Joint Estimation and Inference for Data Integration Problems based on Multiple Multi-layered Gaussian Graphical Models

Subhabrata Majumdar SMAJUMDAR@UFL.EDU
University of Florida Informatics Institute
Gainesville, FL, 32611, USA

George Michailidis* GMICHAIL@UFL.EDU
Department of Statistics and Computer & Information Science & Engineering
University of Florida
Gainesville, FL 32611, USA

Editor:

Abstract

The rapid development of high-throughput technologies has enabled the generation of data from biological or disease processes that span multiple layers, like genomic, proteomic or metabolomic data, and further pertain to multiple sources, like disease subtypes or experimental conditions. In this work, we propose a general statistical framework based on Gaussian 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. We use a combination of neighborhood selection and group-penalized regression to obtain sparse estimates of all model parameters. Following this, we develop a debiasing technique and asymptotic distributions of inter-layer directed edge weights that utilize already computed neighborhood selection coefficients for nodes in the upper layer. Subsequently, we establish global and simultaneous testing procedures for these edge weights. Performance of the proposed methodology is evaluated on synthetic data.

Keywords: Data integration; Gaussian Graphical Models; neighborhood selection; group lasso; high-dimensional asymptotics; multiple testing; false discovery rate

∗. Corresponding Author. Post Address: 205 Griffin Floyd Hall, 1 University Ave, Gainesville, FL, 32611.

1. Introduction

Aberrations in complex biological systems develop in the background of diverse genetic and environmental factors and are associated with multiple complex molecular events. These include changes in the genome, transcriptome, proteome and metabolome, as well as epigenetic effects. Advances in high-throughput profiling techniques have enabled a systematic and comprehensive exploration of the genetic and epigenetic basis of various diseases, including cancer (Lee et al., 2016; Kaushik et al., 2016), diabetes (Yuan et al., 2014; Sas et al., 2018), chronic kidney disease (Atzler et al., 2014), etc. Further, such multi-Omics collections have become available for patients belonging to different, but related disease subtypes, with The Cancer Genome Atlas (TCGA: Tomczak et al. (2015)) being a prototypical one. Hence, there is an increasing need for models that can *integrate* such complex data both *vertically* across multiple modalities and *horizontally* across different disease subtypes.

Figure 1 provides a schematic representation of the horizontal and vertical structure of such heterogeneous multi-modal Omics data as outlined above. A simultaneous analysis of all components in this complex layered structure has been coined in the literature as *data integration*. While it is common knowledge that this will result in a more comprehensive picture of the regulatory mechanisms behind diseases, phenotypes and biological processes in general, there is a dearth of rigorous methodologies that satisfactorily tackle all challenges that stem from attempts to perform data integration (Joyce and Palsson, 2006; Gomez-Cabrero et al., 2014; Gligorijević and Pržulj, 2015). A review of the present approaches towards achieving this goal, which are based mostly on specific case studies, can be found in Gligorijević and Pržulj (2015) and Zhang et al. (2017).

Gaussian Graphical Models (GGM) have been extensively used to model biological networks in the last few years. While the initial work on GGMs focused on estimating undirected edges within a single network through obtaining sparse estimates of the inverse covariance matrix from high-dimensional data (e.g. see references in Bühlmann and van de Geer (2011)), attention has shifted to estimating parameters from more complex structures, including multiple related graphical models and hierarchical multilayer models comprising of both directed and undirected edges. For the first class of problems, Guo et al. (2011) and Xie et al. (2016) assumed perturbations over a common underlying structure to model multiple precision matrices, while Danaher et al. (2014) proposed using fused/group lasso type penalties for the same task. To incorporate prior information on the group structures across several graphs, Ma and Michailidis (2016) proposed the Joint Structural Estimation Method (JSEM), which uses group-penalized neighborhood regression and subsequent refitting for estimating precision matrices. For the second problem, a two-layered structure can be modeled by interpreting directed edges between the two layers as elements of a multitask regression coefficient matrix, while undirected edges inside either layer correspond to the precision matrix of predictors in that layer. While several methods exist in the literature for joint estimation of both sets of parameters (Lee and Liu, 2012; Cai et al., 2012a), only recently Lin et al. (2016a) made the observation that a multi-layer model can, in fact, be decomposed into a series of two-layer problems. Subsequently, they proposed an estimation algorithm and derived theoretical properties of the resulting estimators.

All the above approaches focus either on the horizontal or the vertical dimensions of the full hierarchical structure depicted in Figure 1. Hence, multiple related groups of heteroge-

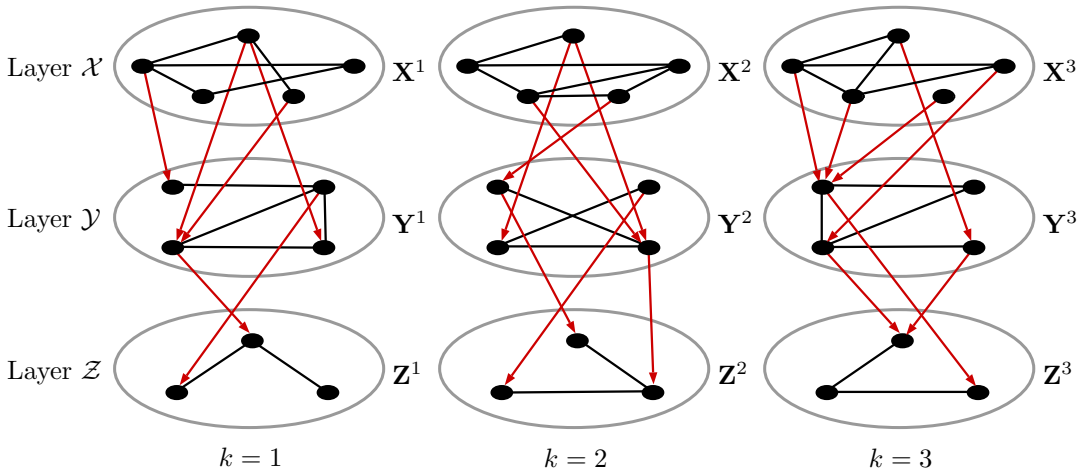

Figure 1: Multiple multilayer graphical models. The matrices $(\mathbf{X}^k, \mathbf{Y}^k, \mathbf{Z}^k), k = 1, 2, 3$ indicate data for each layer and category k. Within-layer connections (black lines) are undirected, while between-layer connections (red lines) go from an upper layer to the successive lower layer. For each type of edges (i.e. within $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ and $\mathcal{X} \to \mathcal{Y}, \mathcal{Y} \to \mathcal{Z}$), there are common edges across some or all k.

neous data sets have to be modeled by analyzing all data in individual layers (i.e. models for $\{\mathbf{X}^k\}$, $\{\mathbf{Y}^k\}$, $\{\mathbf{Z}^k\}$), and then separately analyzing individual hierarchies of datasets (i.e. separate models for $(\mathbf{X}^k, \mathbf{Y}^k, \mathbf{Z}^k), k = 1, 2, 3$). In another line of work, Kling et al. (2015); Zhang et al. (2017) model all undirected edges within all nodes together using penalized log-likelihoods. The advantage of this approach is that it can incorporate feedback loops and connections between nodes in non-adjacent layers. However, it has two considerable caveats. Firstly, it does not distinguish between hierarchies, hence delineating the direction of a connection between two nodes across two different Omics modalities is not possible in such models. Secondly, computation becomes difficult when data from different Omics modalities are considered, since the number of estimable parameters increases at a faster late compared to a hierarchical model.

While there has been some progress for parameter estimation in multilayer models, little is known about the sampling distributions of resulting estimates. Current research on such distributions and related testing procedures for estimates from high-dimensional problems has been limited to single-response regression using lasso (Zhang and Zhang, 2014; Javanmard and Montanari, 2014, 2018+; van de Geer et al., 2014) or group lasso (Mitra and Zhang, 2016) penalties, and partial correlations of single (Cai and Liu, 2016) or multiple (Belilovsky et al., 2016; Liu, 2017) GGMs. From a systemic perspective, testing and identifying downstream interactions that differ across experimental conditions or disease subtypes can offer important insights on the underlying biological process (Mao et al., 2017; Li et al., 2015). In the proposed integrative framework, this can be accomplished by developing a hypothesis testing procedure for entries in the within-layer regression matrices.

The contributions of this paper are two-fold. Firstly, we propose an integrative framework to conduct simultaneous inference for all parameters in multiple multi-layer graphical models, essentially formalizing the structure in Figure 1. We decompose the multi-layer

problem into a series of two-layer problems, propose an estimation algorithm for them based on group penalization, and derive theoretical properties of the estimators. Generalizing to group structures on the model parameters allows us to incorporate prior information, as and when available, on within-layer or between-layer sub-graph components shared across some or all $k = 1, \cdots, K$. For biological processes, such information can stem from experimental or mechanistic knowledge (for example a pathway-based grouping of genes). Secondly, we obtain *debiased* versions of within-layer regression coefficients in this two-layer model, and derive their asymptotic distributions using estimates of model parameters that satisfy generic convergence guarantees. Consequently, we formulate a global test, as well as a simultaneous testing procedure that controls for False Discovery Rate (FDR) to detect important pairwise differences among directed edges between layers.

Our proposed framework for knowledge discovery from heterogeneous data sources is highly flexible. The group sparsity assumptions in our estimation technique can be replaced by other structural restrictions, for example low-rank or low-rank-plus-sparse, as and when deemed appropriate by the prior dependency assumptions across parameters. As long as the resulting estimates converge to corresponding true parameters at certain rates, they can be used by the developed testing methodology.

Organization of paper. We start with the model formulation in Section 2, then introduce our computational algorithm for a two-layer model, and derive theoretical convergence properties of the algorithm and resulting estimates. In section 3, we start by introducing the debiased versions of rows of the regression coefficient matrix estimates in our model, then use already computed parameter estimates that satisfy some general consistency conditions to obtain its asymptotic distribution. We then move on to pairwise testing, and use sparse estimates from our algorithm to propose a global test to detect overall differences in rows of the coefficient matrices, as well as a multiple testing procedure to detect elementwise differences and perform within-row thresholding of estimates in presence of moderate misspecification of the group sparsity structure. Section 4 is devoted to implementation of our methodology. We evaluate the performance of our estimation and testing procedure through several simulation settings, and give strategies to speed up the computational algorithm for high data dimensions. We conclude the paper with a discussion in Section 6. Proofs of all theoretical results, as well as some auxiliary results, are given in the appendix.

Notation. 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 a positive semi-definite matrix $\mathbf{P}$, we denote its smallest and largest eigenvalues by $\Lambda_{\min}(\mathbf{P})$ and $\Lambda_{\max}(\mathbf{P})$, respectively. For any positive integer c, define $\mathcal{I}_c = \{1, \ldots, c\}$. For vectors $\mathbf{v}$ and matrices $\mathbf{M}$, $\|\mathbf{v}\|$, $\|\mathbf{v}\|_1$ or $\|\mathbf{M}\|_1$ and $\|\mathbf{v}\|_\infty$ or $\|\mathbf{M}\|_\infty$ denote euclidean, ℓ_1 and ℓ_∞ norms, respectively. The notation $\text{supp}(\mathbf{A})$ indicates the non-zero edge set in a matrix (or vector) $\mathbf{A}$, i.e. $\text{supp}(\mathbf{A}) = \{(i,j) : (\mathbf{A})_{ij} \neq 0\}$. For any set $\mathcal{S}$, $|\mathcal{S}|$ denotes the number of elements in that set. For positive real numbers A, B we write $A \gtrsim B$ if there exists $c > 0$ independent of model parameters such that $A \geq cB$. We use the ':=' notation to define a quantity for the first time.

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 is to leverage them to estimate the full hierarchical structure of the network -specifically to obtain the undirected edges for the nodes inside a single layer, and the directed edges between two successive layers through jointly estimating $\{\Omega_m^k\}$ and $\{\mathbf{B}_m^k\}$.

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

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

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

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

wherein 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 the theoretical discussion in the remainder 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 (see details in Lin et al. (2016a)). On the other hand, parameters in the very first layer are analogous to $\Omega_x := \{\Omega_x^k\}$, and can be estimated from $\{\mathbf{X}^k\}$ using any method for joint estimation of multiple graphical models (e.g. Guo et al. (2011); Ma and Michailidis (2016)). This provides all building blocks for recovering 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 $\mathcal{I}_K$, and consists of index groups g such that $g \subseteq \mathcal{I}_K, \cup_{g \in \mathcal{G}_x^{ii'}} g = \mathcal{I}_K$. First introduced in Ma and Michailidis (2016), this formulation helps incorporate group structures that are common across some of the precision matrices being modeled. Figure 2 illustrates this through a small example. Subsequently, we use the Joint Structural Estimation Method (JSEM) (Ma and Michailidis, 2016) to estimate Ω_x, which first uses the group

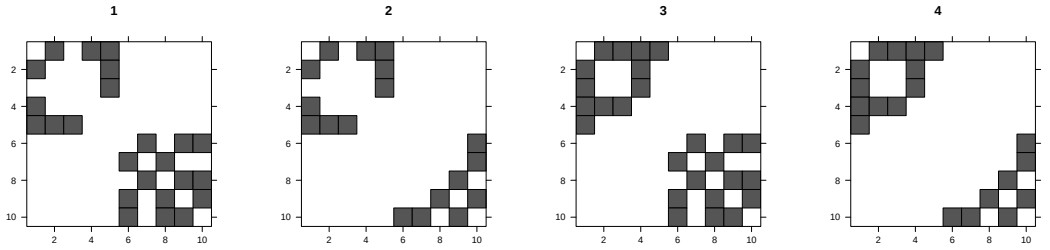

Figure 2: Shared sparsity patterns for four 10×10 precision matrices. for elements $\mathcal{G}_{x,ii'}$ in the upper 5×5 block, matrices (1,2) and (3,4) have the same non-zero support, i.e. $\mathcal{G}_{x,ii'} = \{(1,2),(3,4)\}$. On the other hand, when i, i' are in the lower block, $\mathcal{G}_{x,ii'} = \{(1,3),(2,4)\}$

structure given by $\mathcal{G}_x$ in penalized nodewise regressions (Meinshausen and Bühlmann, 2006) to obtain neighborhood coefficients $\zeta_i = (\boldsymbol{\zeta}_i^1, \ldots, \boldsymbol{\zeta}_i^K)$ of each variable $X_i, i \in \mathcal{I}_p$, then fits a maximum likelihood model over the combined support sets to obtain sparse estimates of the precision matrices:

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

where $\widehat{\mathbf{S}}_x^k := (\mathbf{X}^k)^T \mathbf{X}^k / n$, η_n is a tuning parameter, and $\mathbb{S}_+(\hat{E}_x^k)$ is the set of positive-definite matrices that have non-zero supports restricted to $\hat{E}_x^k$.

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

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

In other words, any arbitrary partition of $\mathcal{I}_p \times \mathcal{I}_q \times \mathcal{I}_K$ can be specified as the sparsity pattern of $\mathcal{B}$.

Denote the neighborhood coefficients of the j^{th} variable in the lower layer by $\boldsymbol{\theta}_j^k$, and $\Theta_j := (\boldsymbol{\theta}_j^1, \ldots, \boldsymbol{\theta}_j^K), \Theta = \{\Theta_j\}$. We obtain sparse estimates of $\mathcal{B}, \Theta$, and subsequently Ω_y, by solving the following group-penalized least square minimization problem that has the

tuning parameters γ_n and λ_n and then refitting:

$$\{\widehat{\mathcal{B}}, \widehat{\Theta}\} = \underset{\mathcal{B},\Theta}{\arg\min} \left\{ \frac{1}{n} \sum_{j=1}^{q} \sum_{k=1}^{K} \|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \mathbf{B}_{-j}^k)\boldsymbol{\theta}_j^k - \mathbf{X}^k \mathbf{B}_j^k\|^2 \right.$$
$$\left. + \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \gamma_n \|\boldsymbol{\theta}_{jj'}^{[g]}\| + \sum_{h \in \mathcal{H}} \lambda_n \|\mathbf{B}^{[h]}\| \right\}, \tag{2.5}$$
$$\widehat{E}_y^k = \{(j, j') : 1 \leq j < j' \leq q, \hat{\theta}_{jj'}^k \neq 0 \text{ OR } \hat{\theta}_{j'j}^k \neq 0\},$$
$$\widehat{\Omega}_y^k = \underset{\Omega_y^k \in \mathbb{S}_+(\hat{E}_y^k)}{\arg\min} \left\{ \mathrm{Tr}(\widehat{\mathbf{S}}_y^k \Omega_y^k) - \log\det(\Omega_y^k) \right\}. \tag{2.6}$$

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

Remark 1. Common sparsity structures across the same layer are incorporated into the regression by the group penalties over the element-wise groups $\boldsymbol{\theta}_{jj'}^{[g]}$, while sparsity pattern overlaps across the different regression matrices $\mathbf{B}^k$ are handled by the group penalties over $\mathbf{B}^{[h]}$, which denote the collection of elements in $\mathcal{B}$ that are in h. Other kinds of structural assumptions on $\mathcal{B}$ or Θ can be handled within the above structure by swapping out the group norms in favor of other appropriate norm-based penalties.

Remark 2. Group sparsity assumptions are not necessary for the JMMLE framework: rather in a practical situation they help leverage additional information regarding interaction of features in and between the layers, *as and when that information is available*. In the vertical direction of the model, i.e. given a fixed k, a framework agnostic of any structural dependency assumptions amounts to element-wise groups in $\mathbf{B}^k$ and Ω_y^k. In JMMLE, this occurs by construction for Θ, and since $\mathcal{H}$ consists of all possible partitions of $\mathcal{I}_p \times \mathcal{I}_q \times \mathcal{I}_K$, it covers the case of element-wise groups as well. On the other hand, the absence of any horizontal (i.e. across k) dependency simply decomposes the problems (2.5) and (2.6) into K independent sub-problems that can be solved separately either by setting $K = 1$ in our framework or by using existing methods, such as Lin et al. (2016a).

2.2.1 Alternating Block 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 algorithm as discussed in detail in Lin et al. (2016a). We choose the initial values $\{\widehat{\mathbf{B}}^{k(0)}\}$ by fitting separate lasso regression models for each j and k:

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

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

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

The steps of our full estimation procedure, coined as the *Joint Multiple Multi-Layer Estimation* (JMMLE) method, are 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

A number of methods have been proposed in the literature to select regularization tuning parameters in ℓ_1-penalized problems. Some approaches rely on traditional criteria like cross-validation, Akaike Information Criterion (AIC) (Danaher et al., 2014) or the Bayesian Information Criterion (BIC) (Lin et al., 2016a; Ma and Michailidis, 2016). A number of studies have proposed their modifications for the case when feature dimensions increase with sample size (Gao et al., 2012; Kim et al., 2012).

As a demonstration, to select the tuning parameter λ_n we use the High-dimensional BIC (HBIC, Kim et al. (2012); Wang et al. (2013)), and for selecting γ_n in the node-wise regression step in the JSEM model (2.4), stick to the use of BIC as done by Ma and Michailidis (2016). Unlike BIC, the penalty term in HBIC scales with the parameter dimensions. As a result, the tuning parameter selected as the minimizer of HBIC asymptotically identifies the oracle estimator in ultra-high dimensional penalized problems (Fan and Tang, 2013; Wang et al., 2013). In our case, we train multiple JMMLE models using Algorithm 1 over

a finite set of values $\lambda_n \in \mathcal{D}_n$, and calculate their HBIC:

$$\text{HBIC}(\lambda_n;\Theta) = \frac{1}{n}\sum_{j=1}^{q}\sum_{k=1}^{K}\|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k\widehat{\mathbf{B}}_{-j,\lambda_n}^k)\boldsymbol{\theta}_j^k - \mathbf{X}^k\widehat{\mathbf{B}}_{j,\lambda_n}^k\|^2 + \log(\log n)\frac{\log(pq)}{n}\sum_{k=1}^{K}\left(\|\mathbf{B}^k\|_0 + |\widehat{E}_{y,\gamma_n^*(\lambda_n)}^k|\right).$$

Following this, we choose the optimal λ_n as the empirical minimizer of HBIC over $\mathcal{D}_n$: $\lambda^* = \arg\min_{\lambda_n\in\mathcal{D}_n}\text{HBIC}(\lambda, \widehat{\Theta}_{\gamma_n^*(\lambda_n)})$.

The step for updating Θ, i.e. (2.10), in our JMMLE algorithm is analogous to the JSEM method Ma and Michailidis (2016), hence we use BIC to select the penalty parameter γ_n. In our setting the BIC for a given γ_n and fixed $\mathcal{B}$ is given by:

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

where γ_n in subscript indicates the corresponding quantity is calculated taking γ_n 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_n as the one with the smallest BIC over a fixed set of values $\mathcal{C}_n$. Thus for a fixed $\lambda_n \equiv \lambda$, our final choice of γ_n will be $\gamma_n^*(\lambda) = \arg\min_{\gamma_n\in\mathcal{C}_n}\text{BIC}(\gamma_n;\widehat{\mathcal{B}}_{\lambda_n})$.

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 body of the paper, while detailed proofs and auxiliary results are delegated to the Appendix.

We introduce some additional notation and define technical conditions that help establish the results that follow. Denote the true values of the parameters by $\Omega_{x0} = \{\Omega_{x0}^k\}, \Omega_{y0} = \{\Omega_{y0}^k\}, \Theta_0 = \{\Theta_{0j}\}, \mathcal{B}_0 = \{\mathbf{B}_0^k\}$. Sparsity levels of individual true parameters are indicated by $s_j := |\text{supp}(\Theta_{0j})|, b_k := |\text{supp}(\mathbf{B}_0^k)|$. Also define $S := \sum_{j=1}^q s_j, B := \sum_{k=1}^K b_k, s := \max_{j\in\mathcal{I}_q} s_j$, and $\mathcal{X} := \{\mathbf{X}^k\}_{k=1}^K, \mathcal{E} := \{\mathbf{E}^k\}_{k=1}^K$.

Definition 1 (Bounded eigenvalues). A positive definite matrix $\Sigma \in \mathbb{M}(b,b)$ is said to have bounded eigenvalues with constants (c_0, d_0) if

$$0 < 1/c_0 \le \Lambda_{\min}(\Sigma) \le \Lambda_{\max}(\Sigma) \le 1/d_0 < \infty$$

Definition 2 (Diagonal dominance). A matrix $\mathbf{M} \in \mathbb{M}(b,b)$ is said to be strictly diagonally dominant if for all $a \in \mathcal{I}_b$,

$$|(\mathbf{M})_{aa}| > \sum_{a'\neq a}|(\mathbf{M})_{aa'}|$$

Also denote $\Delta_0(\mathbf{M}) = \min_a\{|(\mathbf{M})_{aa}| - \sum_{a'\neq a}|(\mathbf{M})_{aa'}|\}$.

Our first result establishes the convergence of Algorithm 1 for fixed realizations of $(\mathcal{X}, \mathcal{E})$.

Theorem 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, and
(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 sub-problems. 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 $\widehat{\Theta}$ and $\widehat{\boldsymbol{\beta}}$ the generic estimators given by

$$\widehat{\Theta}_j = \underset{\Theta_j \in \mathbb{M}(q-1, K)}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|\widehat{\mathbf{E}}_j^k - \widehat{\mathbf{E}}_{-j}^k \boldsymbol{\theta}_j^k\|^2 + \gamma_n \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \|\boldsymbol{\theta}_{jj'}^{[g]}\| \right\}; \quad j \in \mathcal{I}_q, \tag{2.12}$$

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pqK}}{\arg\min} \left\{ -2\boldsymbol{\beta}^T \widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^T \widehat{\boldsymbol{\Gamma}} \boldsymbol{\beta} + \lambda_n \sum_{h \in \mathcal{H}} \|\boldsymbol{\beta}^{[h]}\| \right\}, \tag{2.13}$$

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

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

Using matrix algebra it is easy to see that solving for $\mathcal{B}$ in (2.5) given a fixed $\widehat{\Theta}$ is equivalent to solving (2.13).

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

(E1) The matrices $\Omega_{y0}^k, k \in \mathcal{I}_K$ are diagonally dominant,
(E2) The matrices $\Sigma_{y0}^k, k \in \mathcal{I}_K$ have bounded eigenvalues with constants (c_y, d_y) that are common across k.

Now we are in a position to establish the estimation consistency for (2.12), as well as the consistency of the final estimates $\widehat{\Omega}_y^k$ using their support sets.

Theorem 2. *Consider random $(\mathcal{X}, \mathcal{E})$, any deterministic $\widehat{\mathcal{B}}$ that satisfy the following bound*

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

where $C_\beta = O(1)$ depends only on $\mathcal{B}_0$. Then, for sample size $n \gtrsim \log(pq)$ the following hold: (I) Denote $|g_{\max}| = \max_{g \in \mathcal{G}_y} |g|$. Then for the choice of tuning parameter

$$\gamma_n \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0 \sqrt{\frac{\log(pq)}{n}},$$

where $\mathbb{Q}_0 = O(1)$ depends on the model parameters only, we have

$$\|\widehat{\Theta}_j - \Theta_{0,j}\|_F \leq 12\sqrt{s_j}\gamma_n/\psi, \tag{2.15}$$

$$\sum_{j \neq j', g \in \mathcal{G}_y^{jj'}} \|\hat{\boldsymbol{\theta}}_{jj'}^{[g]} - \boldsymbol{\theta}_{0,jj'}^{[g]}\| \leq 48 s_j \gamma_n/\psi. \tag{2.16}$$

(II) For the choice of tuning parameter $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0\sqrt{\log(pq)/n}$,

$$\frac{1}{K}\sum_{k=1}^{K} \|\widehat{\Omega}_y^k - \Omega_{y0}^k\|_F \leq O\left(\mathbb{Q}_0 \sqrt{\frac{|g_{\max}|S}{K}} \sqrt{\frac{\log(pq)}{n}}\right), \tag{2.17}$$

both with probability $\geq 1 - K(1/p^{\tau_1 - 2} - 12c_1 \exp[-(c_2^2 - 1)\log(pq)] - 2\exp(-c_3 n) - 6c_4 \exp[-(c_5^2 - 1)\log(pq)])$, for some constants $c_1, c_3, c_4 > 0, c_2, c_5 > 1, \tau_1 > 2$.

To prove an equivalent result for the solution of (2.13), we need the following conditions.

(E3) The matrices $(\mathbf{T}^k)^2, k \in \mathcal{I}_K$ are diagonally dominant,
(E4) The matrices $\Sigma_{x0}^k, k \in \mathcal{I}_K$ have bounded eigenvalues with common constants (c_x, d_x).

Given these, we next establish the required consistency results.

Theorem 3. *Assume random $(\mathcal{X}, \mathcal{E})$, and fixed $\widehat{\Theta}$ so that for $j \in \mathcal{I}_q$,*

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

for some $C_\Theta = O(1)$ dependent on Θ_0 only. Then, given the choice of tuning parameter

$$\lambda_n \geq 4\sqrt{|h_{\max}|}\mathbb{R}_0 \sqrt{\frac{\log(pq)}{n}},$$

where $\mathbb{R}_0 = O(1)$ depends on the population parameters only, the following hold

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 48\sqrt{|h_{\max}|} B \lambda_n/\psi_* \tag{2.18}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq 12\sqrt{B}\lambda_n/\psi_* \tag{2.19}$$

$$\sum_{h \in \mathcal{H}} \|\boldsymbol{\beta}^{[h]} - \boldsymbol{\beta}_0^{[h]}\| \leq 48 B \lambda_n/\psi_* \tag{2.20}$$

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \widehat{\boldsymbol{\Gamma}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq 72 B \lambda_n^2/\psi_* \tag{2.21}$$

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

$$\psi_* = \frac{1}{2}\min_k \left[\Lambda_{\min}(\Sigma_{x0}^k)\left(\Delta_0((\mathbf{T}_0^k)^2) - d_k C_\Theta\sqrt{\frac{\log(pq)}{n}}\right)\right],$$

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

Following the choice of tuning parameters in Theorems 2 and 3, $S = o(n/\log(pq))$ and $B = o(n/\log(pq))$ are sufficient conditions on the sparsity of corresponding parameters for the JMMLE estimators to be consistent.

Finally, we ensure that the starting values are satisfactory as previously discussed.

Theorem 4. *Consider the starting values as derived in* (2.7) *and* (2.8). *For sample size* $n \gtrsim \log(pq)$, *and the choice of the tuning parameter*

$$\lambda_n \geq 4c_2 \max_{k\in\mathcal{I}_K}\left\{[\Lambda_{\max}(\Sigma_{x0}^k)\Lambda_{\max}(\Sigma_{y0}^k)]^{1/2}\right\}\sqrt{\frac{\log(pq)}{n}},$$

we have $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\|_1 \leq 64B\lambda_n/\psi^*$ *with probability* $\geq 1 - 6c_1\exp(-(c_2^2-1)\log(pq)) - 2\exp(c_3 n)$. *Also, for* $\gamma_n \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0\sqrt{\log(pq)/n}$ *we have*

$$\|\widehat{\Theta}_j^{(0)} - \Theta_{0,j}\|_F \leq 12\sqrt{s_j}\gamma_n/\psi,$$
$$\sum_{j\neq j', g\in\mathcal{G}_y^{jj'}} \|\hat{\boldsymbol{\theta}}_{jj'}^{[g](0)} - \boldsymbol{\theta}_{0,jj'}^{[g]}\| \leq 48 s_j\gamma_n/\psi,$$

with probability $\geq 1-K(1/p^{\tau_1-2} - 12c_1\exp[-(c_2^2-1)\log(pq)] - 2\exp(-c_3 n) - 6c_4\exp[-(c_5^2-1)\log(pq)])$.

Putting all the pieces together, estimation consistency for the limit points of Algorithm 1 given our choice of starting values follows in a straightforward manner.

Corollary 1. *Assume conditions (E1)-(E4), 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_n

$$\lambda_n \geq 4\max\left[c_2\max_{k\in\mathcal{I}_K}\left\{[\Lambda_{\max}(\Sigma_{x0}^k)\Lambda_{\max}(\Sigma_{y0}^k)]^{1/2}\right\}, \sqrt{|h_{\max}|}\mathbb{R}_0\right]\sqrt{\frac{\log(pq)}{n}},$$

we have

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

with probability $\geq 1 - 18c_1\exp[-(c_2^2-1)\log(pq)] - 4\exp(-c_3 n)$.

(II) For the choice of γ_n

$$\gamma_n \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0\sqrt{\frac{\log(pq)}{n}},$$

(2.15) *and* (2.16) *hold, while for* $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0\sqrt{\log(pq)/n}$, (2.17) *holds, both with probability* $\geq 1 - K(2/p^{\tau_1-2} - 24c_1\exp[-(c_2^2-1)\log(pq)] - 4\exp(-c_3 n) - 12c_4\exp[-(c_5^2-1)\log(pq)])$.

Remark 3. To save computation time for high data dimensions, an initial screening step, e.g. the debiased lasso procedure of Javanmard and Montanari (2014), can be used to first restrict the support set of $\mathbf{B}_j^k$ before obtaining the initial estimates using (2.7). The consistency properties of resulting initial and final estimates follow along the lines of the special case $K = 1$ discussed in Lin et al. (2016a), in conjunction with Theorem 4 and Corollary 1, respectively. We leave the details to the reader.

Remark 4. While the proof of the above results roughly follow roughly similar roadmaps as the $K = 1$ and ℓ_1-penalized case of Lin et al. (2016a) and the joint structural estimation of Ma and Michailidis (2016) and utilize Gaussian concentration inequalities, generalization to an optional grouping structure in $\mathcal{B}$ and $K > 1$ add technical complexity to the proofs. More importantly, we work in presence of minimal assumptions, steering clear of conditions used in previous works, like Incoherence (Lin et al., 2016a) and Uniform Irrepresentability (Ma and Michailidis, 2016) that are often difficult to verify in practice.

3. Hypothesis testing in multilayer models

In this section, we lay out a framework for hypothesis testing in our proposed joint multi-layer structure. Present literature in high-dimensional hypothesis testing either focuses on testing for similarities in the within-layer connections of single-layer networks (Cai and Liu, 2016; Liu, 2017), or coefficients of single response penalized regression (van de Geer et al., 2014; Zhang and Zhang, 2014; Mitra and Zhang, 2016). However, to our knowledge no method is available in the literature to perform testing for *between-layer* connections in a two-layer (or multi-layer) setup.

Denote the i^{th} row of the coefficient matrix $\mathbf{B}^k$ by $\mathbf{b}_i^k$, for $i \in \mathcal{I}_p$. In this section we are generally interested in obtaining asymptotic sampling distributions of $\widehat{\mathbf{b}}_i^k$, and subsequently formulating testing procedures to detect similarities or differences across k in the full vector $\mathbf{b}_i^k$ or its elements. There are two main challenges in doing the above- firstly the need to mitigate the bias of the group-penalized JMMLE estimators, and secondly the dependency among response nodes translating into the need for controlling false discovery rate while simultaneously testing for several element-wise hypotheses concerning the true values $b_{0ij}^k, j \in \mathcal{I}_q$. To this end, in Section 3.1 we first propose a debiased estimator for $\mathbf{b}_i^k$ that makes use of already computed (using JSEM) node-wise 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 assume $K = 2$, and propose asymptotic global tests for detecting differential effects of a variable in the upper layer, i.e. testing for the null hypothesis $H_0^i : \mathbf{b}_{0i}^1 = \mathbf{b}_{0i}^2$, as well as pairwise simultaneous tests across $j \in \mathcal{I}_q$ for detecting the element-wise differences $b_{0ij}^1 - b_{0ij}^2$.

3.1 Debiased estimators and asymptotic normality

Zhang and Zhang (2014) proposed a debiasing procedure for lasso estimates and subsequently calculate confidence intervals for individual coefficients β_j in high-dimensional linear regression: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{M}(n, p)$ and $\epsilon_r \sim N(0, \sigma^2), r \in \mathcal{I}_n$ for some $\sigma > 0$.

Given an initial lasso estimate $\widehat{\boldsymbol{\beta}}^{(\text{init})} \in \mathbb{R}^p$ their debiased estimator was defined as:

$$\hat{\beta}_j^{(\text{deb})} = \hat{\beta}_j^{(\text{init})} + \frac{\mathbf{z}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(\text{init})})}{\mathbf{z}^T\mathbf{x}_j},$$

where $\mathbf{z}_j$ is the vector of residuals from the ℓ_1-penalized regression of $\mathbf{x}_j$ on $\mathbf{X}_{-j}$. With centering around the true parameter value, say β_j^0, and proper scaling this has an asymptotic normal distribution:

$$\frac{\hat{\beta}_j^{(\text{deb})} - \beta_j^0}{\|\mathbf{z}_j\|/|\mathbf{z}_j^T\mathbf{x}_j|} \sim N(0, \sigma^2)$$

Essentially, they obtain the debiasing factor for the j^{th} coefficient by taking residuals from the regularized regression and scale them using the projection of $\mathbf{x}_j$ onto a space approximately orthogonal to it. Mitra and Zhang (2016) later generalized this idea to group lasso estimates. Further, van de Geer et al. (2014) and Javanmard and Montanari (2014) performed debiasing on the entire coefficient vectors.

We start off by defining debiased estimates for individual rows of the coefficient matrices $\mathbf{B}^k$ in our two-layer model:

$$\widehat{\mathbf{c}}_i^k = \widehat{\mathbf{b}}_i^k + \frac{1}{nt_i^k}\left(\mathbf{X}_i^k - \mathbf{X}_{-i}^k\widehat{\boldsymbol{\zeta}}_i^k\right)^T(\mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k); \quad i \in \mathcal{I}_p, k \in \mathcal{I}_K \tag{3.1}$$

where $\widehat{\mathbf{b}}_i^k$ denotes the i^{th} row of $\widehat{\mathbf{B}}^k$, and $t_i^k = (\mathbf{X}_i^k - \mathbf{X}_{-i}^k\widehat{\boldsymbol{\zeta}}_i^k)^T\mathbf{X}_i^k/n$, and $\widehat{\boldsymbol{\zeta}}_i^k, \widehat{\mathbf{B}}^k$ are *generic estimators* of the neighborhood coefficient matrices in the upper layer and within-layer coefficient matrices, respectively. By structure this is similar to the proposal of Zhang and Zhang (2014). However, as seen shortly, minimal conditions need to be imposed on the parameter estimates used in (3.1) for the asymptotic results based on a scaled version of the debiased estimator to go thorugh, and they continue to hold for arbitrary sparsity patterns over k in all of the parameters.

Present methods of debiasing coefficients from regularized regression require specific assumptions on the regularization structure of the main regression, as well as on how to calculate the debiasing factor. While Zhang and Zhang (2014), Javanmard and Montanari (2014) and van de Geer et al. (2014) work on coefficients from lasso regressions, Mitra and Zhang (2016) debias the coefficients of pre-specified groups in the coefficient vector from a group lasso. Current proposals for obtaining the debiasing factor available in the literature include node-wise lasso (Zhang and Zhang, 2014) and a variance minimization scheme with ℓ_∞-constraints (Javanmard and Montanari, 2014). In comparison, we only assume the following generic constraints on the parameter estimates used in our procedure.

(T1) For the upper layer neighborhood coefficients, the following holds for all $k \in \mathcal{I}_K$:

$$\|\widehat{\boldsymbol{\zeta}}^k - \boldsymbol{\zeta}_0^k\|_1 \leq D_\zeta = O\left(\sqrt{\frac{\log p}{n}}\right)$$

where D_ζ depends only on the true values, i.e. $\{\zeta_0^k\}$.

(T2) The lower layer precision matrix estimators satisfy for all $k \in \mathcal{I}_K$

$$\|\widehat{\Omega}_y^k - \Omega_{y0}^k\|_\infty \leq D_\Omega = O\left(\sqrt{\frac{\log(pq)}{n}}\right)$$

where D_Ω depends only on Ω_{y0}.
(T3) For the regression coefficient matrices, the following holds for all $k \in \mathcal{I}_K$:

$$\|\widehat{\mathbf{B}}^k - \mathbf{B}_0^k\|_1 \leq D_\beta = O\left(\sqrt{\frac{\log(pq)}{n}}\right),$$

where D_β depends on $\mathcal{B}_0$ only.

Given these conditions, the following result provides the asymptotic joint distribution of a scaled version of the debiased coefficients. A similar result for fixed design in the context of single-response linear regression can be found in Stucky and van de Geer (2018). However, the authors use the nuclear norm as the loss function while obtaining the debiasing factors and employ the resulting Karush-Kuhn-Tucker (KKT) conditions to derive their results, whereas we leverage bounds on generic parameter estimates combined with the sub-Gaussianity of our random design matrices.

Theorem 5. *Define* $\widehat{s}_i^k = \sqrt{\|\mathbf{X}_i^k - \mathbf{X}_{-i}^k \widehat{\boldsymbol{\zeta}}_i^k\|^2/n}$, *and* $m_i^k = \sqrt{n} t_i^k / \widehat{s}_i^k$. *Consider parameter estimates that satisfy conditions (T1)-(T3). Define the following:*

$$\begin{aligned}
\widehat{\Omega}_y &= \mathrm{diag}(\widehat{\Omega}_y^1, \ldots, \widehat{\Omega}_y^K) \\
\mathbf{M}_i &= \mathrm{diag}(m_i^1, \ldots, m_i^K) \\
\widehat{\mathbf{C}}_i &= \mathrm{vec}(\widehat{\mathbf{c}}_i^1, \ldots, \widehat{\mathbf{c}}_i^K)^T \\
\mathbf{D}_i &= \mathrm{vec}(\mathbf{b}_{0i}^1, \ldots, \mathbf{b}_{0i}^K)^T
\end{aligned}$$

Also assume that conditions (E2), (E4) hold, and the matrices $\Omega_{x0}^k, k \in \mathcal{I}_K$ *are diagonally dominant. Then, for sample size satisfying* $\log p = o(n^{1/2}), \log q = o(n^{1/2})$ *we have*

$$\widehat{\Omega}_y^{1/2} \mathbf{M}_i (\widehat{\mathbf{C}}_i - \mathbf{D}_i) \sim \mathcal{N}_{Kq}(\mathbf{0}, \mathbf{I}) + \mathbf{R}_n \tag{3.2}$$

where $\|\mathbf{R}_n\|_\infty = o_P(1)$.

3.2 Test formulation

We now simply plug in estimators from the JMMLE algorithm in Theorem 5. Doing so is fairly straightforward. Condition (T1) is ensured by the JSEM penalized neighborhood estimators in (2.4) (immediate from Proposition A.1 in Ma and Michailidis (2016)). On the other hand, bounds on total sparsity of the true coefficient matrices: $B = o(\sqrt{n}/\log(pq))$, and lower layer precision matrices: $S = o(n/\log(pq)$, in conjunction with Corollary 1, ensure conditions (T2) and (T3), respectively -all with probability approaching 1 as $(n, p, q) \to \infty$.

An asymptotic joint distribution of debiased versions of the JMMLE regression estimates can then be obtained immediately.

Corollary 2. *Consider the estimates $\widehat{\mathcal{B}}$ and $\widehat{\Omega}_y$ obtained from Algorithm 1, and upper layer neighborhood coefficients from solving the node-wise regression in (2.4). Suppose that $\log(pq)/\sqrt{n} \to 0$, and the sparsity conditions $B = o(\sqrt{n}/\log(pq)), S = o(n/\log(pq))$ are satisfied. Then, with the same notations as in Theorem 5 we have*

$$\widehat{\Omega}_y^{1/2}\mathbf{M}_i(\widehat{\mathbf{C}}_i - \mathbf{D}_i) \sim \mathcal{N}_{Kq}(\mathbf{0}, \mathbf{I}) + \mathbf{R}_{1n} \tag{3.3}$$

where $\|\mathbf{R}_{1n}\|_\infty = o_P(1)$.

We are now ready to formulate asymptotic global and simultaneous testing procedures based on Corollary 2. In this paper, we restrict our attention to testing for pairwise differences only. Specifically, we set $K = 2$, and are interested in testing whether there are overall and elementwise differences between individual rows of the true coefficient matrices, i.e. $\mathbf{b}_{0i}^1$ and $\mathbf{b}_{0i}^2$.

When $\mathbf{b}_{0i}^1 = \mathbf{b}_{0i}^2$, it is immediate from Corollary 2 that a scaled version of the vector of estimated differences $\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2$ follows a q-variate multivariate normal distribution. Consequently, we formulate a global test for detecting differential overall downstream effect of the i^{th} covariate in the upper layer.

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

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

where $\widehat{\Sigma}_y^k = (\widehat{\Omega}_y^k)^{-1}, k = 1, 2$.
3. Reject H_0^i if $D_i \geq \chi^2_{q,1-\alpha}$.

Besides controlling the type-I error at a specified level, the above testing procedure maintains rate optimal power.

Theorem 6. *Consider the global test given in Algorithm 2, performed using parameter estimates satisfying conditions (T1)-(T3). Define $\boldsymbol{\delta} := \mathbf{b}_{0i}^1 - \mathbf{b}_{0i}^2$. Further, assume that either of the following sufficient conditions are satisfied.*

(I) The following bound holds: $D_\Omega \leq \Delta_0(\Omega_{y0}^k), k \in \mathcal{I}_K$;

(II) For every $j \in \mathcal{I}_q, k \in \mathcal{I}_K$, we have $\sum_{j'=1}^q |\sigma_{y0,jj'}^k|^q \leq c_0(p)$ for some $q \in [0, 1)$ and positive-valued function $c_0(\cdot)$.

Denote $\sigma_{x0,i,-i}^k = Var(X_i^k - \mathbb{X}_{-i}^k \boldsymbol{\zeta}_{0,i}^k)$. Then, the power of the global test is given by

$$K_q \left(\chi^2_{q,1-\alpha} + n\boldsymbol{\delta}^T \left(\frac{\Sigma_{y0}^1}{\sigma_{x0,i,-i}^1} + \frac{\Sigma_{y0}^2}{\sigma_{x0,i,-i}^2} \right)^{-1} \boldsymbol{\delta} \right) + o(1)$$

where K_q is the cumulative distribution function of the χ_q^2 distribution. Consequently, for $\|\boldsymbol{\delta}\| > O(n^{-1/2})$, $P(H_0^i$ is rejected $) \to 1$ as $(n, p, q) \to \infty$.

The conditions (I) or (II) above are needed to derive upper bounds for $\|\widehat{\Sigma}_y^k - \Sigma_{y0}^k\|_\infty$ using those for $\|\widehat{\Omega}_y^k - \Omega_{y0}^k\|_\infty$. While (I) imposes a potentially more stringent bound on the estimation error of Ω_y, (II) restricts the power calculations to a uniformity class of covariance matrices (Bickel and Levina, 2008; Cai et al., 2012b).

Remark 5. While the formulation of the testing procedure till now broadly gives parallel results as Zhang and Zhang (2014) and Mitra and Zhang (2016), it does so without assuming any specific penalty function or (group) sparsity conditions (such as strong group sparsity in Mitra and Zhang (2016)). Instead we only require the standard finite-sample bounds (T1)-(T3) that are satisfied by both sparse and non-sparse estimators in a high-dimensional setting (Basu et al., 2019).

3.3 Control of False Discovery Rate

Given that the null hypothesis is rejected, we 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 all $j \in \mathcal{I}_q$. Here we use the test statistic

$$d_{ij} = \frac{\widehat{c}_{ij}^1 - \widehat{c}_{ij}^2}{\sqrt{\hat{\sigma}_{jj}^1/(m_i^1)^2 + \hat{\sigma}_{jj}^2/(m_i^2)^2}} \tag{3.4}$$

with $\hat{\sigma}_{jj}^k$ being the j^{th} diagonal element of $\widehat{\Sigma}_y^k, k = 1, 2$.

For the purpose of simultaneous testing, we consider tests with a common rejection threshold τ, i.e. for $j \in \mathcal{I}_q$, H_0^{ij} is rejected if $|d_{ij}| > \tau$. We denote $\mathcal{H}_0^i = \{j : b_{0,ij}^1 = b_{0,ij}^2\}$ 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 $|d_{ij}| \geq \hat{\tau}$.

To ensure that this procedure maintains FDR and FDP asymptotically at a pre-specified level $\alpha \in (0,1)$, we need some dependence conditions on true correlation matrices in the lower layer. Following Liu and Shao (2014), we consider the following two types of dependencies:

(D1) Define $r^k_{jj'} = \sigma^k_{y0,jj'}/\sqrt{\sigma^k_{y0,jj}\sigma^k_{y0,j'j'}}$ for $j, j' \in \mathcal{I}_q, k = 1,2$. Suppose there exists $0 < r < 1$ such that $\max_{1\le j<j'\le q} |r^k_{jj'}| \le r$, and for every $j \in \mathcal{I}_q$,

$$\sum_{j'=1}^{q} \mathbb{I}\left\{|r^k_{jj'}| \ge \frac{1}{(\log q)^{2+\theta}}\right\} \le O(q^\rho)$$

for some $\theta > 0$ and $0 < \rho < (1-r)/(1+r)$.

(D1*) Suppose there exists $0 < r < 1$ such that $\max_{1\le j<j'\le q} |r^k_{jj'}| \le r$, and for every $j \in \mathcal{I}_q$,

$$\sum_{j'=1}^{q} \mathbb{I}\left\{|r^k_{jj'}| > 0\right\} \le O(q^\rho)$$

for some $0 < \rho < (1-r)/(1+r)$.

Originally proposed by Liu and Shao (2014), the above dependency conditions are meant to control the amount of correlation amongst the test statistics. Condition (D1) allows each variable to be highly correlated with at most $O(q^\rho)$ other variables and weakly correlated with others, while (D1*) limits the number of variables to have *any* correlation with it to $O(q^\rho)$. Note that (D1*) is a stronger condition, and can be seen as the limiting condition of (D1) as $q \to \infty$.

Theorem 7. *Suppose* $\mu_j = b^1_{0,ij} - b^2_{0,ij}, \sigma^2_j = \sigma^1_{y0,jj}/\sigma^1_{x0,i,-i} + \sigma^2_{y0,jj}/\sigma^2_{x0,i,-i}$. *Assume the following holds as* $(n,q) \to \infty$

$$\left|\left\{j \in \mathcal{I}_q : |\mu_j/\sigma_j| \ge 4\sqrt{\log q/n}\right\}\right| \to \infty \tag{3.5}$$

Next, consider conditions (D1) and (D1). If (D1) is satisfied, then the following holds when* $\log q = O(n^\xi), 0 < \xi < 3/23$*:*

$$\frac{FDP(\hat\tau)}{(|\mathcal{H}^i_0|/q)\alpha} \xrightarrow{P} 1; \quad \lim_{n,q\to\infty} \frac{FDR(\hat\tau)}{(|\mathcal{H}^i_0|/q)\alpha} = 1 \tag{3.6}$$

Further, if (D1) is satisfied, then* (3.6) *holds for* $\log q = o(n^{1/3})$.

The condition (3.5) is essential for FDR control in a diverging parameter space (Liu and Shao, 2014; Liu, 2017).

Remark 6. Based on the FDR control procedure in Algorithm 3, we can perform *within-row thresholding* in the matrices $\widehat{\mathbf{B}}^k$ to tackle group misspecification.

$$\begin{aligned}
\hat\tau^k_i &:= \inf\left\{\tau \in \mathbb{R} : 1 - \Phi(\tau) \le \frac{\alpha}{2q} \max\left(\sum_{j\in\mathcal{I}_q} \mathbb{I}(|\sqrt{\hat\omega^k_{jj}} m^k_i \hat c^k_{ij}| \ge \tau), 1\right)\right\} \\
\hat b^{k,\text{thr}}_{ij} &= \hat b^k_{ij} \mathbb{I}\left(|\sqrt{\hat\omega^k_{jj}} m^k_i \hat c^k_{ij}| \ge \hat\tau^k_i\right)
\end{aligned} \tag{3.7}$$

Even without group misspecification, this helps identify directed edges between layers that have high nonzero values. Similar post-estimation thresholdings have been proposed in the context of multitask regression (Obozinski et al., 2011; Majumdar and Chatterjee, 2018) and neighborhood selection (Ma and Michailidis, 2016). However, our procedure is the first one to provide explicit guarantees on the amount of false discoveries while doing so.

Remark 7. Following (3.5), a sufficient condition on the sparsity of $\mathcal{B}_0$ for FDR to be asymptotically controlled at some specified level is $B = o(n^{\zeta}/\log q)$ if (D1) is satisfied, and $B = o(n^{1/3}/\log q)$ if (D1*) is satisfied. In comparison, our results for the global testing procedure require $B = o(\sqrt{n}/\log(pq))$, and point estimation requires $B = o(n/\log(pq))$. In finite samples settings, the stricter sparsity requirements translate to higher sample sizes being needed (given the same (p, q)) for our testing procedures to have satisfactory performances compared to estimation only (See Sections 4.1 and 4.2).

In recent work, Javanmard and Montanari (2018+) showed that the $o(\sqrt{n}/\log p)$ bound on the sparsity of the true coefficient vector required to construct confidence intervals from debiased lasso coefficient estimates (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014) can be weakened to $o(n/(\log p)^2)$ when the random design precision matrix is known, or is unknown but satisfies certain sparsity assumptions. Similar relaxations may be possible in our case. For example, the machinery in Liu (2017), which performs simultaneous testing in multiple (single layer) GGMs using slightly modified FDR thresholds, can be useful in obtaining (3.6) for $\log q = o(n^{1/2})$ under (D1), (D1*) or other suitable dependency assumptions.

4. Performance evaluation

Next, we evaluate the performance of our proposed JMMLE algorithm and the hypothesis testing framework in a two-layer simulation setup (Sections 4.1 and 4.2, respectively), and also introduce some computational techniques that significantly accelerate calculations for high data dimensions (Section 4.3).

4.1 Simulation 1: estimation

As a first step towards obtaining a two-layer structure with horizontal (across k) integration and inter-layer directed edges, we generate the precision matrices $\{\Omega^k_{x0}\}$ and $\{\Omega^k_{y0}\}$ using a dependency structure across k that was first used in the simulation study of Ma and Michailidis (2016). We set $K = 5$, and set different shared sparsity patterns across k inside the lower $p/2 \times p/2$ block of the upper layer precision matrices, and outside the block. In our notation, this gives the following elementwise group structure:

$$\mathcal{G}_{x,ii'} = \begin{cases} \{(1,2),(3,4),5\} \text{ if } i \le p/2 \text{ or } j \le p/2 \\ \{(1,3,5),(2,4)\} \text{ otherwise} \end{cases}$$

The schematic in Figure 3 illustrates this structure. We set an off-diagonal element inside each of these common blocks (i.e. A, B, C and α, β in the figure) to be non-zero with probability $\pi_x \in \{5/p, 30/p\}$, then generate the values of all non-zero elements independently from the uniform distribution in the interval $[-1, 0.5] \cup [0.5, 1]$. The precision matrices Ω^k_{x0}

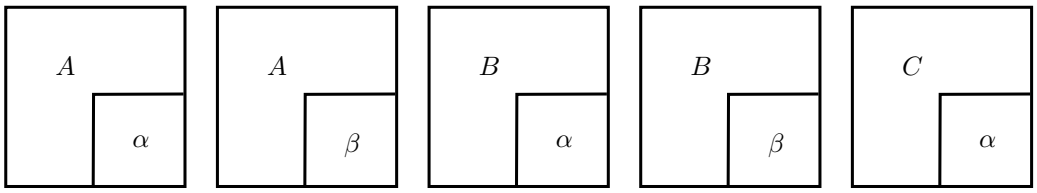

Figure 3: Shared sparsity patterns across k for the precision matrices $\{\Omega^k_{x0}\}$ and $\{\Omega^k_{y0}\}$

are generated by putting together the corresponding common blocks, their positive definiteness ensured by setting all diagonal elements to be $1 + |\Lambda_{\min}(\Omega^k_{x0})|$. Then, we get elements in the covariance matrix as

$$\sigma^k_{x0,ii'} = (\bar{\Omega}^k_{x0})_{ii'} / \sqrt{(\bar{\Omega}^k_{x0})_{ii}(\bar{\Omega}^k_{x0})_{i'i'}}, \text{ where } \bar{\Omega}^k_{x0} = (\Omega^k_{x0})^{-1},$$

and generate rows of $\mathbf{X}^k$ independently from $\mathcal{N}(0, \Sigma^k_{x0})$. We obtain Σ^k_{y0} and then $\mathbf{E}^k$ using the same setup but with the number of variables being q and setting off-diagonal elements non-zero with probability $\pi_y \in \{5/q, 30/q\}$. To obtain the matrices $\mathbf{B}^k_0$, for a fixed $(i, j), i \in \mathcal{I}_p, j \in \mathcal{I}_q$, we set $b^k_{0,ij}$ non-zero across all k with probability $\pi \in \{5/p, 30/p\}$, generate the non-zero groups independently from $\text{Unif}\{[-1, 0.5] \cup [0.5, 1]\}$, and set $\mathbf{Y}^k = \mathbf{X}^k\mathbf{B}^k_0 + \mathbf{E}^k, k \in \mathcal{I}_K$. Finally, we generate 150 such independent two-layer datasets for each of the following model settings:

- Set $\pi_x = \pi = 5/p, \pi_y = 5/q$, and

$$(p, q, n) \in \{(60, 30, 100), (30, 60, 100), (200, 200, 150), (300, 300, 150)\};$$

- Set $\pi_x = \pi = 30/p, \pi_y = 30/q$, and $(p, q, n) \in \{(200, 200, 100), (200, 200, 200)\}$.

We use the following arrays of tuning parameters to train Algorithm 1-

$$\gamma_n \in \{0.3, 0.4, ..., 1\}\sqrt{\frac{\log q}{n}}; \quad \lambda_n \in \{0.4, 0.6, ..., 1.8\}\sqrt{\frac{\log p}{n}},$$

using the one-step version (Section 4.3) instead of the full algorithm to save computation time.

We use the following performance metrics to evaluate our estimates $\widehat{\mathcal{B}} = \{\widehat{\mathbf{B}}^k\}$:

- True positive Rate-

$$\text{TPR}(\widehat{\mathbf{B}}_k) = \frac{|\operatorname{supp}(\hat{\mathbf{B}}^k) \cap \operatorname{supp}(\mathbf{B}^k_0)|}{|\operatorname{supp}(\mathbf{B}^k_0)|}; \quad \text{TPR}(\widehat{\mathcal{B}}) = \frac{1}{K}\sum_{k=1}^{K} \text{TP}(\widehat{\mathbf{B}}_k).$$

- True negative Rate-

$$\text{TNR}(\widehat{\mathbf{B}}_k) = \frac{|\operatorname{supp}^c(\hat{\mathbf{B}}^k) \cap \operatorname{supp}^c(\mathbf{B}^k_0)|}{|\operatorname{supp}^c(\mathbf{B}^k_0)|}; \quad \text{TNR}(\widehat{\mathcal{B}}) = \frac{1}{K}\sum_{k=1}^{K} \text{TNR}(\widehat{\mathbf{B}}_k).$$

(π_x, π_y)	(p, q, n)	Method	TPR	TNR	MCC	RF
$(5/p, 5/q)$	(60,30,100)	JMMLE	0.97(0.02)	0.99(0.003)	0.96(0.014)	0.24(0.033)
		Separate	0.96(0.018)	0.99(0.004)	0.93(0.014)	0.22(0.029)
	(30,60,100)	JMMLE	0.97(0.013)	0.99(0.002)	0.96(0.008)	0.27(0.024)
		Separate	0.99(0.009)	0.99(0.003)	0.93(0.017)	0.18(0.021)
	(200,200,150)	JMMLE	0.98(0.011)	1.0(0)	0.99(0.005)	0.16(0.025)
		Separate	0.99(0.001)	0.99 (0.001)	0.88(0.009)	0.18(0.007)
	(300,300,150)	JMMLE	1.0(0.001)	1.0(0)	0.99(0.001)	0.14 (0.015)
		Separate	1.0(0.001)	0.99(0.001)	0.84(0.01)	0.21(0.007)
$(30/p, 30/q)$	(200,200,100)	JMMLE	0.97(0.017)	1.0(0)	0.98(0.008)	0.21(0.032)
		Separate	0.32(0.01)	0.99(0.001)	0.49(0.009)	0.85(0.06)
	(200,200,200)	JMMLE	0.99(0.006)	1.0(0)	0.99(0.007)	0.13(0.016)
		Separate	0.97(0.004)	0.98(0.001)	0.93(0.002)	0.19(0.07)

Table 1: Table of outputs for estimation of regression matrices, giving empirical mean and standard deviation (in brackets) of each evaluation metric over 150 replications.

- Matthews Correlation Coefficient-

$$\mathrm{TP}(\widehat{\mathbf{B}}_k) = |\operatorname{supp}(\hat{\mathbf{B}}^k) \cap \operatorname{supp}(\mathbf{B}_0^k)|; \quad \mathrm{TN}(\widehat{\mathbf{B}}_k) = |\operatorname{supp}^c(\hat{\mathbf{B}}^k) \cap \operatorname{supp}^c(\mathbf{B}_0^k)|,$$

$$\mathrm{FP}(\widehat{\mathbf{B}}_k) = |\operatorname{supp}^c(\mathbf{B}_0^k)| - \mathrm{TN}(\widehat{\mathbf{B}}_k); \quad \mathrm{FN}(\widehat{\mathbf{B}}_k) = |\operatorname{supp}(\mathbf{B}_0^k)| - \mathrm{TP}(\widehat{\mathbf{B}}_k),$$

$$\mathrm{MCC}(\widehat{\mathbf{B}}_k) = \frac{\mathrm{TP}(\widehat{\mathbf{B}}_k)\mathrm{TN}(\widehat{\mathbf{B}}_k) - \mathrm{FP}(\widehat{\mathbf{B}}_k)\mathrm{FN}(\widehat{\mathbf{B}}_k)}{\sqrt{(\mathrm{TP}(\widehat{\mathbf{B}}_k) + \mathrm{FP}(\widehat{\mathbf{B}}_k))(\mathrm{TP}(\widehat{\mathbf{B}}_k) + \mathrm{FN}(\widehat{\mathbf{B}}_k))(\mathrm{TN}(\widehat{\mathbf{B}}_k) + \mathrm{FP}(\widehat{\mathbf{B}}_k))(\mathrm{TN}(\widehat{\mathbf{B}}_k) + \mathrm{FN}(\widehat{\mathbf{B}}_k))}},$$

$$\mathrm{MCC}(\widehat{\mathcal{B}}) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{MCC}(\widehat{\mathbf{B}}_k).$$

- Relative error in Frobenius norm-

$$\mathrm{RF}(\widehat{\mathcal{B}}) = \frac{1}{K}\sum_{k=1}^{K} \frac{\|\hat{\mathbf{B}}^k - \mathbf{B}_0^k\|_F}{\|\mathbf{B}_0^k\|_F}.$$

We use the same metrics to evaluate the precision matrix estimates $\widehat{\Omega}_y^k$ as well, with TPR and TNR calculations confined to off-diagonal entries.

Tables 1 and 2 summarize the results. For estimation of $\mathcal{B}$, we compare our results to the method in Lin et al. (2016a) that estimates parameters in each of the K two-layer structure separately, while for estimation of Ω_y, we compare them with the results in Lin et al. (2016a) and using the single-layer JSEM (Ma and Michailidis, 2016) that estimates Ω_y assuming structured sparsity patterns and centered matrices $\mathbf{Y}^k$, but not the data in the upper layer, i.e. $\mathcal{X}$.

Our joint method has higher average MCC across all data settings than the separate method for the estimation of $\mathcal{B}$, although TPR and TNR values are similar, except for

(π_x, π_y)	(p, q, n)	Method	TPR	TNR	MCC	RF
$(5/p, 5/q)$	(60,30,100)	JMMLE	0.76(0.018)	0.90(0.006)	0.61(0.024)	0.32(0.008)
		Separate	0.77(0.031)	0.92(0.007)	0.56(0.03)	0.51(0.017)
		JSEM	0.24(0.013)	0.8(0.003)	0.05(0.015)	1.03(0.002)
	(30,60,100)	JMMLE	0.7(0.018)	0.94(0.002)	0.55(0.018)	0.3(0.005)
		Separate	0.76(0.041)	0.89(0.015)	0.59(0.039)	0.49(0.014)
		JSEM	0.13(0.005)	0.9(0.001)	0.03(0.007)	1.04(0.001)
	(200,200,150)	JMMLE	0.68(0.017)	0.98(0)	0.48(0.013)	0.26(0.002)
		Separate	0.78(0.019)	0.97(0.001)	0.55(0.012)	0.6(0.007)
		JSEM	0.05(0.002)	0.97(0)	0.02(0.002)	1.01(0)
	(300,300,150)	JMMLE	0.71(0.014)	0.98(0)	0.44(0.008)	0.25(0.002)
		Separate	0.71(0.017)	0.98(0.001)	0.51(0.011)	0.59(0.005)
		JSEM	0.04(0.002)	0.98(0)	0.02(0.002)	1.01(0)
$(30/p, 30/q)$	(200,200,100)	JMMLE	0.77(0.016)	0.98(0)	0.46(0.013)	0.31(0.003)
		Separate	0.57(0.027)	0.44(0.007)	0.04(0.008)	0.84(0.002)
		JSEM	0.05(0.002)	0.97(0)	0.01(0.002)	1.01(0)
	(200,200,200)	JMMLE	0.76(0.018)	0.98(0)	0.55(0.015)	0.27(0.004)
		Separate	0.73(0.023)	0.94(0.003)	0.39(0.017)	0.62(0.011)
		JSEM	0.05(0.002)	0.97(0)	0.03(0.003)	1.01(0)

Table 2: Table of outputs for estimation of lower layer precision matrices over 150 replications.

$p = 200, q = 200, n = 100$ where JMMLE has a much higher average TPR. For estimation of Ω_y, incorporating information from the upper layer vastly improves performance, as demonstrated by the differences in performances between JMMLE and JSEM. For the 4 data settings with lower sparsity ($\pi_x = \pi = 5/p, \pi_y = 5/q$), JMMLE is either slightly conservative or has similar TPR and TNR compared to the separate method while estimating Ω_y. However, JMMLE does better in both of the higher sparsity settings. Lastly, for the estimation of both $\mathcal{B}$ and Ω_y, JMMLE gives more accurate estimates across the methods, as evident from the lower average RF values across all data settings.

4.1.1 Effect of heterogeneity

We repeat the above setups to check the performance of JMMLE in presence of within-group misspecification. For this task, we first set individual elements inside a non-zero group to be zero with probability 0.2 while generating the data, then pass the JMMLE estimates $\widehat{\mathbf{B}}^k$ through the FDR controlling thresholds as given in (3.7). The results are summarized in Tables 3 and 4. Across the simulation settings, values of all metrics are very close to the correctly specified counterparts in Table 1. Thus, the thresholding step proves largely effective. Also, in all cases the empirical FDR for estimating entries in $\mathcal{B}$ is below 0.2. The performance is slightly worse than the correctly specified cases when estimating Ω_y. This is expected, as the estimates $\widehat{\Omega}_y$ are obtained from neighborhood coefficients that are calculated based on the *pre-thresholding* coefficient estimates.

(π_x, π_y)	(p, q, n)	TPR($\widehat{\mathcal{B}}$)	TNR($\widehat{\mathcal{B}}$)	MCC($\widehat{\mathcal{B}}$)	RF($\widehat{\mathcal{B}}$)
$(5/p, 5/q)$	(60,30,100)	0.98 (0.01)	0.99 (0.002)	0.89 (0.017)	0.29 (0.014)
	(30,60,100)	0.94 (0.022)	0.99 (0.003)	0.93 (0.016)	0.31 (0.028)
	(200,200,150)	0.99 (0.002)	0.99 (0)	0.98 (0.004)	0.17 (0.007)
	(300,300,150)	0.99 (0.001)	1 (0)	0.99 (0.002)	0.15 (0.006)
$(30/p, 30/q)$	(200,200,100)	0.99 (0.006)	1 (0)	0.98 (0.005)	0.2 (0.014)
	(200,200,200)	0.99 (0.009)	1 (0)	0.98 (0.005)	0.15 (0.017)
(π_x, π_y)	(p, q, n)	TPR($\widehat{\Omega}_y$)	TNR($\widehat{\Omega}_y$)	MCC($\widehat{\Omega}_y$)	RF($\widehat{\Omega}_y$)
$(5/p, 5/q)$	(60,30,100)	0.71 (0.024)	0.90 (0.005)	0.64 (0.024)	0.34 (0.008)
	(30,60,100)	0.7 (0.019)	0.94 (0.002)	0.59 (0.014)	0.3 (0.004)
	(200,200,150)	0.62 (0.012)	0.98 (0)	0.43 (0.009)	0.27 (0.003)
	(300,300,150)	0.69 (0.013)	0.98 (0)	0.39 (0.008)	0.26 (0.02)
$(30/p, 30/q)$	(200,200,100)	0.78 (0.024)	0.98 (0)	0.43 (0.012)	0.31 0.003)
	(200,200,200)	0.69 (0.026)	0.98 (0.001)	0.5 (0.02)	0.29 (0.004)

Table 3: Table of outputs for joint estimation in presence of group misspecification

(π_x, π_y)	(p, q, n)	FDR
$(5/p, 5/q)$	(60,30,100)	0.19 (0.077)
	(30,60,100)	0.08 (0.064)
	(200,200,150)	0.04 (0.016)
	(300,300,150)	0.02 (0.007)
$(30/p, 30/q)$	(200,200,100)	0.03 (0.019)
	(200,200,200)	0.03 (0.016)

Table 4: Table of outputs giving empirical FDR for estimating $\mathcal{B}$ using JMMLE in presence of group misspecification

4.2 Simulation 2: testing

We slightly change the data generating model to evaluate our proposed global testing and FDR control procedure. We set $K = 2$, then generate the $\mathbf{B}_0^1$ by first randomly assigning each of its element to be non-zero with probability π, then drawing values of those elements from Unif$\{[-1, -0.5] \cup [0.5, 1]\}$ independently. After this we generate a matrix of differences $\mathbf{D}$, where $(\mathbf{D})_{ij}, i \in \mathcal{I}_p, j \in \mathcal{I}_q$ takes values –1, 1, 0 with probabilities 0.1, 0.1 and 0.8, respectively. Finally we set $\mathbf{B}_0^2 = \mathbf{B}_0^1 + \mathbf{D}$. We set identical sparsity structures for the pairs of precision matrices $\{\Omega_{x0}^1, \Omega_{x0}^2\}$ and $\{\Omega_{y0}^1, \Omega_{y0}^2\}$. We use 150 replications of the above setup to calculate empirical power of global tests, as well as empirical power and FDR of simultaneous tests. To get the empirical sizes of global tests we use estimators obtained from applying JMMLE on a separate set of data generated setting all elements of $\mathbf{D}$ to 0. The type-I error of global tests is controlled at level 0.05, while FDR is set at 0.2 obtained by calculating the respective thresholds.

Table 5 reports the empirical mean and standard deviations (in brackets) of all relevant quantities computed from debiased coefficients obtained from JMMLE and separate estimation. We report outputs for all combinations of data dimensions and sparsity used in

(π_x, π_y)	(p, q, n)	Method	Global test		Simultaneous test	
			Power	Size	Power	FDR
$(5/p, 5/q)$	(60,30,100)	JMMLE	0.98 (0.016)	0.07 (0.011)	0.94 (0.023)	0.24 (0.027)
		Separate	0.99 (0.007)	0.12 (0.02)	0.91 (0.025)	0.34(0.038)
		SepLasso	0.99 (0.007)	0.11 (0.02)	0.91 (0.025)	0.33(0.038)
	(60,30,200)	JMMLE	0.99 (0.014)	0.07 (0.014)	0.97 (0.013)	0.22 (0.032)
		Separate	0.99 (0.005)	0.08 (0.014)	0.94 (0.019)	0.26(0.031)
		SepLasso	0.99 (0.004)	0.08 (0.014)	0.94 (0.019)	0.26(0.033)
	(30,60,100)	JMMLE	0.98 (0.024)	0.07 (0.014)	0.92 (0.027)	0.24 (0.035)
		Separate	1 (0)	0.07 (0.015)	0.86 (0.036)	0.25(0.039)
		SepLasso				
	(30,60,200)	JMMLE	0.99 (0.019)	0.08 (0.016)	0.96 (0.023)	0.24 (0.038)
		Separate	1 (0)	0.06 (0.013)	0.9 (0.038)	0.21(0.035)
		SepLasso	1 (0)	0.06 (0.012)	0.91 (0.038)	0.21(0.034)
	(200,200,150)	JMMLE	0.99 (0.006)	0.06 (0.003)	0.84 (0.011)	0.22 (0.007)
		Separate	1 (0)	0.2 (0.008)	0.93 (0.006)	0.46(0.009)
		SepLasso	1 (0)	0.2 (0.008)	0.93 (0.006)	0.46(0.009)
	(300,300,150)	JMMLE	0.99 (0.004)	0.07 (0.009)	0.54 (0.031)	0.34 (0.016)
		Separate	1 (0)	0.27 (0.01)	0.79 (0.007)	0.58(0.008)
		SepLasso	1 (0)	0.27 (0.01)	0.79 (0.007)	0.58(0.008)
	(300,300,300)	JMMLE	0.99 (0.003)	0.03 (0.002)	0.99 (0.003)	0.12 (0.006)
		Separate	1 (0)	0.16 (0.005)	0.99 (0.004)	0.4 (0.007)
		SepLasso	1 (0)	0.16 (0.005)	0.99 (0.004)	0.4 (0.007)
$(30/p, 30/q)$	(200,200,100)	JMMLE	0.99 (0.005)	0.112 (0.003)	0.41 (0.008)	0.52 (0.007)
		Separate	1 (0)	0.47 (0.008)	0.75 (0.007)	0.71(0.004)
		SepLasso	1 (0)	0.47 (0.008)	0.75 (0.007)	0.71(0.004)
	(200,200,200)	JMMLE	0.99 (0.004)	0.09 (0.004)	0.96 (0.006)	0.27 (0.008)
		Separate	1 (0)	0.42 (0.011)	0.98 (0.005)	0.63(0.006)
		SepLasso	1 (0)	0.42 (0.011)	0.98 (0.005)	0.63(0.006)
	(200,200,300)	JMMLE	0.99 (0.002)	0.06 (0.003)	0.99 (0.004)	0.19 (0.008)
		Separate	1 (0)	0.27 (0.01)	0.99 (0.004)	0.52 (0.009)
		SepLasso	1 (0)	0.27 (0.01)	0.99 (0.004)	0.52 (0.009)

Table 5: Table of outputs for global and simultaneous hypothesis testing.

Section 4.1, and also for increased sample sizes in each setting until a satisfactory FDR is reached. As expected from the theoretical analysis, higher sample sizes than those used in Section 4.1 result in increased power for both global and simultaneous tests, and decreased size and FDR for all but one ($p = 30, q = 60$) of the settings. While separate estimation has slightly higher power in global testing, our joint method gives better results everywhere else.

4.3 Computation

We now discuss some observations and strategies that speed up the JMMLE algorithm and reduce computation time significantly, especially for higher number of features in either layer.

Block update and refit $\mathbf{B}^k$ in each iteration. Similar to the case of $K = 1$ (Lin et al., 2016a), we use block coordinate descent *within* each $\mathbf{B}^k$. This means instead of the full update step (2.9) we perform the following steps in each iteration to speed up convergence:

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

where $\mathbf{r}_1^{k(t)} = \widehat{\mathbf{E}}_{-1}^{k(t)} \widehat{\boldsymbol{\theta}}_1^{k(t)}$, and

$$\mathbf{r}_j^{k(t)} = \sum_{j'=1}^{j-1} \hat{\mathbf{e}}_j^{k(t+1)} \hat{\theta}_{jj'}^{k(t)} + \sum_{j'=j+1}^{q} \hat{\mathbf{e}}_j^{k(t)} \hat{\theta}_{jj'}^{k(t)}$$

for $j \geq 2$. Further, when starting from the initializer of the coefficient matrix given in (2.7), the support set of coefficient estimates becomes constant after only a few (< 10) iterations of our algorithm, after which it refines the values inside the same support until overall convergence. This process speeds up significantly if a refitting step is added *inside each iteration* after the matrices $\widehat{\mathbf{B}}^k$ are updated:

$$\left\{\widetilde{\mathbf{B}}_j^{k(t+1)}\right\}_{k=1}^K = \underset{\substack{\mathbf{b}^k \in \mathbb{R}^p \\ k \in \mathcal{I}_K}}{\arg\min} \left\{ \frac{1}{n} \sum_{j=1}^q \sum_{k=1}^K \|\mathbf{Y}_j^k + \mathbf{r}_j^{k(t)} - \mathbf{X}^k \mathbf{B}_j^k\|^2 + \lambda \sum_{h \in \mathcal{H}} \|\mathbf{B}_{-j}^{[h]}\| \right\};$$

$$\widehat{\mathbf{B}}_j^{k(t+1)} = \left[(\mathbf{X}_{\mathcal{S}_{jk}}^k)^T (\mathbf{X}_{\mathcal{S}_{jk}}^k) \right]^- (\mathbf{X}_{\mathcal{S}_{jk}}^k)^T \mathbf{Y}_j^k$$

where $\mathcal{S}_{jk} = \text{supp}(\widetilde{\mathbf{B}}_j^{k(t+1)})$.

One-step estimator. Algorithm 1, even after the above modifications, is computation-intensive. The reason behind this is the full tuning and updating of the lower layer neighborhood estimates $\{\widehat{\Theta}_j\}$ in each iteration. In practice, the algorithm speeds up significantly without compromising on estimation accuracy if we dispense of the Θ update step in all, but the last iteration. More precisely, we consider the following one-step version of the original algorithm.

Algorithm 4. (The one-step JMMLE Algorithm)
1. Initialize $\widehat{\mathcal{B}}$ using (2.7).
2. Initialize $\widehat{\Theta}$ using (2.8).
3. Update $\widehat{\mathcal{B}}$ as:

$$\widehat{\mathcal{B}}^{(t+1)} = \underset{\substack{\mathbf{B}^k \in \mathbb{M}(p,q) \\ k \in \mathcal{I}_K}}{\arg\min} \left\{ \frac{1}{n} \sum_{j=1}^q \sum_{k=1}^K \|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \mathbf{B}_{-j}^k) \widehat{\boldsymbol{\theta}}_j^{k(0)} - \mathbf{X}^k \mathbf{B}_j^k\|^2 + \lambda_n \sum_{h \in \mathcal{H}} \|\mathbf{B}^{[h]}\| \right\}$$

(p,q,n)	Method	TPR($\widehat{\mathcal{B}}$)	TNR($\widehat{\mathcal{B}}$)	MCC($\widehat{\mathcal{B}}$)	RF($\widehat{\mathcal{B}}$)
(60,30,100)	Full	0.982 (0.013)	0.994 (0.003)	0.959 (0.014)	0.23 (0.021)
	One step	0.971 (0.02)	0.996 (0.003)	0.965 (0.014)	0.242 (0.033)
(30,60,100)	Full	0.966 (0.015)	0.991 (0.003)	0.954 (0.008)	0.269 (0.026)
	One step	0.968 (0.013)	0.992 (0.002)	0.957 (0.008)	0.265 (0.024)
(p,q,n)	Method	TPR($\widehat{\Omega}_y$)	TNR($\widehat{\Omega}_y$)	MCC($\widehat{\Omega}_y$)	RF($\widehat{\Omega}_y$)
(60,30,100)	Full	0.756 (0.019)	0.907 (0.005)	0.616 (0.021)	0.318 (0.007)
	One step	0.764 (0.018)	0.904 (0.006)	0.678 (0.024)	0.321 (0.008)
(30,60,100)	Full	0.695 (0.016)	0.943 (0.002)	0.552 (0.015)	0.304 (0.005)
	One step	0.696 (0.018)	0.943 (0.002)	0.552 (0.018)	0.304 (0.005)

Table 6: Comaprison of evaluation metrics for full and one-step versions of the JMMLE algorithm.

4. Continue till convergence to obtain $\widehat{\mathcal{B}} = \{\widehat{\mathbf{B}}^k\}$.
5. Obtain $\widehat{\mathbf{E}}^k := \mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k, k \in \mathcal{I}_K$. Update $\widehat{\Theta}$ as:

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

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

Compared to one-step algorithms based on first order approximation of the objective function (Zou and Li, 2008; Taddy, 2017), we let $\mathcal{B}$ converge completely, then use these solutions to recover the support set of the precision matrices. The estimation accuracy of Ω_y depends on the solution $\widehat{\mathcal{B}}$ used to solve the sub-problem (2.12) (Theorem 2 and Lemmas 1 and 2). Thus, letting $\mathcal{B}$ converge first ensures that the solutions $\widehat{\Theta}$ and $\widehat{\Omega}_y$ obtained subsequently are of a better quality compared to a simple early stopping of the JMMLE algorithm.

(p,q,n)	Method	Comp. time (min)
(60,30,100)	Full	6.1
	One-step	0.7
(30,60,100)	Full	22.4
	One-step	2.7

Table 7: Comaprison of computation times (averaged over 150 replications) for full and one-step versions of the JMMLE algorithm.

We compared the performance of both versions of our algorithm for the two data settings with smaller feature dimensions. Computations were performed on the HiperGator supercomputer[1], in parallel across 8 cores of an Intel E5-2698v3 2.3GHz processor with 2GB RAM per core, the parallelization being done across the range of values for λ_n within each replication. As seen in Table 6, performance is indistinguishable across all the metrics, but

1. https://www.rc.ufl.edu/services/hipergator

	RMSSPE	NZ($\widehat{\mathcal{B}}$)	NZ($\widehat{\Omega}_y$)
JMMLE	14.38 (3.27)	0.014 (0.004)	0.077 (0.009)
Separate			
JSEM	-	-	

Table 8: Performance metric comparison over 100 random splits of the real data

the one-step algorithm saves a significant amount of computation time compared to the full version (Table 7).

5. Real data example

We now apply the proposed methodology to a gene expression dataset containing mRNA and RNAseq expression values for 4000 genes, divided into 88 pathways, for $n_1 = 262$ ER-positive (ER+) and $n_2 = 76$ ER-negative breast cancer patients. **more info needed on data**

Our objective here is to (a) obtain mRNA-mRNA, mRNA-RNAseq and RNAseq-RNAseq networks for the ER+ and ER- groups while incorporating pathway information, (b) test for differential strengths of mRNA-RNAseq connections between the two sample groups. To this end, we take mRNA and RNAseq expression data as the top and bottm layers (X and Y), respectively, and consider pathway-wise groups. For comparison purposes, we apply JMMLE and the separate estimation method Lin et al. (2016a) for estimating $\mathcal{B}$ and Ω_y, and JSEM for estimating Ω_y. For comparing estimation performances of the methods, we use the following performance metrics calculated over 100 random 80:20 train-test splits of samples within each group:

- Root Mean Squared Scaled Prediction Error-

$$\text{RMSSPE}(\widehat{\mathcal{B}}, \widehat{\Omega}_y) = \left[\sum_{k=1}^{K} \frac{1}{n_k} \text{Tr}\left(\mathbf{Y}_k - \mathbf{X}_k\widehat{\mathbf{B}}_k)^T(\mathbf{Y}_k - \mathbf{X}_k\widehat{\mathbf{B}}_k)\widehat{\Omega}_y^k\right)\right]^{1/2}$$

- Proportion of non-zero coefficients in $\widehat{\mathcal{B}}$-

$$\text{NZ}(\widehat{\mathbf{B}}_k) = \frac{|\,\text{supp}(\hat{\mathbf{B}}_k)|}{pq}; \quad \text{NZ}(\widehat{\mathcal{B}}) = \frac{1}{K}\sum_{k=1}^{K}\text{NZ}(\widehat{\mathbf{B}}_k)$$

- Proportion of non-zero coefficients in off-diagonal entries of $\widehat{\Omega}_y$-

$$\text{NZ}(\widehat{\Omega}_y^k) = \frac{|\,\text{supp}(\hat{\Omega}_y^k) - q|}{q^2}; \quad \text{NZ}(\widehat{\Omega}_y) = \frac{1}{K}\sum_{k=1}^{K}\text{NZ}(\widehat{\Omega}_y^k)$$

To summarize within-layer and between-layer interactions, we consider the 10 highest entries in $\widehat{\mathbf{B}}_k, \widehat{\Omega}_y^k; k = 1, 2$ in terms of absolute value. Table 9 gives their magnitudes, as well as the corresponding mRNA-RNAseq and RNAseq-RNAseq pairs. For the sake of

Sample group	ER+ ($k=1$)			ER- ($k=2$)		
	Value	mRNA	RNAseq	Value	mRNA	RNAseq
	-5.87	TAF9_3	TRA2B_1	-11.32	TAF9_3	TRA2B_1
	-5.7	KCNN3_3	THOC7_1	-6.01	TAF9_3	UQCRQ_1
	4.9	SQRDL_3	COX6A1_1	-5.38	TAF9_3	TAF9_1
	4.35	SQRDL_3	ATP5G3_1	5.17	SQRDL_3	COX6A1_1
Conections	-4.34	KCNN3_3	PABPN1_1	5.14	SQRDL_3	ACTR3_1
in $\widehat{\mathcal{B}}$	4.31	SQRDL_3	ACTR3_1	4.55	SQRDL_3	SSU72_1
	-4.21	KCNN3_3	SNRPD2_1	-4.52	KCNN3_3	THOC7_1
	3.98	CYP7B1_3	ECH1_1	4.4	UNG_3	COX6A1_1
	3.88	SQRDL_3	SSU72_1	-4.18	TAF9_3	ATP5J_1
	3.87	CYP7B1_3	FTH1_1	4.17	CYP7B1_3	FTH1_1
	Value	RNASeq1	RNAseq2	Value	RNAseq1	RNAseq2
	-0.27	ECH1_1	PIGY_1	-0.16	THOC7_1	PABPN1_1
	-0.25	THOC7_1	PABPN1_1	-0.12	RBBP4_1	PABPN1_1
	-0.22	COX6A1_1	SF3B5_1	-0.11	NAPA_1	CD63_1
	-0.21	PCBP1_1	SH3GL1_1	-0.11	SOD1_1	SNRPD3_1
Conections	-0.21	ECH1_1	DDX42_1	-0.1	EIF3I_1	TXNL4A_1
in $\widehat{\Omega}_y$	-0.19	EXOSC2_1	QARS_1	-0.1	PCBP1_1	SH3GL1_1
	-0.19	QARS_1	PIGY_1	-0.1	TAF9_1	COX7C_1
	-0.18	ECH1_1	SDHC_1	-0.1	ECH1_1	HNRNPA1L2_1
	-0.18	PABPN1_1	VAMP8_1	-0.1	KARS_1	FUNDC1_1
	-0.18	EIF3I_1	TXNL4A_1	-0.1	ECH1_1	PIGY_1
	Value	mRNA1	mRNA2	Value	mRNA1	mRNA2
	-0.32	GP1BB_3	COX6A2_3	-0.19	PTPRC_3	ITGAL_3
	-0.32	PTTG1_3	PTTG2_3	-0.17	PTPRC_3	CD2_3
	-0.3	ABCA8_3	C7_3	-0.16	PDCD1_3	ICOS_3
	-0.3	PTPRC_3	CD2_3	-0.16	PDCD1_3	CD2_3
Conections	-0.29	ABCA8_3	FXYD1_3	-0.16	PTPRC_3	CYBB_3
in $\widehat{\Omega}_x$	-0.28	PDCD1_3	ICOS_3	-0.15	GP1BB_3	COX6A2_3
	-0.26	PDCD1_3	CD2_3	-0.15	PTPRC_3	IL2RG_3
	-0.25	CD6_3	PDCD1_3	-0.15	PTPRC_3	CTSS_3
	-0.25	PTPRC_3	PTGER4_3	-0.14	PTPRC_3	PTGER4_3
	-0.25	LAT_3	PDCD1_3	-0.14	LCP2_3	CYBB_3

Table 9: Top 10 within-layer and between-layer connections obtained by JMMLE.

comparison, we also report the same numbers and mRNA-mRNA pairs from the analysis of only the top layer using JSEM (Ma and Michailidis, 2016). **give citations for some connections**

After applying our debiasing procedure and performing the global test, 23 mRNA-s were determined to have significant differences in the corresponding rows across sample groups, i.e. between $\widehat{\mathbf{b}}_i^1$ and $\widehat{\mathbf{b}}_i^2$. Within connections of these mRNAs, 957 total mRNA-RNAseq connections were determined by the simultaneous testing procedure to have significant dif-

(a)

mRNA	Statistic
DCTN2_3	17015.2
ST8SIA1_3	13514.6
FUT5_3	8315.7
XPA_3	7194.2
RETSAT_3	5676.0
TAF4B_3	5385.8
CYP7B1_3	4189.6
UNG_3	3709.1
RAD23A_3	2793.8
TAF9_3	2427.1

(b)

mRNA	RNAseq	Statistic
DCTN2_3	EIF4A1_1	2560.6
DCTN2_3	ARPC4_1	2021.8
ST8SIA1_3	EIF4A1_1	1948.2
DCTN2_3	PAIP1_1	1922.3
DCTN2_3	SNX5_1	1825.1
DCTN2_3	SUMO3_1	1817.6
DCTN2_3	CETN2_1	1779.2
DCTN2_3	SF3B4_1	1755.8
ST8SIA1_3	ARPC4_1	1516.5
ST8SIA1_3	PAIP1_1	1453.9

Table 10: Hypothesis testing outputs from real data analysis: (a) top-10 mRNAs and their global test statistic (D_i) values, (b) top-10 mRNA-RNAseq pairs and their simultaneous test statistic (d_{ij}) values

ferences between their corresponding coefficients, i.e. between $\hat{b}^1_{ij}$ and $\hat{b}^2_{ij}$. **some more references of these connections?**

6. Discussion

This work introduces an integrative framework for knowledge discovery in multiple multi-layer Gaussian Graphical Models. We exploit *a priori* known structural similarities across parameters of the multiple models to achieve estimation gains compared to separate estimation. More importantly, we derive results on the asymptotic distributions of generic estimates of the multiple regression coefficient matrices in this complex setup, and perform global and simultaneous testing for pairwise differences within the between-layer edges.

6.1 Performance improvement

The JMMLE algorithm due to the incorporation of prior information about sparsity patterns improves on the theoretical convergence rates of the estimation method for *single* multi-layer GGMs (i.e. $K = 1$) introduced in Lin et al. (2016a). With our initial estimates, the method of Lin et al. (2016a) achieves the following convergence rates for the estimation of $\mathcal{B}$ and Ω_y, respectively (using Corollary 4 therein):

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_F \leq \sum_{k=1}^{K} O\left(\sqrt{\frac{b_k \log(pq)}{n}}\right),$$
$$\sum_{k=1}^{K} \|\widehat{\Omega}^k_y - \Omega^k_{y0}\|_F \leq O\left(K\sqrt{\frac{(S+q)\log(pq)}{n}}\right).$$

In comparison, JMMLE has the following rates:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_F \leq O\left(\sqrt{\frac{|h_{\max}|B\log(pq)}{n}}\right),$$
$$\sum_{k=1}^{K}\|\widehat{\Omega}_y^k - \Omega_{y0}^k\|_F \leq O\left(\sqrt{\frac{KS|g_{\max}|\log(pq)}{n}}\right).$$

For $\mathcal{B}$, joint estimation outperforms separate estimation when group sizes are small, so that $(|h_{\max}|B)^{1/2} < \sum_k b_k^{1/2}$. The estimation gain for Ω_y is more substantial, especially for higher values of q. This is corroborated by our simulation outputs (Tables 1 and 2), where the joint estimates perform better for both sets of parameters, but the differences between RF errors obtained from joint and separate estimates tend to be lower for $\widehat{\Omega}_y$ than $\widehat{\mathcal{B}}$.

6.2 Extensions

There are two immediate extensions of our hypothesis testing framework.

(I) In recent work, Liu (2017) proposed a framework to test for structural similarities and differences across multiple *single layer* GGMs. For K GGMs with precision matrices $\Omega^k = (\omega^k_{ii'})_{i,i'\in\mathcal{I}_p}$, a test for the partial correlation coefficients $\rho^k_{ii'} = -\omega^k_{ii'}/\sqrt{\omega^k_{ii}\omega^k_{i'i'}}$ using residuals from pK separate penalized neighborhood regressions is developed, one for each variable of each GGM. To incorporate structured sparsity across k, our simultaneous regression techniques for all neighborhood coefficients (i.e. (2.4) and (2.12)) can be used instead, to perform testing on the between-layer edges. Theoretical properties of this procedure can be derived using results in Liu (2017), possibly with adjustments for our neighborhood estimates to adhere to the rate conditions for the constants a_{n1}, a_{n2} therein to account for a diverging (p, q, n) setup.
(II) For $K > 2$, detection of the following sets of inter-layer edges can be scientifically significant:

$$\mathcal{B}_1 = \left\{(i,j) : \sum_{1\leq k<k'\leq K}\left(b^k_{0,ij} - b^{k'}_{0,ij}\right)^2 > 0; i\in\mathcal{I}_p, j\in\mathcal{I}_q\right\}$$
$$\mathcal{B}_2 = \left\{(i,j) : b^1_{0,ij} = \cdots, b^K_{0,ij} \neq 0\right\}$$
$$\mathcal{B}_3 = \left\{(i,j) : b^1_{0,ij} = \cdots, b^K_{0,ij} = 0\right\}$$

e.g. detection of gene-protein interactions that are present, but may have different or same weights across k ($\mathcal{B}_1$ and $\mathcal{B}_2$, respectively), and that are absent for all k ($\mathcal{B}_3$). The asymptotic result in Theorem 5 continues to hold in this situation, and an extension of the global test (Algorithm 2) is immediate. However, extending the FDR control procedure requires a technically more involved approach.

The strength of our proposed debiased estimator (3.1) is that only generic estimates of relevant model parameters that satisfy general rate conditions are necessary to obtain a valid asymptotic distribution. This translates to a high degree of flexibility in choosing

the method of estimation. Our formulation based on sparsity assumptions (Section 2.2) is a specific way (motivated by applications in Omics data integration) to obtain the necessary estimates. Sparsity may not be an assumption that is required or even valid in complex hierarchical structures from different domains of application. For different two-layer components in such multi-layer setups, low-rank, group-sparse or sparse methods (or a combination thereof) can be plugged into our alternating algorithm. Results analogous to those in Section 2.3 need to be established for the corresponding estimators. However, as long as these estimators adhere to the convergence conditions (T1)-(T3), Theorem 5 can be used to derive the asymptotic distributions of between-layer edges.

Finally, extending our framework to non-Gaussian data is of interest. As seen for the $K = 1$ case in Lin et al. (2016a), their alternating block algorithm continues to give comparable results under shrunken or truncated empirical distributions of Gaussian errors. Similar results may be possible in the general case, and improvements can come from modifying different parts of the estimation algorithm. For example, the estimation of the precision matrices based on restricted support sets using log-likelihoods in (2.6) can be replaced by methods like nonparanormal estimation (Liu et al., 2009) or regularized score matching (Lin et al., 2016b).

References

D. Atzler, E. Schwedhelm, and T. Zeller. Integrated genomics and metabolomics in nephrology. *Nephrol. Dial. Transplant.*, 29(8):1467–1474, 2014.

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 2015.

S. Basu, X. Li, and G. Michailidis. Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions. *IEEE Trans. Sig. Proc.*, 67:1207–1222, 2019.

E. Belilovsky, G. Varoquaux, and M. Blaschko. Testing for differences in Gaussian graphical models: Applications to brain connectivity. In *NIPS Proceedings*, pages 595–603, 2016.

R. Bellman. Some inequalities for the square root of a positive definite matrix. *Linear Algebra Appl.*, 1(3):321–324, 1968.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6): 2577–2604, 2008.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer, 2011.

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

T. T. Cai, H. Li, W. Liu, and J. Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, 2012a.

T. T. Cai, W. Liu, and X. Luo. A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607, 2012b.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc. B*, 76(2):373–397, 2014.

Y. Fan and C. Y. Tang. Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc. B*, 75(3):531–552, 2013.

X. Gao, D. Q. Pu, Y. Wu, and H. Xu. Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. *Stat. Sinica*, 22:1123–1146, 2012.

V. Gligorijević and N. Pržulj. Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface*, 12(112):20150571, 2015.

D. Gomez-Cabrero, I. Abugessaisa, D. Maier, et al. Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, 8(Suppl. 2):l1, 2014.

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

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, 2014.

A. Javanmard and A. Montanari. De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. *Ann. Statist.*, To appear, 2018+. https://arxiv.org/abs/1508.02757.

A. R. Joyce and B. Palsson. The model organism as a system: integrating *omics* data sets. *Nat. Rev. Mol. Cell Biol.*, 7:198–210, 2006.

A. K. Kaushik, A. Shojaie, K. Panzitt, et al. Inhibition of the hexosamine biosynthetic pathway promotes castration-resistant prostate cancer. *Nat. Commun.*, 7:11612, 2016.

Y. Kim, S. Kwon, and H. Choi. Consistent Model Selection Criteria on High Dimensions. *J. Mach. Learn. Res.*, 13:1037–1057, 2012.

T. Kling, P. Johansson, J. Sanchez, et al. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucl. Acid Res.*, 43(15):e98, 2015.

J. Lee, H. J. Kee, S. Min, et al. Integrated omics-analysis reveals Wnt-mediated NAD+ metabolic reprogramming in cancer stem-like cells. *Oncotarget*, 26(7(30)):48562–48576, 2016.

W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Mult. Anal.*, 111:241–255, 2012.

H. Li, N. Pouladi, I. Achour, et al. eQTL networks unveil enriched mRNA master integrators downstream of complex disease-associated SNPs. *J. Biomed. Inform.*, 58:226–234, 2015.

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

L. Lin, M. Drton, and A. Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10:806–854, 2016b.

H. Liu, J. Lafferty, and L. Wasserman. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

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

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

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 2012.

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

S. Majumdar and S. Chatterjee. Non-convex penalized multitask regression using data depth-based penalties. *Stat*, 7:e174, 2018.

Y. Mao, S.-W. Kao, L. Chen, et al. The essential and downstream common proteins of amyotrophic lateral sclerosis: A protein-protein interaction network analysis. *PLoS One*, 12(3):e0172246, 2017.

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

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

G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support Union Recovery in High-dimensional Multivariate Regression. *Ann. Statist.*, 39:1–47, 2011.

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

K. M. Sas, J. Lin, T. M. Rajendiran, et al. Shared and distinct lipid-lipid interactions in plasma and affected tissues in a diabetic mouse model. *J. Lipid Res.*, 59(2):173–183, 2018.

B. Stucky and S. van de Geer. Asymptotic Confidence Regions for High-Dimensional Structured Sparsity. *IEEE Trans. Signal Process.*, 66(8):2178–2190, 2018.

M. Taddy. One-Step Estimator Paths for Concave Regularization. *J. Comp. Graph. Stat.*, 26(3):525–536, 2017.

K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.*, 19:A68–A77, 2015.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

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

J. M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra Appl.*, 11:3–5, 1975.

L. Wang, Y. Kim, and R. Li. Calibrating Nonconvex Penalized Regression in Ultra-high Dimension. *Ann. Statist.*, 41:2505–2536, 2013.

Y. Xie, Y. Liu, and W. Valdar. Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika*, 103(3):493–511, 2016.

W. Yuan, Y. Xia, C. G. Bell, et al. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nat. Commun.*, 5(5):5719, 2014.

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

Y. Zhang, Z. Ouyang, and H. Zhao. A statistical framework for data integration through graphical models with application to cancer genomics. *Ann. Appl. Stat.*, 11(1):161–184, 2017.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36:1509–1533, 2008.

Appendix

Appendix A. Proofs of main results

Proof of Theorem 1. The theorem is a generalization of Theorem 1 in Lin et al. (2016a). The proof follows directly from the proof of that theorem, substituting $(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}), (B^*, \Theta_\epsilon^*)$ and $(B^\infty, \Theta_\epsilon^\infty)$ therein with $(\widehat{\mathcal{B}}^{(t)}, \widehat{\Theta}_y^{(t)}), (\mathcal{B}_0, \Theta_{y0})$ and $(\mathcal{B}^\infty, \Theta_y^\infty)$, respectively, and their corresponding variations as required.

□

We use the following condition extensively while deriving the results that follow.

Condition 3 (Restricted eigenvalues). A symmetric matrix $\mathbf{M} \in \mathbb{M}(b, b)$ is said to satisfy the restricted eigenvalue or RE condition with parameters $\psi, \phi > 0$, denoted as curvature and tolerance, respectively, if

$$\boldsymbol{\theta}^T \mathbf{M} \boldsymbol{\theta} \geq \psi \|\boldsymbol{\theta}\|^2 - \phi \|\boldsymbol{\theta}\|_1^2$$

for all $\boldsymbol{\theta} \in \mathbb{R}^b$. In short, this is denoted by $\mathbf{M} \sim RE(\psi, \phi)$.

Starting from Bickel et al. (2009), different versions of the RE conditions have been proposed and used in high-dimensional analysis (Loh and Wainwright, 2012; Basu and Michailidis, 2015; Ma and Michailidis, 2016; van de Geer and Bühlmann, 2009) to ensure that a covariance matrix satisfies a somewhat relaxed positive-definiteness condition.

Proof of Theorem 2. The proof strategy is as follows. We first show that given fixed $(\mathcal{X}, \mathcal{E})$, and some conditions on $\widehat{\mathbf{E}}^k := \mathbf{Y}^k - \mathbf{X}^k \widehat{\mathbf{B}}^k, k \in \mathcal{I}_K$, the bounds in Theorem 2 hold. We then show that for random $(\mathcal{X}, \mathcal{E})$, those conditions hold with probability approaching 1.

Lemma 1. *Assume fixed $\mathcal{X}, \mathcal{E}$ and deterministic $\widehat{\mathcal{B}} = \{\widehat{\mathbf{B}}^k\}$, and the following conditions.*
(A1) *For $k \in \mathcal{I}_K$,*

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

with $C_\beta = O(1)$ is non-negative and depends on $\mathcal{B}_0$ only.
(A2) *For all $j \in \mathcal{I}_q$,*

$$\frac{1}{n} \left\| (\widehat{\mathbf{E}}_{-j}^k)^T \widehat{\mathbf{E}}^k \mathbf{T}_{0,j}^k \right\|_\infty \leq \mathbb{Q}\left(C_\beta, \Sigma_{x0}^k, \Sigma_{y0}^k\right) \sqrt{\frac{\log(pq)}{n}},$$

where $\mathbb{Q}\left(C_\beta, \Sigma_{x0}^k, \Sigma_{y0}^k\right) = O(1)$ is non-negative and depends on $\mathcal{B}_0, \Sigma_{x0}^k$ and Σ_{y0}^k only.
(A3) *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 $\psi = \min_k \psi^k, \phi = \max_k \phi^k$.*

Then the following hold
(I) Given the choice of tuning parameter

$$\gamma_n \geq 4\sqrt{|g_{\max}|}\mathbb{Q}_0 \sqrt{\frac{\log(pq)}{n}}; \quad \mathbb{Q}_0 := \max_{k \in \mathcal{I}_K} \mathbb{Q}\left(C_\beta, \Sigma_{x0}^k, \Sigma_{y0}^k\right)$$

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

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

$$|\operatorname{supp}(\widehat{\Theta}_j)| \leq 128 s_j/\psi \tag{A.3}$$

(II) For the choice of tuning parameter $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0\sqrt{\log(pq)/n}$,

$$\frac{1}{K}\sum_{k=1}^{K}\|\widehat{\Omega}_y^k - \Omega_{y0}^k\|_F \leq O\left(\mathbb{Q}_0\sqrt{\frac{|g_{\max}|S}{K}}\sqrt{\frac{\log(pq)}{n}}\right) \tag{A.4}$$

Condition (A1) holds by assumption. When $\mathcal{X}$ and $\mathcal{E}$ are random, the following proposition ensures that (A2) and (A3) hold with probabilities approaching to 1.

Lemma 2. *Consider deterministic* $\widehat{\mathcal{B}}$ *satisfying assumption (A1), and conditions (E1), (E2) from the main paper. 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*

$$\psi^k = \frac{\Lambda_{\min}(\Sigma_{x0}^k)}{2}; \quad \phi^k = \frac{\psi^k \log p}{n} + 2C_\beta c_2[\Lambda_{\max}(\Sigma_{x0}^k)\Lambda_{\max}(\Sigma_{y0}^k)]^{1/2}\frac{\log(pq)}{n}$$

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* $j \in \mathcal{I}_q$

$$\left\|\frac{1}{n}(\widehat{\mathbf{E}}_{-j}^k)^T\widehat{\mathbf{E}}^k\mathbf{T}_{0,j}^k\right\|_\infty \leq \mathbb{Q}\left(C_\beta, \Sigma_{x0}^k, \Sigma_{y0}^k\right)\sqrt{\frac{\log(pq)}{n}}$$

with probability $\geq 1 - 1/p^{\tau_1 - 2} - 12c_1\exp[-(c_2^2-1)\log(pq)] - 6c_4\exp[-(c_5^2-1)\log(pq)], c_4 > 0, c_5 > 1, \tau_1 > 2$, *where*

$$\mathbb{Q}\left(C_\beta, \Sigma_{x0}^k, \Sigma_{y0}^k\right) = \left[2C_\beta^2 V_x^k + 4C_\beta c_2[\Lambda_{\max}(\Sigma_{x0}^k)\Lambda_{\max}(\Sigma_{y0}^k)]^{1/2}\right]\sqrt{\frac{\log(pq)}{n}} + c_5\left[\Lambda_{\max}(\Sigma_{y0}^k)\sigma_{y0,j,-j}^k\right]^{1/2}\sqrt{\frac{\log q}{\log(pq)}}$$

with $\sigma_{y0,j,-j}^k = Var(E_j - \mathbb{E}_{-j}\boldsymbol{\theta}_{0,j})$, *and*

$$V_x^k = \sqrt{\frac{\log 4 + \tau_1\log p}{c_x^k n}}; \quad c_x^k = \left[128(1 + 4\Lambda_{\max}(\Sigma_{x0}^k))^2\max_i(\sigma_{x0,ii}^k)^2\right]^{-1}$$

We prove the main theorem by putting together Lemma 1 and Lemma 2. □

Proof of Theorem 3. The strategy is the same as in Theorem 2. We first establish the theorem statements hold for fixed $\mathcal{X}, \mathcal{E}$ in the presence of certain regularity conditions, and then show that those conditions are satisfied with probability approaching 1 when $\mathcal{X}$ and $\mathcal{E}$ are random.

Lemma 3. *Assume fixed* $(\mathcal{X}, \mathcal{E})$*, and deterministic* $\widehat{\Theta} = \{\widehat{\Theta}_j\}$*, so that*
(B1) *For* $j \in \mathcal{I}_q$*,*

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

for some $C_\Theta = O(1)$ *dependent on* Θ_0 *only.*
(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:*

$$\left\|\widehat{\boldsymbol{\gamma}}^k - \widehat{\mathbf{\Gamma}}^k \boldsymbol{\beta}_0\right\|_\infty \leq \mathbb{R}(C_\Theta, \Sigma^k_{x0}, \Sigma^k_{y0}) \sqrt{\frac{\log(pq)}{n}}.$$

where $\mathbb{R}(C_\Theta, \Sigma^k_{x0}, \Sigma^k_{y0}) = O(1)$ *depends on* Θ_0, Σ^k_{x0} *and* Σ^k_{y0} *only.*
(B3) $\widehat{\mathbf{\Gamma}} \sim RE(\psi_*, \phi_*)$ *with* $Kpq\phi_* \leq \psi_*/2$*.*

Then, given the choice of the tuning parameter

$$\lambda_n \geq 4\sqrt{|h_{\max}|}\mathbb{R}_0 \sqrt{\frac{\log(pq)}{n}}; \quad \mathbb{R}_0 := \max_{k \in \mathcal{I}_K} \mathbb{R}\left(C_\Theta, \Sigma^k_{x0}, \Sigma^k_{y0}\right)$$

the following holds

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 48\sqrt{|h_{\max}|} B\lambda_n/\psi^* \tag{A.5}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq 12\sqrt{B}\lambda_n/\psi^* \tag{A.6}$$

$$\sum_{h \in \mathcal{H}} \|\boldsymbol{\beta}^{[h]} - \boldsymbol{\beta}_0^{[h]}\| \leq 48B\lambda_n/\psi^* \tag{A.7}$$

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \widehat{\mathbf{\Gamma}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq 72B\lambda_n^2/\psi^* \tag{A.8}$$

Condition (B1) holds by assumption. Next, we verify that conditions (B2) and (B3) hold with high probability given fixed $\widehat{\Theta}$.

Lemma 4. *Consider deterministic* $\widehat{\Theta}$ *satisfying assumption (B1). Also assume conditions (E3), (E4) from the main body of the paper. Then, for sample size* $n \gtrsim \log(pq)$*,*

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

$$\psi_* = \min_k \psi^k \left(\min_i \psi_k^j - d_k C_\Theta \sqrt{\frac{\log(pq)}{n}}\right), \phi_* = \max_k \phi^k \left(\min_i \phi_k^j + d_k C_\Theta \sqrt{\frac{\log(pq)}{n}}\right)$$

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

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

$$\mathbb{R}\left(C_\Theta, \Sigma^k_{x0}, \Sigma^k_{y0}\right) = c_2 \left\{ d_k C_\Theta \sqrt{\frac{\log(pq)}{n}} [\Lambda_{\max}(\Sigma^k_{x0})\Lambda_{\max}(\Sigma^k_{y0})]^{1/2} + \left[\frac{\Lambda_{\max}(\Sigma^k_{x0})}{\Lambda_{\min}(\Sigma^k_{y0})}\right]^{1/2} \right\}$$

The theorem follows straighforwardly by putting together the results from Lemmas 3 and 4. □

Proof of Theorem 4. The first part is immediate from the proof of part I of Theorem 4 in Lin et al. (2016a). By choice of λ_n, we now have

$$\|\widehat{\mathbf{B}}^{k(0)} - \mathbf{B}_0^k\|_1 = O\left(\sqrt{\frac{\log(pq)}{n}}\right),$$

so we can apply Theorem 2 to prove the bounds on $\{\widehat{\Theta}_j^{(0)}\}$. □

Proof of Theorem 5. Define the following:

$$\widehat{\mathbf{D}}_i = \text{vec}(\widehat{\mathbf{b}}_i^1, \ldots, \widehat{\mathbf{b}}_i^K); \quad \mathbf{R}_i^k = \mathbf{X}_i^k - \mathbf{X}_{-i}^k \widehat{\boldsymbol{\zeta}}_i^k; k \in \mathcal{I}_K$$

Then, from (3.1) we have

$$\mathbf{M}_i(\widehat{\mathbf{C}}_i - \widehat{\mathbf{D}}_i)^T = \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T \widehat{\mathbf{E}}^1 \\ \vdots \\ \frac{1}{\widehat{s}_i^K}(\mathbf{R}_i^K)^T \widehat{\mathbf{E}}^K \end{bmatrix} \tag{A.9}$$

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

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

Putting them back in (A.9) and using $t_i^k = (\mathbf{R}_i^k)^T \mathbf{X}_i^k / n$, we get

$$\begin{aligned} \mathbf{M}_i(\widehat{\mathbf{C}}_i - \widehat{\mathbf{D}}_i)^T &= \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T \mathbf{E}^1 \\ \vdots \\ \frac{1}{\widehat{s}_i^K}(\mathbf{R}_i^K)^T \mathbf{E}^K \end{bmatrix} + \mathbf{M}_i(\mathbf{D}_i - \widehat{\mathbf{D}}_i)^T \\ &\quad + \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T \mathbf{X}_{-i}^1(\mathbf{B}_{0,-i}^1 - \widehat{\mathbf{B}}_{-i}^1) \\ \vdots \\ \frac{1}{\widehat{s}_i^K}(\mathbf{R}_i^K)^T \mathbf{X}_{-i}^K(\mathbf{B}_{0,-i}^K - \widehat{\mathbf{B}}_{-i}^K) \end{bmatrix} \\ \Rightarrow \widehat{\Omega}_y^{1/2} \mathbf{M}_i(\widehat{\mathbf{C}}_i - \mathbf{D}_i)^T &= \frac{\widehat{\Omega}_y^{1/2}}{\sqrt{n}} \begin{bmatrix} \frac{1}{\widehat{s}^1}(\mathbf{R}_i^1)^T \mathbf{E}^1 \\ \vdots \\ \frac{1}{\widehat{s}^K}(\mathbf{R}_i^K)^T \mathbf{E}^K \end{bmatrix} + \frac{\widehat{\Omega}_y^{1/2}}{\sqrt{n}} \begin{bmatrix} \frac{1}{\widehat{s}_i^1}(\mathbf{R}_i^1)^T \mathbf{X}_{-i}^1(\mathbf{B}_{0,-i}^1 - \widehat{\mathbf{B}}_{-i}^1) \\ \vdots \\ \frac{1}{\widehat{s}_i^K}(\mathbf{R}_i^K)^T \mathbf{X}_{-i}^K(\mathbf{B}_{0,-i}^K - \widehat{\mathbf{B}}_{-i}^K) \end{bmatrix} \end{aligned} \tag{A.10}$$

At this point, we drop k and 0 in the subscripts since there is no ambiguity, and establish the following:

Lemma 5. *Given conditions (T1) and (T2), the following holds for sample size n such that $n \gtrsim \log(pq)$:*

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

$$\|\mathbf{S}_{1n}\|_\infty \leq \frac{D_\Omega^{1/2}(2+D_\zeta)c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_e)]^{1/2}\sqrt{\log(pq)}}{\sqrt{\sigma_{x,i,-i}} - n^{-1/4} - D_\zeta\sqrt{V_x}} = O\left(\frac{\log(pq)}{\sqrt{n}}\right) \tag{A.11}$$

with probability $\geq 1 - 6c_1\exp[-(c_2^2-1)\log(pq)] - 1/p^{\tau_1-2} - \kappa_i/\sqrt{n}$, where $\kappa_i := Var[(X_i - \mathbb{X}_{-i}\boldsymbol{\zeta}_{0,-i})^2]$.

Additionally, given condition (T3) we have

$$\begin{aligned}&\left\|\frac{1}{\sqrt{n}\widehat{s}_i}\mathbf{R}_i^T\mathbf{X}_{-i}(\mathbf{B}_{-i} - \widehat{\mathbf{B}}_{-i})\widehat{\Omega}_y^{1/2}\right\|_\infty \\ &\leq \frac{D_\beta(\Lambda_{\min}(\Sigma_y)^{1/2} + D_\Omega^{1/2})}{\sigma_{x,i,-i} - n^{-1/2} - D_\zeta\sqrt{V_x}}\left[c_7\sqrt{(\sqrt{\sigma_{x,i,-i}}\Lambda_{\max}(\Sigma_{x,-i}))\log p} + \sqrt{n}D_\zeta V_x\right] = O\left(\frac{\log(pq)}{\sqrt{n}}\right)\end{aligned} \tag{A.12}$$

holds with probability $\geq 1 - 6c_6\exp[-(c_7^2-1)\log(pq)] - 1/p^{\tau_1-2} - \kappa_i/\sqrt{n}$ for some $c_6 > 0, c_7 > 1$.

Given Lemma 5, the first and second summands on the right hand side of (A.10) are bounded above by applying each of (A.11) and (A.12) K times. This completes the proof. □

Proof of Theorem 6. From (A.10) and Lemma 5 we have that

$$(\widehat{\Omega}_y^k)^{1/2}m_i^k(\widehat{\mathbf{c}}_i^k - \mathbf{b}_{0i}^k) \sim \mathcal{N}_q(\mathbf{0},\mathbf{I}) + \mathbf{S}_{2n}^k, \tag{A.13}$$

where $\|\mathbf{S}_{2n}^k\|_\infty = o_P(1)$. We next obtain the following lemma:

Lemma 6. *Drop k in superscripts and 0 in subscripts. Given condition (T1), the following holds with probability $\geq 1 - 6c_6\exp[-(c_7^2-1)\log(p-1)] - 1/p^{\tau_2-2} - \kappa_i/\sqrt{n}, \tau_2 > 2$:*

$$\begin{aligned}\left|\frac{m_i}{\sqrt{n}} - \sqrt{\sigma_{x,i,-i}}\right| \leq \delta_i := \sqrt{\frac{\log 4 + \tau_2}{c_i n}} + \frac{D_\zeta + 1}{\sqrt{\sigma_{x,i,-i}} - n^{-1/2} - D_\zeta\sqrt{V_x}} \times \\ \left[c_7[(\sigma_{x,i,-i}\Lambda_{\max}(\Sigma_{x,-i})]^{1/2}\sqrt{\frac{\log p}{n}} + D_\zeta V_x\right],\end{aligned} \tag{A.14}$$

where $c_i = [128(1+4\sigma_{x,i,-i})^2(\sigma_{x,i,-i})^2]^{-1}$, and the sample size satisfies $n \gtrsim \log p$.

We also provide the following general result:

Lemma 7. *Consider two positive definite matrices $\mathbf{A}, \mathbf{A}_1 \in \mathbb{M}(a,a)$. Then, for $\delta > 0$, we have*

$$\|\mathbf{A} - \mathbf{A}_1\|_\infty \leq \delta \Rightarrow \|\mathbf{A}^{1/2} - \mathbf{A}_1^{1/2}\|_\infty \leq \delta^{1/2}.$$

Applying Lemma 7 it follows immediately from assumption (T2) that

$$\left\|\widehat{\Omega}_y^{1/2} - \Omega_y^{1/2}\right\|_\infty \leq D_\Omega^{1/2} \tag{A.15}$$

Using Lemma 6 in conjunction with (A.15) we now have

$$\begin{aligned}\sqrt{n}(\Omega_{y0}^k)^{1/2}\sqrt{\sigma_{x0,i,-i}^k}(\widehat{\mathbf{c}}_i^k - \mathbf{b}_{0i}^k) &\sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}) + \mathbf{S}_{3n}^k \\ \Rightarrow \sqrt{n}\Sigma_i^{-1/2}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2 - \boldsymbol{\delta}) &\sim \mathcal{N}_q\left(\mathbf{0}, \mathbf{I}\right) + \mathbf{S}_{3n},\end{aligned} \tag{A.16}$$

where $\Sigma_i := \Sigma_{y0}^1/\sigma_{x0,i,-i}^1 + \Sigma_{y0}^2/\sigma_{x0,i,-i}^2$ and $\mathbf{S}_{3n} = \mathbf{S}_{3n}^1 - \mathbf{S}_{3n}^2, \|\mathbf{S}_{3n}^k\|_\infty = o_P(1)$. We now break down the left hand side above as

$$\begin{aligned}\sqrt{n}\Sigma_i^{-1/2}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2 - \boldsymbol{\delta}) &= \sqrt{n}\Sigma_i^{-1/2}\widehat{\Sigma}_i^{1/2}\widehat{\Sigma}_i^{-1/2}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2) - \sqrt{n}\Sigma_i^{-1/2}\boldsymbol{\delta} \\ &= (\Sigma_i^{-1/2}\widehat{\Sigma}_i^{1/2} - \mathbf{I}).\sqrt{n}\widehat{\Sigma}_i^{-1/2}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2) + \\ &\sqrt{n}\widehat{\Sigma}_i^{-1/2}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2) - \sqrt{n}\Sigma_i^{-1/2}\boldsymbol{\delta},\end{aligned} \tag{A.17}$$

with

$$\widehat{\Sigma}_i := \frac{n\widehat{\Sigma}_y^1}{(m_i^1)^2} + \frac{n\widehat{\Sigma}_y^2}{(m_i^2)^2}.$$

Next, we obtain the following lemma:

Lemma 8. *Given conditions (T1) and (T2), for the pooled covariance matrix estimate $\widehat{\Sigma}_i$, we have*

$$\left\|\widehat{\Sigma}_i - \Sigma_i\right\|_\infty = o(1),$$

for sample size $n \gtrsim \log p$.

Lemma 7 now implies that $\|\widehat{\Sigma}_i^{1/2} - \Sigma_i^{1/2}\|_\infty = o(1)$. Putting this in the first summand of (A.17), then using (A.16) we get

$$\sqrt{n}\widehat{\Sigma}_i^{-1/2}(\widehat{\mathbf{c}}_i^1 - \widehat{\mathbf{c}}_i^2) - \sqrt{n}\Sigma_i^{-1/2}\boldsymbol{\delta} \sim \mathcal{N}_q\left(\mathbf{0}, \mathbf{I}\right) + \mathbf{S}_{4n},$$

with $\|\mathbf{S}_{4n}\|_\infty = o_P(1)$. The power of the global test follows as a consequence. Finally, the lower bound on the order of $\|\boldsymbol{\delta}\|$ holds because $n\boldsymbol{\delta}^T\Sigma_i^{-1}\boldsymbol{\delta} \geq n\|\boldsymbol{\delta}\|^2\Lambda_{\min}(\Sigma_i^{-1})$, and

$$\Lambda_{\min}(\Sigma_i^{-1}) = \frac{\Lambda_{\max}(\Sigma_{y0}^1)}{\sigma_{x,i,-i}^1} + \frac{\Lambda_{\max}(\Sigma_{y0}^2)}{\sigma_{x,i,-i}^2}.$$

□

Proof of Theorem 7. The proof follows the general structure of Theorem 4.1 in Liu and Shao (2014), with two modifications. Firstly, we replace the bound in equation (12) of Liu and Shao (2014) by a new deviation bound

$$P\left(\left|d_{ij} - \frac{\mu_j}{\sigma_j}\right| \geq t\right) = (1 - \Phi(t))(1 + o(1))$$

for any t, since $(d_{ij} - \mu_j)/\sigma_j \sim N(0,1) + o_P(1)$ from Corollary 2. We replace $G_\kappa(t)$ in all following calculations in Liu and Shao (2014) with $1 - \Phi(t)$. Secondly, we need to ensure that given both Σ^1_{y0} and Σ^2_{y0} satisfy the condition (D1) or (D1*), the pooled covariance matrix $\Sigma^1_{y0}/\sigma^1_{x0,i,-i} + \Sigma^2_{y0}/\sigma^2_{x0,i,-i}$ also does so.

For this, denote $c_k = \sigma^k_{x0,i,-i}, k = 1, 2$. Notice that for any $C_1, C_2 > 0$,

$$\begin{aligned} r^k_{jj'} \geq C_k &\Rightarrow \sigma^k_{y0,jj'} \geq (\sigma^k_{y0,jj}\sigma^k_{y0,j'j'})^{1/2} C_k \\ &\Rightarrow \frac{\sigma^1_{y0,jj'}}{c_1} + \frac{\sigma^2_{y0,jj'}}{c_2} \geq \frac{(\sigma^1_{y0,jj}\sigma^1_{y0,j'j'})^{1/2} C_1}{c_1} + \frac{(\sigma^2_{y0,jj}\sigma^2_{y0,j'j'})^{1/2} C_2}{c_2} \\ &\Rightarrow \frac{\sigma^1_{y0,jj'}/c_1 + \sigma^2_{y0,jj'}/c_2}{(\sigma^1_{y0,jj}\sigma^1_{y0,j'j'})^{1/2}/c_1 + (\sigma^2_{y0,jj}\sigma^2_{y0,j'j'})^{1/2}/c_2} \geq \min\{C_1, C_2\}. \end{aligned}$$

It now follows that (D1) or (D1*) holds for the pooled covariance matrices. □

Appendix B. Proofs of auxiliary results

Proof of Lemma 1. The proof has the same structure as the proof of Theorem 1 in Ma and Michailidis (2016), where consistency of the (single layer) JSEM estimates are established. Part (I) is analogous to part A.1 therein, but the proof strategy is completely different, which we provide in detail next. Our part (II) follows along similar lines as parts A.2 and A.3, incorporating the updated quantities from the first part (A.1). For this part of the proof, we provide an outline and leave details to the reader.

Proof of part (I). In its reparametrized version, (2.12) becomes

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

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

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

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

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

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

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

by assumption (A2). 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

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

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

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

with $\mathbf{D}_j = (\mathbf{d}_j^1, \ldots, \mathbf{d}_j^K)$. Now the RE condition on $\widehat{\mathbf{S}}^k$ means that

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

by assumption (A3).

Combining the above with (B.3), we finally have

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

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 $|\widehat{\mathcal{S}}_j| := |\operatorname{supp}(\widehat{\Theta}_j)|$, notice that if $\hat{\boldsymbol{\theta}}_{jj'}^{[g]} \neq 0$ for some (j', g),

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

using the KKT condition for (2.12) and assumption (A2). The choice of γ_n now ensures that the right hand side is $\geq 3|g|\gamma_n/4$. Hence,

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

using (B.3) and (B.4).

Proof of part (II). 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

$$\tilde{\Omega}_y^k = \operatorname{diag}(\Omega_{y0}^k) + \Omega_{y,E_0^k\cap\hat{E}^k}^k$$

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

$$|\hat{E}^k| \leq \frac{128S}{\psi} \tag{B.5}$$

$$\frac{1}{K}\sum_{k=1}^K\|\tilde{\Omega}_y^k - \Omega_{y0}^k\|_F \leq \frac{12c_y\sqrt{S}\gamma_n}{\sqrt{K}\psi} \tag{B.6}$$

following which, taking $\gamma_n = 4\sqrt{|g_{\max}|}\mathbb{Q}_0\sqrt{\log(pq)/n}$,

$$\Lambda_{\min}(\tilde{\Omega}_y^k) \geq d_y - \frac{48c_y\mathbb{Q}_0\sqrt{|g_{\max}|S}}{\psi}\sqrt{\frac{\log(pq)}{n}} \geq (1-t_1)d_y > 0 \tag{B.7}$$

$$\Lambda_{\max}(\tilde{\Omega}_y^k) \leq c_y + \frac{48c_y\mathbb{Q}_0\sqrt{|g_{\max}|S}}{\psi}\sqrt{\frac{\log(pq)}{n}} \leq c_y + t_1 d_y < \infty \tag{B.8}$$

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

$$n \geq |g_{\max}|S\left[\frac{48c_y\mathbb{Q}_0}{\psi t_1 d_y}\right]^2 \log(pq).$$

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 (B.5)–(B.8) that

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

The proof is now complete by combining this with (B.6) and then applying the Cauchy-Schwarz inequality and the triangle inequality. □

Proof of Lemma 2. We drop the subscript 0 for true values and the superscript k since there is no scope of ambiguity. For part 1, we start with an auxiliary lemma:

Lemma 9. *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 $c_3 > 0$.

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

$$\begin{aligned}\mathbf{v}^T\widehat{\mathbf{S}}\mathbf{v} &= \frac{1}{n}\|\widehat{\mathbf{E}}\mathbf{v}\|^2 \\ &= \frac{1}{n}\|(\mathbf{E} + \mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}}))\mathbf{v}\|^2 \\ &= \mathbf{v}^T\mathbf{S}\mathbf{v} + \frac{1}{n}\|\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\mathbf{v}\|^2 + 2\mathbf{v}^T(\mathbf{B}_0 - \widehat{\mathbf{B}})^T\left(\frac{(\mathbf{X})^T\mathbf{E}}{n}\right)\mathbf{v}\end{aligned} \tag{B.9}$$

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 9 on $\mathbf{S}$. The second summand is greater than or equal to 0. For the third summand,

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

by assumption (A1). Now, we use another lemma:

Lemma 10. *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 (B.9) and get

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

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:

$$\begin{aligned}
\left\|\frac{1}{n}\widehat{\mathbf{E}}_{-j}^T\widehat{\mathbf{E}}\mathbf{T}_{0,j}\right\|_\infty &= \left\|\frac{1}{n}\left[\mathbf{E}_{-j} + \mathbf{X}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)\right]^T\left[\mathbf{E} + \mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\right]\mathbf{T}_{0,j}\right\|_\infty \\
&\leq \left\|\frac{1}{n}\mathbf{E}_{-j}^T\mathbf{E}\mathbf{T}_{0,j}\right\|_\infty + \left\|\frac{1}{n}\mathbf{E}_{-j}^T\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\mathbf{T}_{0,j}\right\|_\infty \\
&+ \left\|\frac{1}{n}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)^T\mathbf{X}^T\mathbf{X}(\mathbf{B}_0 - \widehat{\mathbf{B}})\mathbf{T}_{0,j}\right\|_\infty + \left\|\frac{1}{n}(\mathbf{B}_{0,j} - \widehat{\mathbf{B}}_j)^T\mathbf{X}^T\mathbf{E}\mathbf{T}_{0,j}\right\|_\infty \\
&= \|\mathbf{W}_1\|_\infty + \|\mathbf{W}_2\|_\infty + \|\mathbf{W}_3\|_\infty + \|\mathbf{W}_4\|_\infty \qquad \text{(B.10)}
\end{aligned}$$

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 10 on $\mathbf{E}_{-j}$ and $\mathbf{E}_j - \mathbf{E}_{-j}\boldsymbol{\theta}_{0,j}$ to obtain that for $n \gtrsim \log(q-1)$,

$$\|\mathbf{W}_1\|_\infty \leq c_5\left[\Lambda_{\max}(\Sigma_{y,-j})\sigma_{y,j,-j}\right]^{1/2}\sqrt{\frac{\log(q-1)}{n}} \qquad \text{(B.11)}$$

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

For $\mathbf{W}_2$ and $\mathbf{W}_4$, identical bounds hold:

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

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 $n \gtrsim \log(pq)$,

$$\|\mathbf{W}_2\|_\infty + \|\mathbf{W}_4\|_\infty \leq 4C_\beta c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_y)]^{1/2}\frac{\log(pq)}{n} \tag{B.12}$$

with probability $\geq 1 - 12c_1\exp[-(c_2^2 - 1)\log(pq)]$ by applying Lemma 10 and assumption (A1).

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

$$n \geq 512(1 + 4\Lambda_{\max}(\Sigma_x^k))^4 \max_i(\sigma_{x,ii}^k)^4 \log(4p^{\tau_1}) \tag{B.13}$$

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,

$$\|\mathbf{W}_4\|_\infty \leq \left\|\frac{\mathbf{X}^T\mathbf{X}}{n}\right\|_\infty \|\widehat{\mathbf{B}} - \mathbf{B}_0\|_1^2 \|\mathbf{T}_{0,j}\|_1 \leq 2C_\beta^2 V_x \frac{\log(pq)}{n} \tag{B.14}$$

We now bound the right hand side of (B.10) using (B.11), (B.12) and (B.14) to complete the proof, with the leading term of the sample size requirement being $n \gtrsim \log(pq)$. □

Proof of Lemma 3. The proof follows that of part (I) of Lemma 1, with a different group norm structure. We only point out the differences.

Putting $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in (2.13) 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 part (I) of Lemma 1, with a different deviation bound and choice of λ_n, we get expressions equivalent to (B.3) and (B.4) respectively:

$$\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}\mathbf{b} \leq \frac{3}{2}\sum_{h\in\mathcal{H}} \|\mathbf{b}^{[h]}\| \tag{B.15}$$

$$\frac{\psi^*}{3}\|\mathbf{b}\|^2 \leq \lambda_n \sum_{h\in\mathcal{H}} \|\mathbf{b}^{[h]}\| \leq 4\lambda_n\sqrt{B}\|\mathbf{b}\| \tag{B.16}$$

Furthermore, $\|\mathbf{b}\|_1 \leq \sqrt{|h_{\max}|}\sum_{h\in\mathcal{H}} \|\mathbf{b}^{[h]}\|$. The bounds in (A.5), (A.6), (A.7) and (A.8) now follow.

□

Proof of Lemma 4. For part 1 it is enough to prove that with $\widehat{\Sigma}_x^k := (\mathbf{X}^k)^T\mathbf{X}^k/n$,

$$\widehat{\mathbf{T}}_k^2 \otimes \widehat{\Sigma}_x^k \sim RE(\psi_*^k, \phi_*^k) \tag{B.17}$$

with high enough probability. because then we can take $\psi_* = \min_k \psi_*^k, \phi_* = \max_k \phi_*^k$. The proof of (B.17) follows similar lines of the proof of Proposition 1 in Lin et al. (2016a), 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. (2016a). □

Proof of Lemma 5. To show (A.11) 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,

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

because Ω_x is diagonally dominant implies $\|\boldsymbol{\zeta}_i\|_1 = \sum_{i'\neq i}|\omega_{x,ii'}|/\omega_{x,ii} \le 1$, and using assumption (T1) and (A.15). Applying Lemma 10, the following holds for $n \gtrsim \log(pq)$:

$$\frac{1}{\sqrt{n}}\left\|(\widehat{\Omega}_y^{1/2} - \Omega_y^{1/2})\mathbf{E}^T\mathbf{R}_i\right\|_\infty \le \sqrt{D_\Omega}(2+D_\zeta)c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_e)]^{1/2}\sqrt{\log(pq)} \tag{B.18}$$

with probability $\ge 1 - 6c_1\exp[-(c_2^2-1)\log(pq)]$.

On the other hand,

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

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

$$\left\|\frac{1}{n}\mathbf{X}_{-i}^T\mathbf{X}_{-i}\right\|_\infty \le \left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}\right\|_\infty \le V_x \tag{B.19}$$

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

$$n \ge 512(1+4\Lambda_{\max}(\Sigma_x))^4 \max_i(\sigma_{x,ii})^4\log(4p^{\tau_1}) \tag{B.20}$$

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

$$P\left(|s_i - \sqrt{\sigma_{x,i,-i}}| \geq \epsilon\right) \leq \frac{Var(s_i)}{\epsilon^2} = \frac{\kappa_i}{n\epsilon^2}$$

Taking $\epsilon = n^{-1/4}$, we have $s_i \geq \sqrt{\sigma_{x,i,-i}} - n^{-1/4}$ with probability $\geq 1 - \kappa_i n^{-1/2}$. Then, for n satisfying (B.20) and $\sqrt{\sigma_{x,i,-i}} - n^{-1/4} > D_\zeta \sqrt{V_x}$, we get the bound with the above probability:

$$\frac{1}{\widehat{s}_i} \leq \frac{1}{\sqrt{\sigma_{x,i,-i}} - n^{-1/4} - D_\zeta \sqrt{V_x}} \tag{B.21}$$

Combining (B.18) and (B.21) gives the upper bound for the right hand side of (A.11) with the requisite probability and sample size conditions.

To prove (A.12) we have

$$\frac{1}{n}\|\mathbf{R}_i^T \mathbf{X}_{-i}\|_\infty \leq \frac{1}{n}\|(\mathbf{X}_i - \mathbf{X}_{-i}\boldsymbol{\zeta}_{0,i})^T \mathbf{X}_{-i}\|_\infty + \frac{1}{n}\|\mathbf{X}_{-i}^T \mathbf{X}_{-i}(\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_{0,i})\|_\infty \tag{B.22}$$

Applying Lemma 10, for $n \gtrsim \log(p-1)$ we have

$$\frac{1}{n}\|(\mathbf{X}_i - \mathbf{X}_{-i}\boldsymbol{\zeta}_i)^T \mathbf{X}_{-i}\|_\infty \leq c_7 [\sigma_{x,i,-i}\Lambda_{\max}(\Sigma_{x,-i})]^{1/2} \sqrt{\frac{\log(p-1)}{n}} \tag{B.23}$$

with probability $\geq 1 - 6c_6 \exp[-(c_7^2 - 1)\log(p-1)]$ for some $c_6 > 0, c_7 > 1$. By (B.19), the second term on the right side of (B.22) is bounded above by $D_\zeta V_x$ with probability $\geq 1 - 1/p^{\tau_1 - 2}$ and n satisfying (B.20). The bound of (A.12) now follows by conditions (T2), (T3) and (B.21). Since $\sqrt{\sigma_{x,i,-i}} - n^{-1/4} > D_\zeta \sqrt{V_x}$ implies $\sqrt{\sigma_{x,i,-i}} > D_\zeta \sqrt{V_x}$, and $D_\zeta = O(\sqrt{\log p / n})$, the leading term of the overall sample size requirement is $n \gtrsim \log(pq)$. □

Proof of Lemma 6. We drop k in the superscripts. By definition,

$$\begin{aligned}
\frac{m_i}{\sqrt{n}} &= \frac{1}{\widehat{s}_i} \frac{(\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\boldsymbol{\zeta}}_i)^T \mathbf{X}_i}{n} \\
&= \frac{1}{\widehat{s}_i} \left[\frac{\|\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\boldsymbol{\zeta}}_i\|^2}{n} + \frac{(\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\boldsymbol{\zeta}}_i)^T \mathbf{X}_{-i}\widehat{\boldsymbol{\zeta}}_i}{n} \right] \\
&\leq \widehat{s}_i + \frac{1}{\widehat{s}_i} . \frac{1}{n} \|\mathbf{R}_i^T \mathbf{X}_{-i}\|_\infty \left(\|\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_{0i}\|_1 + \|\boldsymbol{\zeta}_{0i}\|_1 \right) \\
\Rightarrow \left| \frac{m_i}{\sqrt{n}} - \sqrt{\sigma_{x,i,-i}} \right| &\leq |\widehat{s}_i - \sqrt{\sigma_{x,i,-i}}| + \frac{1}{\widehat{s}_i} . \frac{1}{n} \|\mathbf{R}_i^T \mathbf{X}_{-i}\|_\infty \left(\|\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_i\|_1 + \|\boldsymbol{\zeta}_i\|_1 \right)
\end{aligned} \tag{B.24}$$

By applying Lemma 8 in Ravikumar et al. (2011), we have a bound for the first summand on the right hand side:

$$|\widehat{s}_i - \sqrt{\sigma_{x,i,-i}}| \leq \sqrt{\frac{\log 4 + \tau_2}{c_i n}}; \quad c_i = \left[128(1 + 4\sigma_{x,i,-i})^2 \sigma_{x,i,-i}^2\right]^{-1},$$

with probability $1 - 1/p^{\tau_2 - 2}$ for some $\tau_2 > 2$, and $n \geq 512(1 + 4\sigma_{x,i,-i})^4 2\sigma_{x,i,-i}^4 \log(4)$. For the second summand in the right-hand side of (B.24), $1/\widehat{s}_i$ can be bounded using

(B.21), $(1/n)\|\mathbf{R}_i^T\mathbf{X}_{-i}\|_\infty$ can be bounded using derivations following (B.22). Finally, $\|\widehat{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_i\|_1 \le D_\zeta$ from assumption (T1), and $\|\boldsymbol{\zeta}_i\|_1 \le 1$ because Ω_x is diagonally dominant and $|\zeta_{ii'}| = |\omega_{x,ii'}|/\omega_{x,ii}$ for $i' \neq i$. The lemma now follows by putting everything back together in (B.24). □

Proof of Lemma 7. $\|\mathbf{A} - \mathbf{A}_1\|_\infty \le \delta$ implies that $\mathbf{A}_1 + \delta\mathbf{J}_a \ge \mathbf{A}$ and $\mathbf{A} + \delta\mathbf{J}_a \ge \mathbf{A}_1$, where $\mathbf{J}_a \in \mathbb{M}(a, a)$ has all entries 1, and for positive definite matrices $\mathbf{P}, \mathbf{Q}$, $\mathbf{P} \ge \mathbf{Q}$ means $\mathbf{P} - \mathbf{Q}$ is positive definite. Now applying Theorem 1 part (a) in Bellman (1968) we have

$$(\mathbf{A} + \delta\mathbf{J}_a)^{1/2} \ge \mathbf{A}_1^{1/2}; \quad (\mathbf{A}_1 + \delta\mathbf{J}_a)^{1/2} \ge \mathbf{A}^{1/2}.$$

Using the same result, it is easy to prove that

$$\mathbf{A}^{1/2} + \sqrt{\delta}\mathbf{J}_a \ge (\mathbf{A} + \delta\mathbf{J}_a)^{1/2},$$

and the same for $\mathbf{A}_1$. The lemma follows. □

Proof of Lemma 8. We drop k in the superscripts and 0 in subscripts. Note that it is enough to prove

$$\left\| \frac{n\widehat{\Sigma}_y}{(m_i)^2} - \frac{\Sigma_y}{\sigma_{x,i,-i}} \right\|_\infty = o_P(1).$$

For this, consider the decomposition

$$\begin{aligned} \frac{n\widehat{\Sigma}_y}{(m_i)^2} &= \frac{\widehat{\Sigma}_y - \Sigma_y + \Sigma_y}{\sigma_{x,i,-i}} \cdot \frac{\sigma_{x,i,-i}}{(m_i)^2/n} \\ \Rightarrow \frac{n\widehat{\Sigma}_y}{(m_i)^2} - \frac{\Sigma_y}{\sigma_{x,i,-i}} &= \frac{\widehat{\Sigma}_y - \Sigma_y}{(m_i)^2/n} + \frac{\Sigma_y}{\sigma_{x,i,-i}} \left[1 - \frac{\sigma_{x,i,-i}}{(m_i)^2/n}\right] \\ &= \frac{n}{(m_i)^2} \left[\widehat{\Sigma}_y - \Sigma_y + \frac{\Sigma_y}{\sigma_{x,i,-i}} \left(\frac{(m_i)^2}{n} - \sigma_{x,i,-i}\right)\right]. \end{aligned}$$

From Lemma 6 we now have

$$\frac{m_i}{\sqrt{n}} \ge \sqrt{\sigma_{x,i,-i}} - \delta_i \quad \Rightarrow \quad \frac{m_i^2}{n} \ge (\sqrt{\sigma_{x,i,-i}} - \delta)^2 \ge \sigma_{x,i,-i} - \delta_i^2,$$

so that

$$\left\| \frac{n\widehat{\Sigma}_y}{(m_i)^2} - \frac{\Sigma_y}{\sigma_{x,i,-i}} \right\|_\infty \le \frac{\|\widehat{\Sigma}_y - \Sigma_y\|_\infty + \sigma_{x,i,-i}^{-1}\delta_i^2\|\Sigma_y\|_\infty}{\sigma_{x,i,-i} - \delta_i^2}, \tag{B.25}$$

with probability $\ge 1 - 6c_6 \exp[-(c_7^2 - 1)\log(p-1)] - 1/p^{\tau_2 - 2} - \kappa_i/\sqrt{n}$ and for sample size satisfying $n \gtrsim \log p$, $n \ge 512(1 + 4\sigma_{x,i,-i})^4 2(\sigma_{x,i,-i})^4 \log(4)$ and $\sqrt{\sigma_{x,i,-i}} > \max\{\delta_i, n^{-1/4} - D_\zeta\sqrt{V_x}\}$. For the ℓ_∞ norms on the right-hand side, we have

$$\|\Sigma_y\|_\infty = \|\Omega_y^{-1}\|_\infty \le (\Delta_0(\Omega_y))^{-1} \tag{B.26}$$

following Varah (1975). For a bound on $\|\widehat{\Sigma}_y - \Sigma_y\|_\infty$, if condition (II) of Theorem 6 is satisfied then we have

$$\|\widehat{\Sigma}_y - \Sigma_y\|_\infty \leq \tilde{D}_\Omega \tag{B.27}$$

where $\tilde{D}_\Omega = O(D_\Omega)$ and $D_\Omega = O(\tilde{D}_\Omega)$ Bickel and Levina (2008). If condition (I) is satisfied, denote $\epsilon = D_\Omega / \Delta_0(\Omega_y)$. Then

$$\begin{aligned}
\|\widehat{\Sigma}_y - \Sigma_y\|_\infty &= \|\widehat{\Sigma}_y(\Omega_y - \widehat{\Omega}_y)\Sigma_y\|_\infty \\
&\leq \|\widehat{\Sigma}_y\|_\infty \|\Omega_y - \widehat{\Omega}_y\|_\infty \|\Sigma_y\|_\infty \\
&\leq \|(\mathbf{I} + (\Omega_y - \widehat{\Omega}_y)\Sigma_y)^{-1}\|_\infty \|\Sigma_y\|_\infty \epsilon \\
&\leq \frac{\epsilon}{\Delta_0(\Omega_y)} \left[1 + \sum_{t=1}^{\infty} (\|(\Omega_y - \widehat{\Omega}_y)\Sigma_y\|_\infty)^t\right] \\
&\leq \frac{\epsilon}{(1-\epsilon)\Delta_0(\Omega_y)} \\
&= \frac{D_\Omega}{(\Delta_0(\Omega_y) - D_\Omega)\Delta_0(\Omega_y)}
\end{aligned} \tag{B.28}$$

Combining (B.26) with (B.27) or (B.28) as required and putting them back in the right-hand side of (B.25), we get the needed. □

Proof of Lemma 9. This is the same as in Lemma 2 in Appendix B of Lin et al. (2016a) and its proof can be found there. □

Proof of Lemma 10. This is a part of Lemma 3 of Appendix B in Lin et al. (2016a), and is proved therein. □