Joint Estimation and Inference for Multiple Multi-layered Gaussian Graphical Models

Subhabrata Majumdar

Abstract: The rapid development of high-throughput technologies has enabled generation of data from biological processes that span multiple layers, like genomic, proteomic or metabolomic data; and pertain to multiple sources, like disease subtypes or experimental conditions. In this work we propose a general statistical framework based on graphical models for horizontal (i.e. across conditions or subtypes) and vertical (i.e. across different layers containing data on molecular compartments) integration of information in such datasets. We start with decomposing the multi-layer problem into a series of two-layer problems. For each two-layer problem, we model the outcomes at a node in the lower layer as dependent on those of other nodes in that layer, as well as all nodes in the upper layer. Following the biconvexity of our objective function, this estimation problem decomposes into two parts, where we use neighborhood selection and subsequent refitting of the precision matrix to quantify the dependency of two nodes in a single layer, and use group-penalized least square estimation to quantify the directional dependency of two nodes in different layers. Finally, to test for differences in these directional dependencies across multiple sources, we devise a hypothesis testing procedure that utilizes already computed neighborhood selection coefficients for nodes in the upper layer. We establish theoretical results for the validity of this testing procedure and the consistency of our estimates, and also evaluate their performance through simulations.

Keywords: Data integration; Gaussian Graphical Models; Neighborhood selection; Group lasso

1 Introduction

1.1 Notations

We denote scalars by small letters, vectors by bold small letters and matrices by bold capital letters. For any matrix $\mathbf{A}$, $(\mathbf{A})_{ij}$ denote its element in the $(i,j)^{\text{th}}$ position. For $a, b \in \mathbb{N}$, we denote the set of all $a \times b$ real matrices by $\mathbb{M}(a,b)$. For any positive integer c, define $\mathcal{I}_c = \{1, \ldots, c\}$.

2 The Joint Multiple Multilevel Estimation Framework

2.1 Formulation

Suppose there are K independent datasets, each pertaining to an M-layered Gaussian Graphical Model (GGM). The k^{th} model has the following structure:

$$\begin{aligned} &\textit{Layer 1-} && \mathbb{D}_1^k = (D_{11}^k, \ldots, D_{1p_1}^k) \sim \mathcal{N}(0, \Sigma_1^k); \quad k \in \mathcal{I}_K, \\ &\textit{Layer } m \ (1 < m \leq M)\text{-} && \mathbb{D}_m^k = \mathbb{D}_{m-1}^k \mathbf{B}_m^k + \mathbb{E}_m^k, \text{ with } \mathbf{B}_m^k \in \mathbb{M}(p_{m-1}, p_m) \\ & && \text{and } \mathbb{E}_m^k = (E_{m1}^k, \ldots, E_{mp_m}^k) \sim \mathcal{N}(0, \Sigma_m^k); \quad k \in \mathcal{I}_K. \end{aligned}$$

We assume known structured sparsity patterns, denoted by $\mathcal{G}_m$ and $\mathcal{H}_m$, for the parameters of interest in the above model, i.e. the precision matrices $\Omega_m^k := (\Sigma_m^k)^{-1}$ and the regression coefficient matrices $\mathbf{B}_m^k$, respectively. These patterns provide information on horizontal dependencies across k for the corresponding parameters, and our goal here is to leverage them to estimate the full hierarchical structure of the network- specifically to obtain the undirected edges inside nodes of a single layer, and the directed edges between two successive layers through jointly estimating $\{\Omega_m^k\}$ and $\{\mathbf{B}_m^k\}$.

Consider a two-layer model, which is a special case of the above model with $M = 2$:

$$\mathbb{X}^k = (X_1^k, \ldots, X_p^k)^T \sim \mathcal{N}(0, \Sigma_x^k); \tag{2.1}$$

$$\mathbb{Y}^k = \mathbb{X}^k \mathbf{B}^k + \mathbb{E}^k; \quad \mathbb{E}^k = (E_1^k, \ldots, E_p^k)^T \sim \mathcal{N}(0, \Sigma_y^k); \tag{2.2}$$

$$\mathbf{B}^k \in \mathbb{M}(p,q), \quad \Omega_x^k = (\Sigma_x^k)^{-1}; \quad \Omega_y^k = (\Sigma_y^k)^{-1}; \tag{2.3}$$

where we want to estimate $\{(\Omega_x^k, \Omega_y^k, \mathbf{B}^k); k \in \mathcal{I}_K$ from data $\mathcal{Z}^k = \{(\mathbf{Y}^k, \mathbf{X}^k); \mathbf{Y}^k \in \mathbb{M}(n,q), \mathbf{X}^k \in \mathbb{M}(n,p), k \in \mathcal{I}_K\}$ in presence of known grouping structures $\mathcal{G}_x, \mathcal{G}_y, \mathcal{H}$ respectively and assuming $n_k = n$ for all $k \in \mathcal{I}_K$ for simplicity. We focus most of the theoretical discussion in the rest of the paper on jointly estimating $\Omega_y := \{\Omega_y^k\}$ and $\mathcal{B} := \{\mathbf{B}^k\}$. This is because for $M > 2$, within-layer undirected edges of any m^{th} layer $(m > 1)$ and between-layer directed edges from the $(m-1)^{\text{th}}$ layer to the m^{th} layer can be estimated from the corresponding data matrices in a similar fashion. On the other hand, parameters in the first layer are analogous to $\Omega_x := \{\Omega_x^k\}$ that are dependent only on $\{\mathbf{X}^k\}$, so any method for joint estimation of multiple graphical models can be used to estimate them (e.g. Guo et al. (2011); Ma and Michailidis (2016)). This provides all building blocks for estimating the full hierarchical structure of our M-layered multiple GGMs.

2.2 Algorithm

We assume an element-wise group sparsity pattern over k for the precision matrices Ω_x^k:

$$\mathcal{G}_x = \{\mathcal{G}_x^{ii'} : i \neq i'; i, i' \in \mathcal{I}_p\},$$

where each $\mathcal{G}_x^{ii'}$ is a partition of $\mathcal{I}_K$. Subsequently we use the Joint Structural Estimation Method (JSEM) (Ma and Michailidis, 2016) to estimate Ω_x, which first uses the group structure given by $\mathcal{G}_x$ in penalized nodewise regressions (Meinshausen and Bühlmann, 2006) to obtain neighborhood coefficients of each variable $X_i, i \in \mathcal{I}_p$, then fits a graphical lasso model over the combined support sets to obtain sparse estimates of the precision matrices:

$$\begin{aligned}
\widehat{\zeta}_i &= \arg\min_{\zeta_i} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|\mathbf{X}_i^k - \mathbf{X}_{-i}^k \boldsymbol{\zeta}_i^k\|^2 + \sum_{i' \leq i} \sum_{g \in \mathcal{G}_x^{ii'}} \eta_n \|\boldsymbol{\zeta}_{ii'}^{[g]}\| \right\} \\
\widehat{E}_x^k &= \{(i, i') : 1 \leq i < i' \leq p, \hat{\zeta}_{ii'}^k \neq 0 \text{ OR } \hat{\zeta}_{i'i}^k \neq 0\} \\
\widehat{\Omega}_x^k &= \arg\min_{\Omega_x^k \in \mathbb{S}_+(\hat{E}_x^k)} \left\{ \mathrm{Tr}(\widehat{\mathbf{S}}_x^k \Omega_x^k) - \log\det(\Omega_x^k) \right\}
\end{aligned} \tag{2.4}$$

where $\widehat{\mathbf{S}}_x^k := (\mathbf{X}^k)^T \mathbf{X}^k / n_k$.

For the precision matrices Ω_y^k we assume an element-wise sparsity pattern $\mathcal{G}_y$ defined in a similar manner as $\mathcal{G}_x$, while the sparsity pattern $\mathcal{H}$ for $\mathcal{B}$ is more general, each group $h \in \mathcal{H}$ being defined as:

$$h = \{(\mathcal{S}_p, \mathcal{S}_q, \mathcal{S}_K) : \mathcal{S}_p \subseteq \mathcal{I}_p, \mathcal{S}_q \subseteq \mathcal{I}_q, \mathcal{S}_K \subseteq \mathcal{I}_K\}; \quad \bigcup_{h \in \mathcal{H}} h = \mathcal{I}_p \times \mathcal{I}_q \times \mathcal{I}_K$$

We obtain sparse estimates of Ω_y and $\mathcal{B}$ by solving the following group-penalized least square minimization problem:

$$\begin{aligned}
\{\widehat{\mathcal{B}}, \widehat{\Theta}\} = \arg\min_{\mathcal{B}, \Theta} \Bigg\{ & \frac{1}{n} \sum_{j=1}^{q} \sum_{k=1}^{K} \|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \mathbf{B}_{-j}^k) \boldsymbol{\theta}_j^k - \mathbf{X}^k \mathbf{B}_j^k\|^2 \\
& + \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \gamma_n \|\boldsymbol{\theta}_{jj'}^{[g]}\| + \sum_{h \in \mathcal{H}} \lambda_n \|\mathbf{B}^{[h]}\| \Bigg\}
\end{aligned} \tag{2.5}$$

$$\begin{aligned}
\widehat{E}_y^k &= \{(j, j') : 1 \leq j < j' \leq q, \hat{\theta}_{jj'}^k \neq 0 \text{ OR } \hat{\theta}_{j'j}^k \neq 0\} \\
\widehat{\Omega}_y^k &= \arg\min_{\Omega_y^k \in \mathbb{S}_+(\hat{E}_y^k)} \left\{ \mathrm{Tr}(\widehat{\mathbf{S}}_y^k \Omega_y^k) - \log\det(\Omega_y^k) \right\}
\end{aligned} \tag{2.6}$$

The outcome of a node in the lower layer is thus modeled using all other nodes in that layer *and* nodes in the immediate upper layer, with their effects quantified using $\widehat{\boldsymbol{\theta}}_j^k$ and $\widehat{\mathbf{B}}_j^k$, respectively.

2.2.1 Alternating algorithm

The objective function in (2.5) is bi-convex, i.e. convex in $\mathcal{B}$ for fixed Θ, and vice-versa, but not jointly convex in $\{\mathcal{B}, \Theta\}$. Consequently, we use an alternating iterative algorithm to solve for $\{\mathcal{B}, \Theta\}$ that minimizes (2.5) by iteratively cycling between $\mathcal{B}$ and Θ, i.e. holding one set of parameters fixed and solving for the other, then alternating until convergence.

Choice of initial values plays a crucial role in the performance of this alternating algorithm.We choose the initial values $\{\mathbf{B}^{k(0)}\}$ by fitting separate lasso regression models for each column of the coefficient matrices:

$$\widehat{\mathbf{B}}_j^{k(0)} = \underset{\mathbf{B}_j^k \in \mathbb{R}^p}{\arg\min} \|\mathbf{Y}_j^k - \mathbf{X}^k \mathbf{B}_j^k\|^2 + \lambda_n \|\mathbf{B}_j^k\|_1; \quad j \in \mathcal{I}_q, k \in \mathcal{I}_K. \tag{2.7}$$

We obtain initial estimates of $\Theta_j, j \in \mathcal{I}_q$ by performing group-penalized nodewise regression on the residuals $\widehat{\mathbf{E}}^{k(0)} := \mathbf{Y}^k - \mathbf{X}^k \mathbf{B}_j^{k(0)}$:

$$\widehat{\Theta}_j^{(0)} = \underset{\Theta_j}{\arg\min} \frac{1}{n} \sum_{k=1}^{K} \|\widehat{\mathbf{E}}_j^{k(0)} - \widehat{\mathbf{E}}_{-j}^{k(0)} \boldsymbol{\theta}_j^k\|^2 + \gamma_n \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \|\boldsymbol{\theta}_{jj'}^{[g]}\|. \tag{2.8}$$

The steps of our full estimation procedure, which we call the Joint Multiple Multilayer Estimation (JMMLE) method, can thus be summarized in Algorithm 1.

Algorithm 1. (The JMMLE Algorithm)
1. Initialize $\widehat{\mathcal{B}}$ using (2.7).
2. Initialize $\widehat{\Theta}$ using (2.8).
3. Update $\widehat{\mathcal{B}}$ as:

$$\widehat{\mathcal{B}}^{(t+1)} = \underset{\substack{\mathbf{B}^k \in \mathbb{M}(p,q) \\ k \in \mathcal{I}_K}}{\arg\min} \left\{ \frac{1}{n} \sum_{j=1}^{q} \sum_{k=1}^{K} \|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \mathbf{B}_{-j}^k) \widehat{\boldsymbol{\theta}}_j^{k(t)} - \mathbf{X}^k \mathbf{B}_j^k\|^2 + \lambda_n \sum_{h \in \mathcal{H}} \|\mathbf{B}^{[h]}\| \right\} \tag{2.9}$$

4. Obtain $\widehat{\mathbf{E}}^{k(t+1)} := \mathbf{Y}^k - \mathbf{X}^k \mathbf{B}_j^{k(t)}, k \in \mathcal{I}_K$. Update $\widehat{\Theta}$ as:

$$\widehat{\Theta}_j^{(t+1)} = \underset{\Theta_j \in \mathbb{M}(q-1,K)}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|\widehat{\mathbf{E}}_j^{k(t+1)} - \widehat{\mathbf{E}}_{-j}^{k(t+1)} \boldsymbol{\theta}_j^k\|^2 + \gamma_n \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \|\boldsymbol{\theta}_{jj'}^{[g]}\| \right\} \tag{2.10}$$

5. Continue till convergence.
6. Calculate $\widehat{\Omega}_y^k, k \in \mathcal{I}_K$ using (2.6).

2.2.2 Tuning parameter selection

The nodewise regression step in the JSEM model (2.4) uses Bayesian Information Criterion (BIC) for tuning parameter selection. The step for updating $\{\Theta\}$, i.e. (2.10), in our

JMMLE algorithm is analogous to this procedure, hence we use BIC to select the penalty parameter γ_n. In our setting the BIC for a given γ and fixed $\mathcal{B}$ is given by:

$$\text{BIC}(\gamma; \mathcal{B}) = \text{Tr}\left(\mathbf{S}_y^k \widehat{\Omega}_{y,\gamma}^k\right) - \log\det\left(\widehat{\Omega}_{y,\gamma}^k\right) + \frac{\log n}{n} \sum_{k=1}^{K} |\widehat{E}_{y,\gamma}^k|$$

where γ in subscript indicates the corresponding quantity is calculated taking γ as the tuning parameter, and $\mathbf{S}_y^k := (\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}^k)^T(\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}^k)/n$. Every time $\widehat{\Theta}$ is updated in the JMMLE algorithm, we choose the optimal γ as the one with the smallest BIC over a fixed set of values $\mathcal{C}_n$. Thus for a fixed λ, our final choice of γ will be $\gamma^*(\lambda) = \arg\min_{\gamma \in \mathcal{C}_n} \text{BIC}(\gamma; \widehat{\mathcal{B}}_\lambda)$.

We use the High-dimensional BIC (HBIC) to select the other tuning parameter, λ:

$$\text{HBIC}(\lambda; \Theta) = \frac{1}{n} \sum_{j=1}^{q} \sum_{k=1}^{K} \|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \widehat{\mathbf{B}}_{-j,\lambda}^k)\boldsymbol{\theta}_j^k - \mathbf{X}^k \widehat{\mathbf{B}}_{j,\lambda}^k\|^2 +$$

$$\log(\log n) \frac{\log(pq)}{n} \sum_{k=1}^{K} \left(\|\mathbf{B}^k\|_0 + |\widehat{E}_{y,\gamma^*(\lambda)}^k|\right)$$

We choose an optimal λ as the minimizer of HBIC by training multiple JMMLE models using Algorithm 1 over a finite set of values $\lambda \in \mathcal{D}_n$: $\lambda^* = \arg\min_{\lambda \in \mathcal{D}_n} \text{HBIC}(\lambda, \widehat{\Theta}_{\gamma^*(\lambda)})$.

2.3 Properties of JMMLE estimators

We now provide theoretical results ensuring the convergence of our alternating algorithm, as well as the consistency of estimators obtained from the algorithm. We present statements of theorems in the main paper, giving detailed proofs and auxiliary results in the Appendix.

We introduce some notations that help establish the theorems that follow. Denote the true values of the parameters as $\Omega_{x0} = \{\Omega_{x0}^k\}, \Omega_{y0} = \{\Omega_{y0}^k\}, \Theta_0 = \{\Theta_{0j}\}, \mathcal{B}_0 = \{\mathcal{B}_0^k\}$. The notation $\text{supp}(\mathbf{A})$ indicates the non-zero edge set in a matrix (or vector)valued parameter $\mathbf{A} \in \mathbb{M}(A, B)$, i.e. $\text{supp}(\mathbf{A}) = \{(a, b) : (\mathbf{A})_{ab} \neq 0\}$. Sparsity levels of individual true parameters are indicated by $s_j := |\text{supp}(\Theta_j)|, b_k := |\text{supp}(\mathbf{B}^k)|$. Also define $S := \sum_{j=1}^{q} s_j, B := \sum_{k=1}^{K} b_k, s := \max_{j \in \mathcal{I}_q} s_j$. For positive real numbers A, B we write $A \gtrsim B$ if there exists $c > 0$ independent of A, B such that $A \geq cB$.

Our first result establishes the convergence of Algorithm 1 for any fixed realization of $\mathcal{X}$ and $\mathcal{E}$.

Theorem 2.1. *Suppose for any fixed $(\mathcal{X}, \mathcal{E})$, estimates in each iterate of Algorithm 1 are uniformly bounded by some quantity dependent on only p, q and n:*

$$\left\|(\widehat{\mathcal{B}}^{(t)}, \widehat{\Theta}_y^{(t)}) - (\mathcal{B}_0, \Theta_{y0})\right\|_F \leq R(p, q, n); \quad t \geq 1 \tag{2.11}$$

Then any limit point $(\mathcal{B}^\infty, \Theta_y^\infty)$ of the algorithm is a stationary point of the objective function, i.e. a point where partial derivatives along all coordinates are non-negative.

The next steps are to show that for random realizations of $\mathcal{X}$ and $\mathcal{E}$,

(a) Successive iterates lie in this non-expanding ball around the true parameters,

(b) The procedures in (2.7) and (2.8) ensure starting values that lie inside the same ball,

both with probability approaching 1 as $(p, q, n) \to \infty$.

To do so we break down the main problem into two subproblems. Take as $\boldsymbol{\beta} = (\text{vec}(\mathbf{B}^1)^T, \ldots, \text{vec}(\mathbf{B}^K)^T)^T$: any subscript or superscript on $\mathbf{B}$ being passed on to $\boldsymbol{\beta}$. Denote by $\widehat{\boldsymbol{\beta}}$ and $\widehat{\Theta}$ the generic estimators given by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pqK}}{\arg\min} \left\{ -2\boldsymbol{\beta}^T \widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^T \widehat{\boldsymbol{\Gamma}} \boldsymbol{\beta} + \lambda \sum_{h \in \mathcal{H}} \|\boldsymbol{\beta}^{[h]}\| \right\} \tag{2.12}$$

$$\widehat{\Theta}_j = \underset{\Theta_j \in \mathbb{M}(q-1,K)}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|\widehat{\mathbf{E}}_j^k - \widehat{\mathbf{E}}_{-j}^k \boldsymbol{\theta}_j^k\|^2 + \gamma \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \|\boldsymbol{\theta}_{jj'}^{[g]}\| \right\}; \quad j \in \mathcal{I}_q \tag{2.13}$$

where

$$\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} (\widehat{\mathbf{T}}^1)^2 \otimes \frac{(\mathbf{X}^1)^T \mathbf{X}^1}{n} & & \\ & \ddots & \\ & & (\widehat{\mathbf{T}}^K)^2 \otimes \frac{(\mathbf{X}^K)^T \mathbf{X}^K}{n} \end{bmatrix}; \quad \widehat{\boldsymbol{\gamma}} = \begin{bmatrix} (\widehat{\mathbf{T}}^1)^2 \otimes \frac{(\mathbf{X}^1)^T}{n} & & \\ & \vdots & \\ & & (\widehat{\mathbf{T}}^K)^2 \otimes \frac{(\mathbf{X}^K)^T}{n} \end{bmatrix} \begin{bmatrix} \text{vec}(\mathbf{Y}^1) \\ \vdots \\ \text{vec}(\mathbf{Y}^K) \end{bmatrix}$$

with

$$\hat{T}_{jj'}^k = \begin{cases} 1 \text{ if } j = j' \\ -\hat{\theta}_{jj'}^k \text{ if } j \neq j' \end{cases} \tag{2.14}$$

It is easy to see that solving for $\mathcal{B}$ in (2.5) given a fixed $\widehat{\Theta}$ is equivalent to solving (2.12).

We assume the following conditions on the true parameter versions $(\mathbf{T}_0^k)^2$, defined from Θ_0 similarly as (2.14):

(T1) The matrices $(\mathbf{T}^k)^2, k \in \mathcal{I}_K$ are diagonally dominant, i.e.

$$|t_{0,jj}^k| > \sum_{j' \neq j} |t_{0,jj'}^k|$$

for $j \in \mathcal{I}_q, k \in \mathcal{I}_K$.

Now we are in a position to establish the estimation consistency for the solution of (2.12), given random $(\mathcal{X}, \mathcal{E})$ and good enough estimators $\widehat{\Theta}$.

Theorem 2.2. *Assume random $(\mathcal{X}, \mathcal{E})$, and fixed $\widehat{\Theta}$ so that for $j \in \mathcal{I}_q$,*

$$\|\widehat{\Theta}_j - \Theta_{0,j}\|_F \leq v_\Theta = \eta_\Theta \sqrt{\frac{\log q}{n}}$$

for some $\eta_\Theta > 0$ dependent on Θ only. Then, given the choice of tuning parameter

$$\lambda_n \geq 4\sqrt{|h_{\max}|}\mathbb{R}_0\sqrt{\frac{\log(pq)}{n}}; \quad \mathbb{R}_0 := \max_{k\in\mathcal{I}_K} \mathbb{R}\left(v_\Theta, \Sigma_x^k, \Sigma_y^k\right)$$

the following hold

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 48\sqrt{|h_{\max}|}B\lambda/\psi_* \tag{2.15}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq 12\sqrt{B}\lambda/\psi_* \tag{2.16}$$

$$\sum_{h\in\mathcal{H}} \|\boldsymbol{\beta}^{[h]} - \boldsymbol{\beta}_0^{[h]}\| \leq 48B\lambda/\psi_* \tag{2.17}$$

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq 72B\lambda^2/\psi_* \tag{2.18}$$

with probability $\geq 1 - 12c_1\exp(c_2\log(pq) - 2\exp(-c_3 n)$, where $|h_{\max}| = \max_{h\in\mathcal{H}}|h|$ and

$$\psi_* = \frac{1}{2}\min_k\left[\Lambda_{\min}(\Sigma_{x0}^k)\left(\min_j \psi_j^k - d_k v_\Theta\right)\right], \text{ with } \psi_j^k := t_{0,jj}^k - \sum_{j'\neq j} t_{0,jj'}^k$$

and d_k being the maximum degree of $(\mathbf{T}_0^k)^2$.

To prove an equivalent result for the solution of (2.13), as well as the consistency of the final estimates $\widehat{\Omega}_y^k$ using their support sets, we need the following conditions.

(T2) For $k \in \mathcal{I}_K$, Ω_{y0}^k is diagonally dominant, i.e. $|\omega_{y0,jj}| > \sum_{j'\neq j}|\omega_{y0,jj'}|$ for $j \in \mathcal{I}_q$;
(T3) There exist constants c_0, d_0 such that for $k \in \mathcal{I}_K$,

$$0 < 1/c_0 \leq \Lambda_{\min}(\Sigma_{y0}^k) \leq \Lambda_{\max}(\Sigma_{y0}^k) \leq 1/d_0 < \infty$$

Given these, we establish the required consistency results.

Theorem 2.3. *Consider any deterministic $\widehat{\mathcal{B}}$ that satisfy the following bound*

$$\|\widehat{\mathbf{B}}^k - \mathbf{B}_0^k\|_1 \leq v_\beta = \eta_\beta\sqrt{\frac{\log(pq)}{n}}$$

Then, for sample size $n \gtrsim \log(pq)$ and choice of tuning parameter $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0$, there exist constants $c_1, c_2, c_3, c_4, c_5 > 0$ such that the following holds

$$\frac{1}{K}\sum_{k=1}^K \|\widehat{\Omega}_y^k - \Omega_y^k\|_F \leq O\left(\mathbb{Q}_0\sqrt{\frac{|g_{\max}|S}{K}}\right) \tag{2.19}$$

with probability $\geq 1 - 1/p^{\tau_1-2} - 6c_1\exp[-c_2\log(pq)] - 2\exp(-c_3 n) - 6c_4\exp[-c_5\log(pq)]$.

Finally we ensure that the starting values are good enough.

Theorem 2.4. *Consider the starting values as derived in (2.7) and (2.8). For sample size $n \gtrsim \log(pq)$, there exist constants $d_1, d_2, d_3 > 0$ such that for*

$$\lambda \geq 4d_2 \max_{k \in \mathcal{I}_K} \left\{ [\Lambda_{\max}(\Sigma^k_{x0})\Lambda_{\max}(\Sigma^k_{y0})]^{1/2} \right\} \sqrt{\frac{\log(pq)}{n}}$$

we have $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\|_1 \leq 64B\lambda/\psi^*$ with probability $\geq 1 - 6d_1 \exp(-d_2 \log(pq)) - 2\exp(d_3 n)$. Further, for $\gamma \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0$ we have

$$\|\widehat{\Theta}_j - \Theta_{0,j}\|_F \leq 12\sqrt{s_j}\gamma/\psi \tag{2.20}$$

$$\sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\hat{\boldsymbol{\theta}}^{[g]}_{jj'} - \boldsymbol{\theta}^{[g]}_{0,jj'}\| \leq 48 s_j \gamma/\psi \tag{2.21}$$

with probability $\geq 1 - 1/p^{\tau_1 - 2} - 6c_1 \exp[-(c_2^2 - 1)\log(pq)] - 2\exp(-c_3 n) - 6c_4 \exp[-(c_5^2 - 1)\log(pq)]$.

Putting everything together, estimation consistency for the limit points of Algorithm 1 given our choice of starting values is immediate.

Corollary 2.5. *Assume the conditions (T1)-(T3), and starting values $\{\mathcal{B}^{(0)}, \Theta^{(0)}\}$ obtained using (2.7) and (2.8), respectively. Then, for random realizations of $\mathcal{X}, \mathcal{E}$,*

(I) For the choice of λ

$$\lambda \geq 4 \max \left[d_2 \max_{k \in \mathcal{I}_K} \left\{ [\Lambda_{\max}(\Sigma^k_{x0})\Lambda_{\max}(\Sigma^k_{y0})]^{1/2} \right\}, \sqrt{|h_{\max}|}\mathbb{R}_0 \right] \sqrt{\frac{\log(pq)}{n}}$$

we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq \max\left\{ 48\sqrt{|h_{\max}|} + 64 \right\} \frac{B\lambda}{\psi_*}$$

with probability $\geq$ **tbd**.

(II) For $\gamma \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0$, (2.20) and (2.21) hold with probability $\geq$ **tbd**.

(III) For $\gamma = 4\sqrt{|g_{\max}|}\mathbb{Q}_0$, (4.1) holds with probability $\geq$ **tbd**.

3 Hypothesis testing in multilayer models

In this section, we lay out a framework for hypothesis testing in our proposed joint multilayer structure. Present literature in high-dimensional hypothesis testing either focuses on testing for simularities in the within-layer connections of single-layer networks (Cai and Liu, 2016; Liu, 2017), or coefficients of single response penalized regression (Mitra and Zhang, 2016; van de Geer et al., 2014; Zhang and Zhang, 2014). However a method for testing *between-layer* connections in a multilayer setup is yet to be proposed.

There are two main challenges in doing the above: firstly the need for debiased estimators and assumptions on the design matrix required for the same, and secondly the

dependency among response nodes translating into the need for controlling False Discovery Rate (FDR) while simultaneously testing for such hypotheses. In Section 3.1 we propose a debiased estimator for rows of the coefficient matrix estimates $\mathbf{B}^k$ that makes use of already computed (using JSEM) nodewise regression coefficients in the upper layer, and establish asymptotic properties of scaled version of them. Section 3.2 is devoted to pairwise testing, where we asssume $K = 2$, and propose asymptotic global tests for detecting differential effects of a variable in the upper layer, as well as pairwise simultaneous tests for detecting elementwise difference in the coefficient matrices across k.

3.1 Debiased estimators and asymptotic normality

In this setup, define the desparsified estimate of $\mathbf{b}_{0i}^k$ as

$$\widehat{\mathbf{c}}_i^k = \widehat{\mathbf{b}}_i^k + \frac{1}{nt_i^k}\left(\mathbf{X}_i^k - \mathbf{X}_{-i}^k\widehat{\boldsymbol{\zeta}}_i^k\right)^T(\mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k) \tag{3.1}$$

for $k = 1, 2$, where $t_i^k = (\mathbf{X}_i^k - \mathbf{X}_{-i}^k\widehat{\boldsymbol{\zeta}}_i^k)^T\mathbf{X}_{-i}^k/n$. Then we have the asymptotic joint distribution of a scaled version of the debiased coefficients for the i^{th} predictor effect.

Theorem 3.1. *Define* $\widehat{s}_i^k = \sqrt{\|\mathbf{X}_i^k - \mathbf{X}_{-i}^k\widehat{\boldsymbol{\zeta}}_i^k\|^2/n}$, *and* $m_i^k = \sqrt{n}t_i^k/\widehat{s}_i^k$. *Assume the following:*

(C1) *For the X-neighborhood estimators we have* $\|\widehat{\boldsymbol{\zeta}}^k - \boldsymbol{\zeta}_0^k\|_1 \le v_\zeta = \eta_\zeta\sqrt{\frac{\log p}{n}}$.

(C2) *The precision matrix estimators satisfy*

$$\left\|(\widehat{\Omega}_y^k)^{1/2} - (\Omega_y^k)^{1/2}\right\|_\infty \le v_\Omega = \eta_\Omega\sqrt{\frac{\log q}{n}}$$

(C3) $\|\widehat{\mathbf{B}}^k - \mathbf{B}_0^k\|_1 \le v_\beta$, *where* $v_\beta = \eta_\beta\sqrt{\frac{\log(pq)}{n}}$ *with* η_β *depending on* $\mathcal{B}$ *only.*

Then for the debiased estimators in (3.1) *and sample size satisfying* $n \gtrsim \log(pq), \log p = o(n^{1/2}), \log q = o(n^{1/2})$ *we have*

$$\begin{bmatrix}\widehat{\Omega}_y^1 & \\ & \widehat{\Omega}_y^2\end{bmatrix}^{1/2}\begin{bmatrix}m_i^1(\widehat{\mathbf{c}}_i^1 - \mathbf{b}_{0i}^1) & \\ & m_i^2(\widehat{\mathbf{c}}_i^2 - \mathbf{b}_{0i}^2)\end{bmatrix} \sim \mathcal{N}_{2q}(\mathbf{0}, \mathbf{I}) + o_P(1) \tag{3.2}$$

3.2 Test formulation

Based on the theorem, we have the global test for the effect of the i^{th} X-covariate.

Algorithm 2. (Global test for $H_0^i : \mathbf{b}_{0i}^1 = \mathbf{b}_{0i}^2$ at level $\alpha, 0 < \alpha < 1$)

1. Obtain the debiased estimators $\widehat{\mathbf{c}}_i^1, \widehat{\mathbf{c}}_i^2$ using (3.1).
2. Calculate the test statistic

$$D_i = (\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2)^T\left(\frac{(\widehat{\Omega}_y^1)^{-1}}{(m_i^1)^2} + \frac{(\widehat{\Omega}_y^2)^{-1}}{(m_i^2)^2}\right)^{-1}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2)$$

3. Reject H_0^i if $D_i \ge \chi^2_{q,1-\alpha}$.

Given the null hypothesis is rejected, we now consider the multiple testing problem of simultaneously testing for all entrywise differences, i.e. testing

$$H_0^{ij} : b_{0ij}^1 = b_{0ij}^2 \quad \text{vs.} \quad H_1^{ij} : b_{0ij}^1 \neq b_{0ij}^2$$

for $j \in \mathcal{I}_q$. Here we use the test statistic

$$d_{ij} = \frac{\widehat{c}_{ij}^1 - \widehat{c}_{ij}^2}{\sqrt{\tau_{ij}^1/(m_i^1)^2 + \tau_{ij}^2/(m_i^2)^2}} \tag{3.3}$$

with τ_{ij}^k being the $(i,j)^{\text{th}}$ element of $(\widehat{\Omega}_y^k)^{-1}$, for $k = 1, 2$.

Now consider tests where H_0^{ij} is rejected if $|d_{ij}| > \tau$. We denote $\mathcal{H}_0^i = \{j : b_{ij}^1 = b_{ij}^2\}$ and define the false discovery proportion (FDP) and false discovery rate (FDR) for these tests as follows:

$$FDP(\tau) = \frac{\sum_{j \in \mathcal{H}_0^i} \mathbb{I}(|d_{ij}| \geq \tau)}{\max\left\{\sum_{j \in \mathcal{I}_q} \mathbb{I}(|d_{ij}| \geq \tau), 1\right\}} \quad FDR(\tau) = \mathbb{E}[FDP(\tau)]$$

For a pre-specified level α, we choose a threshold that ensures both FDP and FDR $\leq \alpha$ using the Benjamini-Hochberg (BH) procedure. The procedure for FDR control is now given by Algorithm 3.

Algorithm 3. (Simultaneous tests for $H_0^{ij} : b_{0ij}^1 = b_{0ij}^2$ at level $\alpha, 0 < \alpha < 1$)
1. Calculate the pairwise test statistics d_{ij} using (3) for $j \in \mathcal{I}_q$.
2. Obtain the threshold

$$\hat{\tau} = \inf\left\{\tau \in \mathbb{R} : 1 - \Phi(\tau) \leq \frac{\alpha}{2q} \max\left(\sum_{j \in \mathcal{I}_q} \mathbb{I}(|d_{ij}| \geq \tau), 1\right)\right\}$$

3. For $j \in \mathcal{I}_q$, reject H_0^{ij} if $|d_{ij}| \geq \hat{\tau}$.

This procedure maintains FDR and FDP asymptotically at a pre-specified level $\alpha \in (0, 1)$ under weak dependence conditions.

Theorem 3.2. *Suppose* $\mu_j = b_{0,ij}^1 - b_{0,ij}^2, \sigma_j^2 = n\mathbb{E}(\tau_{ij}^1/(m_i^1)^2 + \tau_{ij}^2/(m_i^2)^2)$. *Assume the following holds as* $(n, q) \to \infty$

$$\left|\left\{j \in \mathcal{I}_q : |\mu_j/\sigma_j| \geq 4\sqrt{\log q/n}\right\}\right| \to \infty \tag{3.4}$$

Now Consider the conditions **C1, C1***

If (C1) is satisfied, then the following holds when $\log q = O(n^\xi), 0 < \xi < 3/23$*:*

$$\frac{FDP(\hat{\tau})}{(|\mathcal{H}_0^i|/q)\alpha} \xrightarrow{P} 1; \quad \lim_{n,q \to \infty} \frac{FDR(\hat{\tau})}{(|\mathcal{H}_0^i|/q)\alpha} = 1 \tag{3.5}$$

Further, if (C1) is satisfied, then (3.5) holds for* $\log q = o(n^{1/3})$.

The condition (3.4) is essential for FDR control in a diverging parameter space (Liu, 2017; Liu and Shao, 2014), while the dependence conditions (C1) and (C1*) control the amount of correlation between variables in the Y-layer.

Remark 1. Following Liu and Shao (2014), a version of Algorithm 3, where the null distribution is calibrated using bootstrap instead of normal approximation, gives asymptotic FDR control under (C1*) and $\log q = o(n^{1/2})$. We believe it is possible to obtain (3.5) under the weaker condition (C1) for $\log q = o(n^{1/2})$ by extending the framework of Liu (2017) that performs multiple testing in multiple (single layer) GGMs, with the added advantage of being generalizable to the case of $K > 2$. However, this requires a significant amount of theoretical analysis, and we leave it for future research.

Remark 2. **does within-group thresholding with FDR control for $K = 1$**

4 Misc

To prove the results in this section, we use a reparametrization of the neighborhood coefficients at the lower level. Specifically, notice that for $j \in \mathcal{I}_q, k \in \mathcal{I}_K$, the corresponding summand in $f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta)$ can be rearranged as

$$\begin{aligned}\|\mathbf{Y}_j^k - \mathbf{X}^k\mathbf{B}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k\mathbf{B}_{-j}^k)\boldsymbol{\theta}_j^k\|^2 &= \|\mathbf{Y}_j^k - \mathbf{Y}_{-j}^k\boldsymbol{\theta}_j^k - (\mathbf{X}^k\mathbf{B}_j^k - \mathbf{X}^k\mathbf{B}_{-j}^k\boldsymbol{\theta}_j^k)\|^2 \\ &= \|(\mathbf{Y} - \mathbf{X}\mathbf{B})\mathbf{T}_j^k\|^2\end{aligned}$$

where

$$T_{jj'}^k = \begin{cases} 1 \text{ if } j = j' \\ -\theta_{jj'}^k \text{ if } j \neq j' \end{cases}$$

Thus, with $\mathbf{T}^k := (\mathbf{T}_j^k)_{j \in \mathcal{I}_q}$, we have

$$f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{K}\|(\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}^k)\mathbf{T}_j^k\|^2 = \frac{1}{n}\sum_{k=1}^{K}\|\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}^k)\mathbf{T}^k\|_F^2 = \sum_{k=1}^{K}\mathrm{Tr}(\mathbf{S}^k(\mathbf{T}^k)^2)$$

where $\mathbf{S}^k = (1/n)(\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}^k)(\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}^k)^T$ is the sample covariance matrix.

Theorem 4.1. *Assume fixed $\mathcal{X}, \mathcal{E}$ and deterministic $\widehat{\mathcal{B}} = \{\widehat{\mathbf{B}}^k\}$. Also for $k = 1, \ldots, K$,*
(T1) *$\|\widehat{\mathbf{B}}^k - \mathbf{B}_0^k\|_1 \leq v_\beta$, where $v_\beta = \eta_\beta\sqrt{\frac{\log(pq)}{n}}$ with $\eta_\beta \geq 0$ depending on $\mathcal{B}$ only;*
(T2) *Denote $\widehat{\mathbf{E}}^k = \mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k, k \in \mathcal{I}_K$. Then for all $j \in \mathcal{I}_q$,*

$$\frac{1}{n}\left\|(\widehat{\mathbf{E}}_{-j}^k)^T\widehat{\mathbf{E}}^k\mathbf{T}_{0,j}^k\right\|_\infty \leq \mathbb{Q}\left(v_\beta, \Sigma_x^k, \Sigma_y^k\right)$$

where $\mathbb{Q}\left(v_\beta, \Sigma_x^k, \Sigma_y^k\right)$ is a $O(\sqrt{\log(pq)/n})$ deterministic function of $\mathcal{B}, \Sigma_x^k$ and Σ_y^k.
(T3) *Denote $\widehat{\mathbf{S}}^k = (\widehat{\mathbf{E}}^k)^T\widehat{\mathbf{E}}^k/n$. Then $\widehat{\mathbf{S}}^k \sim RE(\psi^k, \phi^k)$ with $Kq\phi \leq \psi/2$ where $\psi = \min_k \psi^k, \phi = \max_k \phi^k$;*
(T4) *Assumption (A2) holds for Σ_y^k.*

Then, given the choice of tuning parameter

$$\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0; \quad \mathbb{Q}_0 := \max_{k\in\mathcal{I}_K} \mathbb{Q}\left(v_\beta, \Sigma_x^k, \Sigma_y^k\right)$$

the following holds

$$\frac{1}{K}\sum_{k=1}^{K} \|\widehat{\Omega}_y^k - \Omega_y^k\|_F \leq O\left(\mathbb{Q}_0\sqrt{\frac{|g_{\max}|S}{K}}\right)$$

where $|g_{\max}|$ is the maximum group size.

When $\mathcal{X}$ and $\mathcal{E}$ are random, the following propositions ensures that conditions (T2) and (T3) hold with probabilities approaching to 1.

Proposition 4.2. *Consider deterministic $\widehat{\mathcal{B}}$ satisfying assumption (T1). Then for sample size $n \gtrsim \log(pq)$ and $k \in \mathcal{I}_K$,*

1. *$\widehat{\mathbf{S}}^k$ satisfies the RE condition: $\widehat{\mathbf{S}}^k \sim RE(\psi^k, \phi^k)$, where*

$$\psi^k = \frac{\Lambda_{\min}(\Sigma_x^k)}{2}; \quad \phi^k = \frac{\psi^k \log p}{n} + 2v_\beta c_2[\Lambda_{\max}(\Sigma_x^k)\Lambda_{\max}(\Sigma_y^k)]^{1/2}\sqrt{\frac{\log(pq)}{n}}$$

   with probability $\geq 1 - 6c_1\exp[-(c_2^2-1)\log(pq)] - 2\exp(-c_3 n), c_1, c_3 > 0, c_2 > 1$.

2. *The following deviation bound is satisfied for any $j \in \mathcal{I}_q$*

$$\left\|\frac{1}{n}(\widehat{\mathbf{E}}_{-j}^k)^T\widehat{\mathbf{E}}^k\mathbf{T}_{0,j}^k\right\|_\infty \leq \mathbb{Q}\left(v_\beta, \Sigma_x^k, \Sigma_y^k\right)$$

   with probability $\geq 1 - 1/p^{\tau_1-2} - 6c_1\exp[-(c_2^2-1)\log(pq)] - 6c_4\exp[-(c_5^2-1)\log(pq)], c_4 > 0, c_5 > 1$, where

$$\mathbb{Q}\left(v_\beta, \Sigma_x^k, \Sigma_y^k\right) = 2v_\beta^2 V_x^k + 4v_\beta c_2[\Lambda_{\max}(\Sigma_x^k)\Lambda_{\max}(\Sigma_y^k)]^{1/2}\sqrt{\frac{\log(pq)}{n}} + c_5\left[\Lambda_{\max}(\Sigma_{y,-j}^k)\sigma_{y,j,-j}^k\right]^{1/2}\sqrt{\frac{\log q}{n}}$$

   with $\sigma_{y,j,-j}^k = \mathbb{V}(E_j - \mathbb{E}_{-j}\boldsymbol{\theta}_{0,j})$, and

$$V_x^k = \sqrt{\frac{\log 4 + \tau_1 \log p}{c_x^k n}} + \max_i \sigma_{x,ii}^k; \quad c_x^k = \left[128(1 + 4\Lambda_{\max}(\Sigma_x))^2 \max_i(\sigma_{x,ii})^2\right]^{-1}$$

The error bounds for $\widehat{\Omega}_y^k, k \in \mathcal{I}_K$ follow immediately from the above two results.

Corollary 4.3. *Consider any deterministic $\widehat{\mathcal{B}}$ that satisfy the following bound*

$$\|\widehat{\mathbf{B}}^k - \mathbf{B}_0^k\|_1 \leq v_\beta = \eta_\beta \sqrt{\frac{\log(pq)}{n}}$$

Then, for sample size $n \gtrsim \log(pq)$ and choice of tuning parameter $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0$, there exist constants $c_1, c_3, c_4 > 0, c_2, c_5 > 1$ such that the following holds

$$\frac{1}{K}\sum_{k=1}^{K} \|\widehat{\Omega}_y^k - \Omega_y^k\|_F \leq O\left(\mathbb{Q}_0 \sqrt{\frac{|g_{\max}|S}{K}}\right) \tag{4.1}$$

with probability $\geq 1 - 1/p^{\tau_1 - 2} - 6c_1 \exp[-(c_2^2 - 1)\log(pq)] - 2\exp(-c_3 n) - 6c_4 \exp[-(c_5^2 - 1)\log(pq)]$.

Discuss tighter bound compared to vanilla JSEM

After providing the error bounds for solutions to the subproblem (2.13), we concentrate on the subproblem (2.12). Following a similar strategy, we first get error bounds for $\widehat{\boldsymbol{\beta}}$ assuming everything else fixed.

Theorem 4.4. *Assume fixed $\mathcal{X}, \mathcal{E}$, and deterministic $\widehat{\Theta} = \{\widehat{\Theta}_j\}$, so that for $j \in \mathcal{I}_q$,*

(B1) $\|\widehat{\Theta}_j - \Theta_{0,j}\|_F \leq v_\Theta \sqrt{\frac{\log q}{n}}$ *for some v_Θ dependent on Θ.*

(B2) *Denote $\widehat{\mathbf{\Gamma}}^k = (\widehat{\mathbf{T}}^k)^2 \otimes (\mathbf{X}^k)^T\mathbf{X}^k/n, \widehat{\boldsymbol{\gamma}}^k = (\widehat{\mathbf{T}}^k)^2 \otimes (\mathbf{X}^k)^T\mathbf{Y}^k/n$. Then the deviation bound holds:*

$$\left\|\widehat{\boldsymbol{\gamma}}^k - \widehat{\mathbf{\Gamma}}^k \boldsymbol{\beta}_0\right\|_\infty \leq \mathbb{R}(v_\Theta, \Sigma_x^k, \Sigma_y^k)\sqrt{\frac{\log(pq)}{n}}$$

where $\mathbb{R}\left(v_\Theta, \Sigma_x^k, \Sigma_y^k\right)$ is a $O(1)$ deterministic function of Θ, Σ_x^k and Σ_y^k.

(B3) $\widehat{\mathbf{\Gamma}} \sim RE(\psi_*, \phi_*)$ *with $Kpq\phi_* \leq \psi_*/2$.*

Then, given the choice of tuning parameter

$$\lambda_n \geq 4\sqrt{|h_{\max}|}\mathbb{R}_0\sqrt{\frac{\log(pq)}{n}}; \quad \mathbb{R}_0 := \max_{k \in \mathcal{I}_K} \mathbb{R}\left(v_\Theta, \Sigma_x^k, \Sigma_y^k\right)$$

the following holds

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 48\sqrt{|h_{\max}|}s_\beta\lambda_n/\psi^* \tag{4.2}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq 12\sqrt{s}_\beta\lambda_n/\psi^* \tag{4.3}$$

$$\sum_{h \in \mathcal{H}} \|\boldsymbol{\beta}^{[h]} - \boldsymbol{\beta}_0^{[h]}\| \leq 48 s_\beta \lambda_n/\psi^* \tag{4.4}$$

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\widehat{\mathbf{\Gamma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq 72 s_\beta \lambda_n^2/\psi^* \tag{4.5}$$

Next we verify that conditions (B2) and (B3) hold with high probability given fixed $\widehat{\Theta}$.

Proposition 4.5. *Consider deterministic $\widehat{\Theta}$ satisfying assumption (B1). Assume that the matrices $(\widehat{\mathbf{T}}^k)^2, k \in \mathcal{I}_K$ are diagonally dominant. Then for sample size $n \gtrsim \log(pq)$,*

1. $\widehat{\mathbf{\Gamma}}$ *satisfies the RE condition:* $\widehat{\mathbf{\Gamma}} \sim RE(\psi_*, \phi_*)$, *where*

$$\psi_* = \min_k \psi^k \left(\min_i \psi_t^i - dv_\Theta \right), \phi_* = \max_k \phi^k \left(\min_i \phi_t^i + dv_\Theta \right)$$

with probability $\geq 1 - 2\exp(c_3 n), c_3 > 0$.

2. *The deviation bound in (B2) is satisfied with probability* $\geq 1 - 12c_1 \exp[(c_2^2 - 1)\log(pq)], c_1 > 0, c_2 > 1$, *where*

$$\mathbb{R}\left(v_\Theta, \Sigma_x^k, \Sigma_y^k\right) = c_2 \sqrt{\Lambda_{\max}(\Sigma_x^k)} \left(dv_\Theta \Lambda_{\min}(\Sigma_y^k) + \frac{1}{\Lambda_{\min}(\Sigma_y^k)} \right)$$

We now put both the pieces together, and prove that our alternating algorithm results in a solution sequence $\{\widehat{\mathcal{B}}^{(r)}, \widehat{\Theta}^{(r)}\}, r = 1, 2, \ldots$ that lies uniformly within a non-expanding ball around the true parameter values.

A Proofs

Proof of Theorem 4.1. The proof has three parts, where we prove the consistency of the neighborhood regression coefficients, selection of edge sets, and finally the refitting step, respectively. This is the same structure as the proof of Theorem 1 in Ma and Michailidis (2016), where they prove consistency of the (single layer) JSEM estimates. The derivation of the first part is different from that in the JSEM proof, which we shall show in detail (in the proof of Proposition A.1). The second and third parts follow similar lines, incorporating the updated quantities from the first part. For these we provide outlines and leave the details to the reader.

Step 1: consistency of neighborhood regression. The following proposition establishes error bounds for estimated neighborhood coefficients in the Y-network.

Proposition A.1. *Consider the estimation problem in (2.13) and take* $\gamma_n \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0$. *Given the conditions (T2) and (T3) hold, for any solution of (2.13) we shall have*

$$\|\widehat{\Theta}_j - \Theta_{0,j}\|_F \leq 12\sqrt{s_j}\gamma_n/\psi \tag{A.1}$$

$$\sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\hat{\boldsymbol{\theta}}_{jj'}^{[g]} - \boldsymbol{\theta}_{0,jj'}^{[g]}\| \leq 48 s_j \gamma_n/\psi \tag{A.2}$$

Also denote the non-zero support of $\widehat{\Theta}_j$ *by* $\widehat{\mathcal{S}}_j$, *i.e.* $\widehat{\mathcal{S}}_j = \{(j', g) : \hat{\boldsymbol{\theta}}_{jj'}^{[g]} \neq \mathbf{0}\}$. *Then*

$$|\widehat{\mathcal{S}}_j| \leq 128 s_j/\psi \tag{A.3}$$

Step 2: Edge set selection. We denote the selected edge set for the k^{th} Y-network by $\hat{E}^k$. Denote its population version by E_0^k. Further, let

$$\tilde{\Omega}_y^k = \text{diag}(\Omega_y^k) + \Omega_{y, E_0^k \cap \hat{E}^k}^k$$

With similar derivations to the proof of Corollary A.1 in Ma and Michailidis (2016), The following two upper bounds can be established:

$$|\hat{E}^k| \leq \frac{128S}{\psi} \tag{A.4}$$

$$\frac{1}{K}\sum_{k=1}^{K}\|\tilde{\Omega}_y^k - \Omega_y^k\|_F \leq \frac{12c_0\sqrt{S}\gamma_n}{\sqrt{K}\psi} \tag{A.5}$$

following which, taking $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0$,

$$\Lambda_{\min}(\tilde{\Omega}_y^k) \geq d_0 - \frac{48c_0\mathbb{Q}_0\sqrt{|g_{\max}|S}}{\psi} \geq (1-t_1)d_0 > 0 \tag{A.6}$$

$$\Lambda_{\max}(\tilde{\Omega}_y^k) \leq c_0 + \frac{48c_0\mathbb{Q}_0\sqrt{|g_{\max}|S}}{\psi} \leq c_0 + t_1 d_0 < \infty \tag{A.7}$$

with $0 < t_1 < 1$, and the sample size n satisfying

$$n \geq |g_{\max}|S\left[\frac{48c_0Q_0}{\psi t_1 d_0}\right]^2; \quad Q_0 := \sqrt{n}\mathbb{Q}_0$$

Step 3: Refitting. Following the same steps as part A.3 in the proof of Theorem 4.1 in Ma and Michailidis (2016), it can be proven using (A.4)–(A.7) that

$$\sum_{k=1}^{K}\|\widehat{\Omega}_y^k - \tilde{\Omega}_y^k\|^2 \leq O\left(\mathbb{Q}_0^2|g_{\max}|S\right)$$

The proof is now complete by combining this with (A.5) then applying Cauchy-Schwarz inequality and triangle inequality.

□

Proof of Proposition 4.2. We drop the superscript k since there is no scope of ambiguity. For part 1, we start with an auxiliary lemma:

Lemma A.2. *For a sub-gaussian design matrix $\mathbf{X} \in \mathbb{M}(n,p)$ with columns having mean $\mathbf{0}_p$ and covariance matrix Σ_x, the sample covariance matrix $\widehat{\Sigma}_x = \mathbf{X}^T\mathbf{X}/n$ satisfies the RE condition*

$$\widehat{\Sigma}_x \sim RE\left(\frac{\Lambda_{\min}(\Sigma_x)}{2}, \frac{\Lambda_{\min}(\Sigma_x)\log p}{2n}\right)$$

with probability $\geq 1 - 2\exp(-c_3 n)$ for some $c_3 > 0$.

Now denote $\widehat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}$. For $\mathbf{v} \in \mathbb{R}^q$, we have

$$\begin{aligned}
\mathbf{v}^T\widehat{\mathbf{S}}\mathbf{v} &= \frac{1}{n}\|\widehat{\mathbf{E}}\mathbf{v}\|^2 \\
&= \frac{1}{n}\|(\mathbf{E} + \mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}}))\mathbf{v}\|^2 \\
&= \mathbf{v}^T\mathbf{S}\mathbf{v} + \frac{1}{n}\|\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\mathbf{v}\|^2 + 2\mathbf{v}^T(\mathbf{B}_0 - \widehat{\mathbf{B}})^T\left(\frac{(\mathbf{X})^T\mathbf{E}}{n}\right)\mathbf{v}
\end{aligned} \tag{A.8}$$

For the first summand, $\mathbf{v}^T\mathbf{S}^k\mathbf{v} \geq \psi_y\|\mathbf{v}\|^2 - \phi_y\|\mathbf{v}\|_1^2$ with $\psi_y = \Lambda_{\min}(\Sigma_y)/2, \phi_y = \psi_y \log p/n$ by applying Lemma A.2 on $\mathbf{S}$. The second summand is greater than or equal to 0. For the third summand,

$$2\mathbf{v}^T(\mathbf{B}_0 - \widehat{\mathbf{B}})^T \left(\frac{(\mathbf{X})^T\mathbf{E}}{n}\right)\mathbf{v} \geq -2v_\beta \left\|\frac{(\mathbf{X})^T\mathbf{E}}{n}\right\|_\infty \|\mathbf{v}\|_1^2$$

by assumption (T1). Now we use another lemma:

Lemma A.3. *For zero-mean independent sub-gaussian matrices $\mathbf{X} \in \mathbb{M}(n,p), \mathbf{E} \in \mathbb{M}(n,q)$ with parameters (Σ_x, σ_x^2) and (Σ_e, σ_e^2) respectively, given that $n \gtrsim \log(pq)$ the following holds with probability $\geq 1 - 6c_1\exp[-(c_2^2-1)\log(pq)]$ for some $c_1 > 0, c_2 > 1$:*

$$\frac{1}{n}\|\mathbf{X}^T\mathbf{E}\|_\infty \leq c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_e)]^{1/2}\sqrt{\frac{\log(pq)}{n}}$$

Subsequently we collect all summands in (A.8) and get

$$\mathbf{v}^T\widehat{\mathbf{S}}\mathbf{v} \geq \psi_y\|\mathbf{v}\|^2 - \left(\phi_y + 2v_\beta c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_y)]^{1/2}\sqrt{\frac{\log(pq)}{n}}\right)\|\mathbf{v}\|_1^2$$

with probability $\geq 1 - 2\exp(-c_3 n) - 6c_1\exp[-(c_2^2-1)\log(pq)]$. This concludes the proof of part 1.

To prove part 2, we decompose the quantity in question:

$$\begin{aligned}
\left\|\frac{1}{n}\widehat{\mathbf{E}}_{-j}^T\widehat{\mathbf{E}}\mathbf{T}_{0,j}\right\|_\infty &= \left\|\frac{1}{n}\left[\mathbf{E}_{-j} + \mathbf{X}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)\right]^T\left[\mathbf{E} + \mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\right]\mathbf{T}_{0,j}\right\|_\infty \\
&\leq \left\|\frac{1}{n}\mathbf{E}_{-j}^T\mathbf{E}\mathbf{T}_{0,j}\right\|_\infty + \left\|\frac{1}{n}\mathbf{E}_{-j}^T\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\mathbf{T}_{0,j}\right\|_\infty \\
&+ \left\|\frac{1}{n}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)^T\mathbf{X}^T\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\mathbf{T}_{0,j}\right\|_\infty + \left\|\frac{1}{n}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)^T\mathbf{X}^T\mathbf{E}\mathbf{T}_{0,j}\right\|_\infty \\
&= \|\mathbf{W}_1\|_\infty + \|\mathbf{W}_2\|_\infty + \|\mathbf{W}_3\|_\infty + \|\mathbf{W}_4\|_\infty \qquad \text{(A.9)}
\end{aligned}$$

Now

$$\mathbf{W}_1 = \frac{1}{n}\mathbf{E}_{-j}^T(\mathbf{E}_j - \mathbf{E}_{-j}\boldsymbol{\theta}_{0,j})$$

For node j in the y-network, $\mathbb{E}_{-j}$ and $E_j - \mathbb{E}_{-j}\boldsymbol{\theta}_{0,j}$ are the neighborhood regression coefficients and residuals, respectively. Thus they are orthogonal, so we can apply Lemma A.3 on $\mathbf{E}_{-j}$ and $\mathbf{E}_j - \mathbf{E}_{-j}\boldsymbol{\theta}_{0,j}$ to obtain that for $n \gtrsim \log(q-1)$,

$$\|\mathbf{W}_1\|_\infty \leq c_5\left[\Lambda_{\max}(\Sigma_{y,-j})\sigma_{y,j,-j}\right]^{1/2}\sqrt{\frac{\log(q-1)}{n}} \qquad \text{(A.10)}$$

holds with probability $\geq 1 - 6c_4\exp[-(c_5^2-1)\log(pq)]$ for some $c_4 > 0, c_5 > 1$.

The same bounds hold for $\mathbf{W}_2$ and $\mathbf{W}_4$:

$$\|\mathbf{W}_2\|_\infty \leq \left\|\frac{1}{n}\mathbf{E}_{-j}^T\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\right\|_\infty \|\mathbf{T}_{0,j}\|_1 \leq \left\|\frac{1}{n}\mathbf{E}^T\mathbf{X}\right\|_\infty \|\mathbf{B}_0 - \widehat{\mathbf{B}}\|_1\|\mathbf{T}_{0,j}\|_1$$
$$\|\mathbf{W}_4\|_\infty \leq \left\|\frac{1}{n}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)^T\mathbf{X}^T\mathbf{E}\right\|_\infty \|\mathbf{T}_{0,j}\|_1 \leq \left\|\frac{1}{n}\mathbf{E}^T\mathbf{X}\right\|_\infty \|\mathbf{B}_0 - \widehat{\mathbf{B}}\|_1\|\mathbf{T}_{0,j}\|_1$$

Now since Ω_y is diagonally dominant, $|\omega_{y,jj}| \geq \sum_{j\neq j'} |\omega_{y,jj'}|$ for any $j \in \mathcal{I}_q$. Hence

$$\|\mathbf{T}_{0,j}\|_1 = \sum_{j'=1}^{q} |T_{jj'}| = 1 + \sum_{j\neq j'} |\theta_{jj'}| = 1 + \frac{1}{\omega_{y,jj}} \sum_{j\neq j'} |\omega_{y,jj'}| \leq 2$$

so that for $n \gtrsim \log(pq)$,

$$\|\mathbf{W}_2\|_\infty + \|\mathbf{W}_4\|_\infty \leq 4v_\beta c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_y)]^{1/2}\sqrt{\frac{\log(pq)}{n}} \tag{A.11}$$

with probability $\geq 1 - 6c_1\exp[-(c_2^2 - 1)\log(pq)]$ by applying Lemma A.3 and assumption (T1).

Finally for $\mathbf{W}_3$, we apply Lemma 8 of Ravikumar et al. (2011) on the (sub-gaussian) design matrix $\mathbf{X}$ to obtain that for sample size

$$n \geq 512(1 + 4\Lambda_{\max}(\Sigma_x^k))^4 \max_i(\sigma_{x,ii}^k)^4 \log(4p^{\tau_1}) \tag{A.12}$$

we get that with probability $\geq 1 - 1/p^{\tau_1 - 2}, \tau_1 > 2$,

$$\left\|\frac{\mathbf{X}^T\mathbf{X}}{n}\right\|_\infty \leq \sqrt{\frac{\log 4 + \tau_1 \log p}{c_x n}} + \max_i \sigma_{x,ii} = V_x; \quad c_x = \left[128(1 + 4\Lambda_{\max}(\Sigma_x))^2 \max_i(\sigma_{x,ii})^2\right]^{-1}$$

Thus with the same probability,

$$\|\mathbf{W}_4\|_\infty \leq \left\|\frac{\mathbf{X}^T\mathbf{X}}{n}\right\|_\infty \|\widehat{\mathbf{B}} - \mathbf{B}_0\|_1^2\|\mathbf{T}_{0,j}\|_1 \leq 2v_\beta^2 V_x \tag{A.13}$$

We now bound the right hand side of (A.9) using (A.10), (A.11) and (A.13) to complete the proof, with the leading term of the sample size requirement being $n \gtrsim \log(pq)$.

□

Proof of Theorem 4.4. The proof follows that of Proposition A.1, with a different group norm structure. We only point out the differences.

Putting $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in (2.12) we get

$$-2\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^T\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}} + \lambda_n \sum_{h\in\mathcal{H}} \|\widehat{\boldsymbol{\beta}}^{[h]}\| \leq -2\boldsymbol{\beta}_0^T\widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}_0^T\widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}_0 + \lambda_n \sum_{h\in\mathcal{H}} \|\boldsymbol{\beta}_0^{[h]}\|$$

Denote $\mathbf{b} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Then we have

$$\mathbf{b}^T \widehat{\boldsymbol{\Gamma}} \mathbf{b} \leq 2\mathbf{b}^T (\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}} \boldsymbol{\beta}_0) + \lambda_n \sum_{h \in \mathcal{H}} (\|\boldsymbol{\beta}_0^{[h]}\| - \|\boldsymbol{\beta}_0^{[h]} + \mathbf{b}^{[h]}\|)$$

Proceeding similarly as the proof of Proposition A.1, with a different deviation bound and choice of λ_n, we get expressions equivalent to (B.3) and (B.4) respectively:

$$\mathbf{b}^T \widehat{\boldsymbol{\Gamma}} \mathbf{b} \leq \frac{3}{2} \sum_{h \in \mathcal{H}} \|\mathbf{b}^{[h]}\| \tag{A.14}$$

$$\frac{\psi^*}{3} \|\mathbf{b}\|^2 \leq \lambda_n \sum_{h \in \mathcal{H}} \|\mathbf{b}^{[h]}\| \leq 4\lambda_n \sqrt{s_\beta} \|\mathbf{b}\| \tag{A.15}$$

Furthermore, $\|\mathbf{b}\|_1 \leq \sqrt{|h_{\max}|} \sum_{h \in \mathcal{H}} \|\mathbf{b}^{[h]}\|$. The bounds in (4.2), (4.3), (4.4) and (4.5) now follow.

□

Proof of Proposition 4.5. For part 1 it is enough to prove that with $\widehat{\Sigma}_x^k := (\mathbf{X}^k)^T \mathbf{X}^k / n$,

$$\widehat{\mathbf{T}}_k^2 \otimes \widehat{\Sigma}_x^k \sim RE(\psi_*^k, \phi_*^k) \tag{A.16}$$

with high enough probability. because then we can take $\psi_* = \min_k \psi_*^k, \phi_* = \max_k \phi_*^k$. The proof of (A.16) follows similar lines of the proof of Proposition 1 in Lin et al. (2016), only replacing $\Theta_\epsilon, \widehat{\Theta}_\epsilon, \mathbf{X}$ therein with $(\mathbf{T}^k)^2, (\widehat{\mathbf{T}}^k)^2, \mathbf{X}^k$, respectively. We omit the details.

Part 2 follows the proof of Proposition 2 in Lin et al. (2016).

□

Proof of Theorem 3.1. Let us define the following:

$$\begin{aligned}
\widehat{\Omega}_y &= \text{diag}(\widehat{\Omega}_y^1, \widehat{\Omega}_y^2) \\
\mathbf{M}_i &= \text{diag}(m_i^1, m_i^2) \\
\widehat{\mathbf{C}}_i &= \text{diag}(\widehat{\mathbf{c}}_i^1, \widehat{\mathbf{c}}_i^2) \\
\widehat{\mathbf{D}}_i &= \text{diag}(\widehat{\mathbf{b}}_i^1, \widehat{\mathbf{b}}_i^2) \\
\mathbf{D}_i &= \text{diag}(\mathbf{b}_{0,i}^1, \mathbf{b}_{0,i}^2) \\
\mathbf{R}_i^k &= \mathbf{X}_i^k - \mathbf{X}_{-i}^k \widehat{\boldsymbol{\zeta}}_i^k ; k = 1, 2
\end{aligned}$$

Then from (3.1) we have

$$\mathbf{M}_i (\widehat{\mathbf{C}}_i - \widehat{\mathbf{D}}_i)^T = \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{1}{\widehat{s}_i^1} (\mathbf{R}_i^1)^T \widehat{\mathbf{E}}^1 \\ \frac{1}{\widehat{s}_i^2} (\mathbf{R}_i^2)^T \widehat{\mathbf{E}}^2 \end{bmatrix} \tag{A.17}$$

We now decompose $\widehat{\mathbf{E}}^k$:

$$\begin{aligned}
\widehat{\mathbf{E}}^k &= \mathbf{Y}^k - \mathbf{X}^k \widehat{\mathbf{B}}^k \\
&= \mathbf{E}^k + \mathbf{X}^k (\mathbf{B}_0^k - \widehat{\mathbf{B}}^k) \\
&= \mathbf{E}^k + \mathbf{X}_i^k (\mathbf{b}_{0,i}^k - \widehat{\mathbf{b}}_i^k) + \mathbf{X}_{-i}^k (\mathbf{B}_{0,-i}^k - \widehat{\mathbf{B}}_{-i}^k)
\end{aligned}$$

Putting them back in (A.17) and using $t^k = (\mathbf{R}^k)^T\mathbf{X}^k/n$,

$$\mathbf{M}_i(\widehat{\mathbf{C}}_i - \widehat{\mathbf{D}}_i)^T = \frac{1}{\sqrt{n}}\begin{bmatrix}\frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T\mathbf{E}^1\\ \frac{1}{\widehat{s}_i^2}(\mathbf{R}_i^2)^T\mathbf{E}^2\end{bmatrix} + \mathbf{M}_i(\mathbf{D}_i - \widehat{\mathbf{D}}_i)^T + \frac{1}{\sqrt{n}}\begin{bmatrix}\frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T\mathbf{X}_{-i}^1(\mathbf{B}_{0,-i}^1 - \widehat{\mathbf{B}}_{-i}^1)\\ \frac{1}{\widehat{s}_i^2}(\mathbf{R}_i^2)^T\mathbf{X}_{-i}^2(\mathbf{B}_{0,-i}^2 - \widehat{\mathbf{B}}_{-i}^2)\end{bmatrix}$$

$$\Rightarrow \widehat{\Omega}_y^{1/2}\mathbf{M}_i(\widehat{\mathbf{C}}_i - \mathbf{D}_i)^T = \frac{\widehat{\Omega}_y^{1/2}}{\sqrt{n}}\begin{bmatrix}\frac{1}{\widehat{s}^1}(\mathbf{R}_i^1)^T\mathbf{E}^1\\ \frac{1}{\widehat{s}^2}(\mathbf{R}_i^2)^T\mathbf{E}^2\end{bmatrix} + \frac{\widehat{\Omega}_y^{1/2}}{\sqrt{n}}\begin{bmatrix}\frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T\mathbf{X}_{-i}^1(\mathbf{B}_{0,-i}^1 - \widehat{\mathbf{B}}_{-i}^1)\\ \frac{1}{\widehat{s}_i^2}(\mathbf{R}_i^2)^T\mathbf{X}_{-i}^2(\mathbf{B}_{0,-i}^2 - \widehat{\mathbf{B}}_{-i}^2)\end{bmatrix} \tag{A.18}$$

At this point, we drop k in the subscripts, and prove the following:

Lemma A.4. *Given conditions (C1) and (C2), the following holds for sample size n such that $n \gtrsim \log(pq)$ and $\sigma_{x,i,-i} - n^{-1/4} - v_\zeta\sqrt{V_x} > 0$:*

$$\frac{1}{\sqrt{n}\widehat{s}_i}\widehat{\Omega}_y^{1/2}\mathbf{E}^T\mathbf{R}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}) + \mathbf{S}_{1n};$$

$$\|\mathbf{S}_{1n}\|_\infty \leq \frac{v_\Omega(2+v_\zeta)c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_e)]^{1/2}\sqrt{\log(pq)}}{\sigma_{x,i,-i} - n^{-1/4} - v_\zeta\sqrt{V_x}} = O\left(\frac{\log(pq)}{\sqrt{n}}\right) \tag{A.19}$$

with probability larger than or equal to

$$1 - 6c_1e^{-(c_2^2-1)\log pq} - \frac{1}{p^{\tau_1-2}} - \frac{\kappa_i}{\sqrt{n}} \tag{A.20}$$

for some $c_1, c_4 > 0, c_2, c_5 > 1$, and $\kappa_i := \mathbb{V}[(X_i - \mathbb{X}_{-i}\boldsymbol{\zeta}_{0,-i})^2]$. Additionally, given condition (C3)

$$\left\|\frac{1}{\sqrt{n}\widehat{s}_i}\mathbf{R}_i^T\mathbf{X}_{-i}(\mathbf{B}_{0,-i} - \widehat{\mathbf{B}}_{-i})\widehat{\Omega}_y^{1/2}\right\|_\infty$$
$$\leq \frac{v_\beta(\Lambda_{\min}(\Sigma_y)^{1/2} + v_\Omega)}{\sigma_{x,i,-i} - n^{-1/2} - v_\zeta\sqrt{V_x}}\left[c_7\sqrt{(\sigma_{x,i,-i}\Lambda_{\max}(\Sigma_{x,-i}))\log p} + \sqrt{n}v_\zeta V_x\right] = O\left(\frac{\log(pq)}{\sqrt{n}}\right) \tag{A.21}$$

with probability condition (A.20).

Given Lemma A.4, the first and second summands on the right hand side of (A.18) are bounded above by applying each of (A.19) and (A.21) twice, respectively. This completes our proof. □

Proof of Theorem 3.2. □

B Proof of auxiliary results

Proof of Proposition A.1. In its reparametrized version, (2.13) becomes

$$\widehat{\mathbf{T}}_j = \arg\min_{\mathbf{T}_j} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|(\mathbf{Y}^k - \mathbf{X}^k \widehat{\mathbf{B}}^k)\mathbf{T}_j^k\|^2 + \gamma_n \sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\mathbf{T}_{jj'}^{[g]}\| \right\} \tag{B.1}$$

with $\mathbf{T}_{jj'}^{[g]} := (T_{jj'}^k)_{k \in g}$. Now for any $\mathbf{T}_j \in \mathbb{M}(q, K)$ we have

$$\frac{1}{n} \sum_{k=1}^{K} \|(\mathbf{Y}^k - \mathbf{X}^k \widehat{\mathbf{B}}^k)\widehat{\mathbf{T}}_j^k\|^2 + \gamma_n \sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\widehat{\mathbf{T}}_{jj'}^{[g]}\| \leq \frac{1}{n} \sum_{k=1}^{K} \|(\mathbf{Y}^k - \mathbf{X}^k \widehat{\mathbf{B}}^k)\mathbf{T}_j^k\|^2 + \gamma_n \sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\mathbf{T}_{jj'}^{[g]}\|$$

For $\mathbf{T}_j = \mathbf{T}_{0,j}$ this reduces to

$$\sum_{k=1}^{K} (\mathbf{d}_j^k)^T \widehat{\mathbf{S}}^k \mathbf{d}_j^k \leq -2 \sum_{k=1}^{K} (\mathbf{d}_j^k)^T \widehat{\mathbf{S}}^k \mathbf{T}_{0,j}^k + \gamma_n \sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \left(\|\mathbf{T}_{jj'}^{[g]}\| - \|\mathbf{T}_{jj'}^{[g]} + \mathbf{d}_{jj'}^{[g]}\| \right) \tag{B.2}$$

with $\mathbf{d}_j^k := \widehat{\mathbf{T}}_j^k - \mathbf{T}_{0,j}^k$ etc. For the k^{th} summand in the first term on the right hand side, since $d_{jj}^k = 0$, $\widehat{\mathbf{E}}^k \mathbf{d}_j^k = \widehat{\mathbf{E}}_{-j}^k \mathbf{d}_{-j}^k$. Thus

$$\begin{aligned}
\sum_{k=1}^{K} \left| (\mathbf{d}_j^k)^T \widehat{\mathbf{S}}^k \mathbf{T}_{0,j}^k \right| &= \sum_{k=1}^{K} \left| \mathbf{d}_j^k . \frac{1}{n} (\widehat{\mathbf{E}}^k)^T \widehat{\mathbf{E}}^k \mathbf{T}_{0,j}^k \right| \\
&\leq \sum_{k=1}^{K} \left\| \frac{1}{n} (\widehat{\mathbf{E}}_{-j}^k)^T \widehat{\mathbf{E}}^k \mathbf{T}_{0,j}^k \right\|_\infty \|\mathbf{d}_{-j}^k\|_1 \\
&\leq \mathbb{Q}_0 \sqrt{|g_{\max}|} \sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\mathbf{d}_{jj'}^{[g]}\|
\end{aligned}$$

by assumption (T2). For the second term, suppose $\mathcal{S}_{0,j}$ is the support of $\Theta_{0,j}$, i.e. $\mathcal{S}_{0,j} = \{(j', g) : \boldsymbol{\theta}_{jj'}^{[g]} \neq 0\}$. Then

$$\begin{aligned}
\sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \left(\|\mathbf{T}_{jj'}^{[g]}\| - \|\mathbf{T}_{jj'}^{[g]} + \mathbf{d}_{jj'}^{[g]}\| \right) &\leq \sum_{(j',g) \in \mathcal{S}_{0,j}} \left(\|\mathbf{T}_{jj'}^{[g]}\| - \|\mathbf{T}_{jj'}^{[g]} + \mathbf{d}_{jj'}^{[g]}\| \right) - \sum_{(j',g) \notin \mathcal{S}_{0,j}} \|\mathbf{d}_{jj'}^{[g]}\| \\
&\leq \sum_{(j',g) \in \mathcal{S}_{0,j}} \|\mathbf{d}_{jj'}^{[g]}\| - \sum_{(j',g) \notin \mathcal{S}_{0,j}} \|\mathbf{d}_{jj'}^{[g]}\|
\end{aligned}$$

so that by choice of γ_n (B.2) reduces to

$$\begin{aligned}\sum_{k=1}^{K}(\mathbf{d}_j^k)^T\widehat{\mathbf{S}}^k\mathbf{d}_j^k &\le \frac{\gamma_n}{2}\left[\sum_{(j',g)\in\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| + \sum_{(j',g)\notin\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\|\right] + \gamma_n\left[\sum_{(j',g)\in\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| - \sum_{(j',g)\notin\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\|\right] \\ &= \frac{3\gamma_n}{2}\sum_{(j',g)\in\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| - \frac{\gamma_n}{2}\sum_{(j',g)\notin\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| \\ &\le \frac{3\gamma_n}{2}\sum_{j\neq j',g\in\mathcal{G}_y^{jj'}}\|\mathbf{d}_{jj'}^{[g]}\| \qquad \text{(B.3)}\end{aligned}$$

Since the left hand side is ≥ 0, this also implies

$$\sum_{(j',g)\notin\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| \le 3\sum_{(j',g)\in\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| \quad\Rightarrow\quad \sum_{j\neq j',g\in\mathcal{G}_y^{jj'}}\|\mathbf{d}_{jj'}^{[g]}\| \le 4\sum_{(j',g)\in\mathcal{S}_{0,j}}\|\mathbf{d}_{jj'}^{[g]}\| \le 4\sqrt{s_j}\|\mathbf{D}_j\|_F$$

with $\mathbf{D}_j = (\mathbf{d}_j^k)_{k\in\mathcal{I}_K}$. Now the RE condition on $\widehat{\mathbf{S}}^k$ means that

$$\sum_{k=1}^{K}(\mathbf{d}_j^k)^T\widehat{\mathbf{S}}^k\mathbf{d}_j^k \ge \sum_{k=1}^{K}\left(\psi_k\|\mathbf{d}_j^k\|^2 - \phi_k\|\mathbf{d}_j^k\|_1^2\right) \ge \psi\|\mathbf{D}_j\|_F^2 - \phi\|\mathbf{D}_j\|_1^2 \ge (\psi - Kq\phi)\|\mathbf{D}_j\|_F^2 \ge \frac{\psi}{2}\|\mathbf{D}_j\|_F^2$$

by assumption (T3).

Thus we finally have

$$\frac{\psi}{3}\|\mathbf{D}_j\|_F^2 \le \gamma_n\sum_{j\neq j',g\in\mathcal{G}_y^{jj'}}\|\mathbf{d}_{jj'}^{[g]}\| \le 4\gamma_n\sqrt{s_j}\|\mathbf{D}_j\|_F \qquad \text{(B.4)}$$

Since

$$(\mathbf{D}_j)_{j',k} = \widehat{T}_{jj'}^k - T_{0,jj'}^k = \begin{cases} 0 \text{ if } j = j' \\ -(\widehat{\theta}_{jj'}^k - \theta_{0,jj'}^k) \text{ if } j \neq j' \end{cases}$$

The bounds in (A.1) and (A.2) are obtained by replacing the corresponding elements in (B.4).

For the bound on $|\widehat{\mathcal{S}}_j|$, notice that if $\hat{\boldsymbol{\theta}}_{jj'}^{[g]} \neq 0$ for some (j', g),

$$\begin{aligned}\frac{1}{n}\sum_{k\in g}\left|((\widehat{\mathbf{E}}_{-j}^k)^T\widehat{\mathbf{E}}^k(\widehat{\mathbf{T}}_j^k - \mathbf{T}_{0,j}^k))^{j'}\right| &\ge \frac{1}{n}\sum_{k\in g}\left|((\widehat{\mathbf{E}}_{-j}^k)^T\widehat{\mathbf{E}}^k\widehat{\mathbf{T}}_j^k)^{j'}\right| - \frac{1}{n}\sum_{k\in g}\left|((\widehat{\mathbf{E}}_{-j}^k)^T\widehat{\mathbf{E}}^k\mathbf{T}_{0,j}^k)^{j'}\right| \\ &\ge |g|\gamma_n - \sum_{k\in g}\mathbb{Q}(\upsilon_\beta, \Sigma_x^k, \Sigma_y^k)\end{aligned}$$

using the KKT condition for (2.13) and assumption (T2). The choice of γ_n now ensures

that the right hand side is $\geq 3|g|\gamma_n/4$. Hence

$$\begin{aligned}
|\hat{\mathcal{S}}_j| &\leq \sum_{(j',g)\in\widehat{\mathcal{S}}_j} \frac{16}{9n^2|g|^2\gamma_n^2} \sum_{k\in g} \left|((\widehat{\mathbf{E}}^k_{-j})^T\widehat{\mathbf{E}}^k(\widehat{\mathbf{T}}^k_j - \mathbf{T}^k_{0,j}))^{j'}\right|^2 \\
&\leq \frac{16}{9\gamma_n^2}\sum_{k=1}^K \frac{1}{n}\left\|(\widehat{\mathbf{E}}^k_{-j})^T\widehat{\mathbf{E}}^k(\widehat{\mathbf{T}}^k_j - \mathbf{T}^k_{0,j})\right\|^2 \\
&= \frac{16}{9\gamma_n^2}\sum_{k=1}^K (\mathbf{d}^k_j)^T\widehat{\mathbf{S}}^k\mathbf{d}^k_j \\
&\leq \frac{8}{3\gamma_n}\sum_{j\neq j',g\in\mathcal{G}^{jj'}_y} \|\mathbf{d}^{[g]}_{jj'}\| \leq \frac{128 s_j}{\psi}
\end{aligned}$$

using (B.3) and (B.4). □

Proof of Lemma A.2. This is same as Lemma 2 in Appendix B of Lin et al. (2016) and its proof can be found there. □

Proof of Lemma A.3. This is a part of Lemma 3 of Appendix B in Lin et al. (2016), and is proved therein. □

Proof of Lemma A.4. To show (A.19) we have

$$\frac{1}{\sqrt{n}\widehat{s}_i}\widehat{\Omega}_y^{1/2}\mathbf{E}^T\mathbf{R}_i = \frac{1}{\sqrt{n}\widehat{s}_i}(\widehat{\Omega}_y^{1/2} - \Omega_y^{1/2})\mathbf{E}^T\mathbf{R}_i + \frac{1}{\sqrt{n}\widehat{s}_i}\Omega_y^{1/2}\mathbf{E}^T\mathbf{R}_i$$

The second summand is distributed as $\mathcal{N}_q(\mathbf{0},\mathbf{I})$. For the first summand,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\left\|(\widehat{\Omega}_y^{1/2} - \Omega_y^{1/2})\mathbf{E}^T\mathbf{R}_i\right\|_\infty &\leq \frac{1}{\sqrt{n}}\left\|\widehat{\Omega}_y^{1/2} - \Omega_y^{1/2}\right\|_\infty \left\|\mathbf{E}^T\mathbf{R}_i\right\|_1 \\
&\leq \sqrt{n}v_\Omega\frac{1}{n}\left[\|\mathbf{E}^T(\mathbf{X}_i - \mathbf{X}_{-i}\boldsymbol{\zeta}_i)\|_1 + \|\mathbf{E}^T\mathbf{X}_{-i}(\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_{0,i})\|_1\right] \\
&\leq \sqrt{n}v_\Omega\frac{1}{n}\left[\|\mathbf{E}^T\mathbf{X}_i\|_\infty + \|\mathbf{E}^T\mathbf{X}_{-i}\|_\infty\left\{\|\boldsymbol{\zeta}_i\|_1 + \|\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_i\|_1\right\}\right] \\
&\leq \sqrt{n}v_\Omega\left[\frac{1}{n}\|\mathbf{E}^T\mathbf{X}_i\|_\infty + \frac{1+v_\zeta}{n}\|\mathbf{E}^T\mathbf{X}_{-i}\|_\infty\right] \\
&\leq \sqrt{n}v_\Omega(2+v_\zeta).\frac{1}{n}\|\mathbf{E}^T\mathbf{X}\|_\infty
\end{aligned}$$

because Ω_x is diagonally dominant implies $\|\boldsymbol{\zeta}_i\|_1 = \sum_{i'\neq i}|\omega_{x,ii'}|/\omega_{x,ii} \leq 1$, and using assumption (C1). Applying Lemma A.3, the following holds:

$$\frac{1}{\sqrt{n}}\left\|(\widehat{\Omega}_y^{1/2} - \Omega_y^{1/2})\mathbf{E}^T\mathbf{R}_i\right\|_\infty \leq v_\Omega(2+v_\zeta)c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_e)]^{1/2}\sqrt{\log(pq)} \tag{B.5}$$

with probability $\geq 1 - 6c_1\exp[-(c_2^2-1)\log(pq)]$ for some $c_1 > 0, c_2 > 1$.

On the other hand

$$s_i^2 := \frac{1}{n}\|\mathbf{X}_i - \mathbf{X}_{-i}\boldsymbol{\zeta}_i\|^2 \le \widehat{s}_i^2 + \frac{1}{n}\left\|\mathbf{X}_{-i}(\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_{0,i})\right\|^2 \le \widehat{s}_i^2 + \|\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_{0,i}\|_1^2 \left\|\frac{1}{n}\mathbf{X}_{-i}^T\mathbf{X}_{-i}\right\|_\infty$$

which implies $s_i \le \widehat{s}_i + v_\zeta\sqrt{V_x}$. By applying Lemma 8 of Ravikumar et al. (2011),

$$\left\|\frac{1}{n}\mathbf{X}_{-i}^T\mathbf{X}_{-i}\right\|_\infty \le \left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}\right\|_\infty \le V_x \tag{B.6}$$

with probability $\ge 1 - 1/p^{\tau_1 - 2}, \tau_1 > 2$, and

$$n \ge 512(1 + 4\Lambda_{\max}(\Sigma_x))^4 \max_i(\sigma_{x,ii})^4 \log(4p^{\tau_1}) \tag{B.7}$$

On the other hand, by Chebyshev inequality, for any $\epsilon > 0$

$$P\left(|s_i - \sigma_{x,i,-i}| \ge \epsilon\right) \le \frac{\mathbb{V}s_i}{\epsilon^2} = \frac{\kappa_i}{n\epsilon^2}$$

Taking $\epsilon = n^{-1/4}$, we have $s_i \ge \sigma_{x,i,-i} - n^{-1/4}$ with probability $\ge 1 - \kappa_i n^{-1/2}$. Then, for n satisfying (B.6) and $\sigma_{x,i,-i} - n^{-1/4} > v_\zeta\sqrt{V_x}$, we get the bound with the above probability:

$$\frac{1}{\widehat{s}_i} \le \frac{1}{\sigma_{x,i,-i} - n^{-1/4} - v_\zeta\sqrt{V_x}} \tag{B.8}$$

Combining (B.5) and (B.8) gives the upper bound for the right hand side of (A.19) with the requisite probability and sample size conditions.

To prove (A.21) we have

$$\frac{1}{n}\|\mathbf{R}_i^T\mathbf{X}_{-i}\|_\infty \le \frac{1}{n}\|(\mathbf{X}_i - \mathbf{X}_{-i}\boldsymbol{\zeta}_{0,i})^T\mathbf{X}_{-i}\|_\infty + \frac{1}{n}\|\mathbf{X}_{-i}^T\mathbf{X}_{-i}(\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_{0,i})\|_\infty \tag{B.9}$$

Applying Lemma A.3, for $n \gtrsim \log(p-1)$ we have

$$\frac{1}{n}\|(\mathbf{X}_i - \mathbf{X}_{-i}\boldsymbol{\zeta}_i)^T\mathbf{X}_{-i}\|_\infty \le c_7[\sigma_{x,i,-i}\Lambda_{\max}(\Sigma_{x,-i})]^{1/2}\sqrt{\frac{\log(p-1)}{n}} \tag{B.10}$$

with probability $\ge 1 - 6c_6\exp[-(c_7^2 - 1)\log(p-1)]$ for some $c_6 > 0, c_7 > 1$. By (B.6), the second term on the right side of (B.9) is bounded above by $v_\zeta V_x$ with probability $\ge 1 - 1/p^{\tau_1 - 2}$ and n satisfying (B.7). The bound of (A.21) now follows by conditions (C2), (C3) and (B.8). □

References

Cai, T. T. and Liu, W. (2016). Large-Scale Multiple Testing of Correlations. *J. Amer. Stat. Assoc.*, 111(513):229–240.

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

Lin, J., Basu, S., Banerjee, M., and Michailidis, G. (2016). Penalized Maximum Likelihood Estimation of Multi-layered Gaussian Graphical Models. *J. Mach. Learn. Res.*, 17:5097–5147.

Liu, W. (2017). Structural similarity and difference testing on multiple sparse Gaussian graphical models. *Ann. Statist.*, 45(6):2680–2707.

Liu, W. and Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.*, 42(5):2003–2025.

Ma, J. and Michailidis, G. (2016). Joint Structural Estimation of Multiple Graphical Models. *J. Mach. Learn. Res.*, 17:5777–5824.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Mitra, R. and Zhang, C.-H. (2016). The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electron. J. Stat.*, 10:1829–1873.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *Ann. Statist.*, 42:1166–1202.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. *J. R. Statist. Soc. B*, 76:217–242.