ representing the $(k,l)th$ entry of $-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$. The probability can be formulated as

$$\begin{aligned} &P\left\{\left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{ -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right|\right|\right|_\infty \geqslant \epsilon \right\} \\ &= P\left\{\max_k \frac{1}{n\phi}\sum_l\left|-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{ -l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right\}\right| > \epsilon\right\}. \end{aligned}$$

For any row of $-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$, the summation $\sum_l$ may include

$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial\alpha_{E_{s_2}}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial\alpha_{E_{s_2}}}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial[\theta_{E_{s_2}}]_{h_1}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial\alpha_{E_{s_1}}\partial[\theta_{E_{s_2}}]_{h_1}}\right\}\right|,$$

or

$$\left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial\alpha_{E_{s_2}}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial\alpha_{E_{s_2}}}\right\}\right|, \qquad \left|\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial[\theta_{E_{s_2}}]_{h_2}} - E\left\{\frac{\partial^2 - l_c(\boldsymbol{\beta^0})}{\partial[\theta_{E_{s_1}}]_{h_1}\partial[\theta_{E_{s_2}}]_{h_2}}\right\}\right|,$$

where $E_{s_1}, E_{s_2} \in \mathcal{S}_1, h_1, h_2 = 1, \cdots, H$. Therefore, the number of terms included in $\sum_l$ is $|\mathcal{S}_1|$. By **Lemma** 3.2, we have

$$
\begin{aligned}
&P\bigg\{\frac{1}{n\phi}\bigg|-l_{kl}^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1} - E\Big\{-l_{kl}^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1}\Big\}\bigg| \geqslant \epsilon\bigg\}\\
&\leqslant 2\exp\bigg\{-\min\Big\{\frac{C_2 n\phi^2\epsilon^2}{2\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2q_e u_e^2 c_*}\Big\} + \log q_e + 2\log u_e\bigg\},\\
&P\bigg\{\frac{1}{n\phi}\sum_l\bigg|-l_{kl}^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1} - E\Big\{-l_{kl}^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1}\Big\}\bigg| \geqslant \epsilon\bigg\}\\
&\leqslant 2\exp\bigg\{-\min\Big\{\frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^2\,\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|\,q_e u_e^2 c_*}\Big\} + \log q_e + 2\log u_e + \log|\mathcal{S}_1|\bigg\}.
\end{aligned}
$$

Through the Bonferroni inequality, we show that

$$
\begin{aligned}
&P\bigg\{\bigg|\bigg|\bigg|-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1} - E\Big\{-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1}\Big\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\}\\
&\leqslant |\mathcal{S}_1|\,P\bigg\{\frac{1}{n\phi}\sum_l\bigg|-l_{kl}^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1} - E\Big\{-l_{kl}^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1}\Big\}\bigg| \geqslant \epsilon\bigg\}\\
&\leqslant 2\exp\bigg\{-\min\Big\{\frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^2\,\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|\,q_e u_e^2 c_*}\Big\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\bigg\}.
\end{aligned}
$$

Likewise, we can show similar results for another subset $\mathcal{S}^c$ as follows,

$$
\begin{aligned}
&P\bigg\{\bigg|\bigg|\bigg|-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}^c\mathcal{S}_1} - E\Big\{-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}^c\mathcal{S}_1}\Big\}\bigg|\bigg|\bigg|_\infty \geqslant \epsilon\bigg\}\\
&\leqslant 2\exp\bigg\{-\min\Big\{\frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^2\,\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|\,q_e u_e^2 c_*}\Big\} + \log q_e + 2\log u_e + \log|\mathcal{S}_1| + \log|\mathcal{S}^c|\bigg\}.
\end{aligned}
$$

□

Lemma 3.4 For $\forall\ \epsilon > 0$, we show that

$$P\left\{ \left|\left|\left| E\left\{ -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right\} - \left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right) \right|\right|\right|_2 \geqslant \epsilon \right\}$$
$$\leqslant 2\exp\left\{ -\min\left\{ \frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^4 \mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|^2 q_e u_e^2 c_*} \right\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1| \right\}.$$

Proof. As the spectral norm of any matrix $\boldsymbol{A} \in \mathbb{R}^{p_n \times p_n}$ is always less than or equal to its Frobenius norm, namely $|||\boldsymbol{A}|||_2 \leqslant |||\boldsymbol{A}|||_F$, we obtain the bound $P\left(|||\boldsymbol{A}|||_2 \geqslant \epsilon\right) \leqslant P\left(|||\boldsymbol{A}|||_F \geqslant \epsilon\right)$. Using the same notation introduced in **Lemma** 3.3, we have

$$\left|\left|\left| E\left\{ -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right\} - \left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right) \right|\right|\right|_F = \sqrt{\sum_{k,l} \left| E\left\{ -\frac{1}{n\phi} l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right\} - \left(-\frac{1}{n\phi} l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right) \right|^2}$$
$$\leqslant \sum_{k,l} \left| \left(-\frac{1}{n\phi} l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right) - E\left\{ -\frac{1}{n\phi} l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right\} \right|.$$

The upper bound probability becomes

$$P\left\{ \left|\left|\left| E\left\{ -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right\} - \left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right) \right|\right|\right|_F \geqslant \epsilon \right\}$$
$$\leqslant P\left\{ \sum_{k,l} \left| -l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\left\{ -l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \right\} \right| \geqslant n\phi\epsilon \right\}.$$

Since $-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$ is a $|\mathcal{S}_1| \times |\mathcal{S}_1|$ matrix, the number of terms in the summation

$\sum_{k,l}$ is $|\mathcal{S}_1|^2$. By **Lemma** 3.3, we show

$$
\begin{aligned}
&P\bigg\{\frac{1}{n\phi}\bigg|-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\Big\{-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big\}\bigg| \geqslant \epsilon\bigg\} \\
&\leqslant 2\exp\bigg\{-\min\bigg\{\frac{C_2 n\phi^2\epsilon^2}{2\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2q_e u_e^2 c_*}\bigg\} + \log q_e + 2\log u_e\bigg\}, \\
&P\bigg\{\frac{1}{n\phi}\sum_{k,l}\bigg|-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - E\Big\{-l_{kl}^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big\}\bigg| \geqslant \epsilon\bigg\} \\
&\leqslant 2\exp\bigg\{-\min\bigg\{\frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^4\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|^2 q_e u_e^2 c_*}\bigg\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\bigg\},
\end{aligned}
$$

which completes the proof. □

Lemma 3.5 Under **Assumptions** 6 and $u_e^6 \log u_e = o(n\phi^2)$, there exists some constant $\tau_- > 0$ such that $-\frac{1}{n}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$ has minimum eigenvalue lower bounded by τ_- and is invertible with probability tending to one.

Proof. Under **Assumption** 6, each matrix $E\{-l_c^{(2)}(Y_{mj}, \boldsymbol{\beta^0})\}$ has eigenvalues bounded away from zero and infinity. Then, the eigenvalues of the expected joint Hessian matrix $H(\boldsymbol{\beta^0}) = E\Big\{\frac{1}{n\phi}\sum_{m=1}^{n}\sum_{j=1}^{p_n} -l_c^{(2)}(Y_{mj}, \boldsymbol{\beta^0})\Big\} = E\{-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\}$ are also bounded away from zero and infinity. Therefore, its sub-matrix $H^0_{\mathcal{S}_1\mathcal{S}_1} = E\{-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\}$, also has a minimum eigenvalue bounded below by some positive constant $\tau > 0$, namely $\lambda_{\min}(H^0_{\mathcal{S}_1\mathcal{S}_1}) \geqslant \tau$. By the Courant-Fischer variational representation, we have

$$\begin{aligned}
\lambda_{\min}(H^0_{\mathcal{S}_1\mathcal{S}_1}) &= \lambda_{\min}\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} + H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) \\
&= \min_{||v||_2=1} \bigg\{ v^T\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} + H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) v \bigg\} \\
&= \min_{||v||_2=1} \bigg\{ v^T\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) v + v^T\bigg(H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) v \bigg\} \\
&\leqslant y^T\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) y + y^T\bigg(H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) y,
\end{aligned}$$

where y is a unit-norm eigenvector of $-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}$. Therefore, we show that

$$\begin{aligned}
&y^T\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) y \geqslant \lambda_{\min}\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) \geqslant \lambda_{\min}(H^0_{\mathcal{S}_1\mathcal{S}_1}) - y^T\bigg(H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) y, \\
&\lambda_{\min}\bigg(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\bigg) \geqslant \tau - \bigg|\bigg|\bigg| H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \bigg|\bigg|\bigg|_2 .
\end{aligned}$$

By **Lemma** 3.4, choosing any $\epsilon \in (0, \tau)$ ensures that the following probability tends to 0,

$$\begin{aligned}
P\bigg\{ \bigg|\bigg|\bigg| H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \bigg|\bigg|\bigg|_2 \geqslant \epsilon \bigg\} &\leqslant P\bigg\{ \bigg|\bigg|\bigg| H^0_{\mathcal{S}_1\mathcal{S}_1} + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} \bigg|\bigg|\bigg|_F \geqslant \epsilon \bigg\} \\
\leqslant 2\exp\bigg\{ -\min\bigg\{ \frac{C_2 n\phi^2\epsilon^2}{2|\mathcal{S}_1|^4 \mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon}{2|\mathcal{S}_1|^2 q_e u_e^2 c_*} \bigg\} &+ \log q_e + 2\log u_e + 2\log|\mathcal{S}_1| \bigg\}.
\end{aligned}$$

Therefore, $\lambda_{\min}\Big(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\Big) > \tau - \epsilon = \tau_-$ holds with probability tending to 1. □

Lemma 3.6 (**Incoherence Condition**) Under **Assumptions** 6, 8 and $u_e^6 \log u_e =$

$o(n\phi^2)$, the following condition

$$\left|\!\left|\!\left|\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty \leqslant 1 - \frac{1}{3}\xi^2 \tag{3.22}$$

holds with probability tending to one.

Proof. To show (3.22), we need to apply several facts derived from **Lemma** 3.3, **Lemma** 3.4 and **Lemma** 3.5, which are summarized as follows

1. By **Lemma** 3.3, for $\forall\ \epsilon > 0, \forall\ \epsilon' \in (0, \tau)$, and some constant $\tau_- = \tau - \epsilon' > 0$, we have

$$P\left\{\left|\!\left|\!\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1} - H^0_{\mathcal{S}_1\mathcal{S}_1}\right|\!\right|\!\right|_\infty \geqslant \frac{\tau_-\epsilon}{\sqrt{|\mathcal{S}_1|}}\right\}$$
$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2 n\phi^2\tau_-^2\epsilon^2}{2|\mathcal{S}_1|^3\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\tau_-\epsilon}{2\sqrt{|\mathcal{S}_1|^3} q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\right\},$$

$$P\left\{\left|\!\left|\!\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\!\right|\!\right|_\infty \geqslant \frac{\tau\epsilon}{\sqrt{|\mathcal{S}_1|}}\right\}$$
$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2 n\phi^2\tau^2\epsilon^2}{2|\mathcal{S}_1|^3\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\tau\epsilon}{2\sqrt{|\mathcal{S}_1|^3} q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + \log|\mathcal{S}_1| + \log|\mathcal{S}^c|\right\}.$$

2. By **Lemma** 3.4, for $\forall\ \epsilon' > 0$, we have

$$P\left\{\left|\!\left|\!\left| H^0_{\mathcal{S}_1\mathcal{S}_1} - \left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2 \geqslant \epsilon'\right\}$$
$$\leqslant 2\exp\left\{-\min\left\{\frac{C_2 n\phi^2(\epsilon')^2}{2|\mathcal{S}_1|^4\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\epsilon'}{2|\mathcal{S}_1|^2 q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\right\}.$$

3. By **Lemma** 3.5, the following inequality holds

$$\left|\left|\left|\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_2 \leqslant \frac{1}{\tau_-}$$

with the probability tending to one, where $\tau_- = \tau - \epsilon' > 0$.

We decompose the term $\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}$ as

$$\begin{aligned}
&\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\\
&= \left[H^0_{\mathcal{S}^c\mathcal{S}_1} - \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right]\left[\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} - \left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} - \left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right]\\
&= \underbrace{H^0_{\mathcal{S}^c\mathcal{S}_1}\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}}_{\mathcal{I}_1} + \underbrace{H^0_{\mathcal{S}^c\mathcal{S}_1}\left[\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} - \left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right]}_{\mathcal{I}_2}\\
&+ \underbrace{\left[-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right]\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}}_{\mathcal{I}_3}\\
&+ \underbrace{\left[-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right]\left[\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} - \left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right]}_{\mathcal{I}_4}.
\end{aligned}$$

Adding the matrix infinity norm, we obtain $\left|\left|\left|\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \leqslant$ $|||\mathcal{I}_1|||_\infty + |||\mathcal{I}_2|||_\infty + |||\mathcal{I}_3|||_\infty + |||\mathcal{I}_4|||_\infty$. By **Assumption** 8, $|||\mathcal{I}_1|||_\infty \leqslant 1 - \xi$. The second term $\mathcal{I}_2$ can be formulated as

$$\begin{aligned}
\mathcal{I}_2 &= H^0_{\mathcal{S}^c\mathcal{S}_1}\left[\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} - \left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right]\\
&= H^0_{\mathcal{S}^c\mathcal{S}_1}\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1} - \left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right]\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}.
\end{aligned}$$

By **Fact** 1, **Fact** 3, and **Assumption** 8, we show that

$$\begin{aligned}
\left|\left|\left|\mathcal{I}_2\right|\right|\right|_\infty &\leqslant \left|\left|\left|H^0_{\mathcal{S}^c\mathcal{S}_1}\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - H^0_{\mathcal{S}_1\mathcal{S}_1}\right|\right|\right|_\infty \left|\left|\left|\left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \\
&\leqslant (1-\xi)\left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1} - H^0_{\mathcal{S}_1\mathcal{S}_1}\right|\right|\right|_\infty \sqrt{|\mathcal{S}_1|}\left|\left|\left|\left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_2 \\
&\leqslant (1-\xi)\frac{\tau_-\epsilon}{\sqrt{|\mathcal{S}_1|}}\frac{\sqrt{|\mathcal{S}_1|}}{\tau_-} = (1-\xi)\epsilon
\end{aligned}$$

with a probability of at least $1-2\exp\left\{-\min\left\{\frac{C_2 n\phi^2\tau_-^2\epsilon^2}{2|\mathcal{S}_1|^3\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\tau_-\epsilon}{2\sqrt{|\mathcal{S}_1|^3}q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\right\} - 2\exp\left\{-\min\left\{\frac{C_2 n\phi^2(\epsilon')^2}{2|\mathcal{S}_1|^4\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi(\epsilon')}{2|\mathcal{S}_1|^2 q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\right\}$.

Likewise, by **Fact** 1 and **Assumption** 6, the third term $\left|\left|\left|\mathcal{I}_3\right|\right|\right|_\infty$ can be bounded by

$$\begin{aligned}
\left|\left|\left|\mathcal{I}_3\right|\right|\right|_\infty &\leqslant \left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\right|\right|_\infty \left|\left|\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_\infty \\
&\leqslant \left|\left|\left| -\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1} - H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\right|\right|_\infty \sqrt{|\mathcal{S}_1|}\left|\left|\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\right|\right|_2 \leqslant \frac{\tau\epsilon}{\sqrt{|\mathcal{S}_1|}}\frac{\sqrt{|\mathcal{S}_1|}}{\tau} = \epsilon,
\end{aligned}$$

with a probability of at least $1-2\exp\left\{-\min\left\{\frac{C_2 n\phi^2\tau^2\epsilon^2}{2|\mathcal{S}_1|^3\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\tau\epsilon}{2\sqrt{|\mathcal{S}_1|^3}q_e u_e^2 c_*}\right\} + \log q_e + 2\log u_e + \log|\mathcal{S}_1| + \log|\mathcal{S}^c|\right\}$. For $\mathcal{I}_4$, we first show that

$$\begin{aligned}
&\left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} - \left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} \\
&= \left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1} - \left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right]\left(-\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}.
\end{aligned}$$

Taking the infinity norm, it becomes

$$
\begin{aligned}
&\left|\!\left|\!\left|\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty\\
&=\left|\!\left|\!\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1}+\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right]\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty\\
&\leqslant\sqrt{|\mathcal{S}_1|}\cdot\left|\!\left|\!\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left[H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right]\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_2\\
&\leqslant\sqrt{|\mathcal{S}_1|}\cdot\left|\!\left|\!\left|\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_2\cdot\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2\cdot\left|\!\left|\!\left|\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_2\\
&\leqslant\frac{\sqrt{|\mathcal{S}_1|}}{\tau\cdot\tau_-}\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2.
\end{aligned}
$$

Finally, according to **Fact** 1 and **Fact** 2, we can calculate the upper bound of $|||\mathcal{I}_4|||_\infty$ as

$$
\begin{aligned}
|||\mathcal{I}_4|||_\infty &\leqslant \left|\!\left|\!\left|-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\!\right|\!\right|_\infty\left|\!\left|\!\left|\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}-\left(H^0_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty\\
&\leqslant \left|\!\left|\!\left|-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}-H^0_{\mathcal{S}^c\mathcal{S}_1}\right|\!\right|\!\right|_\infty\frac{\sqrt{|\mathcal{S}_1|}}{\tau\cdot\tau_-}\left|\!\left|\!\left|H^0_{\mathcal{S}_1\mathcal{S}_1}-\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)\right|\!\right|\!\right|_2\\
&\leqslant \frac{\tau\epsilon}{\sqrt{|\mathcal{S}_1|}}\frac{\sqrt{|\mathcal{S}_1|}}{\tau\cdot\tau_-}\epsilon'=\frac{\epsilon'}{\tau_-}\epsilon.
\end{aligned}
$$

We choose $\epsilon' < \tau_- = \tau - \epsilon'$, namely $\epsilon' < 0.5\tau$, and choose $\epsilon < \frac{\xi}{3}$ for some constant $\xi \in (0,1)$ defined in **Assumption** 8, then we have

$$
\begin{aligned}
\left|\!\left|\!\left|\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta^0})_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty &\leqslant |||\mathcal{I}_1|||_\infty+|||\mathcal{I}_2|||_\infty+|||\mathcal{I}_3|||_\infty+|||\mathcal{I}_4|||_\infty\\
&\leqslant 1-\xi+(1-\xi)\epsilon+\epsilon+\frac{\epsilon'}{\tau_-}\epsilon\\
&\leqslant 1-\xi+\xi-\frac{1}{3}\xi^2=1-\frac{1}{3}\xi^2
\end{aligned}
$$

with the probability of at least $1-2\exp\Bigg\{-\min\Big\{\frac{C_2 n\phi^2\tau_-^2\epsilon^2}{2|\mathcal{S}_1|^3\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\tau_-\epsilon}{2\sqrt{|\mathcal{S}_1|^3}q_e u_e^2 c_*}\Big\}+\log q_e+ 2\log u_e + 2\log|\mathcal{S}_1|\Bigg\} - 2\exp\Bigg\{ - \min\Big\{\frac{C_2 n\phi^2(\epsilon')^2}{2|\mathcal{S}_1|^4\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi(\epsilon')}{2|\mathcal{S}_1|^2 q_e u_e^2 c_*}\Big\} + \log q_e + 2\log u_e + 2\log|\mathcal{S}_1|\Bigg\} - 2\exp\Bigg\{ - \min\Big\{\frac{C_2 n\phi^2\tau^2\epsilon^2}{2|\mathcal{S}_1|^3\mathcal{K}_2 q_e^2 u_e^4}, \frac{n\phi\tau\epsilon}{2\sqrt{|\mathcal{S}_1|^3}q_e u_e^2 c_*}\Big\} + \log q_e + 2\log u_e + \log|\mathcal{S}_1| + \log|\mathcal{S}^c|\Bigg\}$. □

Proofs of Theorems 3.1

Proof. The objective function is given by

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -l_c(\boldsymbol{\beta}) + n\phi\sum_{s=1}^{w}\rho_\lambda(|\alpha_{E_s}|) + n\phi\sum_{s=1}^{w}\sum_{h=1}^{H}\rho_\lambda(|[\theta_{E_s}]_h|),$$

where $\rho_\lambda(\cdot)$ is the lasso penalty function. Define a function $G(\boldsymbol{\Delta}) = Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}^0)$, where $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$. If we take a closed convex set $\mathcal{A}$ which contains $\mathbf{0}$ and further show that the function $G(\boldsymbol{\Delta})$ is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that $G(\boldsymbol{\Delta})$ has a local minimum inside the convex set $\mathcal{A}$, since $G(\boldsymbol{\Delta})$ is continuous and $G(\mathbf{0}) = \mathbf{0}$. In particular, we define the closed convex set $\mathcal{A} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{0}}||_2 \leqslant \mathcal{C}r_n\}$ and its boundary $\partial\mathcal{A} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{0}}||_2 = \mathcal{C}r_n\}$, where $\mathcal{C}$ is some positive constant and $r_n = \Big\{\frac{p_n \log p_n}{n}\Big\}^{\frac{1}{2}}$.

The function $G(\boldsymbol{\Delta})$ can be written as

$$G(\boldsymbol{\Delta}) = l_c(\boldsymbol{\beta}^0) - l_c(\boldsymbol{\beta}) + n\phi\sum_{s=1}^{w}\Big\{\rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|)\Big\} + n\phi\sum_{s=1}^{w}\sum_{h=1}^{H}\Big\{\rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|)\Big\}.$$

Applying Taylor's expansion to $l_c(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathbf{0}}$, we show that

$$l_c(\boldsymbol{\beta}) = l_c(\boldsymbol{\beta}^0) + l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot l_c^{(2)}(\boldsymbol{\beta}^*) \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0), \qquad (3.23)$$

where $\boldsymbol{\beta}^* = a\boldsymbol{\beta} + (1-a)\boldsymbol{\beta}^0$ for some constant $a \in (0,1)$. Therefore, $l_c(\boldsymbol{\beta}^0) - l_c(\boldsymbol{\beta})$ becomes

$$\begin{aligned} l_c(\boldsymbol{\beta}^0) - l_c(\boldsymbol{\beta}) &= -l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot l_c^{(2)}(\boldsymbol{\beta}^*) \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) \\ &= \underbrace{-l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)}_{\mathcal{I}_1} + \underbrace{\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\} \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)}_{\mathcal{I}_2} \\ &\quad + \underbrace{\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \cdot \Big\{ - l_c^{(2)}(\boldsymbol{\beta}^*) - E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\Big\} \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)}_{\mathcal{I}_3}. \end{aligned} \tag{3.24}$$

As stated in **Assumption** 6, the eigenvalues of $E\Big\{ - l_c^{(2)}(Y_{mj}, \boldsymbol{\beta}^*)\Big\}$ are uniformly bounded away from zero and infinity whenever $\boldsymbol{\beta}^*$ is within an η-neighborhood of $\boldsymbol{\beta}^{\mathbf{0}}$. Consequently, there exist constants $\kappa_-, \kappa_+ > 0$ such that $0 < \kappa_- \leqslant \lambda\Big(E\{ - l_c^{(2)}(Y_{mj}, \boldsymbol{\beta}^*)\}\Big) \leqslant \kappa_+ < \infty$, which further implies $\boldsymbol{v}^T E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\boldsymbol{v} = \mathcal{O}_p(np_n) > 0$, where $\boldsymbol{v}$ is any unit-norm vector. On the boundary $\partial\mathcal{A} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{0}}||_2 = \mathcal{C}r_n\}$ for some constant $\mathcal{C} > 0$, we have

$$\begin{aligned} \mathcal{I}_2 &= \frac{1}{2}\mathcal{C}^2 r_n^2 \cdot \boldsymbol{v}^T E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\boldsymbol{v} = \frac{1}{2}\mathcal{C}^2 r_n^2 \cdot \mathcal{O}_p(np_n) > 0, \\ \mathcal{I}_3 &= \frac{1}{2}\mathcal{C}^2 r_n^2 \cdot \boldsymbol{v}^T \Big\{ - l_c^{(2)}(\boldsymbol{\beta}^*) - E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\Big\}\boldsymbol{v} \\ &= \frac{1}{2}\mathcal{C}^2 r_n^2 \cdot \sum_{r,t} v_r o_p\Big\{ \Big[E\{-l_c^{(2)}(\boldsymbol{\beta}^*)\}\Big]_{rt}\Big\} v_t, \end{aligned}$$

where $\mathcal{I}_3$ is dominated by $\mathcal{I}_2$. For $\mathcal{I}_1$, it can be partitioned into two parts

$$
\begin{aligned}
|\mathcal{I}_1| &= \left| -l_c^{(1)}(\boldsymbol{\beta}^0)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0) \right| = \left| \sum_t \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t]} \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{[t]} \right| \\
&\leqslant \underbrace{\mathcal{C} r_n \sum_{t_1:vertex} \left| \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t_1]} \right| |v_{t_1}|}_{|\mathcal{I}_{1.1}|} + \underbrace{\mathcal{C} r_n \sum_{t_2:edge} \left| \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t_2]} \right| |v_{t_2}|}_{|\mathcal{I}_{1.2}|}.
\end{aligned}
$$

First Term $|\mathcal{I}_{1.1}|$: By **Lemma** 3.1 and Cauchy-Schwarz inequality, the upper bound is

$$
\begin{aligned}
|\mathcal{I}_{1.1}| = \mathcal{C} r_n \sum_{t_1:vertex} \left| \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t_1]} \right| |v_{t_1}| &\leqslant \mathcal{C} r_n \sqrt{\sum_{t_1:vertex} \big[-l_c^{(1)}(\boldsymbol{\beta}^0) \big]_{[t_1]}^2} \sqrt{\sum_{t_1:vertex} v_{t_1}^2} \\
&\leqslant \mathcal{C} r_n p_n \mathcal{O}_p \Big\{ (n s_n^4 \log p_n)^{\frac{1}{2}} \Big\}.
\end{aligned}
$$

Divide $|\mathcal{I}_{1.1}|$ by $\mathcal{I}_2$, we have

$$
\frac{|\mathcal{I}_{1.1}|}{\mathcal{I}_2} \leqslant \frac{\mathcal{C} r_n p_n \mathcal{O}_p \Big\{ (n s_n^4 \log p_n)^{\frac{1}{2}} \Big\}}{\mathcal{C}^2 r_n^2 n p_n \kappa_-} = \frac{\mathcal{O}_p \Big\{ (s_n^4 \log p_n)^{\frac{1}{2}} \Big\}}{\mathcal{C} r_n n^{\frac{1}{2}} \kappa_-} = \frac{\mathcal{O}_p \Big\{ (s_n^4 \log p_n)^{\frac{1}{2}} \Big\}}{\mathcal{C} (p_n \log p_n)^{\frac{1}{2}} \kappa_-}.
$$

Set $r_n = \left\{ \frac{p_n \log p_n}{n} \right\}^{\frac{1}{2}}$ and $s_n^4 = \mathcal{O}(p_n)$, $\mathcal{I}_2$ can dominate $|\mathcal{I}_{1.1}|$ with an appropriate constant $\mathcal{C} > 0$.

Second Term $|\mathcal{I}_{1.2}|$: Note that the # of edge parameters is $(H+1)|\boldsymbol{\mathcal{E}}|$, which is a fixed number. By **Lemma** 3.1 , we have $\max_{E_s, h_1} \left| \frac{\partial -l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial [\theta_{E_s}]_{h_1}} \right| = \max_{E_s} \left| \frac{\partial -l_c(\mathbf{Y}_j, \boldsymbol{\beta}^{\mathbf{0}})}{\partial \alpha_{E_s}} \right| =$

$\mathcal{O}_p\Big\{(ns_n^2u_e^2\log p_n)^{\frac{1}{2}}\Big\}$. By Cauchy–Schwarz inequality, $|\mathcal{I}_{1.2}|$ becomes

$$\begin{aligned}|\mathcal{I}_{1.2}| = \mathcal{C}r_n\sum_{t_2:edge}\Big|\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]_{[t_2]}\Big||v_{t_2}| &\leqslant \mathcal{C}r_n\sqrt{\sum_{t_2:edge}\big[-l_c^{(1)}(\boldsymbol{\beta}^0)\big]^2_{[t_2]}}\sqrt{\sum_{t_2:edge}[v_{t_2}]^2}\\ &\leqslant \mathcal{C}r_np_n\mathcal{O}_p\Big\{(ns_n^2u_e^2\log p_n)^{\frac{1}{2}}\Big\}.\end{aligned}$$

Divide $|\mathcal{I}_{1.2}|$ by $\mathcal{I}_2$, we show that

$$\frac{|\mathcal{I}_{1.2}|}{\mathcal{I}_2} \leqslant \frac{\mathcal{C}r_np_n\mathcal{O}_p\Big\{(ns_n^2u_e^2\log p_n)^{\frac{1}{2}}\Big\}}{\frac{1}{2}\mathcal{C}^2r_n^2np_n\kappa_-} = \frac{2\mathcal{O}_p\Big\{(s_n^2u_e^2\log p_n)^{\frac{1}{2}}\Big\}}{\mathcal{C}r_nn^{\frac{1}{2}}\kappa_-} = \frac{\mathcal{O}_p\Big\{(s_n^2u_e^2\log p_n)^{\frac{1}{2}}\Big\}}{\mathcal{C}(p_n\log p_n)^{\frac{1}{2}}\kappa_-}.$$

Since $r_n = \Big\{\frac{p_n\log p_n}{n}\Big\}^{\frac{1}{2}}$ and $s_n^2u_e^2 = \mathcal{O}(p_n)$,$|\mathcal{I}_{1.2}|$ can be dominated by $\mathcal{I}_2$ with an appropriate $\mathcal{C} > 0$.

Next, we move on to the penalty terms and denote

$$G(\boldsymbol{\Delta}) = l_c(\boldsymbol{\beta}^0) - l_c(\boldsymbol{\beta}) + \underbrace{n\phi\sum_{s=1}^{w}\Big\{\rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|)\Big\}}_{\mathcal{I}_4} + \underbrace{n\phi\sum_{s=1}^{w}\sum_{h=1}^{H}\Big\{\rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|)\Big\}}_{\mathcal{I}_5}.$$

We further decompose $\mathcal{I}_4$, $\mathcal{I}_5$ as follows

$$\begin{aligned}
\mathcal{I}_4 &= n\phi \sum_{s:\alpha^0_{E_s}=0} \left\{\rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|)\right\} + n\phi \sum_{s:\alpha^0_{E_s}\neq 0} \left\{\rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|)\right\} \\
&= \underbrace{n\phi \sum_{s:\alpha^0_{E_s}=0} \left\{\rho_\lambda(|\alpha_{E_s}|)\right\}}_{\mathcal{I}_{4.1}} + \underbrace{n\phi \sum_{s:\alpha^0_{E_s}\neq 0} \left\{\rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|)\right\}}_{\mathcal{I}_{4.2}}, \\
\mathcal{I}_5 &= n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h=0} \left\{\rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|)\right\} + n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h\neq 0} \left\{\rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|)\right\} \\
&= \underbrace{n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h=0} \left\{\rho_\lambda(|[\theta_{E_s}]_h|)\right\}}_{\mathcal{I}_{5.1}} + \underbrace{n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h\neq 0} \left\{\rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|)\right\}}_{\mathcal{I}_{5.2}}.
\end{aligned} \tag{3.25}$$

Since

$$\rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|) = \begin{cases}
\lambda([\theta_{E_s}]_h - [\theta^0_{E_s}]_h), & \text{if } [\theta_{E_s}]_h \geqslant 0, [\theta^0_{E_s}]_h > 0; \\
-\lambda([\theta_{E_s}]_h - [\theta^0_{E_s}]_h), & \text{if } [\theta_{E_s}]_h \leqslant 0, [\theta^0_{E_s}]_h < 0; \\
-\lambda[\theta_{E_s}]_h - \lambda[\theta^0_{E_s}]_h, & \text{if } [\theta_{E_s}]_h < 0, [\theta^0_{E_s}]_h > 0; \\
\lambda[\theta_{E_s}]_h + \lambda[\theta^0_{E_s}]_h, & \text{if } [\theta_{E_s}]_h > 0, [\theta^0_{E_s}]_h < 0;
\end{cases}$$

and

$$\text{when } [\theta_{E_s}]_h < 0, [\theta^0_{E_s}]_h > 0, \left|-\lambda[\theta_{E_s}]_h - \lambda[\theta^0_{E_s}]_h\right| \leqslant \left|\lambda([\theta_{E_s}]_h - [\theta^0_{E_s}]_h)\right|,$$

$$\text{when } [\theta_{E_s}]_h > 0, [\theta^0_{E_s}]_h < 0, \left|\lambda[\theta_{E_s}]_h + \lambda[\theta^0_{E_s}\right|]_h \leqslant \left|\lambda([\theta_{E_s}]_h - [\theta^0_{E_s}]_h)\right|.$$

Following similar steps, the expression $\rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|)$ yields equivalent bounds, leading us to conclude that

$$
\begin{aligned}
|\mathcal{I}_{5.2}| = \left| n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} \left\{ \rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|) \right\} \right| &\leqslant n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} \left| \rho_\lambda(|[\theta_{E_s}]_h|) - \rho_\lambda(|[\theta^0_{E_s}]_h|) \right| \\
&\leqslant n\phi \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} \left| \lambda([\theta_{E_s}]_h - [\theta^0_{E_s}]_h) \right| \leqslant n\phi\lambda \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} \mathcal{C} r_n |[v_s]_h| \\
&\leqslant n\phi\lambda\mathcal{C} r_n \sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} |[v_s]_h| \cdot 1,
\end{aligned}
$$

$$
|\mathcal{I}_{4.2}| \leqslant n\phi \sum_{s:\alpha^0_{E_s} \neq 0} \left| \rho_\lambda(|\alpha_{E_s}|) - \rho_\lambda(|\alpha^0_{E_s}|) \right| \leqslant n\phi\lambda\mathcal{C} r_n \sum_{s:\alpha^0_{E_s} \neq 0} |v_s| \cdot 1.
$$

The # of terms in the summation $\left(\sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} + \sum_{s:\alpha^0_{E_s} \neq 0} \right)$ is $|\mathcal{S}_1|$, which is finite. By Cauchy-Schwarz inequality, the upper bound of $|\mathcal{I}_{4.2}| + |\mathcal{I}_{5.2}|$ can be given as

$$
\begin{aligned}
|\mathcal{I}_{4.2}| + |\mathcal{I}_{5.2}| &\leqslant n\phi\lambda\mathcal{C} r_n \left(\sum_{h=1}^{H} \sum_{s:[\theta^0_{E_s}]_h \neq 0} |[v_s]_h| \cdot 1 + \sum_{s:\alpha^0_{E_s} \neq 0} |v_s| \cdot 1 \right) \\
&\leqslant n\phi\lambda\mathcal{C} r_n \underbrace{\sqrt{\sum ([v_s]_h)^2}}_{\leqslant 1} \underbrace{\sqrt{\sum 1^2}}_{|\mathcal{S}_1| \text{ terms}} \leqslant n\phi\lambda\mathcal{C} r_n |\mathcal{S}_1|^{\frac{1}{2}}.
\end{aligned}
\tag{3.26}
$$

We determine its magnitude with $\mathcal{I}_2$

$$
\frac{|\mathcal{I}_{4.2}| + |\mathcal{I}_{5.2}|}{\mathcal{I}_2} \leqslant \frac{2n\phi\lambda\mathcal{C} r_n |\mathcal{S}_1|^{\frac{1}{2}}}{\mathcal{C}^2 r_n^2 n p_n \kappa_-} = \frac{2\phi\lambda |\mathcal{S}_1|^{\frac{1}{2}}}{\mathcal{C} r_n p_n \kappa_-} \leqslant \frac{2\delta_2 |\mathcal{S}_1|^{\frac{1}{2}}}{\mathcal{C}\kappa_-} \frac{\phi}{p_n} \left(\frac{s_n^4}{p_n} \right)^{\frac{1}{2}} \leqslant \frac{2\delta_2 |\mathcal{S}_1|^{\frac{1}{2}}}{\mathcal{C}\kappa_-} \left(\frac{s_n^4}{p_n} \right)^{\frac{1}{2}}.
$$

The numerator can be dominated by $\mathcal{I}_2$ with an appropriate constant $\mathcal{C} > 0$ as long as λ is chosen from $\delta_2 \left(\frac{s_n^4 \log p_n}{n} \right)^{\frac{1}{2}} \geqslant \lambda \geqslant \delta_1 \left(\frac{\log p_n}{n} \right)^{\frac{1}{2}}$ for any $\delta_1, \delta_2 > 0$.

To conclude, we have demonstrated that $G(\boldsymbol{\Delta}) > 0$ on the boundary $\partial\mathcal{A}$ by choosing $\mathcal{C} > 0$ sufficiently large, thereby completing the proof. $\square$

Proof of Theorem 3.2

Proof. Since the property of sign consistency is regarding the edge parameters, we treat diagonal parameters as known in this theorem. Recall that the composite likelihood estimator $\hat{\boldsymbol{\beta}}$ is obtained by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \Big\{ - l_c(\boldsymbol{\beta}) + n\phi \sum_{s=1}^{w}\sum_{h=1}^{H} \rho_\lambda(|[\theta_{E_s}]_h|) + n\phi \sum_{s=1}^{w} \rho_\lambda(|\alpha_{E_s}|) \Big\}. \tag{3.27}$$

Here $\boldsymbol{\beta}$ excludes diagonal parameters. Similar to Ravikumar and Yu (2008), we introduce another estimator

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}=(\boldsymbol{\beta}_{\mathcal{S}_1},\mathbf{0})} \Big\{ - l_c(\boldsymbol{\beta}) + n\phi \sum_{s=1}^{w}\sum_{h=1}^{H} \rho_\lambda(|[\theta_{E_s}]_h|) + n\phi \sum_{s=1}^{w} \rho_\lambda(|\alpha_{E_s}|) \Big\}, \tag{3.28}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_{\mathcal{S}_1}, \tilde{\boldsymbol{\beta}}_{\mathcal{S}^c})$ with the constraint $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}$. The set of first derivative equations $\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$ can be partitioned into two sets of equations

$$\frac{1}{n\phi} l_c^{(1)}(\tilde{\boldsymbol{\beta}})_{\mathcal{S}_1} = \lambda \tilde{z}_{\mathcal{S}_1}, \tag{3.29}$$

$$\frac{1}{n\phi} l_c^{(1)}(\tilde{\boldsymbol{\beta}})_{\mathcal{S}^c} = \lambda \tilde{z}_{\mathcal{S}^c}, \tag{3.30}$$

where the sub-differential of the absolute value function is defined as

$$\tilde{z} = \frac{\partial |x|}{\partial x} = \begin{cases} 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

If the estimator $\tilde{\boldsymbol{\beta}}$ satisfies conditions (3.29) and (3.30), then $\tilde{\boldsymbol{\beta}}$ is the local minimizer $\hat{\boldsymbol{\beta}}$ of the objective function $Q(\boldsymbol{\beta})$. We show the proof in two steps.

Step 1: Verify $\|\tilde{z}_{\mathcal{S}^c}\|_\infty < 1$.

By Taylor's expansion, we have

$$\frac{1}{n\phi} l_c^{(1)}(\tilde{\boldsymbol{\beta}}) = \frac{1}{n\phi} l_c^{(1)}(\boldsymbol{\beta}^0) + \frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \underbrace{\frac{1}{n\phi}\{l_c^{(2)}(\boldsymbol{\beta}^*) - l_c^{(2)}(\boldsymbol{\beta}^0)\}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}_{R}, \tag{3.31}$$

where $\boldsymbol{\beta}^* = a\boldsymbol{\beta}^0 + (1-a)\tilde{\boldsymbol{\beta}}$ for some constant $a \in [0, 1]$. In block format, we can write

$$\begin{pmatrix} \frac{1}{n\phi} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1} \\ \frac{1}{n\phi} l_c^{(1)}(\boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{S}^c} \end{pmatrix} + \frac{1}{n\phi} \begin{pmatrix} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1} & l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}^c} \\ l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1} & l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}^c} \end{pmatrix} \begin{pmatrix} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)_{\mathcal{S}_1} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} R_{\mathcal{S}_1} \\ R_{\mathcal{S}^c} \end{pmatrix} = \lambda \begin{pmatrix} \tilde{z}_{\mathcal{S}_1} \\ \tilde{z}_{\mathcal{S}^c} \end{pmatrix},$$

where $R_{\mathcal{S}_1}, R_{\mathcal{S}^c}$ are the partition of the residual vector. From the block format, we know that

$$\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)_{\mathcal{S}_1} = \left(\frac{1}{n\phi} l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1} \left(\lambda \tilde{z}_{\mathcal{S}_1} - R_{\mathcal{S}_1} - \frac{1}{n\phi} l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right), \tag{3.32}$$

$$\tilde{z}_{\mathcal{S}^c} = \frac{1}{\lambda}\bigg(\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} + \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\big(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big)_{\mathcal{S}_1} + R_{\mathcal{S}^c}\bigg). \tag{3.33}$$

By equations (3.32) and (3.33), we have

$$\begin{aligned}
\tilde{z}_{\mathcal{S}^c} =& \frac{1}{\lambda}\bigg\{\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} + \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\bigg(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\bigg)^{-1}\bigg(\lambda\tilde{z}_{\mathcal{S}_1} - R_{\mathcal{S}_1} - \frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\bigg) + R_{\mathcal{S}^c}\bigg\} \\
=& \underbrace{\frac{1}{\lambda}\bigg\{\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} - \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\bigg(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\bigg)^{-1}\bigg(\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\bigg)\bigg\}}_{\mathcal{I}_1} \\
&+ \underbrace{\frac{1}{\lambda}\bigg\{R_{\mathcal{S}^c} - \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\bigg(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\bigg)^{-1}R_{\mathcal{S}_1}\bigg\}}_{\mathcal{I}_2} + \underbrace{\frac{1}{\lambda}\bigg\{\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\bigg(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\bigg)^{-1}\lambda\tilde{z}_{\mathcal{S}_1}\bigg\}}_{\mathcal{I}_3}.
\end{aligned}$$

Since $\|\tilde{z}_{\mathcal{S}^c}\|_\infty = \|\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3\|_\infty \leqslant \|\mathcal{I}_1\|_\infty + \|\mathcal{I}_2\|_\infty + \|\mathcal{I}_3\|_\infty$, we start with the upper bound of $\|\mathcal{I}_2\|_\infty$. By Taylor's expansion, we show that each element of R can be written as

$$(R)_u = \frac{1}{n\phi}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{\mathbf{0}})^T\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}),$$

where β_u denotes any off-diagonal parameter and $\boldsymbol{\beta}^{**} = b\boldsymbol{\beta}^0 + (1-b)\boldsymbol{\beta}^*$ for some constant $b \in [0,1]$. Also note that $(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{\mathbf{0}})^T = (1-a)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T$, we can rewrite it as

$$\begin{aligned}
(R)_u =& \frac{1}{n\phi}(1-a)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^T\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}) \\
=& \underbrace{\frac{1}{n\phi}(1-a)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^T\bigg\{\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u} - E\bigg\{\frac{\partial^3 l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\bigg\}\bigg\}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})}_{\mathcal{M}_1} \\
&- \underbrace{\frac{1}{n\phi}(1-a)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^T E\bigg\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\bigg\}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})}_{\mathcal{M}_2}.
\end{aligned}$$

Following the same methodology as in **Theorem** 3.1, we can show $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2 = \mathcal{O}_p\Big\{\Big(\frac{p_n \log p_n}{n}\Big)^{\frac{1}{2}}\Big\}$. Thus, we have $\mathcal{M}_2 \leqslant \frac{1}{n\phi}(1-a)\mathcal{O}_p\Big\{\frac{p_n \log p_n}{n}\Big\}\boldsymbol{v}^T E\Big\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\Big\}\boldsymbol{v}$ and

$$
\begin{aligned}
\big\|(R)_u\big\|_\infty &= \big\|\mathcal{M}_1 - \mathcal{M}_2\big\|_\infty \leqslant \big\|\mathcal{M}_1\big\|_\infty + \big\|\mathcal{M}_2\big\|_\infty, \\
&\leqslant \big\|\mathcal{M}_2\big\|_\infty\big(1+o_p(1)\big) \leqslant \frac{1}{n\phi}\mathcal{O}_p\Big\{\frac{p_n \log p_n}{n}\Big\}\Big|\Big|\Big|E\Big\{\frac{\partial^3 - l_c(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\Big\}\Big|\Big|\Big|_2\big(1+o_p(1)\big).
\end{aligned}
$$

By **Assumptions** 6-7, we have $\Big|\Big|\Big|E\Big\{\frac{\partial^3 - l_c(Y_{mj},\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\partial\beta_u}\Big\}\Big|\Big|\Big|_2 \leqslant \mathcal{W}^*$ for some constant $\mathcal{W}^* > 0$ and show that

$$
\big\|(R)_u\big\|_\infty \leqslant \mathcal{O}_p\Big\{\frac{p_n \log p_n}{n}\Big\}\frac{1}{n\phi}n\phi\mathcal{W}^*\big(1+o_p(1)\big) = \mathcal{O}_p\Big\{\frac{p_n \log p_n}{n}\Big\}\big(1+o_p(1)\big) = o_p(1).
$$

Therefore, by **Lemma** 3.6 and $p_n^2 \log p_n = \mathcal{O}(n)$, the upper bound of $\|\mathcal{I}_2\|_\infty$ can be given as

$$
\begin{aligned}
\|\mathcal{I}_2\|_\infty &= \frac{1}{\lambda}\Big|\Big|\Big|R_{\mathcal{S}^c} - \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\Big(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}R_{\mathcal{S}_1}\Big|\Big|\Big|_\infty \\
&\leqslant \frac{1}{\lambda}\|R_{\mathcal{S}^c}\|_\infty + \frac{1}{\lambda}\Big|\Big|\Big|\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\Big(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\Big)^{-1}\Big|\Big|\Big|_\infty\|R_{\mathcal{S}_1}\|_\infty \\
&\leqslant \frac{1}{\delta_2}\Big(\frac{n}{s_n^4 \log p_n}\Big)^{\frac{1}{2}}\mathcal{O}_p\Big\{\frac{p_n \log p_n}{n}\Big\}\big(1+o_p(1)\big) \\
&+ \frac{1}{\delta_2}\Big(\frac{n}{s_n^4 \log p_n}\Big)^{\frac{1}{2}}(1-\frac{1}{3}\xi^2)\mathcal{O}_p\Big\{\frac{p_n \log p_n}{n}\Big\}\big(1+o_p(1)\big) \\
&= \mathcal{O}_p\Big\{\Big(\frac{p_n^2 \log p_n}{n s_n^4}\Big)^{\frac{1}{2}}\Big\} = o_p(1).
\end{aligned}
$$

Similarly, by **Lemma** 3.1 and **Lemma** 3.6, the rest two terms can be bounded by

$$\begin{aligned}
\|\mathcal{I}_1\|_\infty &= \frac{1}{\lambda}\left\|\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c} - \frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\left(\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right)\right\|_\infty \\
&\leqslant \frac{1}{\lambda}\left\{\left\|\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c}\right\|_\infty + \left|\!\left|\!\left|\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty\left\|\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right\|_\infty\right\} \\
&\leqslant \frac{1}{\lambda}\left\{\epsilon + (1-\frac{1}{3}\xi^2)\epsilon\right\} \leqslant \frac{1}{\lambda}(2-\frac{1}{3}\xi^2)\epsilon, \\
\|\mathcal{I}_3\|_\infty &= \frac{1}{\lambda}\left\|\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\lambda\tilde{z}_{\mathcal{S}_1}\right\|_\infty \leqslant \left|\!\left|\!\left|\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}^c\mathcal{S}_1}\left(\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty \\
&\leqslant 1-\frac{1}{3}\xi^2.
\end{aligned}$$

Therefore, we choose $\epsilon = \frac{1}{6}\xi^2\lambda$ and

$$\begin{aligned}
\|\tilde{z}_{\mathcal{S}^c}\|_\infty \leqslant \|\mathcal{I}_1\|_\infty + \|\mathcal{I}_2\|_\infty + \|\mathcal{I}_3\|_\infty &\leqslant \frac{1}{\lambda}(2-\frac{1}{3}\xi^2)\epsilon + o(1) + 1 - \frac{1}{3}\xi^2 \\
&\leqslant \frac{1}{3}\xi^2 - \frac{1}{18}\xi^4 + 1 - \frac{1}{3}\xi^2 + o(1) < 1.
\end{aligned}$$

Once the first step is completed, it demonstrates that the solution $\tilde{\boldsymbol{\beta}}$ to the constrained problem (3.28) is equivalent to the solution $\hat{\boldsymbol{\beta}}$ to the original unrestricted problem (3.27).

Step 2: Verify $sign(\hat{\boldsymbol{\beta}}_{\mathcal{S}_1}) = sign(\boldsymbol{\beta}^0_{\mathcal{S}_1})$. From equation (3.32) and **Lemma** 3.1, **Lemma** 3.5, we show that for some constant $c^* = \frac{1+\xi^2/6}{\tau_-}\big(1+o_p(1)\big)$,

$$\begin{aligned}
&\left\|(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right\|_\infty \leqslant \left|\!\left|\!\left|\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_\infty\left\|\left(-\lambda\tilde{z}_{\mathcal{S}_1} + R_{\mathcal{S}_1} + \frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right)\right\|_\infty \\
&\leqslant \sqrt{|\mathcal{S}_1|}\left|\!\left|\!\left|\left(-\frac{1}{n\phi}l_c^{(2)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1\mathcal{S}_1}\right)^{-1}\right|\!\right|\!\right|_2\left(\lambda\|\tilde{z}_{\mathcal{S}_1}\|_\infty + \|R_{\mathcal{S}_1}\|_\infty + \left\|\frac{1}{n\phi}l_c^{(1)}(\boldsymbol{\beta}^0)_{\mathcal{S}_1}\right\|_\infty\right) \\
&\leqslant \frac{\sqrt{|\mathcal{S}_1|}}{\tau_-}\left(\lambda + \mathcal{O}_p\left\{\frac{p_n\log p_n}{n}\right\}\big(1+o_p(1)\big) + \frac{1}{6}\xi^2\lambda\right) \\
&\leqslant \frac{\sqrt{|\mathcal{S}_1|}}{\tau_-}(1+\frac{1}{6}\xi^2)\lambda\big(1+o_p(1)\big) = c^*\sqrt{|\mathcal{S}_1|}\lambda \leqslant \min_{u\in\mathcal{S}_1}\left\{\left|\beta_u^0\right|\right\}.
\end{aligned}$$

Based on all the results presented above, we have $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^0)$ with probability tending to one. $\square$

4 Conclusions and Future Work

4.1 Conclusions and Future Work

In this dissertation, we present a covariate-dependent Gaussian graphical model (cdexGGM) for modeling dynamic network structures that vary with covariates via a novel parameterization, ensuring the positive definiteness of precision matrices at different values of covariates. Unlike some existing work, we allow the diagonal elements of the precision matrix to vary with covariates, and we propose to jointly estimate dynamic edge and vertex parameters through a likelihood framework. For networks with small dimensions, we conduct the maximum likelihood estimation and establish the asymptotic distribution of the estimator. This enables us to construct various statistical inference procedures on the underlying dynamic network. In the context of large networks, we employ the composite likelihood method with an ℓ_1 penalty to estimate sparse network structures. We provide the estimation error bound in ℓ_2 norm and establish the model selection consistency of the proposed method. The estimation procedure combines the coordinate descent algorithm and Broyden's method to update the covariate-varying edge and vertex parameters, respectively. The algorithm performs well in high-dimensional settings and yields superior or comparable performance to alternative methods.

Moreover, we also propose an extension of symmetry-constrained GGMs to covariate-dependent contexts, where the precision matrix with known symmetric patterns is modeled as a linear function of covariates, reflecting dynamic colored network structures. With proper assumptions, the overall precision matrix $\mathbf{\Sigma}_m^{-1}$ is guaranteed to be positive definite. We adopt composite likelihood estimation together with an ℓ_1 penalty to capture sparse graph patterns in large-dimensional data. Estimation procedures are developed separately for scenarios with fixed and covariate-dependent diagonal parameters. In the case of fixed diagonal entries, coordinate descent provides an efficient solution for large-scale data. For varying diagonal components, the estimation requires a hybrid approach incorporating both coordinate descent and Broyden's method to accommodate the extra computational complexity. Extensive simulation experiments demonstrate the effectiveness and practical feasibility of our proposed methods. In the context of theoretical results, consistency in both parameter and sign recovery is guaranteed for both fixed and varying diagonal parameters. Evidence from the application to real-world data supports the model's effectiveness and practical relevance in handling covariate-dependent biological networks with known symmetric patterns.

To facilitate broader application, we plan to develop an R package that implements the proposed likelihood-based estimation procedure for covariate-dependent GGMs and colored covariate-dependent GGMs.

4.2 Remarks and Comments

4.2.1 Sensitivity Analysis

We perform a sensitivity analysis on the influenza vaccination dataset to investigate the impact of γ on the estimation results. Specifically, we conduct the analysis across 5 levels of γ, namely $\gamma = 0, 0.25, 0.50, 0.75, 1$, with the tuning parameter λ determined based on minimum EBIC (Chen and Chen, 2008; Foygel and Drton, 2010; Gao and Song, 2010). Table 4.1 summarizes the number of estimated nonzero parameters under each scenario, and Figure 4.1 provides the sparsity structures across 4 levels of γ. Note that the estimation results are identical for $\gamma = 0$ and $\gamma = 0.25$ (e.g., the optimal λ selected by EBIC is the same for both cases). Therefore, we just present the sparsity structure for $\gamma = 0.25$, omitting $\gamma = 0$ for simplicity. Table 4.1 and Figure 4.1 indicate that the number of estimated nonzero parameters declines, leading to greater sparsity in the graph structure as γ increases from 0 to 1. This result reflects the expected property of EBIC as reported in Chen and Chen (2008); Gao and Song (2010). The following (4.1) gives the EBIC for Gaussian graphical models,

$$EBIC_\gamma = -2l_c(\hat{\boldsymbol{\beta}}_c) + df \log n + 4df\gamma \log p_n. \tag{4.1}$$

In this formulation, the first term corresponds to minus twice the composite log-likelihood, capturing the goodness-of-fit for the current model. The second component penalizes model complexity, while the third term promotes sparsity. More specifically, the parameter $\gamma \in [0, 1]$ determines the emphasis on sparsity, with higher values leading to smaller, simpler models. Previous studies (Chen and Chen, 2008, 2012) have demonstrated that the EBIC sacrifices a small degree of sensitivity but effectively manages the false positive instances. This feature makes it especially valuable for variable selection tasks involving moderate sample sizes with a high-dimensional covariate space, as often encountered in genomic studies. The consistency of the composite likelihood-based EBIC in selecting the true underlying model has been established by Gao and Song (2010). In general, as γ becomes larger, some moderate and weak interactions are filtered out, whereas strong interactions between genes can still be captured.

Table 4.1: Sensitivity analysis regarding γ applied to the influenza dataset

Matrix	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$	$\gamma = 1$
Q_0	604	604	500	380	306
Q_1	404	404	312	239	178

This table shows the number of nonzero estimated parameters across different levels of γ in EBIC.

4.2.2 Inference Methodology

In this section, we introduce a set of bootstrap hypothesis tests designed to evaluate off-diagonal elements, namely $H_0 : [\theta^0_{ij}]_h = 0$ versus $H_1 : [\theta^0_{ij}]_h \neq 0$ or $H_0 : \alpha^0_{ij} = 0$ versus $H_0 : \alpha^0_{ij} \neq 0$, and thus can be applied to determine the presence or absence of an edge.

- Bootstrap Wald test (studentized). For $b = 1, \cdots, B$, do:
  - Resample with replacement to generate the bootstrap samples.
  - Fit the proposed cdexGGM using the bootstrap sample and get the estimate $\hat{\theta}^{(b)}_{ij}$.
  - Construct the Wald-type statistic $T^{(b)*} = \frac{\hat{\theta}^{(b)*}_{ij}}{\hat{SE}(\hat{\theta}^{(b)*}_{ij})}$, where $\hat{\theta}^{(b)*}_{ij} = \hat{\theta}^{(b)}_{ij} - \overline{\hat{\theta}}^*_{ij} + \theta^0_{ij}$ and $\hat{SE}(\hat{\theta}^{(b)*}_{ij}) \approx \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)*}_{ij} - \overline{\hat{\theta}}^*_{ij})^2}$, $\overline{\hat{\theta}}^*_{ij} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)*}_{ij}$.

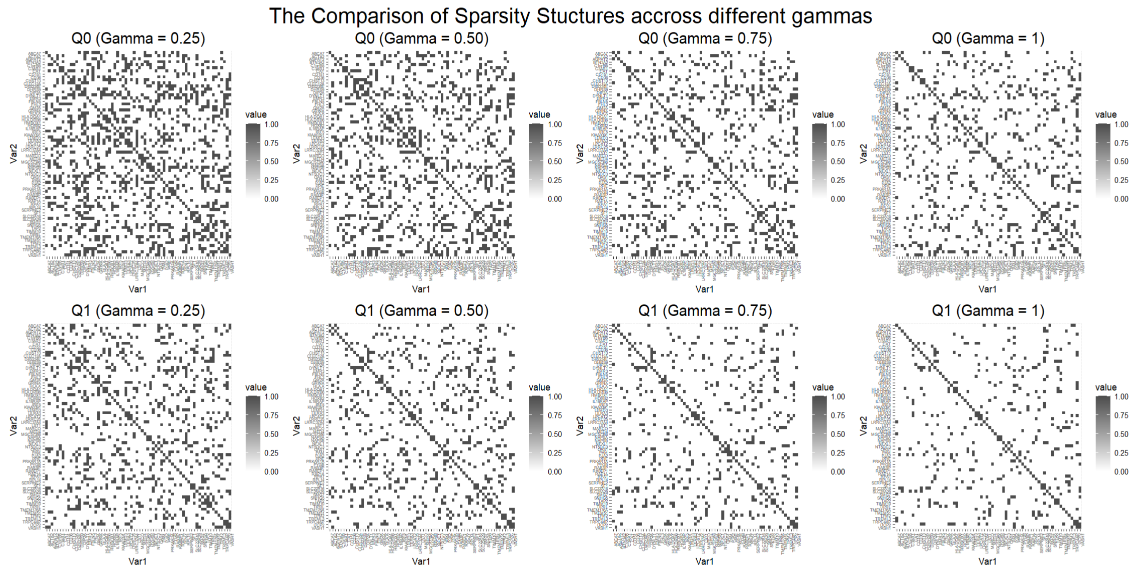


Figure 4.1: The comparison of estimated sparsity patterns across 4 levels of γ. Each subplot has 68 rows and 68 columns, representing the estimated elements of $\boldsymbol{Q}_0, \boldsymbol{Q}_1$ respectively. The corresponding cell is marked in black if the estimated entry is nonzero. Otherwise, it is white.

  - Compute the p-value $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{I}(|T^{*(b)}| \geqslant |T^{obs}|)$, where $T^{obs} = \frac{\hat{\theta}_{ij}^{obs}}{\hat{SE}(\hat{\theta}_{ij}^{*(b)})}$.

- An alternative testing procedure (Centered bootstrap test, see Horowitz (2019)). For $b = 1, \cdots, B$, do:

  - Resample with replacement to generate the bootstrap samples.

- Fit the proposed cdexGGM using the bootstrap sample and get the estimate $\hat{\theta}_{ij}^{(b)}$.
- Center the bootstrap distribution under the null, namely $\hat{\theta}_{ij}^{(b,c)} = \hat{\theta}_{ij}^{(b)} - \bar{\hat{\theta}}_{ij}^{(b)}$, where $\bar{\hat{\theta}}_{ij}^{(b)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_{ij}^{(b)}$.
- Compute the p-value $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{I}(|\hat{\theta}_{ij}^{(b,c)}| \geqslant |\hat{\theta}_{ij}^{obs}|)$, where $\hat{\theta}_{ij}^{obs}$ represents the original estimate.

If a covariate is categorical, we can additionally conduct a two-sample test on the partial correlations from the two groups. This allows us to assess whether the interaction strength differs significantly between groups. The procedure is outlined below:

- Two-sample non-Wald centered test. For $b = 1, \cdots, B$, do:
  - Resample with replacement to generate the bootstrap samples.
  - Fit the proposed cdexGGM using the bootstrap sample and compute the partial correlation matrix for each group.
  - Extract the bootstrap partial correlations $\hat{\rho}_{ij}^{1,(b)}, \hat{\rho}_{ij}^{2,(b)}$ between the (i, j)th pair of the interaction for the two groups, and record their difference $\delta_b = \hat{\rho}_{ij}^{1,(b)} - \hat{\rho}_{ij}^{2,(b)}$.

- Center bootstrap differences by defining $\delta_b^c = \delta_b - \overline{\delta}$, where $\overline{\delta} = \frac{1}{B}\sum_{b=1}^{B} \delta_b$.

- Compute two-sided p-value by exceedance counting:

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{I}(|\delta_b^c| \geqslant |\delta_{obs}|).$$

4.2.3 Comparison of the Influenza Vaccine Analyses

This section provides a comparative overview of the influenza vaccine analyses conducted in the two chapters. Table 4.2 summarizes their key similarities and differences. An upward arrow reflects a rising pattern over the observation period. A value of '0' indicates that the interaction is not present at any time point, and a downward arrow reflects a declining trend across the two weeks.

4.2.4 Remarks

We now present several remarks concerning the Min-Max scaling. The main reason for the transformation of covariates is to ensure the positive definiteness of the precision matrix, given any possible values of covariates. Specifically, the precision matrix $\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x_m})$ is assumed to be a linear combination of $\boldsymbol{Q}_0, \ldots, \boldsymbol{Q}_H$ with coefficients $1 - \sum_{h=1}^{H} \frac{x_m^{(h)}}{H}, x_m^{(1)}, \cdots, x_m^{(H)}$, as shown in (2.2). Assuming that $\boldsymbol{Q}_0, \cdots, \boldsymbol{Q}_H$

Table 4.2: Empirical Results Comparison Between Two Studies

Interactions	CdexGGM	Cdex Colored GGM
DIP2A & SERPINE2	$\uparrow$	$\uparrow$
GRINA & MARCO	$\uparrow$	$\uparrow$
PPIE & RAB3IP	$\uparrow$	$\uparrow$
ABCA7 & TRPC4AP	$\uparrow$	$\uparrow$
UCP2 & VASH1	$\uparrow$	$\uparrow$
PPIE & TREML3	$\uparrow$	$\uparrow$
D4S234E & OAS1	0	0
GRINA & RABEP1	0	0
C3AR1 & OAS1	$\downarrow$	0
KIAA0391 & RABEP1	$\downarrow$	0

are all positive definite matrices, the corresponding precision matrix is ensured to be positive definite for any covariates in the range $[0, 1]$. When the covariates are bounded, one can apply Min-Max scaling to transform the covariates into the range of $[0, 1]$.

To make Min-Max scaling less sensitive to outliers, hypothetical minimum and maximum values can be used instead of the true actual extremes of a given dataset. This results in a more robust transformation while maintaining values within the range $[0, 1]$. In particular, the hypothetical minimum and maximum values can be chosen sufficiently large, ensuring that the transformation remains stable despite adding or removing a data point, thereby preserving the consistency of model estimation and interpretation. In asymptotic theory, all theoretical results remain valid regardless of the type of transformation applied to the covariates, as long as the covariates are transformed to $[0, 1]$.

However, when covariates are unbounded, non-linear transformation techniques such as the quantile transformation can be applied. This strategy starts by determining the CDF of the original feature and then maps the original values to a uniform distribution. Nevertheless, this transformation reshapes the underlying distribution of the covariates and may be more suitable for a nonlinear relationship between the pre-transformed covariates and the underlying graph structure.

Lastly, we add more comments on the varying network structures defined in (2.6). Let $\mathcal{E}_0, \mathcal{E}_1, \cdots, \mathcal{E}_H$ be the true nonzero off-diagonal parameter index sets of $\boldsymbol{Q}_0, \boldsymbol{P}_1, \cdots, \boldsymbol{P}_H$, namely $\mathcal{E}_0 = \{(i, j) : \alpha_{ij}^0 \neq 0, i < j\}$, $\mathcal{E}_h = \{(i, j) : \left[\theta_{ij}^0\right]_h \neq 0, i < j\}$, where $i, j = 1, \cdots, p_n, h = 1, \cdots, H$ and $\alpha_{ij}^0, \left[\theta_{ij}^0\right]_1, \cdots, \left[\theta_{ij}^0\right]_H$ are the true values of $\alpha_{ij}, \left[\theta_{ij}\right]_1, \cdots, \left[\theta_{ij}\right]_H$. Then the parametrization (2.6) can be rewritten as

$$\boldsymbol{K}_m = \boldsymbol{\Sigma}_m^{-1} = \sum_{s: s \in \mathcal{E}_0 \cup V} \alpha_s \boldsymbol{T}^s + \sum_{h=1}^{H} x_m^{(h)} \Big\{ \sum_{u: u \in \mathcal{E}_h \cup V} [\theta_u]_h \boldsymbol{T}^u \Big\}. \tag{4.2}$$

Under this parametrization, the sparsity pattern varies at specific values of the covariates. For example, if we have two covariates $x^{(1)}$ and $x^{(2)}$, the precision matrix is given by $\boldsymbol{Q}_0 + x^{(1)}\boldsymbol{P}_1 + x^{(2)}\boldsymbol{P}_2$. Let $\mathcal{G}_0$, $\mathcal{G}_1$, and $\mathcal{G}_2$ be the graph structures corresponding to $\boldsymbol{Q}_0, \boldsymbol{P}_1$ and $\boldsymbol{P}_2$. There are four possible graph structures: the graph structure is $\mathcal{G}_0$ when the covariates belong to the set $A = \{x^{(1)} = 0, x^{(2)} = 0\}$; the graph structure is $\mathcal{G}_0 + \mathcal{G}_1$ for the set $B = \{x^{(1)} \in (0, 1], x^{(2)} = 0\}$; the graph structure is $\mathcal{G}_0 + \mathcal{G}_2$ for the set $C = \{x^{(1)} = 0, x^{(2)} \in (0, 1]\}$; the graph structure is $\mathcal{G}_0 + \mathcal{G}_1 + \mathcal{G}_2$ for the set $D = \{x^{(1)} \in (0, 1], x^{(2)} \in (0, 1]\}$. We could obtain these possible graph structures as long as sets A, B, C, and D each have a non-zero probability under the covariate distribution. On the other hand, if the covariates are uniformly distributed within the unit square, the graph structure would almost surely be $\mathcal{G}_0 + \mathcal{G}_1 + \mathcal{G}_2$. However, the strengths of the edges keep varying as a linear function of the covariates.

Bibliography

Ahmed, M. M., Dhanasekaran, A. R., Block, A., Tong, S., Costa, A. C. S., Stasko, M., and Gardiner, K. J. (2015). Protein dynamics associated with failed and rescued learning in the ts65dn mouse model of down syndrome. *PloS one*, 10(3):e0119491–.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232–253.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22(2):555–574.

Chojecki, A., Morgen, P., and Kołodziejek, B. (2025). Learning permutation symmetry of a gaussian vector with gips in r. *Journal of Statistical Software*, 112(7):1–38.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Danaher, P., Wang, P., and Witten, D. M. (2014). Joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76(2):373–397.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–1548.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2).

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 64(4):657–680.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 23:20200–22028.

Franco, L. M., Bucasas, K. L., Wells, J. M., Niño, D., Wang, X., Zapata, G. E., Arden, N., Renwick, A., Yu, P., Quarles, J. M., Bray, M. S., Couch, R. B., Belmont, J. W., and Shaw, C. A. (2013). Integrative genomic analysis of the human immune response to influenza vaccination. *eLife*, 2013(2):e00299–e00299.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332.

Gao, X. and Massam, H. (2015). Estimation of symmetry-constrained Gaussian graphical models: Application to clustered dense networks. *Journal of computational and graphical statistics*, 24(4):909–929.

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Gehrmann, H. (2011). Lattices of graphical gaussian models with symmetries. *Symmetry*, 3(3):653–678.

Giampaolo, A., Acampora, D., Zappavigna, V., Pannese, M., D'Esposito, M., Carè, A., Feiella, A., Stornaiuolo, A., Russo, G., Simeone, A., Boncinelli, E., and Peschle, C. (1989). Differential expression of human hox-2 genes along the anterior-posterior axis in embryonic central nervous system. *Differentiation (London)*, 40(3):191–197.

Gibberd, A. J. and Nelson, J. D. B. (2017). Regularized estimation of piecewise constant Gaussian graphical models: The group-fused graphical lasso. *Journal of computational and graphical statistics*, 26(3):623–634.

Graczyk, P., Ishi, H., Kołodziejek, B., and Massam, H. (2022). Model selection in the space of gaussian models invariant by symmetry. *The Annals of statistics*, 50(3):1747–.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (2015). *Handbook of Cluster Analysis*, volume 9 of *Chapman & Hall/CRC handbooks of modern statistical methods*. CRC Press, Milton, 1 edition.

Higuera, C., Gardiner, K. J., and Cios, K. J. (2015). Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one*, 10(6):e0129126.

Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual review of economics*, 11(1):193–224.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical Models with R*. Use R! Springer New York, New York, NY, 1st ed. 2012. edition.

Højsgaard, S. and Lauritzen, S. L. (2007). Inference in graphical gaussian models with edge and vertex symmetries with the grc package for r. *Journal of statistical software*, 23(6):1–26.

Højsgaard, S. and Lauritzen, S. L. (2008). Graphical gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 70(5):1005–1027.

Jensen, S. T., Johansen, S., and Lauritzen, S. L. (1991). Globally convergent algorithms for maximizing likelihood function. *Biometrika*, 78(4):867–877.

John, C. R., Watson, D., Barnes, M. R., Pitzalis, C., and Lewis, M. J. (2020). Spectrum: fast density-aware spectral clustering for single and multi-omic data. *BIOINFORMATICS*, 36(4):1159–1166.

Kolar, M., Parikh, A. P., and Xing, E. P. (2010a). On sparse nonparametric conditional covariance selection. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pages 559–566.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010b). Estimating time-varying networks. *The annals of applied statistics*, 4(1):94–123.

Kolar, M. and Xing, E. P. (2012). Estimating networks with jumps. *Electronic journal of statistics*, 6(none):2069–2106.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43–69.

Lauritzen, S. L. (1996). *Graphical models*. Oxford statistical science series ; 17. Clarendon Press, Oxford.

Li, Q., Gao, X., and Massam, H. (2020). Bayesian model selection approach for coloured graphical gaussian models. *Journal of statistical computation and simulation*, 90(14):2631–2654.

Li, Q., Sun, X., Wang, N., and Gao, X. (2021). Penalized composite likelihood for colored graphical gaussian models. *Statistical analysis and data mining*, 14(4):366–378.

Lin, S. M., Du, P., Huber, W., and Kibbe, W. A. (2008). Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic acids research*, 36(2):e11–e11.

Lindsay, B. (1988). Composite likelihood. *Contemporary Mathematics*, 80:221–239.

Malakooti, N., Pritchard, M. A., Chen, F., Yu, Y., Sgambelloni, C., Adlard, P. A., and Finkelstein, D. I. (2020). The long isoform of intersectin-1 has a role in learning and memory. *Frontiers in behavioral neuroscience*, 14:24–24.

Mallo, M., Wellik, D. M., and Deschamps, J. (2010). Hox genes and regional patterning of the vertebrate body plan. *Developmental biology*, 344(1):7–15.

Massam, H., Li, Q., and Gao, X. (2018). Bayesian precision and covariance matrix estimation for graphical gaussian models with edge and vertex symmetries. *Biometrika*, 105(2):371–388.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of statistics*, 34(3):1436–1462.

Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., and Montana, G. (2014). Estimating time-varying brain connectivity networks from functional mri time series. *NeuroImage (Orlando, Fla.)*, 103:427–443.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems: Vol. 14: Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference*.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019). Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2022). Bayesian covariate-dependent gaussian graphical models with varying structure. *Journal of Machine Learning Research*, 23(242):1–29.

Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362–20120362.

Niu, Y., Ni, Y., Pati, D., and Mallick, B. (2023). Covariate-assisted bayesian graph learning for heterogeneous data. *Journal of the American Statistical Association*, pages 1–25.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer Nature, New York, NY, second edition edition.

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., and Adebiyi, E. (2016). Clustering algorithms: Their application to gene expression data. *Bioinformatics and biology insights*, 10.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel.

Ranciati, S. and Roverato, A. (2024). On the application of gaussian graphical models to paired data problems. *Statistics and Computing*, 34(6).

Ranciati, S., Roverato, A., and Luati, A. (2021). Fused graphical lasso for brain networks with symmetries. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 70(5):1299–1322.

Raskutti, G., Yu, B., Wainwright, M. J., and Ravikumar, P. (2008). Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. 21.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic journal of statistics*, 5:935–980.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

Scott, J. (2019). Social network analysis, fourth edition. *Publisher: SAGE Publications Ltd*.

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K., and Kuijjer, M. L. (2017). Understanding tissue-specific gene regulation. *Cell Reports*, 21(4):1077–1088.

Srivastava, B., Błazejewska, P., Heßmann, M., Bruder, D., Geffers, R., Manuel, S., Gruber, A. D., and Schughart, K. (2009). Host genetic background strongly influences the response to influenza a virus infections. *PloS one*, 4(3):e4857–e4857.

Trammell, R. A. and Toth, L. A. (2008). Genetic susceptibility and resistance to influenza infection and disease in humans and mice. *Expert review of molecular diagnostics*, 8(4):515–529.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Tu, Y.-H., Juan, H.-F., and Huang, H.-C. (2023). Context-dependent gene regulatory network reveals regulation dynamics and cell trajectories using unspliced transcripts. *Briefings in Bioinformatics*, 24(2):bbac633.

Varin, C. (2008). On composite marginal likelihoods. *Advances in statistical analysis : AStA : a journal of the German Statistical Society*, 92(1):1–28.

Wang, J. and Kolar, M. (2014). Inference for sparse conditional precision matrices. arxiv preprint arxiv:1412.7638.

Weighill, D., Ben Guebila, M., Glass, K., Quackenbush, J., and Platig, J. (2022). Predicting genotype-specific gene regulatory networks. *Genome research*, 32(3):524–533.

Yang, J. and Peng, J. (2020). Estimating time-varying graphical models. *Journal of computational and graphical statistics*, 29(1):191–202.

Yin, J., Geng, Z., Li, R., and Wang, H. (2010). Nonparametric covariance model. *Statistica Sinica*, 20(1):469–479.

Zhang, J. and Li, Y. (2023). High-dimensional gaussian graphical regression models with covariates. *Journal of the American Statistical Association*, 118(543):2088–2100. PMID: 38143787.

Zhang, Z., Liu, C., Skogerbø, G., Zhu, X., Lu, H., Chen, L., Shi, B., Zhang, Y., Wang, J., Wu, T., and Chen, R. (2006). Dynamic changes in subgraph preference profiles of crucial transcription factors. *PLoS computational biology*, 2(5):383–391.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of machine learning research*, 7:2541–2563.

Zhong, C., Hau-wen, C., Dmytro, K., Dongwon, L., and Nishanth, S. (2017). Wearing many (social) hats: How different are your different social network personae? *Proceedings of the International AAAI Conference on Web and Social Media*, 11 (1):397–406.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine learning*, 80(2-3):295–319.