A Bayesian Approach to Joint Estimation of Multiple Graphical Models

Peyman Jalali, Kshitij Khare* and George Michailidis

Department of Statistics and Informatics Institute
University of Florida

Abstract

The problem of joint estimation of multiple graphical models from high dimensional data has been studied in the statistics and machine learning literature, due to its importance in diverse fields including molecular biology, neuroscience and the social sciences. This work develops a Bayesian approach that decomposes the model parameters across the multiple graphical models into shared components across subsets of models and edges, and idiosyncratic ones. Further, it leverages a novel multivariate prior distribution, coupled with a pseudo-likelihood that enables fast computations through a robust and efficient Gibbs sampling scheme. We establish strong posterior consistency for model selection, as well as estimation of model parameters under high dimensional scaling with the number of variables growing exponentially with the sample size. The efficacy of the proposed approach is illustrated on both synthetic and real data.

Keywords: Pseudo-likelihood, Gibbs sampling, posterior consistency, Omics data

1 Introduction

The problem of joint estimation of multiple related Gaussian graphical models has attracted a lot of interest in the statistics and machine learning due to its wide application in biomedical studies involving Omics data -e.g. Pierson et al. [2015] and Kling et al. [2015]-, as well as text mining and roll call voting (Guo et al. [2011]). The key idea is to “borrow strength” across the related models and thus enhance the “effective” sample size used for estimation of

*Kshitij Khare was partially supported by NSF grant DMS-119545
the model parameters, which is achieved primarily by the use of various penalty functions. Specifically, Guo et al. [2011], who first formulated the problem, modeled the elements of each inverse covariance matrix as a product of a common across all models component and an idiosyncratic (model specific) component and imposed an $\ell_1$ penalty on each one; thus, when the penalty sets the common component to zero, the corresponding edge is absent across all models, whereas if the common component is not zero, edges can be absent because the penalty sets the idiosyncratic one to zero for selected models. Another set of approaches aims to achieve a certain amount of “fusing” across all models under consideration, thus focusing both of the presence of common edges, as well as their absence across all models simultaneously. Examples of such approaches include Danaher et al. [2014] that employed a group lasso and/or a fused lasso penalty on each edge parameter across all models and Cai et al. [2016] that used a mixed $\ell_1/\ell_\infty$ norm for the same task. Variants of the above approaches with modifications to the penalties have also been explored (Zhu et al. [2014], Majumdar and Michailidis [2018]).

However, in many application settings, shared connectivity patterns across models occurs only for a subset of edges, while the remaining ones exhibit different connectivity patterns in each model. In other settings, subsets of edges share common connectivity patterns across only a subset of models. In both instances, the previously mentioned approaches will exhibit a rather poor performance in discovering these more complex patterns. To address this issue, Ma and Michailidis [2016] proposed a supervised approach based on fusing through a group lasso penalty, wherein the various connectivity patterns across subsets of edges and subsets of models are a priori known. An alternative supervised approach Saegusa and Shojaie [2016] employed a similarity graph penalty for fusing across models, coupled with an $\ell_1$ penalty for obtaining sparse model estimates. The similarity graph is assumed to be a priori known. A Bayesian variant of the latter approach was introduced in Peterson et al. [2015], wherein a Markov random field prior distribution was used to capture model similarity, followed by a spike-and-slab prior distribution on the edge model parameters. Another Bayesian approach was recently developed in Tan et al. [2017] which, similar to Peterson et al. [2015], uses $G$-Wishart prior distributions on the group-wise precision matrices given the sparsity patterns/networks in each group, and then employs a multiplicative model based hierarchical prior distribution on these networks to induce similarity/dependence.

Most of the frequentist approaches reviewed above come with performance guarantees in the form of high probability errors bounds for the model parameters as a function of the number of variables (nodes), sparsity level and sample size. Some recent work has focused on constructing confidence intervals for the difference in the magnitude of the edge parameter across two models that are estimated separately using an $\ell_1$ penalization. On the other
hand, theoretical guarantees based on high-dimensional posterior consistency results are not available for the Bayesian approaches mentioned above. Also, these approaches can suffer from computational scalability/efficiency issues in moderate/high dimensional settings, say in the presence of \( p > 30 \) nodes/variables.

The key objective of this paper is to develop a scalable approach to jointly estimate multiple related Gaussian graphical models that exhibit complex edge connectivity patterns across models for different \textit{subsets of edges}. To that end, we introduce a novel \textit{Subset Specific (S\(^2\)) prior} that for each edge aims to select the subset of models it is common to. We couple it with the Gaussian pseudo-likelihood used in Khare et al. [2015] for estimating a single Gaussian graphical model, that leads to an easy to implement and scalable Gibbs sampling scheme for exploring the posterior distribution. Finally, we establish strong posterior model selection consistency results that can be leveraged for construction of credible intervals for the edge parameters. Intuitively, the proposed framework achieves the objectives set forth in the Ma and Michailidis [2016] work, \textit{without} requiring a priori specification of the shared edge connectivity patterns; thus, the approach can be considered as fully unsupervised.

The remainder of the paper is organized as follows. Section 2 formulates the problem, while Section 3 introduces the \( S\(^2\) \) prior. In Section 4, we combine this prior with a Gaussian pseudo-likelihood to obtain a (pseudo) posterior distribution of the model parameters, and discuss how to sample from the posterior distribution. Section 5 establishes high-dimensional consistency properties for this posterior distribution. Section 6 presents extensive numerical results based on synthetic data for the framework’s performance and comparisons with existing approaches in the literature. Section 7 presents an application to metabolomics data from a case-control study on Inflammatory Bowel Disease.

## 2 Framework for Joint Estimation

Suppose we have data from \( K \) \textit{a priori} defined groups. For each group \( k \) \((k = 1, 2, ..., K)\), let \( \mathcal{Y}_k := \{ \mathbf{y}^{(k)}_i \}_{i=1}^{n_k} \) denote \( p \)-dimensional i.i.d observations from a multivariate normal distribution, with mean \( \mathbf{0} \) and covariance matrix \( (\Omega^k)^{-1} \), which is specific to group \( k \). Based on the discussion in the introductory section, the \( K \) precision matrices \( \{ \Omega^k \}_{k=1}^{K} \) can share common patterns across subsets of the \( K \) models, as delineated next. Our goal is to account for these shared structures.

Let \( \mathcal{P}(K) \) denote the power set of \( \{1, \cdots, K\} \) and for \( k = 1, ..., K \), define \( \vartheta_k \) as follows:

\[
\vartheta_k = \{ r \in \mathcal{P}(K) \setminus \{0\} : k \in r \}, \quad k = 1, ..., K.
\]  

(2.1)
It is easy to check that each $\vartheta_k$ is the collection of subsets which contain $k$, and has 
\[ \sum_{i=0}^{K-1} \binom{K-1}{i} = 2^{K-1} \] members. Denote by $\Psi^r$ the matrix that contains common patterns amongst precision matrices $\{\Omega^j\}_{j \in r}$. Specifically, for any singleton set $r = \{k\}$, the matrix $\Psi^r$ contains a pattern that is unique to group $k$, while for any other set $r$ containing more than a single element, $\Psi^r$ captures edge connectivity patterns (and their magnitudes) that are common across all members in $r$. For example, $\Psi^{123} := \Psi^{\{1,2,3\}}$ contains shared structures in $\Omega^1$, $\Omega^2$, and $\Omega^3$.

Therefore, each precision matrix $\Omega^k$ can be decomposed as
\[
\Omega^k = \sum_{r \in \vartheta_k} \Psi^r, \quad k = 1, \ldots, K, \tag{2.2}
\]
where $\sum_{r \in \vartheta_k} \Psi^r$ accounts for all the structures in $\Omega^k$ which are either unique to group $k$ (i.e. $\Psi^k$) or are shared exclusively between group $k$ and some combination of other groups (i.e. $\sum_{r \in \vartheta_k \setminus \{k\}} \Psi^r$). We further assume that $\Psi^k \in \mathbb{M}_p^+$ for $k = 1, 2, \ldots, K$, where $\mathbb{M}_p^+$ denotes the space of all $p \times p$ matrices with positive diagonal entries. Finally, the diagonal entries of every joint matrix $\Psi^r$, with $r \in \bigcup_{k=1}^K (\vartheta_k \setminus \{k\})$ are set to zero; in other words, the diagonals entries of $\Omega^k$ are contained in the corresponding $\Psi^k$.

To illustrate the notation, consider the case of $K = 3$ groups. Then, each precision matrix is decomposed as
\[
\begin{align*}
\Omega^1 &= \Psi^1 + \Psi^{12} + \Psi^{13} + \Psi^{123} \\
\Omega^2 &= \Psi^2 + \Psi^{12} + \Psi^{23} + \Psi^{123} \\
\Omega^3 &= \Psi^3 + \Psi^{13} + \Psi^{23} + \Psi^{123}
\end{align*}
\]
where the $\Psi^1$, $\Psi^2$, and $\Psi^3$ matrices contain group specific patterns, the $\Psi^{12}$, $\Psi^{13}$, $\Psi^{23}$ matrices contain patterns shared across pairs of models (for subsets of the edges) and finally matrix $\Psi^{123}$ contains patterns shared across all models/groups.

### 2.1 Identifiability Considerations

A moment of reflection shows that the model decomposition (2.2) is not unique. For example, for any arbitrary matrix $X$, the model (2.2) is equivalent to $\Omega^k = \sum_{r \in \vartheta_k} \Phi^r$ with $\Phi^r = \Psi^r + X$ and $\Phi^k = \Psi^k - \frac{1}{2^{K-1}-1}X$. Hence, without imposing appropriate identifiability constraints, meaningful estimation of the model parameters is not feasible.

In order to address this issue, we first rewrite the element-wise representation of model
\[ \omega^k_{ij} = \sum_{r \in \Theta_k} \psi^r_{ij}, \quad 1 \leq i < j \leq p, 1 \leq k \leq K, \]

(2.3)

where \( \omega^k_{ij} \) and \( \psi^r_{ij} \) are the \( ij \)th coordinates of the matrices \( \Omega^k \) and \( \Psi^r \), respectively. We only consider the upper off-diagonal entries due to the symmetry of the precision matrix and thus define vectors \( \theta_{ij} \) for every \( 1 \leq i < j \leq p \), as follows

\[ \theta_{ij} = \{ \psi^r_{ij} \}_{r \in \mathcal{P}(K)\setminus\{0\}}, \]

(2.4)

where each \( \theta_{ij} \) has \( 2^K - 1 \) distinct parameters. For identifiability purposes we require that each vector \( \theta_{ij} \) has at most one non-zero element. Note that under this constraint, if an edge \((i, j)\) is shared amongst many groups, the non-zero element will be allocated to the maximal set \( s \in \cup_{k=1}^K (\theta_k \setminus \{ k \}) \), while all subsets of \( s \) will be allocated a zero value. Further, the magnitude of all \( \{ \omega^k_{ij} \}_{k \in s} \) will be the same. As an example, consider again the case of \( K = 3 \) groups and an edge \((i, j)\) shared amongst all three groups. In this case, the edge will be allocated to the \( \Psi^{123} \) component and not to any other components, such as \( \Psi^{12} \) or \( \Psi^{13} \). Hence \( \Psi^{123}_{ij} \) will be non-zero, but

\[ \Psi^{12}_{ij} = \Psi^{13}_{ij} = \Psi^{23}_{ij} = \Psi^{i1}_{ij} = \Psi^{i2}_{ij} = \Psi^{i3}_{ij} = 0. \]

Next, we construct a novel prior distribution that respects the introduced identifiability constraint.

### 3 Subset Specific (\( S^2 \)) Prior Distribution

For any generic symmetric \( p \times p \) matrix \( A \), define

\[ \alpha = (a_{12}, a_{13}, \ldots, a_{p-1p}) , \delta_A = (a_{11}, \ldots, a_{pp}) , \]

where due to the symmetric nature of \( A \), the vector \( \alpha \) contains all the off-diagonal elements, while \( \delta_A \) the diagonal ones. In particular, \( \psi^r \) is the vectorized version of the off-diagonal elements of \( \Psi^r \).

Using the above notation, define \( \Theta \) to be the vector obtained by combining the vectors \( \{ \psi^r, r \in \mathcal{P}(K)\setminus\{0\}\} \). To illustrate, for \( K = 3 \) groups, \( \Theta \) is given by

\[ \Theta = (\psi^{123}, \psi^{23}, \psi^{13}, \psi^{12}, \psi^{3}, \psi^{2}, \psi^{1})' . \]

(3.1)
In view of (2.4), it can be easily seen that that $\Theta$ is a rearrangement of the vector $(\theta_{12}, \theta_{13}, \ldots, \theta_{p-1,p})'$. Thus, according to the location of the zero coordinates in $\theta_{ij}$ (2^K possibilities), there are $2^{K(p-1)}$ possible sparsity patterns across the K groups that $\Theta$ can have. Let $\ell$ be a generic sparsity pattern for $\Theta$ and denote the set of all the $2^{K(p-1)}$ sparsity patterns by $\mathcal{L}$. To illustrate, consider $K = 2$ groups and $p = 3$ variables. In this case, each matrix has 3 off-diagonal edges ($\{ij : 1 \leq i < j \leq p\} = \{12, 13, 23\}$). Now, assume edge 12 is shared between the two groups, edge 13 is unique to group 2, and edge 23 is absent in both groups. In this case, $\Theta$ is given by $\Theta = ((\psi_{12}^2, 0, 0), (0, \psi_{13}^2, 0), (0, 0, 0))'$; and the sparsity pattern extracted from $\Theta$ is as follows:

$$\ell = ((1, 0, 0), (0, 1, 0), (0, 0, 0))'.$$

For every sparsity pattern $\ell$, let $d_\ell$ be the density (number of non-zero entries) of $\ell$, and $M_\ell$ be the space where $\Theta$ varies, when restricted to follow the sparsity pattern $\ell$. We specify the hierarchical prior distribution $S^2$ as follows:

$$\pi(\Theta | \ell) = \frac{|\Lambda_{\ell\ell}|^{\frac{1}{2}}}{(2\pi)^{\frac{d_\ell}{2}}} \exp \left( -\frac{\Theta' \Lambda \Theta}{2} \right) I(\Theta \in M_\ell), \quad (3.2)$$

$$\pi(\ell) \propto \begin{cases} 
q_1^{d_\ell} (1 - q_1)^{(p-1) - d_\ell} & d_\ell \leq \tau, \\
q_2^{d_\ell} (1 - q_2)^{(p-1) - d_\ell} & d_\ell > \tau,
\end{cases} \quad (3.3)$$

where $\Lambda$ is a diagonal matrix whose entries determine the amount of shrinkage imposed on the corresponding elements in $\Theta$, $\Lambda_{\ell\ell}$ is a sub-matrix of $\Lambda$ obtained after removing the rows and columns corresponding to the zeros in $\Theta \in M_\ell$, and $q_1$ and $q_2$ are edge inclusion probabilities, respectively, for the case of sparse ($d_\ell \leq \tau$) and dense ($d_\ell > \tau$) $\Theta$. Later in our theoretical analysis, we specify values for $q_1$, $q_2$, and the threshold $\tau$. Let $\Theta_\ell$ be the vector containing the non-zero coordinates of $\Theta \in M_\ell$. Then, the prior in (3.2) corresponds to putting an independent normal prior on each entry of $\Theta_\ell$.

Using the prior distribution posited in (3.2) and (3.3), we derive the marginal prior distribution on $\Theta$, as follows

$$\pi(\Theta) = \sum_{\ell \in \mathcal{L}} \pi(\Theta | \ell) \pi(\ell) = \sum_{\ell \in \mathcal{L}} \left\{ \frac{|\Lambda_{\ell\ell}|^{\frac{1}{2}}}{(2\pi)^{\frac{d_\ell}{2}}} \exp \left( -\frac{\Theta' \Lambda \Theta}{2} \right) I(\Theta \in M_\ell) \right\} \pi(\ell)$$

$$= \sum_{\ell \in \mathcal{L}} u_\ell \frac{|\Lambda_{\ell\ell}|^{\frac{1}{2}}}{(2\pi)^{\frac{d_\ell}{2}}} \exp \left( -\frac{\Theta'_\ell \Lambda_{\ell\ell} \Theta_\ell}{2} \right) I(\Theta_\ell \in \mathbb{R}^{d_\ell}), \quad (3.4)$$
where
\[ u_\ell = \frac{q_1 d_\ell (1 - q_1)^{d_\ell} I_{\{d_\ell \leq \tau\}} + q_1 d_\ell (1 - q_1)^{d_\ell} I_{\{d_\ell \leq \tau\}}}{\sum_{\ell \in \mathcal{L}} \left[ q_1 d_\ell (1 - q_1)^{d_\ell} I_{\{d_\ell \leq \tau\}} + q_1 d_\ell (1 - q_1)^{d_\ell} I_{\{d_\ell \leq \tau\}} \right]} . \] (3.5)

In other words, \( \pi(\Theta) \) can be regarded as a mixture of \( 2^{K_\ell(p-1)} \) multivariate normal distributions of dimensions \( d_\ell \) that is obtained after projecting a larger dimension \( \frac{K_\ell(p-1)}{2}(2^K - 1) \) multivariate normal distribution into the union of all the subspaces \( \mathcal{M}_\ell \); namely, \( \bigcup_{\ell \in \mathcal{L}} \mathcal{M}_\ell \).

Note that the \( S^2 \) prior induces sparsity on \( \Theta \), which will be helpful for model selection purposes. Further, the prior respects the identifiability constraint by forcing at least \( \frac{K_\ell(p-1)}{2}(2^K - 2) \) parameters to be exactly equal to zero. In addition to forcing sparsity, the diagonal entries of \( \Lambda_\ell \) enforce shrinkage to the corresponding elements in \( \Theta_\ell \). We shall discuss the selection of these shrinkage parameters later in Section 4.2.

Note that the vector \( \Theta \) only incorporates the off-diagonal entries of \( \Psi \) matrices. Regarding the diagonal entries, for every \( k \in \{1, \ldots, K\} \), we let \( \delta_{\Psi,k} \) be the vector comprising of the diagonal elements of the matrix \( \Psi^k \) and define \( \Delta \) to be the vector of all diagonal vectors \( \delta_{\Psi,k} \), i.e.
\[ \Delta = (\delta_{\Psi,1}, \ldots, \delta_{\Psi,K}) . \] (3.6)

We assign an independent Exponential(\( \gamma \)) prior on each coordinate of \( \Delta \) (diagonal element of the matrices \( \Psi_k, k = 1, \ldots, K \)), i.e.,
\[ \pi (\Delta) \propto \exp \left( -\gamma 1' \Delta \right) I_{\mathbb{R}^{Kp}}(\Delta) . \] (3.7)

The selection of the hyperparameter \( \gamma \) is also discussed in Section 4.2. Since the diagonal entries of every joint matrix \( \Psi^r \), with \( r \in \cup_{k=1}^{K} (\partial_k \setminus \{k\}) \) are set to zero, the specification of the prior is now complete.

### 4 The Bayesian Joint Network Selector (BJNS)

Estimation of the model parameters \((\Theta, \Delta)\) is based on a pseudo-likelihood approach. The pseudo-likelihood, which is based on the regression interpretation of the entries of \( \Omega \), can be regarded as a weight function and as long as the product of the pseudo-likelihood and the prior density is integrable over the parameter space, one can construct a (pseudo) posterior distribution and carry out Bayesian inference. The main advantage of using a pseudo-likelihood, as opposed to a full Gaussian likelihood, is that it allows for an easy to implement sampling scheme from the posterior distribution and in addition provides more robust results under deviations from the Gaussian assumption, as illustrated by work in the frequentist
domain Khare et al. [2015], Peng et al. [2009]. Note that the pseudo-likelihood approach does not respect the positive definite constraint on the precision matrices under consideration, but since our primary goal is estimating the skeleton of the underlying graphs this mitigates this issue. Further, accurate estimation of the magnitude of the estimated edges can be accomplished through a refitting step of the model parameters restricted to the skeleton, as shown in Ma and Michailidis [2016]. We will also establish high-dimensional sparsity selection and estimation consistency for our procedure later in Section 5.

Let $S^k$ denote the sample covariance matrix of the observations in the $k^{th}$ group. Based on the above discussion, we employ the CONCORD pseudo-likelihood introduced in Khare et al. [2015],

$$
\exp \left\{ n \sum_{j=1}^{p} \log \omega_{jj}^k - \frac{n}{2} \text{tr} \left[ (\Omega^k)^2 S^k \right] \right\}, \quad k = 1, \ldots, K,
$$

and the model specification (2.2) to construct the joint pseudo-likelihood function for $K$ precision matrices, as follows,

$$
\prod_{k=1}^{K} \exp \left\{ n \sum_{j=1}^{p} \log \psi_{jj}^k - \frac{n}{2} \text{tr} \left[ \left( \sum_{r \in \vartheta_k} \Psi_r \right)^2 S^k \right] \right\}.
$$

(4.1)

Since we have parametrized the $S^2$ prior in terms of $(\Theta, \Delta)$, we will rewrite the above pseudo-likelihood function in terms of $(\Theta, \Delta)$, as well. Some straightforward algebra shows that

$$
\text{tr} \left[ \left( \sum_{r \in \vartheta_k} \Psi_r \right)^2 S^k \right] = \left( \Theta' \Delta' \right) \left( \begin{array}{cc}
\Upsilon & A \\
A' & D
\end{array} \right) \left( \Theta \Delta \right),
$$

(4.2)

where, $\Upsilon$ is a $\frac{p(p-1)(2K-1)}{2} \times \frac{p(p-1)(2K-1)}{2}$ symmetric matrix whose entries are either zero or a linear combination of $\{ s_{ij}^k \}_{1 \leq i \leq j \leq p}$; $D$ is a $Kp \times Kp$ diagonal matrix with entries $\{ s_{ii}^k \}_{1 \leq i \leq p}$; $a$ is a $\frac{p(p-1)(2K-1)}{2} \times 1$ vector whose entries are depend on $\Delta$ and $\{ s_{ij}^k \}_{1 \leq i < j \leq p}$; and finally $A$ is a $\frac{p(p-1)(2K-1)}{2} \times Kp$ matrix such that $A\Delta = a$. To see the algebraic details of the equality in (4.2), structures of $\Upsilon$ and $a$, we refer the reader to Supplemental section S1.

Now, letting $Y := (\{ y_1^{1i} \}_{i=1}^{m}, \ldots, \{ y_{K}^{1i} \}_{i=1}^{n_K})$ and by applying Bayes’ rule, the posterior distribution of $(\Theta, \Delta)$ is given by
\[ \pi \{(\Theta, \Delta) | \mathcal{Y}\} \propto \exp \left\{ n \mathbf{1}' \log(\Delta) - \frac{n}{2} \left[ \left( \begin{array}{c} \Theta' \\
 \end{array} \right) \left( \begin{array}{c} \mathbf{Y} \\
 \end{array} \right) \left( \begin{array}{c} \mathbf{A}' \\
 \end{array} \right) \left( \begin{array}{c} \Theta \\
 \end{array} \right) \right] \right\} \times \exp \left( -\frac{\Theta' \Lambda \Theta}{2} \right) \sum_{\ell \in \mathcal{L}} \left\{ \left| \Lambda_{\ell\ell} \right|^\frac{3}{2} I_{(\Theta \in \mathcal{M}_\ell)} \right\} \right) \] 

(4.3)

Moreover, the conditional posterior distribution of \( \Theta \) given \( \Delta \) is given by

\[ \pi \{\Theta | \Delta, \mathcal{Y}\} \propto \exp \left\{ -\frac{1}{2} \left[ \Theta' (n \mathbf{Y} + \Lambda) \Theta + 2n \Theta' \mathbf{a} \right] \right\} \times \sum_{\ell \in \mathcal{L}} \left\{ \left| \Lambda_{\ell\ell} \right|^\frac{3}{2} I_{(\Theta \in \mathcal{M}_\ell)} \right\} \left\{ q_1^{d_s}(1 - q_1)^{(p)} - d_t I_{(d_s \leq \tau)} + q_1^{d_s}(1 - q_1)^{(p)} - d_t I_{(d_s \leq \tau)} \right\} \times \exp (-\gamma \mathbf{1}' \Delta) \right. \]]

(4.4)

while that of \( \Delta \) given \( \Theta \) by

\[ \pi \{\Delta | \Theta, \mathcal{Y}\} \propto \prod_{k=1}^{K} \prod_{i=1}^{p} (\psi_{ii}^k)^n \exp \left\{ -\frac{n}{2} s_{ii}^k (\psi_{ii}^k)^2 - \left( \gamma + n \sum_{j \neq i} \omega_{ij}^k s_{ij}^k \right) \psi_{ii}^k \right\} \] 

(4.5)

where \( \omega_{ij}^k = \sum_{r \in \mathcal{P}_i} \psi_{ij}^r \), for \( 1 \leq i < j \leq p \) and \( 1 \leq k \leq K \).

Next, we discuss a Gibbs sampling algorithm to generate approximate samples from the above posited posterior distribution.

### 4.1 Gibbs Sampling Scheme for BJNS

Generating exact samples from the multivariate distribution in (4.3) is not computationally feasible. We instead generate approximate samples from (4.3), by computing the full conditional distribution of the vectors \( \theta_{ij} \), \( 1 \leq i < j \leq p \) and that of the diagonal entries \( \psi_{ii}^k \), \( 1 \leq i \leq p \), \( 1 \leq k \leq K \).

As discussed earlier, each \( \theta_{ij} \) contains \( 2^K - 1 \) elements of which at most one is non-zero. For ease of exposition, let \( \theta_{l_{ij}} \) denote the \( l^{th} \) element of \( \theta_{ij} \) for \( l = 1, ..., 2^K - 1 \) (based on (2.4), every \( \theta_{l_{ij}} \) represents a \( \psi_{ij}^r \), for some \( r \in \mathcal{P}(K) \)). Using the same notation for the shrinkage parameters (diagonal elements of \( \Lambda \)), let \( \lambda_{l_{ij}} \) be the shrinkage parameter corresponding to \( \theta_{l_{ij}} \). Since there are \( 2^K \) possibilities for the location of the zeros in each \( \theta_{ij} \), one can see \( \theta_{ij} \) as an element in one of the disjoint spaces \( \mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_{2^K - 1} \), where \( \mathcal{M}_0 \) is the singleton set
consisting of the zero vector of length $2^K - 1$ and $\mathbb{M}_l$ ($l = 1, \ldots, 2^K - 1$) is the space spanned by the $l^{th}$ unit vector of length $2^K - 1$.

Denote by $\Theta_{-(ij)}$ the sub vector of $\Theta$ obtained by removing $\theta_{ij}$. Define $\Upsilon_{(ij)(ij)}$ to be the sub-matrix of $\Upsilon$ obtained by removing the rows and columns with indices corresponding to the zero elements of $\theta_{ij}$ inside $\Theta$. Similarly, let $\Sigma_{(ij)(-(ij))}$ be the sub-matrix of $\Sigma$ obtained by removing the rows with indices corresponding to zero elements of $\theta_{ij}$ inside $\Theta$ and columns with indices corresponding to zero elements of $\Theta_{-(ij)}$ inside $\Theta$. Further, let $a_{ij}$ be the sub-vector of $a$ whose indices match those of $\theta_{ij}$ inside $\Theta$. Then, the conditional posterior distribution of $\theta_{ij}$ given $\Theta_{-(ij)}$ and $\Delta$ is given by

$$
\pi \left\{ \theta_{ij} | \Theta_{-(ij)}, \Delta, \Upsilon \right\} 
\propto \exp \left\{ -\frac{1}{2} \left[ \theta'_{ij} (n \Upsilon + \Lambda)_{(ij)(ij)} \theta_{ij} + 2n\theta'_{ij} (a_{ij} + \Upsilon_{(ij)(-(ij))} \Theta_{-(ij)}) \right] \right\} \sum_{l=0}^{2^K-1} I_{\mathbb{M}_l} (\theta_{ij}).
$$

Note that since $\theta_{ij}$ has at most one non-zero element, all cross products in $\theta'_{ij} (n \Upsilon + \Lambda)_{(ij)(ij)} \theta_{ij}$ are equal to zero, i.e.

$$
\theta'_{ij} (n \Upsilon + \Lambda)_{(ij)(ij)} \theta_{ij} = \sum_{l=1}^{2^K-1} \theta'^2_{l,ij} [n \Upsilon_{(ij)(ij)}]_{ll} = \sum_{l=1}^{2^K-1} \theta'^2_{l,ij} \{ n \Upsilon_{(ij)(ij)} \}_{ll} + \lambda_{l,ij}
$$

where $[\Upsilon_{(ij)(ij)}]_{ll}$ is the $l^{th}$ diagonal element of matrix $\Upsilon_{(ij)(ij)}$, for $l = 1, \ldots, 2^K - 1$. Hence, denoting the univariate normal probability density function by $\phi$, we get

$$
\pi \left\{ \theta_{ij} | \Theta_{-(ij)}, \Delta, \Upsilon \right\} 
\propto \exp \left\{ -\frac{1}{2} \sum_{l=1}^{2^K-1} \{ \theta'^2_{l,ij} \{ n \Upsilon_{(ij)(ij)} \}_{ll} + \lambda_{l,ij} \} \right\} \sum_{l=0}^{2^K-1} I_{\mathbb{M}_l} (\theta_{ij})
$$

$$
= I_{\theta_{ij} \in \mathbb{M}_0} + \sum_{l=1}^{2^K-1} c_{l,ij}\phi \left\{ \theta_{l,ij} \mid (\mu_{l,ij}, \nu^2_{l,ij}) \right\} I_{\theta_{ij} \in \mathbb{M}_l}, \quad 1 \leq i < j \leq p,
$$

with

$$
\mu_{l,ij} = -\frac{n (a_{ij} + \Upsilon_{(ij)(-(ij))} \Theta_{-(ij)})_{ll}}{n \Upsilon_{(ij)(ij)} + \lambda_{l,ij}},
$$

$$
\nu^2_{l,ij} = \frac{1}{n \Upsilon_{(ij)(ij)} + \lambda_{l,ij}},
$$

$$
c_{l,ij} = \sqrt{2\pi \nu^2_{l,ij}} \exp \left\{ \frac{\mu^2_{l,ij}}{2\nu^2_{l,ij}} \right\}.
$$

(4.6)
for $l = 0, \ldots, 2^K - 1$, and $1 \leq i < j \leq p$. Hence, for $1 \leq i < j \leq p$ we can write,

$$\pi \{ \theta_{ij} | \Theta_{-(ij)}, \Delta, \mathcal{Y} \} = \frac{I_{\theta_{ij} \in \mathcal{M}_0} + \sum_{l=1}^{2^K-1} c_{l,ij} \phi \{ \theta_{l,ij}; \left( \mu_{l,ij}, \nu_{l,ij}^2 \right) \}}{1 + \sum_{l=1}^{2^K-1} c_{l,ij}},$$

(4.7)

The above density is a mixture of univariate normal densities and the cost of generating samples from this density is comparable to that of generating from a univariate normal distribution.

In view of (4.5), one can also easily see that conditional on $\Theta$, the diagonal entries $\psi_{ii}^k$ ($1 \leq k \leq K$, $1 \leq i \leq p$) are a posteriori independent and their conditional posterior density given $\Theta$ is as follows,

$$\pi \{ \psi_{ii}^k | \Theta, \mathcal{Y} \} \propto \exp \left\{ n \log (\psi_{ii}^k) - \frac{n}{2} s_{ii}^k (\psi_{ii}^k)^2 - b_{ii}^k \psi_{ii}^k \right\},$$

(4.8)

where $b_{ii}^k = \gamma + n \sum_{j \neq i} \left( \sum_{r \in \vartheta_k} s_{ij}^r \right) s_{ii}^k$, for $1 \leq i < j \leq p$ and $1 \leq k \leq K$.

Note that the density in (4.8) is not a standard density but an efficient algorithm to sample from this density is provided in Supplemental section S2.

4.2 Selection of Shrinkage Parameters

Let $\theta$ and $\delta$ be generic elements of $\Theta$ and $\Delta$ and let $\lambda$ and $\gamma$ be their corresponding shrinkage parameters. Selecting appropriate values for the latter is an important task. In other Bayesian analysis of high dimensional models, shrinkage parameters are usually generated based on an appropriate prior distribution; see Park and Casella [2008], Kyung et al. [2010], Hans [2009]) for regression analysis and Wang et al. [2012] for graphical models. We assign independent gamma priors on each shrinkage parameter $\lambda$ or $\gamma$; specifically, $\lambda, \gamma \sim \text{Gamma}(r, s)$, for some hyper-parameters $r$ and $s$. The amount of shrinkage imposed on each element $\theta$ and $\delta$ can be calculated by considering the posterior distribution of $\lambda$ and $\gamma$, given $(\Theta, \Delta)$. Straightforward algebra shows that

$$\lambda | (\Theta, \Delta) \sim \text{Gamma}(r + 0.5, 0.5 \theta^2 + s),$$

$$\gamma | (\Theta, \Delta) \sim \text{Gamma}(r + 1, |\delta| + s).$$

Thus, we shrink $\theta$ and $\gamma$ on average by $\mathbb{E}\{ \lambda | (\Theta, \Delta) \} = \frac{r + 0.5}{0.5 \theta^2 + s}$ and $\mathbb{E}\{ \gamma | (\Theta, \Delta) \} = \frac{r + 1}{|\delta| + s}$, respectively. That is, our approach selects the shrinkage parameters $\lambda$ and $\gamma$ based on the current values of $\theta$ and $\delta$ in a way that larger (smaller) entries are regularized more (less). The selection of the hyper-parameters $r$ and $s$ is also an important task and can significantly
affect performance. Based on numerical evidence from synthetic data, we set \( r = 10^{-2} \) and \( s = 10^{-6} \); a similar suggestion is also made in Wang et al. [2012].

Finally, the construction of the Gibbs sampler proceeds as follows: matrices \( \{ \Psi^k \}_{k=1}^K \) are initialized as the identity matrix, while \( \{ \Psi^r \}_{r \in \mathcal{P}(K), k \mid r \mid > 1} \) at zero. Then, in each iteration of the Markov Chain Monte Carlo chain, we update the vectors \( \theta_{ij} \) in (2.4) and the diagonal entries \( \psi_{ii}^k \), one at a time, using their full conditional posterior densities given in (4.7) and (4.8), respectively. Algorithm 1 describes one iteration of the resulting Gibbs Sampler.

**Algorithm 1** Gibbs Sampler for BJNS

1: procedure BJNS  
   \( \triangleright \) Input \( Y, \Theta, \Delta \)
2: \hspace{1em} for \( i = 1, \ldots, p - 1 \) do
3: \hspace{2em} for \( j = i + 1, \ldots, p \) do
4: \hspace{3em} for \( l = 1, \ldots, 2^K - 1 \) do
5: \hspace{4em} \( \lambda \leftarrow \text{Gamma}(r + 0.5, 0.5(\theta^2_{ij} + s)) \)
6: \hspace{4em} \( \mu \leftarrow -\frac{n(a_{ij} + \text{avg}(\theta_{ij}))}{n(\text{avg}(\theta_{ij})) + \lambda} \)
7: \hspace{4em} \( \nu^2 \leftarrow 1 - \frac{n(\text{avg}(\theta_{ij}))}{n(\text{avg}(\theta_{ij})) + \lambda} \)
8: \hspace{4em} \( c_l \leftarrow \sqrt{2\pi \nu^2} \exp \left\{ \frac{\mu^2}{2\nu^2} \right\} \)
9: \hspace{3em} end for
10: \hspace{2em} end for
11: \hspace{1em} end for
12: \hspace{1em} for \( k = 1, \ldots, K \) do
13: \hspace{2em} \( \gamma \leftarrow \text{Gamma}(r + 1, |\psi_{ii}^k| + s) \)
14: \hspace{2em} \( b \leftarrow \gamma + n \sum_{j \neq i} \left( \sum_{r \in \vartheta_k} \psi_{ij}^r \right) s_{ij}^k \)
15: \hspace{2em} Update \( \psi_{ii}^k \) using Algorithm 1 in Supplemental section S2
16: \hspace{2em} end for
17: \hspace{1em} end for
18: \hspace{1em} for \( k = 1, \ldots, K \) do
19: \hspace{2em} \( \gamma \leftarrow \text{Gamma}(r + 1, |\psi_{pp}^k| + s) \)
20: \hspace{2em} \( b \leftarrow \gamma + n \sum_{j \neq p} \left( \sum_{r \in \vartheta_k} \psi_{pj}^r \right) s_{pj}^k \)
21: \hspace{2em} Update \( \psi_{pp}^k \) using Algorithm 1 in Supplemental section S2
22: \hspace{2em} end for
23: \hspace{1em} end for
24: \hspace{1em} return \( \Theta, \Delta \)  \( \triangleright \) Return the set of updated parameters
25: end procedure
Remark 1. Note that although BJNS is a completely unsupervised approach, available prior knowledge on shared patterns across the groups can be easily incorporated by removing redundant components $\Psi^r$. Further, prior information could be incorporated through appropriate specification of the edge selection probabilities $q_1, q_2$.

Remark 2. The Gibbs sampler described in Algorithm 1 does not involve any matrix inversion which is critical for its computational efficiency.

4.3 Procedure for Sparsity Selection

Note that the conditional posterior probability density of the off-diagonal elements of $\theta_{ij}$ is a mixture density that puts all of its mass on the events $\{\theta_{ij} : |\theta_{ij}| \leq 1\}$, where $|\theta_{ij}|$ is the number of non-zero coordinates of $\theta_{ij}$. This property of BJNS allows for model selection, in the sense that in every iteration of the Gibbs sampler one can check whether $\theta_{ij} = 0$ or which element of $\theta_{ij}$ (there could be at most one non-zero element) is non-zero. Finally, in the end of the procedure, we choose the event with the highest frequency during sampling. In addition, credible intervals can be constructed, by using the empirical quantiles of the values corresponding to the frequency distribution during sampling.

5 High Dimensional Sparsity Selection Consistency and Convergence Rates

Next, we establish theoretical properties for BJNS. Let $\{\Psi^{r,0}, r \in \bigcup_{k=1}^{K} \vartheta_k\}$ be the true values of the matrices in (2.2), so that $\Omega^{k,0} = \sum_{r \in \vartheta_k} \Psi^{r,0}$ corresponds to the true decomposition of each precision matrix $k = 1, ..., K$. Using a similar ordering as in (3.1), we define $\Theta^0$ to be the vectorized version of the off-diagonal elements of the true matrices $\{\Psi^{r,0}, r \in \bigcup_{k=1}^{K} \vartheta_k\}$.

The following assumptions are made to obtain the results.

Assumption 1. (Accurate Diagonal estimates) There exist estimates $\{\hat{\psi}_{ii}^k\}_{1 \leq i \leq p, 1 \leq k \leq K}$ such that for any $\eta > 0$, there exists a constant $C > 0$, such that

$$\max_{1 \leq i \leq p, 1 \leq k \leq K} \|\hat{\psi}_{ii}^k - \psi_{ii}^k\| \leq C \left(\sqrt{\frac{\log p}{n}}\right),$$

with probability at least $1 - O(n^{-\eta})$.

Note that our main objective is to accurately estimate the off-diagonal entries of all the matrices present in the model (2.2). Hence, as commonly done for pseudo-Likelihood based
high-dimensional consistency proofs -see Khare et al. [2015], Peng et al. [2009]- we assume the existence of accurate estimates for the diagonal elements through Assumption 5.1. One way to get the accurate estimates of the diagonal entries is discussed in Lemma 4 of Khare et al. [2015]. Denote the resulting estimates of the vectors $\Delta$ and $a$ by $\hat{\Delta}$ and $\hat{a}$, respectively. We now consider the accuracy of the estimates of the off-diagonal entries obtained after running the BJNS procedure with the diagonal entries fixed at $\hat{\Delta}$.

**Assumption 2.** $d t \sqrt{\frac{\log p}{n}} \to 0$, as $n \to \infty$.

This assumption essentially states that the number of variables $p$ has to grow slower than $e^{(\frac{1}{d^2 t})}$. Similar assumptions have been made in other high dimensional covariance estimation methods e.g. Banerjee et al. [2014], Banerjee and Ghosal [2015], Bickel and Levina [2008], and Xiang et al. [2015].

**Assumption 3.** There exists $c > 0$, independent of $n$ and $K$ such that

$$
\mathbb{E} \left[ \exp (\alpha' y^k) \right] \leq \exp (c \alpha' \alpha).
$$

While building our method, we assumed that the data in each group comes from a multivariate normal distribution. The above assumption allows for deviations from Gaussianity. Hence, Theorem 1 below will show that the BJNS procedure is robust (in terms of consistency) under misspecification of the data generating distribution, as long as its tails are sub-Gaussian.

**Assumption 4.** (Bounded eigenvalues). There exists $\tilde{\epsilon}_0 > 0$, independent of $n$ and $K$, such that for all $1 \leq k \leq K$,

$$
\tilde{\epsilon}_0 \leq \text{eig}_{\text{min}}(\Sigma^k) \leq \text{eig}_{\text{max}}(\Sigma^k) \leq \frac{1}{\tilde{\epsilon}_0}.
$$

This is a standard assumption in high dimensional analysis to obtain consistency results, see for example Bühlmann and Van De Geer [2011]. Henceforth, we let $\epsilon_0 = \frac{c^2 \tilde{\epsilon}_0}{K}$, $a_1 = \frac{\epsilon_0^2}{768 K}$, $a_2 = \frac{8c^2}{K \epsilon_0}$, $a_3 = \frac{16c^2}{\lambda}$.

**Assumption 5.** (Signal Strength). Let $s$ be the smallest non-zero entry (in magnitude) in the vector $\Theta_0$. We assume $\frac{1}{\alpha_1 n^{a_2 d_k \log p}} \to 0$.

This is again a standard assumption. Similar assumptions on the appropriate signal size can be found in Khare et al. [2015], Peng et al. [2009].

**Assumption 6.** (Decay rate of the edge probabilities). Let $q_1 = p^{-a_2 d_k}$, $q_2 = p^{-a_3 n}$, and $\tau_n = \frac{\tilde{\epsilon}_0}{4c} \frac{n}{\log p}$, for some constant $c$. 

This can be interpreted as a priori penalizing matrices with too many non-zero entries. We have faster rate $q_2$ for the case of super dense matrices. Similar assumptions are common in the literature, see for example Narisetty et al. [2014] and Cao et al. [2016].

We now establish the main posterior consistency result. In particular, we show that the posterior mass assigned to the true model converges to one in probability (under the true model).

**Theorem 1.** (Strong Selection Consistency) Based on the joint posterior distribution given in (4.3), and under Assumptions 1-6, the following holds,

$$
\pi \left\{ \Theta \in S_t | \hat{\Delta}, Y \right\} \xrightarrow{P} 1, \quad \text{as} \quad n \to \infty.
$$

(5.2)

Our next result establishes estimation consistency of the BJNS procedure for $\Theta$ and also provides a corresponding rate of convergence.

**Theorem 2.** (Estimation Consistency Rate) Let $R_n$ be the maximum value (in magnitude) in the vector $\Theta_0$ and assume that $R_n$ can not grow at a rate faster than $\sqrt{n \log p}$. Then, based on the joint posterior distribution given in (4.3), and under Assumptions 1-6, there exists a large enough constant $G$ (not depending on $n$), such that the following holds,

$$
E_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G \sqrt{\frac{d_t \log p}{n}} | \hat{\Delta}, Y \right) \right] \to 0 \quad \text{as} \quad n \to \infty.
$$

(5.3)

The proofs of the above results are provided in the Supplement, section S3.

### 6 Simulation Studies

In this section, we present three simulation studies to evaluate the performance of BJNS. In the first study, we illustrate the performance of BJNS in two scenarios with four precision matrices, each. In the second simulation, we compare the performance of BJNS with other methodologies, such as Glasso, where the Graphical lasso by Friedman et al. [2008] is applied to each graphical model separately, joint estimation by Guo et al. [2011], denoted by JEM-G, the Group Graphical Lasso denoted by GGL by Danaher et al. [2014], and the Joint Structural Estimation Method denoted by JSEM, by Ma and Michailidis [2016]. In the third simulation, we demonstrate the numerical scalability of BJNS, when the number of precision matrices $K$ is relatively large.

Throughout, for any $K$ precision matrices $\{\Omega^k\}_{k=1}^K$ of dimensions $p \times p$, we generate data
as follows,
\[ y^{k}_{i} \sim N_{p}\left(0, \left(\Omega^{k}\right)^{-1}\right), \quad i = 1, \ldots, n_{k}, k = 1, \ldots, K. \]

We assess the model performance using three well known accuracy measures, specificity (SP), sensitivity (SE) and Matthews Correlation Coefficients (MCC) defined as:

\[
\begin{align*}
\text{SP} &= \frac{TN}{TN + FP}, \\
\text{SE} &= \frac{TP}{TP + FN}, \\
\text{MCC} &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\end{align*}
\] (6.1)

where, TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives, respectively. Larger values of any of the above metrics indicates a better sparsity selection produced by the underlying method.

Recall that BJNS estimates the vectors \( \theta_{ij} \) so that only a single coordinate is non-zero; i.e. regardless of the number of networks \( K \), BJNS estimates at most \( p(p-1)/2 \) off-diagonal parameters and enforces at least \( (2^{K-1} - 1)p(p-1) \) of the remaining off-diagonal parameters to be zero. The final decision for the sparsity of each vector \( \theta_{ij} \) in (2.4), is made based on majority voting. For every experiment, we base our inference on 2000 samples that are generated from the MCMC chain after removing 2000 burn-in samples. We ensure the robustness of the presented results, by repeating each experiment over 100 replicate data sets and taking the average of the accuracy measures across the replicates.

Finally, the implementation of BJNS using the Rcpp and Rcpp Armadillo libraries (Eddelbuettel and Sanderson [2014], Eddelbuettel and François [2011]) together with avoidance of matrix inversion operations as previously discussed contributes to its computational efficiency.

### 6.1 Simulation 1: Four groups \( (K = 4) \)

We consider two challenging scenarios to examine the performance of BJNS for the case of simultaneously estimating four inverse covariance matrices \( (K = 4) \).

(i) Four precision matrices \( \Omega^{1}, \Omega^{2}, \Omega^{3}, \) and \( \Omega^{4} \), with different degrees of shared structures.

We let the first matrix \( \Omega^{1} \) to be an AR(2) model with \( \omega_{jj}^{1} = 1 \), for \( j = 1, \ldots, p; \omega_{jj+1}^{1} = \omega_{j+1j}^{1} = 0.5 \), for \( j = 1, \ldots, p - 1 \); and \( \omega_{jj+2}^{1} = \omega_{j+2j}^{1} = 0.25 \), for \( j = 1, \ldots, p - 2 \).

To construct \( \Omega^{2} \) we randomly replace \( \frac{p}{4} \) non zero edges from \( \Omega^{1} \) with zeros and replace \( \frac{p}{4} \) zero edges, at random, with numbers generated from \([-0.6, -0.4] \cup [0.4, 0.6]\). We construct \( \Omega^{3} \) by randomly removing \( \frac{p}{2} \) edges shared between \( \Omega^{1} \) and \( \Omega^{2} \) and then using \([-0.6, -0.4] \cup [0.4, 0.6] \), we randomly add \( \frac{p}{2} \) other edges that are present in neither \( \Omega^{1} \)
nor $\Omega^2$. Finally, we construct $\Omega^4$ by removing the remaining $2p - 3 - \frac{3p}{4}$ edges that are common in $\Omega^1$ and $\Omega^2$ and randomly add $2p - 3 - \frac{3p}{4}$ edges that are not present in any of the other graphs. The resulting matrix $\Omega^4$ has nothing in common with the other precision matrices. Therefore, the true relation between the four networks is as follows,

$$\begin{align*}
\Omega^1 &= \Psi^1 + \Psi^{12} + \Psi^{123} \\
\Omega^2 &= \Psi^2 + \Psi^{12} + \Psi^{123} \\
\Omega^3 &= \Psi^3 + \Psi^{123} \\
\Omega^4 &= \Psi^4,
\end{align*}$$

(6.2)

where, $\Psi^1$, $\Psi^2$, $\Psi^3$, and $\Psi^4$ account for the edges that are unique to their corresponding groups and $\Psi^{12}$ and $\Psi^{123}$ contains the edges that are common between the four groups. A heat map plot of the true precision matrices with $p = 50$ is given figure 1.

Figure 1: Heat maps of the precision matrices of the four groups

(ii) In most real data applications, the underlying networks tend to follow sparsity patterns that are completely random and do not necessarily follow a certain structure. For this reason, we would like to examine the performance of BJNS under unstructured settings. We will specifically consider estimating four precision matrices with complete random sparsity patterns with signals generated from $[-0.6, -0.4] \cup [0.4, 0.6]$. We take the level of sparsity in each matrix to be 95%, half of which is unique to the matrices and the other half is shared between all four of them. The true relationship between the networks is as follows,

$$\Omega^k = \Psi^k + \Psi^{1234}, \quad k = 1, 2, 3, 4.$$  

(6.3)

where, $\Psi^1$, $\Psi^2$, $\Psi^3$, and $\Psi^4$ account for the edges that are unique to their corresponding groups and $\Psi^{1234}$ contains the edges that are common between all of the four groups.

For each of the above settings, we run the following combinations of $p$ and $n$:

- $p = 200$, and $n = 50, 100, 150, 200, 250$, and $300$,
- $p = 500$, and $n = 300, 400$, and $500$. 

17
The estimations are based on the full decomposition model, which for the case of \(K = 4\) is given as,

\[
\begin{align*}
\Omega^1 &= \Psi^1 + \Psi^{12} + \Psi^{13} + \Psi^{14} + \Psi^{123} + \Psi^{124} + \Psi^{134} + \Psi^{1234}, \\
\Omega^2 &= \Psi^2 + \Psi^{12} + \Psi^{23} + \Psi^{24} + \Psi^{123} + \Psi^{124} + \Psi^{234} + \Psi^{1234}, \\
\Omega^3 &= \Psi^3 + \Psi^{13} + \Psi^{23} + \Psi^{34} + \Psi^{123} + \Psi^{134} + \Psi^{234} + \Psi^{1234}, \\
\Omega^4 &= \Psi^4 + \Psi^{14} + \Psi^{24} + \Psi^{34} + \Psi^{124} + \Psi^{134} + \Psi^{234} + \Psi^{1234}. \\
\end{align*}
\tag{6.4}
\]

Tables 1 and 2 summarize the average of the accuracy measures for both settings, across 100 repetitions. As can be seen, the values of the accuracy measures tend to be much higher for the joint effects (namely \(\Psi^{12}\) and \(\Psi^{123}\) in table 1 and \(\Psi^{1234}\) in table 2), which implies that BJNS is borrowing strength across the distinct samples, to provide more robust estimates of the joint edges. In addition, high values of specificity, regardless of the sample size, shows a surprisingly low false positive rate.

Another key strength of BJNS is its ability to select the correct underlying model from the above full representation. As described in the Gibbs sampler in section 4.1, model selection takes place at the level of \(\theta_{ij}\)s (for \(1 \leq i < j \leq p\)) in (2.4) which in the case of \(K = 4\), are vectors of length \(2^K - 1 = 15\),

\[
\theta_{ij} = (\psi_{ij}^1, \psi_{ij}^2, \psi_{ij}^3, \psi_{ij}^4, \psi_{ij}^{12}, \psi_{ij}^{13}, \psi_{ij}^{14}, \psi_{ij}^{23}, \psi_{ij}^{24}, \psi_{ij}^{123}, \psi_{ij}^{124}, \psi_{ij}^{134}, \psi_{ij}^{234}, \psi_{ij}^{1234})'.
\tag{6.5}
\]

for every \(1 \leq i < j \leq p\), BJNS aims at detecting whether \(\theta_{ij}\) is a zero vector or which of it’s components is non-zero (recall that \(\theta_{ij}\) has at most one non-zero element). For example, in the case of the first simulation, if an edge \((ij)\) is common across all four networks, then \(\psi_{ij}^{1234}\) must be the non-zero element in the corresponding \(\theta_{ij}\). That is, we must have that

\[
\theta_{ij} = \begin{pmatrix} 0_{14} \\ \psi_{ij}^{1234} \end{pmatrix}.
\]

Thus, based on the way we estimate \(\theta_{ij}\)s, high values of the accuracy measures for the matrices that are present in the true models (6.2) and (6.3), automatically indicates that the model is performing well in selecting the true sparsity patterns of all \(\theta_{ij}\)s. In view of the full model (6.4), high values of the accuracy measures for the inverse covariance matrices \(\{\Omega^k\}_{k=1}^4\) is another indication of strong selection capability of BJNS.

Further, note that the model selection properties of BJNS to check convergence of the MCMC algorithm. This is accomplished by tracking the ratio of correctly selected \(\theta_{ij}\)s,
|       | MC%     | SP%     | SE%     |
|-------|---------|---------|---------|
| $\Omega_1$ | $n = 50$ | $n = 100$ | $n = 150$ | $n = 200$ | $n = 250$ | $n = 300$ |
| $\Psi_1$ | 60 (0.020) | 79 (0.016) | 86 (0.012) | 88 (0.011) | 89 (0.010) | 89 (0.009) |
| $\Psi_2$ | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) |
| $\Psi_3$ | 67 (0.021) | 90 (0.014) | 97 (0.009) | 99 (0.004) | 100 (0.002) | 100 (0.001) |
| $\Omega_2$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_4$ | 51 (0.021) | 69 (0.016) | 78 (0.013) | 81 (0.014) | 80 (0.011) | 82 (0.012) |
| $\Psi_5$ | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) |
| $\Psi_6$ | 59 (0.024) | 82 (0.018) | 92 (0.014) | 95 (0.011) | 95 (0.010) | 97 (0.010) |
| $\Omega_3$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_7$ | 43 (0.021) | 64 (0.019) | 73 (0.015) | 70 (0.015) | 80 (0.014) | 80 (0.014) |
| $\Psi_8$ | 98 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) |
| $\Psi_9$ | 51 (0.024) | 78 (0.022) | 89 (0.016) | 95 (0.015) | 96 (0.011) | 96 (0.010) |
| $\Omega_4$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{10}$ | 19 (0.020) | 47 (0.020) | 54 (0.020) | 72 (0.015) | 75 (0.016) | 81 (0.012) |
| $\Psi_{11}$ | 98 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) | 99 (0.001) |
| $\Psi_{12}$ | 21 (0.020) | 53 (0.024) | 61 (0.023) | 84 (0.016) | 87 (0.014) | 95 (0.010) |
| $\Psi_{13}$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{14}$ | 30 (0.048) | 52 (0.047) | 66 (0.051) | 75 (0.043) | 76 (0.040) | 81 (0.037) |
| $\Psi_{15}$ | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) |
| $\Psi_{16}$ | 33 (0.056) | 57 (0.058) | 71 (0.067) | 82 (0.047) | 85 (0.044) | 89 (0.040) |
| $\Psi_{17}$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{18}$ | 31 (0.051) | 56 (0.043) | 71 (0.035) | 75 (0.037) | 74 (0.033) | 76 (0.035) |
| $\Psi_{19}$ | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) |
| $\Psi_{20}$ | 37 (0.067) | 69 (0.067) | 87 (0.041) | 92 (0.040) | 94 (0.038) | 97 (0.024) |
| $\Psi_{21}$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{22}$ | 25 (0.041) | 58 (0.040) | 73 (0.033) | 82 (0.025) | 84 (0.023) | 84 (0.022) |
| $\Psi_{23}$ | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) |
| $\Psi_{24}$ | 23 (0.038) | 60 (0.046) | 79 (0.039) | 92 (0.028) | 94 (0.024) | 96 (0.022) |
| $\Psi_{25}$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{26}$ | 24 (0.027) | 56 (0.023) | 64 (0.021) | 81 (0.017) | 84 (0.013) | 90 (0.011) |
| $\Psi_{27}$ | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) |
| $\Psi_{28}$ | 14 (0.017) | 42 (0.024) | 52 (0.022) | 76 (0.020) | 80 (0.017) | 90 (0.015) |
| $\Psi_{29}$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{30}$ | 36 (0.041) | 57 (0.034) | 70 (0.021) | 78 (0.029) | 80 (0.030) | 83 (0.026) |
| $\Psi_{31}$ | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) |
| $\Psi_{32}$ | 34 (0.043) | 59 (0.040) | 74 (0.038) | 82 (0.036) | 83 (0.039) | 88 (0.029) |
| $\Psi_{33}$ | $n = 300$ | $n = 400$ | $n = 500$ |
| $\Psi_{34}$ | 50 (0.029) | 71 (0.022) | 80 (0.020) | 88 (0.018) | 89 (0.014) | 90 (0.013) |
| $\Psi_{35}$ | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) | 100 (0.000) |
| $\Psi_{36}$ | 34 (0.028) | 60 (0.028) | 72 (0.029) | 83 (0.027) | 85 (0.020) | 88 (0.022) |
Table 2: Summary results for simulation ii

| Simulation | $p = 200$ | $p = 500$ |
|------------|-----------|-----------|
| $\Omega^1$ |           |           |
| MC%        | 17 (0.009) | 61 (0.004) |
| SP%        | 88 (0.004) | 99 (0.000) |
| SE%        | 39 (0.015) | 47 (0.004) |
| $\Omega^2$ |           |           |
| MC%        | 18 (0.011) | 59 (0.004) |
| SP%        | 88 (0.004) | 99 (0.000) |
| SE%        | 39 (0.016) | 46 (0.004) |
| $\Omega^3$ |           |           |
| MC%        | 18 (0.009) | 61 (0.005) |
| SP%        | 88 (0.004) | 99 (0.000) |
| SE%        | 39 (0.014) | 48 (0.005) |
| $\Omega^4$ |           |           |
| MC%        | 17 (0.009) | 61 (0.005) |
| SP%        | 88 (0.004) | 99 (0.000) |
| SE%        | 38 (0.013) | 47 (0.005) |
| $\Psi^1$   |           |           |
| MC%        | 7 (0.013)  | 30 (0.008) |
| SP%        | 98 (0.001) | 100 (0.000) |
| SE%        | 9 (0.012)  | 11 (0.005) |
| $\Psi^2$   |           |           |
| MC%        | 7 (0.012)  | 29 (0.008) |
| SP%        | 98 (0.001) | 100 (0.000) |
| SE%        | 9 (0.011)  | 11 (0.004) |
| $\Psi^3$   |           |           |
| MC%        | 8 (0.013)  | 30 (0.008) |
| SP%        | 98 (0.001) | 100 (0.000) |
| SE%        | 9 (0.013)  | 11 (0.005) |
| $\Psi^4$   |           |           |
| MC%        | 7 (0.014)  | 30 (0.009) |
| SP%        | 98 (0.001) | 100 (0.000) |
| SE%        | 9 (0.013)  | 11 (0.005) |
| $\Psi^{1234}$ |         |           |
| MC%        | 21 (0.020) | 68 (0.007) |
| SP%        | 99 (0.001) | 100 (0.000) |
| SE%        | 16 (0.015) | 49 (0.009) |
denoted by \( \kappa \) and defines as

\[
\kappa = \frac{\# \theta_{ij}s \text{ that are correctly selected}}{\binom{p(p-1)}{2}}
\]

In addition to accuracy assessment, \( \kappa \) helps studying the number of iterations that on average it takes for the Gibbs sampler to converge. Figure 2 describes the trace plot and the histogram of \( \kappa \) for 4000 iterations for the above two simulations for the \( p = 200, n = 300 \) settings. The MCMC trace plots of \( \kappa \) show that the Gibbs sampler converges fairly quickly. Moreover, the histograms of \( \kappa \) indicate a high proportion of correctly selected \( \theta_{ij}s \).

Figure 2: Convergence and the distribution of \( \kappa \) during the MCMC samplings

6.2 Comparison With Existing Methods

In this section, we compare the performance of BJNS with Glasso, JEM-G, GGL and JSEM, in two settings involving 6 networks. We also discuss a pure computational strategy to scale up BJNS to handle efficiently an even larger number of networks.

6.2.1 Comparison With Existing Methods

We consider a scenario with \( K = 6 \) graphs each with \( p = 200 \) variables (see Figure 3), where we first generate the adjacency matrices corresponding to three distinct \( p \)-dimensional networks, so that the adjacency matrices in each column of the plot in Figure 3 are common. Next, we replace the connectivity structure of the bottom right diagonal block of size \( p/2 \) by \( p/2 \) in each adjacency matrix with that of an other two distinct \( p/2 \)-dimensional networks, so that graphical models in each column exhibit the same connectivity pattern except in the bottom right diagonal block of their adjacency matrices. The resulting true matrix
decompositions of the networks is as follows:

\[ \Omega^1 = \Psi^{12} + \Psi^{135}, \quad \Omega^3 = \Psi^{34} + \Psi^{135}, \quad \Omega^5 = \Psi^{56} + \Psi^{135}, \]
\[ \Omega^2 = \Psi^{12} + \Psi^{246}, \quad \Omega^4 = \Psi^{34} + \Psi^{246}, \quad \Omega^6 = \Psi^{56} + \Psi^{246}. \] (6.6)

Note that by replacing the connectivity structure among the second half of the nodes, the relationships between the first half and the second half of the nodes are also altered. In summary, these sparsity patterns illustrate the common components across different subsets of edges, as well as differences. The sparsity level for all networks is set to 92%, while the proportion of common zeros (no edge present) across all six networks is about 60%. Given the adjacency matrices, we then construct the inverse covariance matrices with the nonzero off-diagonal entries in each \( \Omega^k \) being uniformly generated from the \([-0.6, -0.4] \cup [0.4, 0.6]\) interval. To implement JESM, we supply the sparsity patterns defined according to the pattern in 3. We also study the effect of mis-specification to JSEM in particular, by adding an additional \( \rho = 4\% \) of edges to the networks. Adding \( \rho = 4\% \) of edges results in having 60% of the information in the sparsity patterns being correct for JSEM.

At each level of pattern mis-specification, we generated \( n_k = 200, 300 \) independent samples for each \( k = 1, \ldots, K \) and examined the finite sample performance of different methods in identifying the true graphs and estimating the precision matrices at the optimal choice of tuning parameters. Table 3 show the deviance measures between the estimated and the true precision matrices based on 50 replications for varying levels of sample size and pattern mis-specification.

When \( \rho = 0 \) (no mis-specifications), JSEM, which benefits from knowing 100% of the information about which sets of edges can be fused across which subsets of networks, achieves a good balance between false positives and false negatives and yields the highest MCC score, as expected. BJNS is also very competitive and its overall performance is significantly better than all other unsupervised methods. Glasso and GGL tend to perform well in controlling false negatives and JEM-G is constantly among the best two models in terms of the SP score. With \( \rho = 0.04 \) and sparsity level of 12%, naturally the overall performance of all methods decrease. However, JSEM suffers more as it only has correctly specified information about shared patterns for only 60% of the edges. On the other hand, BJNS proves robust and achieves a good balance between specificity and sensitivity, thus outperforming competing methods. In this setting, Glasso is also competitive due to the overall heterogeneity of the 6 networks.
6.2.2 A Computational Strategy for Speeding Up BJNS for large $K$

As presented in Algorithm 1, the Gibbs sampler updates all the $p(p-1)/2$ vectors $\theta_{ij}$ based on their full conditional distributions. Although the conditional posterior distribution of $\theta_{ij}s$ is a mixture of univariate normal densities, all $2^K$ mixture probabilities $c_{i,j}$ given in (4.6) still need to be calculated, which in turn involves matrix-vector multiplications. Hence with increasing $K$, the computation complexity of the full decomposition (2.2) grows quickly.

Next, we discuss a strategy that starts by examining all pairwise decompositions to identify inactive pairwise components and the higher matrices $r$ of such pairwise components. In the first step, $\binom{K}{2}$ pairwise joint models are considered; namely, for any pair $1 \leq k_1 < k_2 \leq K$, we examine

$$\Omega^{k_1} = \Psi^{k_1} + \Psi^{k_1k_2},$$
$$\Omega^{k_2} = \Psi^{k_2} + \Psi^{k_1k_2},$$

(6.7)

Subsequently, we remove all the pairwise matrices $\Psi^{k_1k_2}$ that are considered “inactive”, i.e have significantly fewer edges compared to other pairwise matrices. Once the “inactive” pairwise components are identified, we then remove any higher order component $r$ that contains them. Next, we run the resulting reduced model and count the number of edges present in each component $r$ and calculate the number of edges present in the estimated matrices. We then further reduce the model if any matrix component seem to be inactive (has significantly smaller number of edges) and finally run BJNS one last time with the resulting reduced model.

The proposed purely computational strategy is illustrated on the setting given in Figure 3, which involves $K = 6$ groups with $p = 200$ variables and $n_k = 200, 300$ samples per group.
Table 3: Summary of model comparisons based on average of deviance measures across the 6 networks, for the case of random sparsity patterns

| ρ   | Glasso | JEM-G | GGL  | JSEM  | BJNS  |
|-----|--------|-------|------|-------|-------|
|     | MC%    | SP%   | SE%  |       |       |
| 0   | 47 (0.009) | 94 (0.003) | 60 (0.010) | 61 (0.009) | 57 (0.010) |
|     | (0.009) | (0.010) | (0.010) | (0.009) | (0.010) |
| 0.04| 40 (0.008) | 93 (0.003) | 46 (0.009) | 46 (0.008) | 40 (0.010) |
|     | (0.010) | (0.008) | (0.008) | (0.009) | (0.010) |
|     | 40 (0.008) | 93 (0.003) | 46 (0.009) | 46 (0.008) | 51 (0.010) |
|     | (0.010) | (0.008) | (0.008) | (0.009) | (0.010) |
|     | 54 (0.009) | 93 (0.003) | 72 (0.009) | 75 (0.009) | 77 (0.010) |
|     | (0.010) | (0.008) | (0.008) | (0.009) | (0.010) |
|     | 54 (0.008) | 93 (0.003) | 72 (0.009) | 75 (0.009) | 77 (0.010) |
|     | (0.010) | (0.008) | (0.008) | (0.009) | (0.010) |
|     | 42 (0.008) | 90 (0.003) | 42 (0.008) | 42 (0.008) | 51 (0.010) |
|     | (0.007) | (0.006) | (0.006) | (0.007) | (0.006) |

Note that the full decomposition would involve $r$-tuple interaction components, $r = 2, \cdots, 6$, which renders computations expensive, since each vector $\theta_{ij}$ would have length $2^{K-1} = 63$. However, as can be seen by the design of the simulation, most components $r$ are “inactive”. For instance, in the case of (6.6), only 5 out of the 63 components in the full model are non-zero (active).

We start by first studying the $\binom{6}{2}$ pairwise interaction components. Figure 4a shows the number of edges in the pairwise matrices $\Psi_{k_1 k_2}$, $1 \leq k_1 < k_2 \leq 6$.

Based on this plot, one can see that matrices $\Psi_{14}$, $\Psi_{16}$, $\Psi_{23}$, $\Psi_{25}$, $\Psi_{35}$, and $\Psi_{45}$ have significantly smaller number of edges compared to other pairwise matrices. Hence, in the next step we remove all of these matrices and any higher order components, whose super scripts contain those of $\Psi_{14}$, $\Psi_{16}$, $\Psi_{23}$, $\Psi_{25}$, $\Psi_{35}$, and $\Psi_{45}$.

Doing so results in a reduced model with components, $\Psi_{12}$, $\Psi_{13}$, $\Psi_{15}$, $\Psi_{24}$, $\Psi_{26}$, $\Psi_{34}$, $\Psi_{35}$, $\Psi_{46}$, $\Psi_{56}$, $\Psi_{135}$, and $\Psi_{246}$. Thus, in the second step, we run BJNS with a decomposition that is based on the above components; the edge count for these components are shown in plot 4b. From which, it is clear that matrices $\Psi_{13}$, $\Psi_{15}$, $\Psi_{24}$, $\Psi_{26}$, $\Psi_{34}$, $\Psi_{35}$, $\Psi_{46}$, $\Psi_{56}$, $\Psi_{135}$, and $\Psi_{246}$ are redundant and should be removed from the model. This further model reduction achieves the true decomposition given in 6.6. The results of BJNS running on the resulting reduced model can be read off from the Table 4. As can be seen, BJNS outperforms JSEM, which has been
Figure 4: Bar plots of the edge count of the pairwise joint components (a) and components of the reduced model (b)

supplied with complete information about the sparsity patterns.

Table 4: Comparisons between JSEM and BJNS when employing the step wise computational strategy; the results are based on 50 replications

|            | JSEM | BJNS | JSEM | BJNS |
|------------|------|------|------|------|
|            | n = 200 |      | n = 300 |      |
| MC%        | 61 (0.009) | 69 (0.011) | 73 (0.008) | 83 (0.007) |
| SP%        | 99 (0.001) | 99 (0.001) | 99 (0.001) | 100 (0.001) |
| SE%        | 46 (0.009) | 56 (0.010) | 63 (0.008) | 76 (0.010) |

6.2.3 Computational Cost of BJNS

Lastly, as presented in Table 5, we investigate the computational cost associated with the above strategy, across varying values of $p$ and $n$. Each experiment was repeated 5 times and all computations were done sequentially using one processor (CPU). Note that around 60% of the time in each experiment was spent on the first step which is investigating all the $\binom{6}{2}$ pairwise models. Since, the pairwise models are ran independently, one can use parallel computing and reduce the computational time in table 5 potentially by 50%. Finally, as can be seen in the last row of the table, the memory usage of BJNS is not necessarily large and that is due to the fact that the algorithm does not involve any matrix inversions or generation of samples from multivariate distributions.
Table 5: accuracy and cost of BJNS for varying values of \( p \)

| \( p \) = 200 | \( p \) = 500 | \( p \) = 700 | \( p \) = 1000 |
|-----|-----|-----|-----|
| \( n \) = 300 | \( n \) = 750 | \( n \) = 1050 | \( n \) = 1500 |
| MC% | 83  | 84  | 85  | 85  |
| SP% | 100 | 100 | 100 | 100 |
| SE% | 76  | 78  | 78  | 78  |
| hours | 2.9h | 38.9h | 83.3h | 214.5h |
| GigaBytes | 0.25gb | 0.5gb | 0.6gb | 0.9gb |

7 An Application of BJNS to Metabolomics Data

In this section, we employ the proposed methodology to obtain networks across four groups of patients that participated in the Integrative Human Microbiome Project. The data were downloaded from the Metabolomics Workbench [www.metabolomicsworkbench.org](http://www.metabolomicsworkbench.org) (Study ID ST000923) and correspond to measurements of 428 primary and secondary metabolites and lipids from stool samples of 542 subjects, partitioned in the following groups: inflammatory bowel disease (IBD) patients (males \( n_1 = 202 \) and females \( n_2 = 208 \)) and non-IBD controls (males \( n_3 = 72 \) and females \( n_4 = 70 \)), Groups 1-4, respectively. Since there are two factors in the study design, the following model was fitted to the data.

\[
\begin{align*}
\Omega_1 &= \Psi_1 + \Psi_{12} + \Psi_{13} + \Psi_{1234}, \\
\Omega_2 &= \Psi_2 + \Psi_{12} + \Psi_{24} + \Psi_{1234}, \\
\Omega_3 &= \Psi_3 + \Psi_{13} + \Psi_{34} + \Psi_{1234}, \\
\Omega_4 &= \Psi_4 + \Psi_{24} + \Psi_{34} + \Psi_{1234}.
\end{align*}
\]

The results are shown in the next Table (set1: 289 lipids in red and set2: 139 metabolites in blue, set12: interaction edges between set1 and set2), for both the final estimates of the precision matrices, as well as the components in the proposed decomposition. It is interesting to note that a large number of edges are shared across all groups, indicating common patterns. Further, the component shared between male and female IBD patients has a fairly large number of edges, indicating that the disease status exhibits commonalities across both males and females.

Table 6: Number of edges in each matrix

| \( \Omega \) | \( \Psi \) | set1 | set2 | set12 |
|-----|-----|-----|-----|-----|
| \( \Omega^1 \) | \( \Psi^1 \) | 941 | 264 | 154 |
| \( \Omega^2 \) | \( \Psi^2 \) | 930 | 269 | 155 |
| \( \Omega^3 \) | \( \Psi^3 \) | 797 | 243 | 124 |
| \( \Omega^4 \) | \( \Psi^4 \) | 799 | 244 | 124 |
| \( \Psi_{12} \) | \( \Psi_{4} \) | 47 | 2 | 10 |
| \( \Psi_{13} \) | \( \Psi_{12} \) | 29 | 5 | 11 |
| \( \Psi_{1234} \) | \( \Psi_{12} \) | 18 | 2 | 5 |
| \( \Psi_{13} \) | \( \Psi_{12} \) | 13 | 2 | 5 |
| \( \Psi_{24} \) | \( \Psi_{12} \) | 161 | 34 | 34 |
| \( \Psi_{34} \) | \( \Psi_{12} \) | 81 | 14 | 13 |
| \( \Psi_{1234} \) | \( \Psi_{12} \) | 88 | 16 | 13 |

26
The next plot shows a network map of the common component shared across all groups. The primary and secondary metabolites are depicted in red, while the lipids in blue. Not surprisingly, primary metabolites (those involved in cellular growth, development and reproduction) form a fairly strongly connected network. On the other hand, the connectivity between lipids (whose functions include storing energy, signaling and acting as structural components of cell membranes) to the metabolites is not particularly strong. On the other hand, different fairly strongly connected subnetworks amongst lipids are present, including dicylglycerols (DAG) with tricylglycerols (TAG) that are main constituents of animal and vegetable fat (upper right corner of the plot) and various phospholipids (upper left corner of the plot). In general, the results reveal interesting patterns that can be used to understand progression of IBD disease.

Figure 5: Network plot of the edges shared between the four groups (Ψ^{1234}); names of the metabolites and the lipids appear in red and blue, respectively.
References

Sayantan Banerjee and Subhashis Ghosal. Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162, 2015.

Sayantan Banerjee, Subhashis Ghosal, et al. Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics*, 8(2):2111–2137, 2014.

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445, 2016.

Xuan Cao, Kshitij Khare, and Malay Ghosh. Posterior graph selection and estimation consistency for high-dimensional bayesian dag models. *arXiv preprint arXiv:1611.01205*, 2016.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

Dirk Eddelbuettel and Romain Franois. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08. URL http://www.jstatsoft.org/v40/i08/.

Dirk Eddelbuettel and Conrad Sanderson. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014. URL http://dx.doi.org/10.1016/j.csda.2013.02.005.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.
Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):803–825, 2015.

Teresa Kling, Patrik Johansson, José Sanchez, Voichita D Marinescu, Rebecka Jörnsten, and Sven Nelander. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic acids research*, 43(15):e98–e98, 2015.

Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research*, 17(1):5777–5824, 2016.

Subhabrata Majumdar and George Michailidis. Joint estimation and inference for data integration problems based on multiple multi-layered gaussian graphical models. *arXiv preprint arXiv:1803.03348*, 2018.

Naveen Naidu Narisetty, Xuming He, et al. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

Emma Pierson, Daphne Koller, Alexis Battle, Sara Mostafavi, GTEx Consortium, et al. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS computational biology*, 11(5):e1004220, 2015.

Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Takumi Saegusa and Ali Shojaie. Joint estimation of precision matrices in heterogeneous populations. *Electronic journal of statistics*, 10(1):1341, 2016.
Linda SL Tan, Ajay Jasra, Maria De Iorio, Timothy MD Ebbels, et al. Bayesian inference for multiple gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, 11(4):2222–2251, 2017.

Hao Wang et al. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

Ruoxuan Xiang, Kshitij Khare, Malay Ghosh, et al. High dimensional posterior convergence rates for decomposable graphical models. *Electronic Journal of Statistics*, 9(2):2828–2854, 2015.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.
Supplemental Document for “A Bayesian Approach to Joint Estimation of Multiple Graphical Models”

S 1 The Structures of $\Upsilon$ and $a$

We first note that,

$$\text{tr} \left[ \left( \sum_{r \in \vartheta_k} \Psi^r \right)^2 \right] = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \left[ \left( \sum_{r \in \vartheta_k} \Psi^r \right) y_{ik}^k \right]' \left( \sum_{r \in \vartheta_k} \Psi^r \right) y_{ik}^k$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n} \left[ \sum_{r \in \vartheta_k} \Psi^r y_{ik}^k \right]^2 = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{r \in \vartheta_k} \left[ \Psi^r y_{ik}^k \right]^2$$

$$+ \frac{2}{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{r \in \vartheta_k} \sum_{s \in \vartheta_k, r \neq s} \left[ \Psi^r y_{ik}^k \right] \left[ \Psi^s y_{ik}^k \right] ,$$

(S 1.1)

where $\{y_{ik}^k\}_{i=1}^{n_k}$ denote $p$-dimensional observations for group $k$. Next, for any $1 \leq j \leq p$, $1 \leq k \leq K$, and $r \in \vartheta_k$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{r \in \vartheta_k} \left[ \Psi^r y_{ik}^k \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{r \in \vartheta_k} \left( \sum_{l=1}^{p} \psi_{jl} y_{il}^k \right)^2$$

$$= \sum_{r \in \vartheta_k} \sum_{l=1}^{p} (\psi_{jl}^k)^2 s_{ll}^k + 2 \sum_{r \in \vartheta_k} \sum_{l=1}^{p} \sum_{m=1}^{p} (\psi_{jl}^k \psi_{jm}^k) s_{lm}^k .$$

(S 1.2)

Similarly, for any $1 \leq j \leq p$, $1 \leq k \leq K$, and $(r \neq s) \in \vartheta_k$,

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{r \in \vartheta_k, s \in \vartheta_k, r \neq s} \left[ \Psi^r y_{ik}^k \right] \left[ \Psi^s y_{ik}^k \right] = \sum_{r \in \vartheta_k} \sum_{l=1}^{p} \left( \psi_{jl}^r \psi_{jl}^s \right) s_{ll}^k$$

$$+ 2 \sum_{r \in \vartheta_k} \sum_{s \in \vartheta_k, r \neq s} \sum_{l=1}^{p} \sum_{m=1}^{p} (\psi_{jl}^r \psi_{jm}^s) s_{lm}^k .$$

(S 1.3)

Thus, by combining (S 1.1), (S 1.2), and (S 1.3), we have that
\[
\text{tr}
\left[
\left(\sum_{r \in \mathcal{E}_k} \Psi^r\right)^2 S^k
\right]
= 
\left(\Theta' \Delta'\right)
\left(\begin{array}{cc}
\Upsilon & A' \\
A' & D
\end{array}\right)
\left(\begin{array}{c}
\Theta \\
\Delta
\end{array}\right),
\] (S 1.4)

The matrix \( \Upsilon \) is as follows

\[
\Upsilon =
\begin{pmatrix}
B_1 + B_2 + B_3 & B_2 + B_3 & B_1 + B_3 & B_1 + B_2 & B_3 & B_2 & B_1 \\
B_2 + B_3 & B_2 + B_3 & B_3 & B_2 & B_3 & B_2 & 0 \\
B_1 + B_3 & B_3 & B_1 + B_3 & B_1 & B_3 & 0 & B_1 \\
B_1 + B_2 & B_2 & B_1 & B_1 + B_2 & 0 & B_2 & B_1 \\
B_3 & B_3 & B_3 & 0 & B_3 & 0 & 0 \\
B_2 & B_2 & 0 & B_2 & 0 & B_2 & 0 \\
B_1 & 0 & B_1 & B_1 & 0 & 0 & B_1 
\end{pmatrix},
\] (S 1.5)

where, \( B_k \)s are \( \frac{p(p-1)}{2} \times \frac{p(p-1)}{2} \) symmetric matrices. To understand the structure of the matrices \( B_k \)s, we index it’s rows and columns as \((12, 13, ..., p - 1p)\). Then,

\[
B_{k(ab,cd)} =
\begin{cases}
 s_{aa}^k + s_{bb}^k & \text{if } a = b \& c = d, \\
 s_{ac}^k & \text{if } b = d \& a \neq c, \quad \text{for } 1 \leq a < b \leq p, \text{ and } 1 \leq c < d \leq p. \\
 s_{bd}^k & \text{if } a = c \& b \neq d, \\
 0 & \text{if } a \neq b \& c \neq d,
\end{cases}
\] (S 1.6)

For further illustration, when \( p = 5 \), \( B_k \) is as follows,

\[
\begin{pmatrix}
s_{11}^k + s_{22}^k & s_{23}^k & s_{24}^k & s_{25}^k & s_{13}^k & s_{14}^k & s_{15}^k & 0 & 0 & 0 \\
 s_{34}^k & s_{11}^k + s_{33}^k & s_{34}^k & s_{35}^k & s_{12}^k & 0 & 0 & s_{14}^k & s_{15}^k & 0 \\
 s_{23}^k & s_{34}^k & s_{11}^k + s_{44}^k & s_{35}^k & 0 & s_{12}^k & 0 & s_{13}^k & 0 & s_{25}^k \\
 s_{25}^k & s_{35}^k & s_{45}^k & s_{11}^k + s_{55}^k & 0 & 0 & s_{12}^k & 0 & s_{13}^k & s_{24}^k \\
 s_{13}^k & s_{12}^k & 0 & 0 & s_{12}^k & s_{13}^k & s_{34}^k & s_{35}^k & s_{24}^k & 0 \\
 s_{14}^k & 0 & s_{12}^k & 0 & s_{34}^k & s_{22}^k + s_{44}^k & s_{45}^k & s_{23}^k & 0 & s_{25}^k \\
 s_{15}^k & 0 & 0 & s_{12}^k & s_{15}^k & s_{45}^k & s_{22}^k + s_{55}^k & 0 & s_{23}^k & s_{24}^k \\
 0 & s_{14}^k & s_{13}^k & 0 & s_{24}^k & s_{23}^k & 0 & s_{33}^k + s_{44}^k & s_{45}^k & s_{55}^k \\
 0 & s_{15}^k & 0 & s_{13}^k & s_{25}^k & 0 & s_{23}^k & s_{45}^k & s_{33}^k + s_{55}^k & s_{54}^k \\
 0 & 0 & s_{15}^k & s_{14}^k & 0 & s_{25}^k & s_{24}^k & s_{34}^k & s_{45}^k & s_{54}^k + s_{55}^k 
\end{pmatrix}
\]
Also, the vector $\mathbf{a}$ for the special case of $K = 3$, is given as

$$
\mathbf{a} = \begin{pmatrix}
\mathbf{a}^1 + \mathbf{a}^2 + \mathbf{a}^3 \\
\mathbf{a}^2 + \mathbf{a}^3 \\
\mathbf{a}^1 + \mathbf{a}^3 \\
\mathbf{a}^1 + \mathbf{a}^2 \\
\mathbf{a}^3 \\
\mathbf{a}^2 \\
\mathbf{a}^1
\end{pmatrix},
$$

(S 1.7)

with $\mathbf{a}_k$ being the following vector,

$$
\mathbf{a}^k = (s_{12}^k(\psi_{11}^k + \psi_{22}^k), \ldots, s_{1p}^k(\psi_{11}^k + \psi_{pp}^k), \ldots, s_{p-1p}^k(\psi_{p-1p-1}^k + \psi_{pp}^k))', \quad k = 1, 2, 3.
$$

S 2 An Algorithm for Generating Samples from Posterior Conditional Density of $\psi_{ii}^k$

Using the fact that it has a unique mode at

$$
\frac{-b_i^k + \sqrt{(b_i^k)^2 + 4n^2\sigma_{ii}^k}}{2n\sigma_{ii}^k},
$$

one can use a discretization technique, as described in Algorithm 2, to generate samples from it.

However, we have observed in the numerical work undertaken that the density (4.8) exhibits a high pick at its mode. As a result, one can simply approximate it using a degenerate density with a point mass at $\frac{-b_i^k + \sqrt{(b_i^k)^2 + 4n^2\sigma_{ii}^k}}{2n\sigma_{ii}^k}$. This approximation allows faster implementation of the algorithm without much sacrificing on its accuracy.

S 3 Proofs of Theorems 1 and 2

By Assumption 3 and Hanson-Wright inequality from Rudelson et al. [2013], there exists a $c > 0$, independent of $n$ and $K$, such that

$$
P \left\{ \max_{i,j,k} \| s_{ij}^k - \sigma_{ij}^k \| < c \sqrt{\frac{\log p}{n}} \right\} \geq 1 - \frac{1}{p^2},
$$
Algorithm 2 Generating samples from density in (4.8)

1: procedure Generating samples from density in (4.8) \(\triangleright\) Input \(n, \sigma_{ii}^k, b_i^k\)
2: \(\text{mode} \leftarrow \frac{-b_i^k + \sqrt{(b_i^k)^2 + 4n^2\sigma_{ii}^k}}{2n\sigma_{ii}^k}\)
3: \(S \leftarrow \text{seq}(0, 6 \times \text{mode}, 0.001)\) \(\triangleright\) sequence from 0 to 6 \(\times\) mode with increments 0.001
4: for \(t \in S\) do
5: \(p_t \leftarrow \exp\left\{n \log t - \frac{n}{2} s_{ii}^k t^2 - b_i^k t\right\}\)
6: end for
7: Set sum = \(\sum_{t \in S} p_t\)
8: if sum \(\leftarrow \infty\) then
9: for \(t \in S\) do
10: \(p_t \leftarrow \frac{\exp\left\{n \log t - \frac{n}{2} s_{ii}^k t^2 - b_i^k t\right\}}{\exp\left\{n \log \text{mode} - \frac{n}{2} s_{ii}^k \text{mode}^2 - b_i^k \text{mode}\right\}}\)
11: end for
12: Set sum = \(\sum_{t \in S} p_t\)
13: end if
14: \(x \leftarrow \text{sample}(1, S, \text{probs} \propto \{p_t : t \in S\})\)
15: return \(x\) \(\triangleright\) Output \(x\)
16: end procedure

and,
\[
P\left\{\max_{i,j,k} \left\| \Omega_{i,j,k}^k - \sigma_{ii}^k \right\| < c \sqrt{\log p \over n} \right\} \geq 1 - \frac{1}{p^2}.
\]

Define the events \(C_{1,n}, C_{2,n}\) as
\[
C_{1,n} = \left\{ \max_{i,j,k} \left\| \sigma_{ii}^k - \sigma_{ij}^k \right\| < c \sqrt{\log p \over n} \right\}, \quad (S \, 3.1)
\]
\[
C_{2,n} = \left\{ \max_{i,j,k} \left\| \Omega_{i,j,k}^k - \sigma_{ij}^k \right\| < c \sqrt{\log p \over n} \right\}, \quad (S \, 3.2)
\]

for the next series of lemmas, we restrict ourself to the event \(C_{1,n} \cap C_{2,n}\).

The next two lemmas prove important properties of the matrix \(Y\) and its components \(B_k, k = 1, \ldots, K\).

**Lemma S 1.** For any \(k = 1, \ldots, K\), the following holds
\[
eig_{\text{min}}(S^k) \leq \eig_{\text{min}}(B^k) \leq \eig_{\text{max}}(B^k) \leq \eig_{\text{max}}(S^k). \quad (S \, 3.3)
\]

**Proof.** Let \(y = \text{y} (\Omega^k)\) be a vectorized version of \(\Omega^k\) obtained by shifting the corresponding diagonal entry at the bottom of each column of \(\Omega^k\) and then stacking the columns on top of
each other. Let $P^i$ be the $p \times p$ permutation matrix such that $P^iz = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_p, z_i)$ for every $z \in \mathbb{R}^p$. It follows by the definition of $y$ that

$$y = y\left(\Omega^k\right) = \left((P^1\Omega^k_1)' , (P^2\Omega^k_2)' , \ldots, (P^p\Omega^k_p)'ight)' .$$

Let $x \in \mathbb{R}^{\frac{p(p+1)}{2}}$ be the symmetric version of $y$ obtained by removing all $\omega_{k,ij}$ with $i > j$. More precisely,

$$x = (\omega^k_{11}, \omega^k_{12}, \omega^k_{22}, \ldots, \omega^k_{p-1p}, \ldots, \omega^k_{pp})' .$$

Let $\tilde{P}$ be the $p^2 \times \frac{p(p+1)}{2}$ matrix such that every entry of $\tilde{P}$ is either zero or one, exactly one entry in each row of $\tilde{P}$ is equal to 1, and $y = \tilde{P}x$.

Now, define $\omega^k = (\omega^k_{12}, \omega^k_{13}, \ldots, \omega^k_{p-1p})'$ and $\delta^k = (\omega^k_{11}, \omega^k_{22}, \ldots, \omega^k_{pp})'$ and let $Q$ be the $\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ permutation matrix for which

$$x = Q\left(\begin{array}{c} \omega^k \\ \delta^k \end{array}\right) .$$

Let $\tilde{\Sigma}^k$ be a $p^2 \times p^2$ block diagonal matrix with $p$ diagonal blocks, the $i$th block is equal to $\tilde{\Sigma}^k_i := P^iS^kP^i'$. It follows that

$$\text{tr}\left[(\Omega^k)^2 S^k\right] = \sum_{i=1}^p \Omega^k_iS^k\Omega^k_i = \sum_{i=1}^p \Omega^k_iP^i'P^iS^kP^iP^i\Omega^k_i = \sum_{i=1}^p \Omega^k_iP^i'\left(P^iS^kP^i'\right)P^i\Omega^k_i = y'\tilde{\Sigma}^k y = x'\tilde{P}'\tilde{\Sigma}^k\tilde{P}x = \left(\omega^k', \delta^k\right) Q'\tilde{P}'\tilde{\Sigma}^k\tilde{P}Q\left(\begin{array}{c} \omega^k \\ \delta^k \end{array}\right) .$$

There also exist appropriate matrices $A^k$ and $D^k$ such that

$$\text{tr}\left[(\Omega^k)^2 S^k\right] = \left(\omega^k', \delta^k\right) \left(\begin{array}{cc} B^k & A^k \\ A^k & D^k \end{array}\right) \left(\begin{array}{c} \omega^k \\ \delta^k \end{array}\right) ,$$

therefore, we must have

$$Q'\tilde{P}'\tilde{\Sigma}^k\tilde{P}Q = \left(\begin{array}{cc} B^k & A^k \\ A^k & D^k \end{array}\right) .$$

Now, since $\tilde{P}Q$ is orthogonal, we conclude that the eigenvalues of $\tilde{\Sigma}^k$ and $\left(\begin{array}{cc} B^k & A^k \\ A^k & D^k \end{array}\right)$ are the same. Moreover, the diagonal blocks of $\tilde{\Sigma}^k$ all have the same eigenvalues as $S^k$, and $B^k$
can be regarded as a principal sub matrix of $\tilde{\Sigma}^k$, hence, we have that
\[
\text{eig}_{\text{min}}(S^k) = \text{eig}_{\text{min}}(\tilde{\Sigma}^k) \leq \text{eig}_{\text{min}}(B^k) \leq \text{eig}_{\text{max}}(B^k) \leq \text{eig}_{\text{max}}(\tilde{\Sigma}^k) = \text{eig}_{\text{max}}(S^k).
\]
\[\square\]

**Lemma S 2.** Let $\ell \in \mathcal{L}$ be any sparsity pattern/model with $d_\ell < \frac{\varepsilon_0}{4\ell} \sqrt{\frac{n}{\log p}}$, then the sub matrix $Y_{\ell\ell}$ of $Y$, obtained by taking out all the rows and columns corresponding to the zero coordinates in $\Theta \in \mathcal{M}_\ell$, is positive definite. Specifically,
\[
\frac{3K\varepsilon_0}{4} \leq \text{eig}_{\text{min}}(Y_{\ell\ell}) \leq \text{eig}_{\text{max}}(Y_{\ell\ell}) \leq \frac{3K}{2\varepsilon_0}, \quad \forall \ell \in \mathcal{L}. \quad (\text{S 3.4})
\]

**Proof.** For ease of exposition, we show this result holds for the case of $K = 3$. The proof for a general case will follow exactly from the same argument. Let $x$ be a $d_\ell \times 1$ vector in $\mathbb{R}^{d_\ell}$ and partition $x$ as
\[
x = \left( x'_\ell, x'_{\ell,123}, x'_{\ell,13}, x'_{\ell,12}, x'_{\ell,13}, x'_{\ell,2}, x'_{\ell,1} \right),
\]
then, by making similar partitions on each block of $Y_{\ell\ell}$ (see $Y$ in (S 1.5)), we have that
\[
x'Y_{\ell\ell}x = \left( x'_\ell, x'_{\ell,123}, x'_{\ell,13}, x'_{\ell,12}, x'_{\ell,13}, x'_{\ell,2}, x'_{\ell,1} \right)' \begin{pmatrix} B_1^* & B_2^* & B_3^* \end{pmatrix} \begin{pmatrix} x'_\ell, x'_{\ell,123}, x'_{\ell,13}, x'_{\ell,12}, x'_{\ell,13}, x'_{\ell,2}, x'_{\ell,1} \end{pmatrix},
\]
where,
\[
B_j^* = \begin{pmatrix}
B_{1}^{j,123} & B_{1}^{j,123} & B_{1}^{j,123} & B_{1}^{j,123} \\
B_{1}^{j,13} & B_{1}^{j,13} & B_{1}^{j,13} & B_{1}^{j,13} \\
B_{1}^{j,12} & B_{1}^{j,12} & B_{1}^{j,12} & B_{1}^{j,12} \\
B_{1}^{j,1} & B_{1}^{j,1} & B_{1}^{j,1} & B_{1}^{j,1}
\end{pmatrix}.
\]
Let $B_{j}^{k,0}$ denote the population version of $B_{j}^{K}$. Since, we are restricted to $C_{1,n} \cap C_{2,n}$, $\|B_{j}^{K} - B_{j}^{0}\| \leq cd_\ell \sqrt{\frac{\log p}{n}}$, hence
\[
eig_{\min}(\nabla_{\ell\ell}) \geq \sum_{k=1}^{K} \eig_{\min}(B^k) = \sum_{k=1}^{K} \inf_{|x|=1} x'B^k x
\]
\[
\geq \sum_{k=1}^{K} \left[ \inf_{|x|=1} x'B^k,0 x - \inf_{|x|=1} x' \left( B^k - B^k,0 \right) x \right]
\]
\[
\geq \sum_{k=1}^{K} \left[ \inf_{|x|=1} x'B^k,0 x \right] - \sum_{k=1}^{K} \|B^k - B^k,0\|_2
\]
\[
\geq \sum_{k=1}^{K} \left[ \inf_{|x|=1} x'B^k,0 x \right] - Kd\ell c \sqrt{\log p \over n}
\]

hence, by Lemma S 2,

\[
eig_{\min}(\nabla_{\ell\ell}) \geq K\varepsilon_0 - Kcd\ell \sqrt{\log p \over n}
\]
\[
\geq K \left( \varepsilon_0 - c\tau_n \sqrt{\log p \over n} \right) = {3K\varepsilon_0 \over 4}.
\]

Similarly one can show that

\[
eig_{\max}(\nabla_{\ell\ell}) \leq {3K \over 2\varepsilon_0}.
\]

By Lemma S 2, the value of the threshold \(\tau_n\) which we used in building our hierarchical prior in (3.3) is given as \(\tau_n = {\varepsilon_0 \over 4c} \sqrt{\log p \over n}\). Hence by Assumption 2, we can write \(d_t \leq \tau_n\), for any sufficiently large \(n\).

**Lemma S 3.** Let, \(\nabla\), and \(a\) be according to (S 1.5), (S 1.7), and let \(\Theta^0\) be the true value of \(\Theta\) in (3.1). Then for large enough \(n\), there exists a constant \(c_0\) such that

\[
\|\nabla\Theta^0 + \hat{a}\|_{\max} \leq c_0 \sqrt{\log p \over n}.
\]

**Proof.** Note that by the triangular inequality,

\[
\|\nabla\Theta^0 + \hat{a}\|_{\max} \leq \|\nabla\Theta^0 + a\|_{\max} + \|\hat{a} - a\|_{\max},
\]

where, \(\hat{a}\) is the estimate of \(a\) provided by Assumption 1.
Now, in view of (S 1.5), (S 1.7), and (3.1), one can easily check that
\[
\Upsilon \Theta^0 + a = \begin{pmatrix}
B_1 \omega^{1,0} + a^1 + B_2 \omega^{2,0} + a^2 + B_3 \omega^{3,0} + a^3 \\
B_2 \omega^{2,0} + a^2 + B_3 \omega^{3,0} + a^3 \\
B_1 \omega^{1,0} + a^1 + B_3 \omega^{3,0} + a^3 \\
B_1 \omega^{1,0} + a^1 + B_2 \omega^{2,0} + a^2 \\
B_3 \omega^{3,0} + a^3 \\
B_2 \omega^{2,0} + a^2 \\
B_1 \omega^{1,0} + a^1
\end{pmatrix},
\]
where \( \omega^{1,0} = (\psi^{1,0} + \psi^{12,0} + \psi^{13,0} + \psi^{123,0}) \), \( \omega^{2,0} = (\psi^{2,0} + \psi^{12,0} + \psi^{23,0} + \psi^{123,0}) \), and \( \omega^{3,0} = (\psi^{3,0} + \psi^{13,0} + \psi^{23,0} + \psi^{123,0}) \). Furthermore,
\[
B_k \omega^{k,0} + a^k = \begin{pmatrix}
\Omega_{k,1}^{k,0} S_2 + \Omega_{k,2}^{k,0} S_1 \\
\Omega_{k,1}^{k,0} S_3 + \Omega_{k,3}^{k,0} S_1 \\
\vdots \\
\Omega_{k,p-1}^{k,0} S_p + \Omega_{k,p}^{k,0} S_{p-1}
\end{pmatrix}, \quad k = 1, 2, 3.
\]
Now, we rewrite \( \Upsilon \Theta^0 + a \) as,
\[
\Upsilon \Theta^0 + a = \begin{pmatrix}
I_{p(p-1)/2} & I_{p(p-1)/2} & I_{p(p-1)/2} \\
I_{p(p-1)/2} & I_{p(p-1)/2} & 0 \\
I_{p(p-1)/2} & 0 & I_{p(p-1)/2} \\
0 & I_{p(p-1)/2} & I_{p(p-1)/2} \\
0 & 0 & I_{p(p-1)/2} \\
0 & 0 & 0 \\
0 & 0 & I_{p(p-1)/2}
\end{pmatrix}
\begin{pmatrix}
B_3 \omega^{3,0} + a^3 \\
B_2 \omega^{2,0} + a^2 \\
B_1 \omega^{1,0} + a^1
\end{pmatrix}.
\]
The norm of the matrix in the right hand side of the above equation is equal to \( \sqrt{K(2^K - 1)} \), hence, by restricting to the event \( C_{1,n} \cap C_{2,n} \), we have that
\[ \| \mathbf{Y} \mathbf{\Theta}^0 + \hat{\mathbf{a}} \|_{\max} \leq \sqrt{\sum_{k=1}^{K} \| \mathbf{B}_k \mathbf{\omega}^{k,0} + \mathbf{a}_k \|_{\max}^2} \]

\[ \leq \sqrt{K} \max_{1 \leq k \leq K} \left( \Omega_{ij}^{k,0'} \mathbf{S}_j^k \right)^2 \]

\[ \leq \sqrt{K} \max_{1 \leq i < j \leq p} \left| \Omega_{ij}^{k,0'} (\mathbf{S}_j^k - \Sigma_j^k) \right| \]

\[ \leq 2c \sqrt{K} \frac{\sqrt{\log p}}{n}. \]  \hspace{1cm} (S 3.7)

Moreover, by (S 1.7), and Assumption 1, it is easy to see that

\[ \| \hat{\mathbf{a}} - \mathbf{a} \|_{\max} \leq 2KC \| \mathbf{Y} \|_{\max} \sqrt{\frac{\log p}{n}} \]

also, \( \| \mathbf{Y} \|_{\max} \leq \max_{\ell \in \mathcal{L}} \| \mathbf{Y}_{\ell \ell} \|_{\max} \) hence, by applying Lemma S 2, we have that

\[ \| \hat{\mathbf{a}} - \mathbf{a} \|_{\max} \leq \frac{3CK^2}{\varepsilon_0} \sqrt{\frac{\log p}{n}}. \] \hspace{1cm} (S 3.8)

Therefore, by combining (S 3.6), (S 3.7), and (S 3.8),

\[ \| \mathbf{Y} \mathbf{\Theta}^0 + \hat{\mathbf{a}} \|_{\max} \leq \left( 2c \sqrt{K} + \frac{3CK^2}{\varepsilon_0} \right) \sqrt{\frac{\log p}{n}}, \]

thus, the conclusion follows by letting \( c_0 = 2c \sqrt{K} + \frac{3CK^2}{\varepsilon_0} \). \( \square \)

For simplicity in writing, we denote the ratio of the posterior probabilities of any sparsity pattern/model \( \ell \) and the true sparsity pattern/model \( t \), by \( PR (\ell, t) \), i.e.

\[ PR (\ell, t) = \frac{P \{ \ell | \Delta, \mathbf{y} \}}{P \{ t | \Delta, \mathbf{y} \}}, \quad \text{for any sparsity pattern } \ell \neq t. \] \hspace{1cm} (S 3.9)

**Lemma S 4.** Under Assumption 1, the ratio of the posterior probabilities of any sparsity pattern/model \( \ell \) and the true sparsity pattern/model \( t \) satisfies:
\[ PR(\ell, t) = \frac{P(\ell | \hat{\Delta}, Y)}{P(t | \hat{\Delta}, Y)} = \frac{\left[q_1^{d\ell}(1 - q_1)^{(\ell)} - d\ell I_{\{\ell \leq \tau\}} + q_1^{d\ell}(1 - q_1)^{(\ell)} - d\ell I_{\{\ell \leq \tau\}}\right]}{\left[q_1^{d_{\hat{\ell}}}(1 - q_1)^{(\ell)} - d\ell I_{\{\ell \leq \tau\}} + q_1^{d\ell}(1 - q_1)^{(\ell)} - d\ell I_{\{\ell \leq \tau\}}\right]} \] (S 3.10)

\[ \times \frac{|\Lambda_{\ell\ell}|^{\frac{1}{2}}}{|\Lambda_{\ell\ell}|^{\frac{1}{2}}} |(nY + \Lambda)_{\ell\ell}|^{\frac{1}{2}} \exp \left\{ \frac{n^2}{2} \hat{a}_\ell (nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell \right\} \]

**Proof.** We note that

\[ P(\ell | \hat{\Delta}, Y) = P(\Theta \in M_{\ell} | \hat{\Delta}, Y) = \int_{M_{\ell}} \pi(\Theta | \hat{\Delta}, Y) d\Theta, \]

hence, in view of (4.4),

\[ P(\ell | \hat{\Delta}, Y) = C_0 \left[q_1^{d\ell}(1 - q_1)^{(\ell)} - d\ell I_{\{\ell \leq \tau\}} + q_1^{d\ell}(1 - q_1)^{(\ell)} - d\ell I_{\{\ell \leq \tau\}}\right] \]

\[ \times \frac{|\Lambda_{\ell\ell}|^{\frac{1}{2}}}{|\Lambda_{\ell\ell}|^{\frac{1}{2}}} |(nY + \Lambda)_{\ell\ell}|^{\frac{1}{2}} \exp \left\{ \frac{n^2}{2} \hat{a}_\ell (nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell \right\}, \]

where the last equality is achieved using the properties of the multivariate normal distribution. \( \square \)

In the next series of lemmas, we will show that for any sparsity pattern \( \ell \in \mathcal{L} \), the posterior probability ratio \( PR(\ell, t) \) is approaching zero, as \( n \) goes to \( \infty \). Specifically, we consider four cases of underfitted (\( \ell \subset t \)), overfitted (\( t \subset \ell \) with \( d_\ell < \tau_n \)), unrealistically overfitted (\( t \subset \ell \) with \( d_\ell > \tau_n \)), and non-inclusive (\( t \not\subset \ell \) and \( \ell \not\subset t \)).

**Lemma S 5.** Suppose \( \ell \subset t \) then, under Assumptions 1-6,

\[ PR(\ell, t) \to 0, \quad \text{as} \quad n \to \infty. \] (S 3.11)

**Proof.** By Assumption 2, \( d_t \leq \tau \), hence \( d_\ell \leq d_t < \tau \). Now,

\[ PR(\ell, t) = \frac{\|\Lambda_{\ell\ell}\|^{\frac{1}{2}}}{\|\Lambda_{\ell\ell}\|^{\frac{1}{2}}} \left( \frac{q_1}{1 - q_1} \right)^{d_{\hat{\ell}} - d_t} |(nY + \Lambda)_{\ell\ell}|^{\frac{1}{2}} \exp \left\{ \frac{n^2}{2} \hat{a}_\ell (nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell \right\} \]

\[ = \frac{\|\Lambda_{\ell\ell}\|^{\frac{1}{2}}}{\|\Lambda_{\ell\ell}\|^{\frac{1}{2}}} \left( \frac{q_1}{1 - q_1} \right)^{d_{\hat{\ell}} - d_t} |(nY + \Lambda)_{\ell\ell}|^{\frac{1}{2}} \exp \left\{ \frac{n^2}{2} \hat{a}_\ell (nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell \right\} \]

\[ \exp \left\{ -\frac{n^2}{2} [\hat{a}_\ell - n^{-1}(nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell]' (nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell - n^{-1} \hat{a}_\ell (nY + \Lambda)^{-1}_{\ell\ell} \hat{a}_\ell \right\}, \]

10
that is,

\[
PR(\ell, t) \leq \frac{\|\Lambda_{t\ell}\|^2}{\|\Lambda_{tt}\|^2} \left( \frac{q_1}{1-q_1} \right)^{d_t-d_{\ell}} \frac{\|nY + \Lambda\|_{tt}}{\|nY + \Lambda\|_{tt}}^{\frac{1}{2}} \exp \left\{ -\frac{n^2\|\hat{a}_{t\ell} - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} \hat{a}_{t\ell}\|^2}{2\lambda_{\text{max}}(nY + \Lambda)_{tt}} \right\},
\]

Now, by the triangular inequality,

\[
\|\hat{a}_{t\ell} - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} \hat{a}_{t\ell}\| \\
\geq \|a_{t\ell} - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} a_{t\ell}\| \\
- \|(-\hat{a}_{t\ell} - a_{t\ell}) - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} (\hat{a}_{t\ell} - a_{t\ell})\| \\
\geq \|(-Y\Theta^0 + a)_{t\ell} - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} (Y\Theta^0 + a)_{t\ell}\| \\
- \|(-\hat{a}_{t\ell} - a_{t\ell}) - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} (\hat{a}_{t\ell} - a_{t\ell})\|.
\]

(S 3.12)

Now, by appropriately partitioning \(Y\), we can write \((Y\Theta^0)_{t\ell} = Y_{t\ell}\Theta^0_{t\ell} + Y_{t\ell}e^{\Theta^0}_{t\ell}\) and \((Y\Theta^0)_{t} = Y_{tt}\Theta^0_{tt} + Y_{t\ell}e^{\Theta^0}_{t\ell}\). Hence, for large enough \(n\),

\[
\|(-Y\Theta^0 + a)_{t\ell} - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} (Y\Theta^0 + a)_{t\ell}\| \\
\geq \|(-Y\Theta^0 + a)_{t\ell} - nY_{t\ell} (nY + \Lambda)^{-1}_{t\ell} (Y\Theta^0 + a)_{t\ell}\| \\
\geq \|\frac{1}{n} (nY + \Lambda)_{t\ell} \Theta^0_{t\ell} - \frac{\lambda_{\text{min}}(Y_{t\ell})}{\lambda_{\text{max}}(nY + \Lambda)_{tt}} \Theta^0_{t\ell}\| \\
\geq \|\frac{1}{n} (nY + \Lambda)_{t\ell} \Theta^0_{t\ell} - \frac{2\lambda_{\text{min}}^2(\Lambda_{t\ell} \Theta^0_{t\ell})}{n\lambda_{\text{max}}^2(\Lambda_{tt}(nY + \Lambda))_{tt}}\| \\
\geq \frac{1}{2} \|\frac{1}{n} (nY + \Lambda)_{t\ell} \Theta^0_{t\ell}\| \\
\geq \frac{1}{2} \frac{n\lambda_{\text{min}}(Y_{t\ell})}{\lambda_{\text{max}}(nY + \Lambda)_{tt}} s_n \sqrt{(d_t - d_{\ell})} \\
\geq \frac{1}{2} \frac{n\lambda_{\text{max}}(Y)_{tt}}{\lambda_{\text{max}}(nY + \Lambda)_{tt}} s_n \sqrt{(d_t - d_{\ell})} \\
\geq \frac{1}{2} \varepsilon_0 s_n \sqrt{(d_t - d_{\ell})}
\]

(S 3.13)
Moving onto the second term in the right hand side of (S 3.12),

\[
\| (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} - n\mathbf{Y}_{\ell\ell} (n\mathbf{Y} + \Lambda)_{\ell\ell}^{-1} (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \|
\leq \| (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \| + \| n\mathbf{Y}_{\ell\ell} (n\mathbf{Y} + \Lambda)_{\ell\ell}^{-1} (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \|
\leq \| (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \| + \frac{\text{neig}_{\max} (\mathbf{Y}_{\ell\ell}) \| (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \|}{\text{eig}_{\min} (n\mathbf{Y} + \Lambda)_{\ell\ell}}
\leq \| (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \| + \frac{2\| (\mathbf{Y}\mathbf{\Theta}^0 + a)_{\ell} \|}{\varepsilon_0^2}
\leq c_0 \sqrt{\frac{\log p}{n}} \left( \sqrt{d_t - d_\ell} + \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right),
\]

where the last equality was achieved by Lemma S 3. Further, regarding the third term in right hand side of (S 3.12) we can write

\[
\| (\hat{a}_{\ell} - a_{\ell}) - n\mathbf{Y}_{\ell\ell} (n\mathbf{Y} + \Lambda)_{\ell\ell}^{-1} (\hat{a}_{\ell} - a_{\ell}) \| \leq \| \hat{a}_{\ell} - a_{\ell} \| + \frac{\text{neig}_{\max} (\mathbf{Y}_{\ell\ell}) \| (\hat{a}_{\ell} - a_{\ell}) \|}{\text{eig}_{\min} (n\mathbf{Y} + \Lambda)_{\ell\ell}}
\leq 3CK^2 \varepsilon_0 \sqrt{\frac{\log p}{n}} \left( \sqrt{d_t - d_\ell} + \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right),
\]

hence, by combining (S 3.12), (S 3.13), (S 3.14), and (S 3.15), for sufficiently large \( n \), we have that

\[
\| \hat{a}_{\ell} - n\mathbf{Y}_{\ell\ell} (n\mathbf{Y} + \Lambda)_{\ell\ell}^{-1} \hat{a}_{\ell} \| \geq \frac{1}{2} \varepsilon_0 s_n \sqrt{(d_t - d_\ell)}
- c_0 \sqrt{\frac{\log p}{n}} \left( \sqrt{d_t - d_\ell} + \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right)
- \frac{3CK^2}{\varepsilon_0} \sqrt{\frac{\log p}{n}} \left( \sqrt{d_t - d_\ell} + \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right)
\geq \frac{1}{2} \varepsilon_0 s_n \sqrt{(d_t - d_\ell)}
- \left( c_0 + \frac{3CK^2}{\varepsilon_0} \right) \sqrt{\frac{\log p}{n}} \left( \sqrt{d_t - d_\ell} + \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right)
\geq \frac{1}{2} \varepsilon_0 s_n - \left( c_0 + \frac{3CK^2}{\varepsilon_0} \right) \sqrt{\frac{\log p}{n}} \left( \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right),
\]

in view of Assumption 5, \( \frac{1}{2} \varepsilon_0 s_n \left( c_0 + \frac{3CK^2}{\varepsilon_0} \right) \sqrt{\frac{\log p}{n}} \left( \frac{2\sqrt{d_\ell}}{\varepsilon_0^2} \right) \rightarrow \infty \), as \( n \rightarrow \infty \), hence, for all large \( n \), we
can write,

$$\|\hat{a}_\ell - n Y_{\ell t} (n Y + \Lambda)_{\ell t}^{-1} \hat{a}_\ell\| \geq \frac{1}{4} \varepsilon_0 s_n$$

Now, once again by Lemma S 3

$$PR(\ell, t) \leq \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{\ell t}\|^{\frac{1}{2}}} \left(2q_1\right)^{d_\ell - d_t} \frac{d_\ell - d_t}{d_\ell} \exp \left\{- \frac{n^2 1}{64} \varepsilon_0^2 s_n^2 \varepsilon_0^{-1} \right\}$$

$$= \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{\ell t}\|^{\frac{1}{2}}} \left(\sqrt{n} \frac{q_1}{q_1} \exp \left\{- \frac{n^2 1}{384K} s_n^2 \right\} \right)^{d_\ell - d_t}$$

by Assumption 5, for all large $n$,

$$PR(\ell, t) \leq \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{\ell t}\|^{\frac{1}{2}}} \left(\sqrt{n} \frac{q_1}{q_1} \exp \left\{- \log n - 2a_1 d_{t\log_p} \right\} \right)^{d_\ell - d_t}$$

$$= \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{\ell t}\|^{\frac{1}{2}}} \left(\frac{p^{-2a_2 d_{t}}}{\sqrt{n q_1}} \right)^{d_\ell - d_t} \rightarrow 0 \text{ as } n \to \infty.$$

**Lemma S 6.** Suppose $\ell \supset t$, and $d_\ell < \tau_n$ then, under Assumptions 1-6,

$$PR(\ell, t) \to 0, \quad \text{as } n \to \infty.$$

**Proof.** In this case,

$$PR(\ell, t) \leq \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{\ell t}\|^{\frac{1}{2}}} \left(\frac{2q_1}{q_1} \exp \left\{- \frac{n^2 1}{64} \varepsilon_0^2 s_n^2 \varepsilon_0^{-1} \right\} \right)^{d_\ell - d_t}$$

Now, we note that

$$(n Y + \Lambda)_{\ell t} \Theta^\ell + n \hat{a}_\ell = n (Y \Theta^\ell + a)_{\ell t} + \Lambda_{\ell t} \Theta^\ell_{\ell t} + n (\hat{a} - a)_{\ell t}$$
each entry of the above vector in absolute value is smaller than
\[ nc \sqrt{\frac{\log p}{n}} + \frac{\|\Lambda\|_{\text{max}}}{\epsilon_0} + \frac{3CK^2}{\epsilon_0} \sqrt{\frac{\log p}{n}} \leq 2nc \sqrt{\frac{\log p}{n}}, \]
hence, by Lemma S 2,
\[ \left( (n\mathbf{Y} + \Lambda)_{tt} \Theta_{\ell}^0 + n\hat{a}_{\ell} \right)' \left( (n\mathbf{Y} + \Lambda)_{\ell\ell} \Theta_{\ell}^0 + n\hat{a}_{\ell} \right) \leq \frac{1}{nK\epsilon_0} \frac{4n^2c^2 \log p}{n} = \frac{4c^2d_{\ell} \log p}{K\epsilon_0}. \]
Hence,
\[ PR(\ell, t) \leq \frac{\|\Lambda_{tt}\|_2}{\|\Lambda_{tt}\|_2} \exp \left( \frac{2c^2}{K\epsilon_0} d_{\ell} \log p \right) = \frac{2^{d_{\ell}} \|\Lambda_{tt}\|_2}{\|\Lambda_{tt}\|_2} \left( \frac{nK\epsilon_0}{2} \right)^{\frac{d_{\ell}}{2}}. \]
(S 3.16)
Note that by Assumption 6, for all large \( n \), \( q_1 = p \frac{\kappa^2 d_{\ell}}{\kappa \epsilon_0} \leq p \frac{\kappa^2}{\kappa \epsilon_0} \). Therefore, if \( d_{\ell} > 2d_t \),
\[ q_1^{d_{\ell}-d_t} p \frac{\kappa^2 d_{\ell}}{\kappa \epsilon_0} \leq q_1^{\frac{d_{\ell}}{2}} \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}}{2}} \leq \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}}{2}} \leq \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}-d_t}{2}} \]
and, in the case of \( d_t < d_{\ell} < 2d_t \),
\[ q_1^{d_{\ell}-d_t} p \frac{\kappa^2 d_{\ell}}{\kappa \epsilon_0} \leq q_1^{\frac{d_{\ell}}{2}} \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}}{2}} \leq \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}}{2}} \leq \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}-d_t}{2}} \]
thus, by (S 3.16), (S 3.16), and (S 3.18) we have that
\[ PR(\ell, t) \leq \frac{2^{d_{\ell}} \|\Lambda_{tt}\|_2}{\|\Lambda_{tt}\|_2} \left( \frac{nK\epsilon_0}{2} \right)^{\frac{d_{\ell}}{2}} \left( p \frac{\kappa^2}{\kappa \epsilon_0} \right)^{\frac{d_{\ell}-d_t}{2}} \rightarrow 0 \quad \text{as} \quad n \rightarrow 0. \]

\[ \square \]

**Lemma S 7.** Suppose \( \ell \supset t \), and \( d_{\ell} > \tau_n \) then, under Assumptions 1-6,
\[ PR(\ell, t) \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty. \]
Proof. When $\ell \supset t$,

$$PR(\ell, t) \leq \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}} q_{2}^{d_{t}} (1 - q_{2})^{(\frac{r}{2}) - d_{t}}}{\|\Lambda_{tt}\|^{\frac{1}{2}} q_{1}^{d_{t}} (1 - q_{1})^{(\frac{r}{2}) - d_{t}}} \times \frac{1}{\|\Lambda_{\ell t}\|^{\frac{1}{2}}} \exp \left\{ \frac{1}{2} \left[ (n \Upsilon + \Lambda)_{\ell t} \Theta_{\ell}^{0} + n \tilde{a}_{\ell} \right]' (n \Upsilon + \Lambda)_{\ell t}^{-1} \left[ (n \Upsilon + \Lambda)_{\ell t} \Theta_{\ell}^{0} + n \tilde{a}_{\ell} \right] \right\},$$

similar to the argument in Lemma S 6, each entry of the vector $(n \Upsilon + \Lambda)_{\ell t} \Theta_{\ell}^{0} + n \tilde{a}_{\ell}$, in absolute value, is smaller than $2nc\sqrt{\log p/n}$. Now, since $\Upsilon$ is non-negative definite (note that in the case of $d_{\ell} > \tau_{n}$, $\Upsilon$ is not necessarily positive definite) we have that $\text{eig}_{\min}(n \Upsilon + \Lambda)_{\ell t} \geq \text{eig}_{\min}(\Lambda) = \|\Lambda\|_{\min}$, hence for large enough $n$

$$PR(\ell, t) \leq \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{tt}\|^{\frac{1}{2}}} \left( \frac{q_{2}}{1 - q_{2}} \right)^{d_{t} - d_{t}} \times \frac{q_{2}^{d_{t}} (1 - q_{2})^{(\frac{r}{2}) - d_{t}}}{q_{1}^{d_{t}} (1 - q_{1})^{(\frac{r}{2}) - d_{t}}} \exp \left\{ \frac{4c^{2}n^{2}d_{\ell} \log p}{n\|\Lambda\|_{\min}} \right\},$$

now, since the function $q^{d_{t}} (1 - q)^{(\frac{r}{2}) - d_{t}}$ is globally maximized at $\hat{q} = \frac{d_{t}}{q_{2}}$ and $q_{2} < q_{1} < \hat{q}$,

$$\frac{q_{2}^{d_{t}} (1 - q_{2})^{(\frac{r}{2}) - d_{t}}}{q_{1}^{d_{t}} (1 - q_{1})^{(\frac{r}{2}) - d_{t}}} \leq 1,$$

hence,

$$PR(\ell, t) \leq \frac{\|\Lambda_{\ell t}\|^{\frac{1}{2}}}{\|\Lambda_{tt}\|^{\frac{1}{2}}} \left(2q_{2}\right)^{d_{t} - d_{t}} \exp \left\{ \frac{4c^{2}nd_{\ell} \log p}{\|\Lambda\|_{\min}} \right\},$$

since $d_{\ell} > \tau$, and by Assumption 2 $d_{t} < \frac{\tau}{2}$, we have that,

$$d_{\ell} - d_{t} \geq \frac{d_{\ell}}{2},$$

15
hence,

\[
PR(\ell, t) \leq \frac{2^{d_\ell} 2^{d_t}}{\|\Lambda_{tt}\|^{1/2}} \left(2q_2\right)^{d_\ell} \exp\left\{\frac{4c^2 nd_\ell \log p}{\|\Lambda\|_{\min}}\right\}
\]

\[
\leq \frac{2^{d_\ell} ||\Lambda_{tt}\|^{1/2}}{\|\Lambda_{tt}\|^{1/2}} \left[q_2 \exp\left\{\frac{8c^2 n \log p}{\|\Lambda\|_{\min}}\right\}\right]^{d_\ell}
\]

\[
\leq \frac{2^{d_\ell} 2^{d_t}}{\|\Lambda_{tt}\|^{1/2}} \left[p^{-\frac{8nc^2}{\|\Lambda\|_{\min}}}\right]^{d_\ell}
\]

\[
\leq \frac{2^{d_\ell} 2^{d_t}}{\|\Lambda_{tt}\|^{1/2}} \left[p^{-\frac{8nc^2}{\|\Lambda\|_{\min}}}\right]^{d_\ell} \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty.
\]

\[
\square
\]

Now, let,

\[
f_n = \max_{\ell \in \mathcal{L}} \left\{2^{d_\ell} \frac{2^{d_\ell} ||\Lambda_{tt}\|^{1/2}}{\|\Lambda_{tt}\|^{1/2}}, 2^{d_t} \frac{2^{d_t} ||\Lambda_{tt}\|^{1/2}}{\|\Lambda_{tt}\|^{1/2}}, 2^{d_\ell} \frac{2^{d_t} ||\Lambda_{tt}\|^{1/2}}{\|\Lambda_{tt}\|^{1/2}}\right\} \times \max \left\{\left(\frac{p}{\sqrt{n}}\right)^{1/2}, \left(p^{-\frac{8nc^2}{\|\Lambda\|_{\min}}}\right)^{1/2}, \left(p^{-\frac{4c^2}{\|\Lambda\|_{\min}}} \frac{1}{nK^0}\right)^{1/2}\right\}.
\] (S 3.19)

**Lemma S 8.** Let \(\ell \in \mathcal{L}\) such that \(\ell \not\subset t\), \(t \not\subset \ell\), and \(\ell \neq t\), then under Assumptions 1-6, for sufficiently large \(n\),

\[
PR(\ell, t) \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty.
\]

**Proof.** Denote the shared part between \(\ell\) and \(t\) by \(h = \ell \cap t\). Then,

\[
PR(\ell, t) = PR(\ell, h) \times PR(h, t),
\]

Since, \(h \subset \ell\) and \(h \subset t\), by Lemma S 5, S 6, and S 7, we have that

\[
PR(\ell, t) \leq f_m^{d_\ell-d_h} f_n^{d_t-d_h} = f_m^{d_\ell+d_t-2d_h} \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty.
\]

\[
\square
\]

**Corollary 1.** Let \(\ell \in \mathcal{L}\) such that \(\ell \not\subset t\), \(t \not\subset \ell\), and \(\ell \neq t\). Denote the total number of
disagreements by \( D(\ell, t) \). Then, under Assumptions 1-6, for sufficiently large \( n \),

\[
PR(\ell, t) \leq f^{D(\ell, t)}_n, \quad \forall \ell \in \mathcal{L}.
\]

**Proof.** The proof is straightforward application of Lemmas S 5, S 6, S 7 and S 8. \( \square \)

**Proof of Theorem 1.**

\[
1 - P \left\{ \Theta \in \mathcal{M}_t | \hat{\Delta}, \mathcal{Y} \right\} = \sum_{\ell \neq t} PR(\ell, t)
\]

\[
= \sum_{\ell \neq t} \sum_{j=1}^{\mathcal{G}} PR(\ell, t) I_{(D(\ell, t)=j)}
\]

\[
\leq \sum_{j=1}^{\mathcal{G}} \left( \binom{p}{2} \right)^j f_n^j
\]

\[
\leq \sum_{j=1}^{\mathcal{G}} \binom{p}{2} f_n^j
\]

\[
\leq \sum_{j=1}^{\mathcal{G}} (p^2 f_n)^j
\]

\[
\leq \frac{p^2 f_n}{1 - p^2 f_n} \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty.
\]

The last two inequalities follow from the fact that \( p^2 f_n < 1 \) and \( p^2 f_n \rightarrow 0 \). Which follows from (S 3.19) and choice of \( \varepsilon_0 = \frac{\delta^2}{K \varepsilon_0} \). \( \square \)

**Proof of Theorem 2.** For simplicity in notation, let \( \varepsilon_n = \sqrt{\frac{d_k \log n}{n}} \). First note that for any constant \( G \),

\[
\mathbb{E}_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G\varepsilon_n | \hat{\Delta}, \mathcal{Y} \right) \right]
\]

\[
= \sum_{\ell \in \mathcal{L}} \mathbb{E}_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G\varepsilon_n | \ell, \hat{\Delta}, \mathcal{Y} \right) P \left( \ell | \hat{\Delta}, \mathcal{Y} \right) \right]
\]

\[
\leq \mathbb{E}_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G\varepsilon_n | t, \hat{\Delta}, \mathcal{Y} \right) \right] + \mathbb{E}_0 \left[ \sum_{\ell \neq t} P \left( \ell | \hat{\Delta}, \mathcal{Y} \right) \right].
\]

By Theorem 1, it is sufficient to prove \( \mathbb{E}_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G\varepsilon_n | t, \hat{\Delta}, \mathcal{Y} \right) \right] \rightarrow 0 \) as \( n \rightarrow \infty \).
First note that
\[
P \left( \| \Theta - \Theta^0 \|_2 > G\epsilon_n |t, \hat{\Delta}, \mathcal{Y} \right) = P \left( \| \Theta_t - \Theta^0_t \|_2 > G\epsilon_n |t, \hat{\Delta}, \mathcal{Y} \right),
\]
now, from (4.4), it is easy to see that
\[
\left( \Theta_t |t, \hat{\Delta}, \mathcal{Y} \right) \sim MVN \left[ \mathcal{M}, (n\mathbf{Y} + \Lambda)^{-1}_{tt} \right],
\]
(S 3.21)
with \( \mathcal{M} = -n(n\mathbf{Y} + \Lambda)_{tt}^{-1} \). Hence,
\[
\mathbb{E}_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G\epsilon_n |t, \hat{\Delta}, \mathcal{Y} \right) \right] \leq \mathbb{E}_0 \left[ P \left( \| \Theta_t - \Theta^0_t \|_2 > \frac{G\epsilon_n}{2} |t, \hat{\Delta}, \mathcal{Y} \right) \right] + \mathbb{E}_0 \left[ P \left( \| \mathcal{M}_t - \Theta^0_t \|_2 > \frac{G\epsilon_n}{2} |t, \hat{\Delta}, \mathcal{Y} \right) \right].
\]
(S 3.22)

Now,
\[
P \left( \| \Theta_t - \mathcal{M}_t \|_2 > \frac{G\epsilon_n}{2} |t, \hat{\Delta}, \mathcal{Y} \right) \leq P \left( \max_{1 \leq l \leq d_t} | (\Theta_t - \mathcal{M}_t)_l | > \frac{G\epsilon_n}{2\sqrt{d_t}} |t, \hat{\Delta}, \mathcal{Y} \right)
\]
\[
\leq \sum_{l=1}^{d_t} P \left( | (\Theta_t - \mathcal{M}_t)_l | > \frac{G\epsilon_n}{2\sqrt{d_t}} |t, \hat{\Delta}, \mathcal{Y} \right),
\]
denoting the \( l \)-th element of \((\Theta_t - \mathcal{M}_t)\) by \((\Theta_t - \mathcal{M}_t)_l\) and in view of (S 3.21),
\[
(\Theta_t - \mathcal{M}_t)_l \sim N(0, \rho_l),
\]
where by Lemma S 2, \( \rho_l \leq \text{eig}_{\max} (n\mathbf{Y} + \Lambda)^{-1}_{tt} = \frac{1}{\text{eig}_{\min} (n\mathbf{Y}_{tt})} \leq \frac{2}{nK\varepsilon_0} \). Therefore,
\[
P \left( \| \Theta_t - \mathcal{M}_t \|_2 > \frac{G\epsilon_n}{2} |t, \hat{\Delta}, \mathcal{Y} \right) \leq \sum_{l=1}^{d_t} P \left( | (\Theta_t - \mathcal{M}_t)_l | > \frac{G\epsilon_n}{2\sqrt{d_t}} |t, \hat{\Delta}, \mathcal{Y} \right)
\]
\[
\leq d_t P \left( | (\Theta_t - \mathcal{M}_t)_1 | > \frac{G\epsilon_n}{2\sqrt{d_t}} |t, \hat{\Delta}, \mathcal{Y} \right)
\]
\[
\leq d_t P \left( | Z | > \frac{\sqrt{nK\varepsilon_0 G\epsilon_n}}{2\sqrt{2d_t}} \right)
\]
\[
\leq d_t e^{-\frac{nK\varepsilon_0 G^2\epsilon_n^2}{8d_t}},
\]
where the last inequality was achieved using the Mills ratio inequality. Moving onto the second term in (S 3.21), and recalling that we are restricted to the even \( C_{1,n} \cap C_{2,n} \), for all
large \( n \) we have that

\[
\| \mathcal{M}_t - \Theta^0_t \|_2 = \| n (n \Upsilon + \Lambda)_{tt}^{-1} a_t + \Theta^0_t \|_2 \\
\leq \frac{1}{\text{eig}_\min (\Upsilon + \frac{1}{n} \Lambda)_{tt}} \| a_t + \left( \Upsilon + \frac{1}{n} \Lambda \right)_{tt} \Theta^0_t \|_2 \\
\leq \frac{4}{3K\varepsilon_0} \| \Upsilon_{tt} \Theta^0_t + a_t + \frac{1}{n} \Lambda_{tt} \Theta^0_t \|_2 \\
\leq \frac{4}{3K\varepsilon_0} \left( \sqrt{d_t} \| \Upsilon_{tt} \Theta^0_t + a_t \|_{\text{max}} + \| \frac{1}{n} \Lambda_{tt} \Theta^0_t \|_2 \right) \\
\leq \frac{4}{3K\varepsilon_0} \left( c_0 \sqrt{d_t \log p} \frac{n}{n} + \sqrt{d_t \log p} \| \Lambda \|_{\text{max}} \right) \\
\leq \frac{4}{3K\varepsilon_0} \left( c_0 \sqrt{d_t \log p} + \sqrt{d_t \log p} \| \Lambda \|_{\text{max}} \right) \\
= \frac{4(c_0 + \| \Lambda \|_{\text{max}})}{3K\varepsilon_0} \sqrt{d_t \log p} \frac{n}{n},
\]

where we used Lemma S 2 to get the second inequality, used Lemma S 3 with \( c_0 = 2c\sqrt{K} + \frac{3CK^2}{\varepsilon_0} \) to get the fourth inequality and the last inequality is a direct application of the assumption. Hence, by taking \( G = \frac{4(c_0 + \| \Lambda \|_{\text{max}})}{3K\varepsilon_0} \), we have that

\[
P \left( \| \mathcal{M}_t - \Theta^0_t \|_2 > \frac{G\varepsilon_n}{2} | t, \hat{\Delta}, \Upsilon \right) = P \left( \| \mathcal{M}_t - \Theta^0_t \|_2 > \frac{G\varepsilon_n}{2} \right) \\
\leq 1 - P \left( C_{1,n} \cap C_{2,n} \right) \to 0.
\]

(S 3.24)

Thus, by combining (S 3.21), (S 3.22), (S 3.23), and (S 3.24), we have that

\[
\mathbb{E}_0 \left[ P \left( \| \Theta - \Theta^0 \|_2 > G\varepsilon_n | \hat{\Delta}, \Upsilon \right) \right] \\
\leq d_t e^{-\frac{nK\varepsilon_0 G^2 \varepsilon_n^2}{8d_t}} + 0 = d_t e^{-\frac{32c_0^2 \log p}{9K\varepsilon_0}} \to 0 \quad \text{as} \quad n \to \infty,
\]

(S 3.25)

which completes the proof. \( \square \)