A Bayesian Subset Specific Approach to Joint Selection of Multiple Graphical Models

Peyman Jalali, Kshitij Khare and George Michailidis

Wells Fargo; and University of Florida

A. Useful Expressions in the Derivation of the Pseudo-Likelihood Function

A.1 The structures of some matrices and vectors in (4.11)

We first note that

$$\begin{aligned}\sum_{k=1}^{K}\text{tr}\left[\left(\sum_{r\in\vartheta_k}\mathbf{\Psi}^r\right)^2\mathbf{S}^k\right] &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\left[\left(\sum_{r\in\vartheta_k}\mathbf{\Psi}^r\right)\mathbf{y}_{i:}^k\right]'\left[\left(\sum_{r\in\vartheta_k}\mathbf{\Psi}^r\right)\mathbf{y}_{i:}^k\right] \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i=1}^{n}\left[\sum_{r\in\vartheta_k}\mathbf{\Psi}_{j:}^r\mathbf{y}_{i:}^k\right]^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{r\in\vartheta_k}\left[\mathbf{\Psi}_{j:}^r\mathbf{y}_{i:}^k\right]^2 \\ &+ \frac{2}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{r\in\vartheta_k}\sum_{\substack{s\in\vartheta_k\\ r\neq s}}\left[\mathbf{\Psi}_{j:}^r\mathbf{y}_{i:}^k\right]\left[\mathbf{\Psi}_{j:}^s\mathbf{y}_{i:}^k\right],\end{aligned} \tag{A.1}$$

where $\left\{\mathbf{y}_{i:}^k\right\}_{i=1}^{n_k}$ denote p-dimensional observations for group k. Next, for any $1 \leq j \leq p$,

$1 \leq k \leq K$, and $r \in \vartheta_k$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\vartheta_k}\left[\mathbf{\Psi}^r_{j:}\mathbf{y}^k_{i:}\right]^2 &= \frac{1}{n}\sum_{r\in\vartheta_k}\sum_{i=1}^{n}\left(\sum_{l=1}^{p}\psi^r_{jl}\mathbf{y}^k_{il}\right)^2 \\ &= \sum_{r\in\vartheta_k}\sum_{l=1}^{p}\left(\psi^r_{jl}\right)^2 s^k_{ll} + 2\sum_{r\in\vartheta_k}\sum_{\substack{l=1\\ l\neq m}}^{p}\sum_{m=1}^{p}\left(\psi^r_{jl}\psi^r_{jm}\right)s^k_{lm}.\end{aligned} \tag{A.2}$$

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\vartheta_k}\sum_{\substack{s\in\vartheta_k\\ r\neq s}}\left[\mathbf{\Psi}^r_{j:}\mathbf{y}^k_{i:}\right]\left[\mathbf{\Psi}^s_{j:}\mathbf{y}^k_{i:}\right] &= \sum_{r\in\vartheta_k}\sum_{l=1}^{p}\left(\psi^r_{jl}\psi^s_{jl}\right)s^k_{ll} \\ &\quad + 2\sum_{\substack{r\in\vartheta_k\\ r\neq s}}\sum_{\substack{s\in\vartheta_k}}\sum_{\substack{l=1\\ l\neq m}}^{p}\sum_{m=1}^{p}\left(\psi^r_{jl}\psi^s_{jm}\right)s^k_{lm}.\end{aligned} \tag{A.3}$$

Thus, by combining (A.1), (A.2), and (A.3), we have that

$$\text{tr}\left[\left(\sum_{r\in\vartheta_k}\mathbf{\Psi}^r\right)^2\mathbf{S}^k\right] = \begin{pmatrix}\mathbf{\Theta}' & \mathbf{\Delta}'\end{pmatrix}\begin{pmatrix}\mathbf{\Upsilon} & \mathbf{A}\\ \mathbf{A}' & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{\Theta}\\ \mathbf{\Delta}\end{pmatrix}, \tag{A.4}$$

The matrix $\boldsymbol{\Upsilon}$ is as follows

$$\boldsymbol{\Upsilon} = \begin{pmatrix} \mathbf{B}^1+\mathbf{B}^2+\mathbf{B}^3 & \mathbf{B}^2+\mathbf{B}^3 & \mathbf{B}^1+\mathbf{B}^3 & \mathbf{B}^1+\mathbf{B}^2 & \mathbf{B}^3 & \mathbf{B}^2 & \mathbf{B}^1 \\ \mathbf{B}^2+\mathbf{B}^3 & \mathbf{B}^2+\mathbf{B}^3 & \mathbf{B}^3 & \mathbf{B}^2 & \mathbf{B}^3 & \mathbf{B}^2 & \mathbf{0} \\ \mathbf{B}^1+\mathbf{B}^3 & \mathbf{B}^3 & \mathbf{B}^1+\mathbf{B}^3 & \mathbf{B}^1 & \mathbf{B}^3 & \mathbf{0} & \mathbf{B}^1 \\ \mathbf{B}^1+\mathbf{B}^2 & \mathbf{B}^2 & \mathbf{B}^1 & \mathbf{B}^1+\mathbf{B}^2 & \mathbf{0} & \mathbf{B}^2 & \mathbf{B}^1 \\ \mathbf{B}^3 & \mathbf{B}^3 & \mathbf{B}^3 & \mathbf{0} & \mathbf{B}^3 & \mathbf{0} & \mathbf{0} \\ \mathbf{B}^2 & \mathbf{B}^2 & \mathbf{0} & \mathbf{B}^2 & \mathbf{0} & \mathbf{B}^2 & \mathbf{0} \\ \mathbf{B}^1 & \mathbf{0} & \mathbf{B}^1 & \mathbf{B}^1 & \mathbf{0} & \mathbf{0} & \mathbf{B}^1 \end{pmatrix}, \tag{A.5}$$

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

$$\mathbf{B}^k_{(ab,cd)} = \begin{cases} s^k_{aa} + s^k_{bb} & \text{if } a = b \;\&\; c = d, \\ s^k_{ac} & \text{if } b = d \;\&\; a \neq c, \\ s^k_{bd} & \text{if } a = c \;\&\; b \neq d, \\ 0 & \text{if } a \neq b \;\&\; c \neq d, \end{cases} \quad \text{for} \quad 1 \leq a < b \leq p, \text{ and } 1 \leq c < d \leq p. \tag{A.6}$$

For further illustration, when $p = 5$, $\mathbf{B}^k$ is as follows,

$$\begin{pmatrix}
s_{11}^k + s_{22}^k & s_{23}^k & s_{24}^k & s_{25}^k & s_{13}^k & s_{14}^k & s_{15}^k & 0 & 0 & 0 \\
s_{23}^k & s_{11}^k + s_{33}^k & s_{34}^k & s_{35}^k & s_{12}^k & 0 & 0 & s_{14}^k & s_{15}^k & 0 \\
s_{24}^k & s_{34}^k & s_{11}^k + s_{44}^k & s_{45}^k & 0 & s_{12}^k & 0 & s_{13}^k & 0 & s_{15}^k \\
s_{25}^k & s_{35}^k & s_{45}^k & s_{11}^k + s_{55}^k & 0 & 0 & s_{12}^k & 0 & s_{13}^k & s_{14}^k \\
s_{13}^k & s_{12}^k & 0 & 0 & s_{22}^k + s_{33}^k & s_{34}^k & s_{35}^k & s_{24}^k & s_{25}^k & 0 \\
s_{14}^k & 0 & s_{12}^k & 0 & s_{34}^k & s_{22}^k + s_{44}^k & s_{45}^k & s_{23}^k & 0 & s_{25}^k \\
s_{15}^k & 0 & 0 & s_{12}^k & s_{35}^k & s_{45}^k & s_{22}^k + s_{55}^k & 0 & s_{23}^k & s_{24}^k \\
0 & s_{14}^k & s_{13}^k & 0 & s_{24}^k & s_{23}^k & 0 & s_{33}^k + s_{44}^k & s_{45}^k & s_{35}^k \\
0 & s_{15}^k & 0 & s_{13}^k & s_{25}^k & 0 & s_{23}^k & s_{45}^k & s_{33}^k + s_{55}^k & s_{34}^k \\
0 & 0 & s_{15}^k & s_{14}^k & 0 & s_{25}^k & s_{24}^k & s_{35}^k & s_{34}^k & s_{44}^k + s_{55}^k
\end{pmatrix},$$

Also, the vector $\boldsymbol{a}$ for the special case of $K = 3$, is given as

$$\boldsymbol{a} = \begin{pmatrix}
\boldsymbol{a}^1 + \boldsymbol{a}^2 + \boldsymbol{a}^3 \\
\boldsymbol{a}^2 + \boldsymbol{a}^3 \\
\boldsymbol{a}^1 + \boldsymbol{a}^3 \\
\boldsymbol{a}^1 + \boldsymbol{a}^2 \\
\boldsymbol{a}^3 \\
\boldsymbol{a}^2 \\
\boldsymbol{a}^1
\end{pmatrix}, \tag{A.7}$$

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

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

B. Posterior Densities and the Gibbs Sampler for BJNS

B.1 Form of joint and conditional posterior densities for $\boldsymbol{\Theta}$ and $\boldsymbol{\Delta}$

Letting $\mathcal{Y} := \left(\{\mathbf{y}_{i:}^1\}_{i=1}^{n_1}, ..., \{\mathbf{y}_{i:}^K\}_{i=1}^{n_K}\right)$ and by applying Bayes' rule, the posterior distribution of $(\boldsymbol{\Theta}, \boldsymbol{\Delta})$ is given by

$$\begin{aligned}\pi\left\{(\boldsymbol{\Theta}, \boldsymbol{\Delta})\,|\mathcal{Y}\right\} \propto \exp\left\{n\mathbf{1}'\log\left(\boldsymbol{\Delta}\right) - \frac{n}{2}\left[\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{\Delta}'\end{pmatrix}\begin{pmatrix}\boldsymbol{\Upsilon} & \mathbf{A}\\ \mathbf{A}' & \mathbf{D}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Theta}\\ \boldsymbol{\Delta}\end{pmatrix}\right]\right\}\\ \times \exp\left(-\frac{\boldsymbol{\Theta}'\boldsymbol{\Lambda}\boldsymbol{\Theta}}{2}\right)\\ \sum_{\boldsymbol{\ell}\in\mathcal{L}}\left\{I_{(\boldsymbol{\Theta}\in\mathcal{M}_\ell)}\left[q_1^{d_{\boldsymbol{\ell}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{\ell}}}I_{\{d_{\boldsymbol{\ell}}\le\tau\}} + q_2^{d_{\boldsymbol{\ell}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{\ell}}}I_{\{d_{\boldsymbol{\ell}}>\tau\}}\right]\right\}\\ \times \exp\left(-\gamma\mathbf{1}'\boldsymbol{\Delta}\right).\end{aligned} \tag{B.8}$$

Moreover, the conditional posterior distribution of $\boldsymbol{\Theta}$ given $\boldsymbol{\Delta}$ is given by

$$\begin{aligned}\pi\left\{\boldsymbol{\Theta}|\boldsymbol{\Delta}, \mathcal{Y}\right\} \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\Theta}'\left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)\boldsymbol{\Theta} + 2n\boldsymbol{\Theta}'\boldsymbol{a}\right]\right\} \times\\ \sum_{\boldsymbol{\ell}\in\mathcal{L}}\left\{I_{(\boldsymbol{\Theta}\in\mathcal{M}_{\boldsymbol{\ell}})}\left[q_1^{d_{\boldsymbol{\ell}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{\ell}}}I_{\{d_{\boldsymbol{\ell}}\le\tau\}} + q_2^{d_{\boldsymbol{\ell}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{\ell}}}I_{\{d_{\boldsymbol{\ell}}>\tau\}}\right]\right\},\end{aligned} \tag{B.9}$$

while that of $\boldsymbol{\Delta}$ given $\boldsymbol{\Theta}$ by

$$\pi\left\{\boldsymbol{\Delta}|\boldsymbol{\Theta}, \mathcal{Y}\right\} \propto \prod_{k=1}^{K}\prod_{i=1}^{p}\left(\psi_{ii}^k\right)^n \exp\left\{-\frac{n}{2}s_{ii}^k\left(\psi_{ii}^k\right)^2 - \left(\gamma + n\sum_{j\neq i}\omega_{ij}^k s_{ij}^k\right)\psi_{ii}^k\right\}, \tag{B.10}$$

where $\omega_{ij}^k = \sum_{r\in\vartheta_k}\psi_{ij}^r$, for $1 \le i < j \le p$ and $1 \le k \le K$.

B.2 Derivation of joint and conditional posterior distribution of θ_{ij}

Define $\boldsymbol{\Upsilon}_{(ij)(ij)}$ to be the diagonal block of $\boldsymbol{\Upsilon}$ obtained by restricting to the rows and columns with indices corresponding to the elements of $\boldsymbol{\theta}_{ij}$ inside $\boldsymbol{\Theta}$. Similarly, let $\boldsymbol{\Upsilon}_{(ij)(-(ij))}$ be the sub-matrix of $\boldsymbol{\Upsilon}$ obtained by restricting to the rows with indices corresponding to the elements of $\boldsymbol{\theta}_{ij}$ inside $\boldsymbol{\Theta}$, and columns with indices corresponding to the elements of $\boldsymbol{\Theta}_{-(ij)}$ inside $\boldsymbol{\Theta}$. For example, let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_{12} = (\psi_{12}^1, \psi_{12}^2, \psi_{12}^{12}), \boldsymbol{\theta}_{13} = (\psi_{13}^1, \psi_{13}^2, \psi_{13}^{12}), \boldsymbol{\theta}_{23} = (\psi_{23}^1, \psi_{23}^2, \psi_{23}^{12}))'$, then, $\boldsymbol{\Upsilon}_{(12,12)}$ is the 3×3 diagonal block of $\boldsymbol{\Upsilon}$ located at the top left corner, while $\boldsymbol{\Upsilon}_{(12),(-(12))}$ is the 3×6 block of Υ located at the top right corner.

Further, let $\boldsymbol{a}_{ij}$ be the sub-vector of $\boldsymbol{a}$ whose indices match those of $\boldsymbol{\theta}_{ij}$ inside $\boldsymbol{\Theta}$. Then, the conditional posterior distribution of $\boldsymbol{\theta}_{ij}$ given $\boldsymbol{\Theta}_{-(ij)}$ and $\boldsymbol{\Delta}$, denoted as $\pi\{\boldsymbol{\theta}_{ij}|.\}$, is given by

$$\pi\{\boldsymbol{\theta}_{ij}|.\} \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}_{ij}'\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{(ij)(ij)}\boldsymbol{\theta}_{ij} + 2n\boldsymbol{\theta}_{ij}'\left(\boldsymbol{a}_{ij} + \boldsymbol{\Upsilon}_{(ij)(-(ij))}\boldsymbol{\Theta}_{-(ij)}\right)\right]\right\}$$
$$\times\left(I_{\boldsymbol{\theta}_{ij}=\mathbf{0}} + cI_{\cup_{l=1}^{2^K-1}\mathbb{M}_l}(\boldsymbol{\theta}_{ij})\right),$$

where $c = q_1/(1-q_1)$ if the number of non-zero elements in $\boldsymbol{\Theta}_{-(ij)}$ is less than τ, $c = q_2/(1-q_2)$ if the number of non-zero elements in $\boldsymbol{\Theta}_{-(ij)}$ is greater than τ, and $c = \frac{q_2^{\tau+1}(1-q_2)^{\binom{p}{2}-\tau-1}}{q_1^{\tau}(1-q_1)^{\binom{p}{2}-\tau}}$ if the number of non-zero elements in $\boldsymbol{\Theta}_{-(ij)}$ is equal to τ. Note that since $\boldsymbol{\theta}_{ij}$ has at most one non-zero element, all cross products in $\boldsymbol{\theta}_{ij}'\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{(ij)(ij)}\boldsymbol{\theta}_{ij}$ are equal to zero, i.e.

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

where $\left[\boldsymbol{\Upsilon}_{(ij)(ij)}\right]_{ll}$ is the lth diagonal element of matrix $\boldsymbol{\Upsilon}_{(ij)(ij)}$, for $l = 1, ..., 2^K - 1$. Hence,

denoting the univariate normal probability density function by ϕ, we get

$$\pi\left\{\boldsymbol{\theta}_{ij}|.\right\} \propto \exp\left\{-\frac{1}{2}\sum_{l=1}^{2^K-1}\left(\theta_{l,ij}^2\{n\left[\boldsymbol{\Upsilon}_{(ij)(ij)}\right]_{ll}+\lambda_{l,ij}\}+2n\theta_{l,ij}\left[\boldsymbol{a}_{ij}+\boldsymbol{\Upsilon}_{(ij)(-(ij))}\boldsymbol{\Theta}_{-(ij)}\right]_l\right)\right\}$$
$$\times\left(I_{\boldsymbol{\theta}_{ij}=\mathbf{0}}+cI_{\cup_{l=1}^{2^K-1}\mathbb{M}_l}(\boldsymbol{\theta}_{ij})\right)=I_{\boldsymbol{\theta}_{ij}\in\mathbf{M}_0}+\sum_{l=1}^{2^K-1}c_{l,ij}\phi\{\theta_{l,ij};\left(\mu_{l,ij},\nu_{l,ij}^2\right)\}I_{\boldsymbol{\theta}_{ij}\in\mathbf{M}_l},$$

with

$$\begin{aligned}\mu_{l,ij}&=-\frac{n\left(\boldsymbol{a}_{ij}+\boldsymbol{\Upsilon}_{(ij)(-(ij))}\boldsymbol{\Theta}_{-(ij)}\right)_l}{n\left[\boldsymbol{\Upsilon}_{(ij)(ij)}\right]_{ll}+\lambda_{l,ij}},\\ \nu_{l,ij}^2&=\frac{1}{n\left[\boldsymbol{\Upsilon}_{(ij)(ij)}\right]_{ll}+\lambda_{l,ij}},\qquad c_{l,ij}=c\sqrt{2\pi\nu_{l,ij}^2}\exp\left\{\frac{\mu_{l,ij}^2}{2\nu_{l,ij}^2}\right\},\end{aligned}\tag{B.11}$$

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

$$\pi\left\{\boldsymbol{\theta}_{ij}|.\right\}=\pi\left\{\boldsymbol{\theta}_{ij}|\boldsymbol{\Theta}_{-(ij)},\boldsymbol{\Delta},\mathcal{Y}\right\}=\frac{I_{\boldsymbol{\theta}_{ij}\in\mathbf{M}_0}+\sum_{l=1}^{2^K-1}c_{l,ij}\phi\{\theta_{l,ij};\left(\mu_{l,ij},\nu_{l,ij}^2\right)\}I_{\boldsymbol{\theta}_{ij}\in\mathbf{M}_l}}{1+\sum_{l=1}^{2^K-1}c_{l,ij}},\tag{B.12}$$

The above density is a mixture of univariate normal densities.

B.3 Gibbs Sampling Scheme for BJNS

B.3.1 Gibbs Sampling Algorithm

Procedure 1. Gibbs Sampler for BJNS

Input $\mathcal{Y}, \boldsymbol{\Theta}, \boldsymbol{\Delta}$

For $i=1$ to $i=p-1$

For $j=i+1$ to $j=p$

For $l = 1$ to $2^K - 1$

$\lambda \leftarrow$ Gamma$(r + 0.5, 0.5(\theta^2_{l,ij} + s))$

$\mu \leftarrow -\frac{n\left(\boldsymbol{a}_{ij} + \boldsymbol{\Upsilon}_{(ij)(-(ij))}\boldsymbol{\Theta}_{-(ij)}\right)_l}{n\left[\boldsymbol{\Upsilon}_{(ij)(ij)}\right]_{ll} + \lambda}$

$\nu^2 \leftarrow \frac{1}{n\left[\boldsymbol{\Upsilon}_{(ij)(ij)}\right]_{ll} + \lambda}$

$c_l \leftarrow \sqrt{2\pi\nu^2}\exp\left\{\frac{\mu^2}{2\nu^2}\right\}$

End For

$\boldsymbol{\theta}_{ij} \leftarrow 0_{(2^K-1)\times 1}$

$l \leftarrow$ sample $\left(1, \left\{0, 1, ..., 2^K - 1\right\}, \text{probs} \propto \left\{1, c_{1,ij}..., c_{2^K-1,ij}\right\}\right)$

If $l \neq 0$

$\boldsymbol{\theta}_{l,ij} \leftarrow N\left(\mu, \nu^2\right)$

End If

End For

For $k = 1$ to $k = K$

$\gamma \leftarrow$ Gamma$(r + 1, |\psi^k_{ii}| + s)$

$b \leftarrow \gamma + n\sum_{j\neq i}\left(\sum_{r\in\vartheta_k}\psi^r_{ij}\right)s^k_{ij}$

Update ψ^k_{ii} using Algorithm 2 below

End For

End For

For $k = 1$ to $k = K$

$\gamma \leftarrow \text{Gamma}(r+1, |\psi_{pp}^k| + s)$

$b \leftarrow \gamma + n \sum_{j \neq p} \left(\sum_{r \in \vartheta_k} \psi_{pj}^r \right) s_{pj}^k$

Update ψ_{pp}^k using Procedure 2 below

End For

Output $\mathbf{\Theta}, \mathbf{\Delta}$

The conditional posterior density of ψ_{ii}^k given $\mathbf{\Theta}$ is as follows.

$$\pi \left\{ \psi_{ii}^k | \mathbf{\Theta}, \mathcal{Y} \right\} \propto \exp \left\{ n \log \left(\psi_{ii}^k \right) - \frac{n}{2} s_{ii}^k \left(\psi_{ii}^k \right)^2 - b_i^k \psi_{ii}^k \right\}, \tag{B.13}$$

where $b_i^k = \gamma + n \sum_{j \neq i} \left(\sum_{r \in \vartheta_k} \psi_{ij}^r \right) s_{ij}^k$, for $1 \leq i < j \leq p$ and $1 \leq k \leq K$. Note that the density in (B.13) is not a standard density. For generating samples from posterior conditional density of ψ_{ii}^k, we first note that it has a unique mode at

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

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

Procedure 2. Generating samples from density in (19)

Input n, σ_{ii}^k, b_i^k

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

$S \leftarrow$ seq$(0, 6 \times$ mode$, 0.001)$ sequence from 0 to $6 \times$ mode with increments 0.001

For$t \in S$

$p_t \leftarrow \exp\left\{n \log t - \frac{n}{2} s_{ii}^k t^2 - b_i^k t\right\}$

End For

Set sum $= \sum_{t \in S} p_t$

Ifsum $\leftarrow \infty$

For$t \in S$ $\qquad p_t \leftarrow \frac{\exp\left\{n \log t - \frac{n}{2} s_{ii}^k t^2 - b_i^k t\right\}}{\exp\left\{n \log \text{mode} - \frac{n}{2} s_{ii}^k \text{mode}^2 - b_i^k \text{mode}\right\}}$

End For

Set sum $\leftarrow \sum_{t \in S} p_t$

End If

$x \leftarrow$ sample $(1, S, \text{probs} \propto \{p_t : t \in S\})$

Output x

However, we have observed in the numerical work undertaken that the density (19) 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.

C. Hyperparameter Sensitivitiy

We have discussed the interpretation, effect and possible choices of the hyperparameters r, s, q_1, q_2, τ in Section 4. We now illustrate some of the facts discussed in Section 4 using a simulation experiment. We consider two networks with $p = 100$ variables. The corresponding true precision matrices have 8% non-zero off-diagonal entries, and 70% of the non-zero locations are common in both matrices. The average percentage of non-zero edges in the estimated networks for different values of r, s, q_1, q_2 are provided in the Table C.1. The results in the table demonstrate that the BJNS estimates are less sensitive to the choice of hyperparameters as the sample size increases, and that larger values of q_1 and q_2 lead to sparser networks.

Table C.1: Hyper-parameter sensitivity analysis: Average number of edges in the estimated networks for different sample sizes and hyperparameter values (two networks with $p = 100$). True average is 396.

	$n = 20$	$n = 100$	$n = 500$	$n = 1000$
$r = 10^{-4}, s = 10^{-8}, q_1 = q_2 = 0.5$	111.5	221	404.5	427.5
$r = 10^{-4}, s = 10^{-8}, q_1 = q_2 = 1/p$	27	109	388	407.5
$r = 10^{-4}, s = 10^{-6}, q_1 = q_2 = 0.5$	322	331	461	486
$r = 10^{-4}, s = 10^{-4}, q_1 = q_2 = 0.5$	1491	631.5	619.5	651

D. Additional Simulation Results

D.1 BJNS with four groups ($K = 4$)

We consider two challenging scenarios to examine the performance of BJNS for the case of simultaneously selecting the sparsity pattern for four precision matrices ($K = 4$). In the first scenario, we consider four positive definite 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: $\mathbf{\Omega}^k = \mathbf{\Psi}^k + \mathbf{\Psi}^{1234}, \quad k = 1, 2, 3, 4$, where $\mathbf{\Psi}^1$, $\mathbf{\Psi}^2$, $\mathbf{\Psi}^3$, and $\mathbf{\Psi}^4$ account for the edges that are unique to their corresponding groups and $\mathbf{\Psi}^{1234}$ contains the edges that are common between all of the four groups.

In the second simulation setting, we consider four positive definite precision matrices $\mathbf{\Omega}^1$, $\mathbf{\Omega}^2$, $\mathbf{\Omega}^3$, and $\mathbf{\Omega}^4$, with different degrees of shared structures. We let the first matrix $\mathbf{\Omega}^1$ to be an AR(2) model with $\omega^1_{jj} = 1$, for $j = 1, ..., p$; $\omega^1_{jj+1} = \omega^1_{j+1j} = 0.5$, for $j = 1, ..., p-1$; and $\omega^1_{jj+2} = \omega^1_{j+2j} = 0.25$, for $j = 1, ..., p-2$. To construct $\mathbf{\Omega}^2$ we randomly replace $\frac{p}{4}$ non zero edges from $\mathbf{\Omega}^1$ 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 $\mathbf{\Omega}^3$ by randomly removing $\frac{p}{2}$ edges shared between $\mathbf{\Omega}^1$ and $\mathbf{\Omega}^2$ 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 $\mathbf{\Omega}^1$ nor $\mathbf{\Omega}^2$. Finally, we construct $\mathbf{\Omega}^4$ by removing the remaining $2p - 3 - \frac{3p}{4}$ edges

that are common in $\mathbf{\Omega}^1$ and $\mathbf{\Omega}^2$ and randomly add $2p - 3 - \frac{3p}{4}$ edges that are not present in any of the other graphs. The resulting matrix $\mathbf{\Omega}^4$ has nothing in common with the other precision matrices. A heat map plot of the true precision matrices with $p = 50$ is given in Figure G.1 in the Supplement. Formally, the true relation between the four networks is as follows.

$$\mathbf{\Omega}^1 = \mathbf{\Psi}^1 + \mathbf{\Psi}^{12} + \mathbf{\Psi}^{123} \qquad \mathbf{\Omega}^3 = \mathbf{\Psi}^3 + \mathbf{\Psi}^{123} \; \mathbf{\Omega}^2 = \mathbf{\Psi}^2 + \mathbf{\Psi}^{12} + \mathbf{\Psi}^{123} \qquad \mathbf{\Omega}^4 = \mathbf{\Psi}^4,$$

where, $\mathbf{\Psi}^1$, $\mathbf{\Psi}^2$, $\mathbf{\Psi}^3$, and $\mathbf{\Psi}^4$ account for the edges that are unique to their corresponding groups and $\mathbf{\Psi}^{12}$ and $\mathbf{\Psi}^{123}$ contains the edges that are common between the four groups.

For each of the above settings, we run the following combinations of p and n: (i) $p = 200$, and $n = 50, 100, 150, 200, 250, \text{and}, 300$, and (ii) $p = 500$, and $n = 300, 400, \text{and}, 500$. Estimation is based on the full decomposition model given by

$$\mathbf{\Omega}^i = \mathbf{\Psi}^i + \sum_{j \neq i} \mathbf{\Psi}^{ij} + \sum_{j \neq k \neq i} \mathbf{\Psi}^{ijk} + \mathbf{\Psi}^{1234} \tag{D.14}$$

for $i = 1, 2, 3, 4$. Figures D.3 and D.2 below summarize the average of the accuracy measures for both scenarios, across 100 repetitions. Note that in these figures, in addition to MCC, we also provide Specificity (SP) and Sensitivity (SE) defined by SP $= \frac{\text{TN}}{\text{TN+FP}}$, SE $= \frac{\text{TP}}{\text{TP+FN}}$. As can be seen, the values of the accuracy measures tend to be much higher for the joint effects (namely $\mathbf{\Psi}^{1234}$ in Figure D.3 and $\mathbf{\Psi}^{12}$ and $\mathbf{\Psi}^{123}$ in Figure D.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 FP rate.

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

$$\boldsymbol{\theta}_{ij} = \left(\psi_{ij}^{1}, \psi_{ij}^{2}, \psi_{ij}^{3}, \psi_{ij}^{4}, \psi_{ij}^{12}, \psi_{ij}^{13}, \psi_{ij}^{14}, \psi_{ij}^{23}, \psi_{ij}^{24}, \psi_{ij}^{123}, \psi_{ij}^{124}, \psi_{ij}^{134}, \psi_{ij}^{234}, \psi_{ij}^{1234}\right)'. \tag{D.15}$$

for every $1 \leq i < j \leq p$, BJNS aims at detecting whether $\boldsymbol{\theta}_{ij}$ is a zero vector or which of its components is non-zero (recall that $\boldsymbol{\theta}_{ij}$ has at most one non-zero element). For example, if an edge (ij) is common across all four networks, then ψ_{ij}^{1234} must be the non-zero element in the corresponding $\boldsymbol{\theta}_{ij}$. That is, we must have $\boldsymbol{\theta}'_{ij} = (\mathbf{0}'_{14}, \psi_{ij}^{1234})$. Therefore, including the case where all elements of $\boldsymbol{\theta}_{ij}$ are zero, there are a total of 16 ($= 2^K$) possible sparsity patterns for $\boldsymbol{\theta}_{ij}$. Hence, for each $\boldsymbol{\theta}_{ij}$, $1 \leq i < j \leq p$ and in each iteration of the Gibbs sampler, our model generates 16 probabilities (one per sparsity pattern) and randomly selects a sparsity pattern using the generated probabilities.

As mentioned previously, the final choice of the subset for each edge (i, j) is made using majority voting. In particular, we compute the proportion of iterations for which the element of $\boldsymbol{\theta}_{ij}$ corresponding to a given subset of $\{1, 2, \cdots, K\}$ is chosen to be non-zero. The subset with the highest proportion is then chosen, and the (i, j) entries of the precision matrices corresponding to the elements not in the chosen subset are set to zero. These proportions provide crucial uncertainty quantification and indicate the degree of decisiveness in the estimated spar-

sity patterns. We closely studied this issue by extracting these proportions in various simulation scenarios. We observed that the highest proportions are indeed almost always close to one, indicating a high degree of decisiveness. Just as an example, in the case of the above simulation scenario, with $p = 200$ and $n = 300$, the distribution of the highest proportions is narrowly concentrated around a value that is very close to one. A plot of the distribution of the highest proportions for all the $\binom{200}{2} = 19900$ variable pairs in this specific case is given in Figure D.1. In view of the full model (D.14), high values of the accuracy measures for the inverse covariance matrices $\{\boldsymbol{\Omega}^k\}_{k=1}^4$ is another indication of strong selection capability of BJNS.

Fast MCMC convergence: Note that the model selection properties of BJNS can be used to check convergence of the MCMC algorithm. This is accomplished by tracking the proportion of correctly selected $\boldsymbol{\theta}_{ij}$s (denoted by κ). In addition to accuracy assessment, κ helps studying the number of iterations that on average it takes for the Gibbs sampler to converge. Figure G.2 in Supplemental Section G describes the trace plot and the histogram of κ for 4000 iterations for the above two simulations for the $p = 200$, $n = 300$ settings. The MCMC trace plots of κ show that the Gibbs sampler converges fairly quickly. Moreover, the histograms of κ indicate a high proportion of correctly selected $\boldsymbol{\theta}_{ij}$s.

In addition, the trace plots of three sample edges (a negative, a zero and a positive edge) are provided in Figure G.3, which further confirm the reasonably quick convergence of our Gibbs sampler.

Figure D.1: Uncertainty quantification: Distribution of the highest proportions (among $2^4 = 16$ subsets) for each of the $\binom{200}{2}$ variable pairs in a simulation replication with $p = 200$. One can see that the highest proportions are all densely concentrated around a value that is very close to one, indicating a strong degree of decisiveness/certainty for sparsity selection.

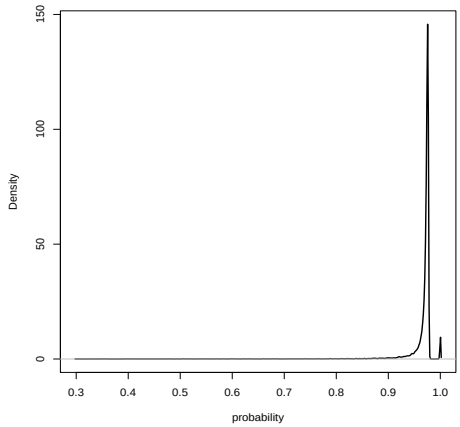
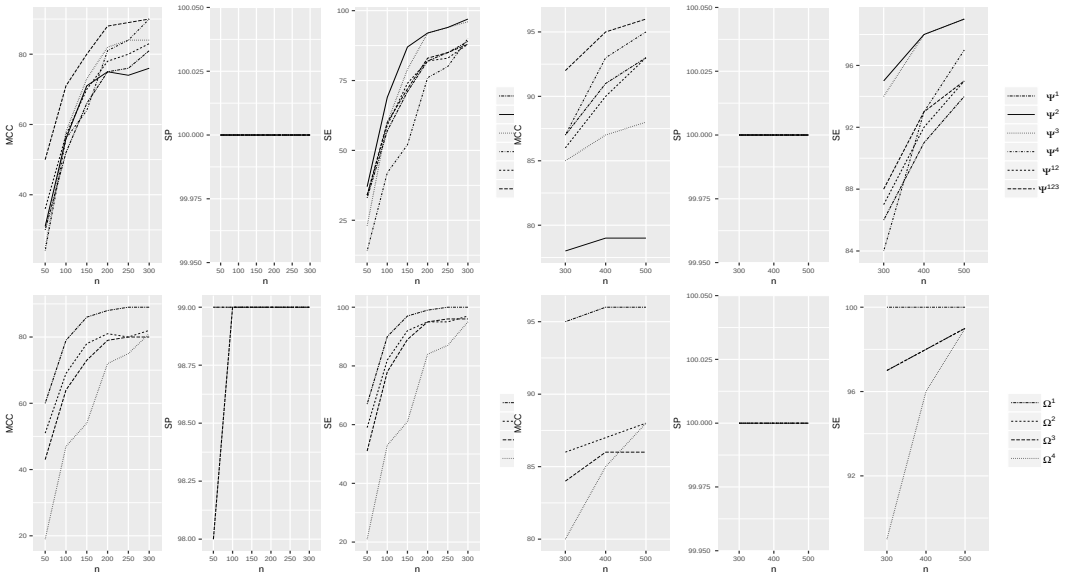
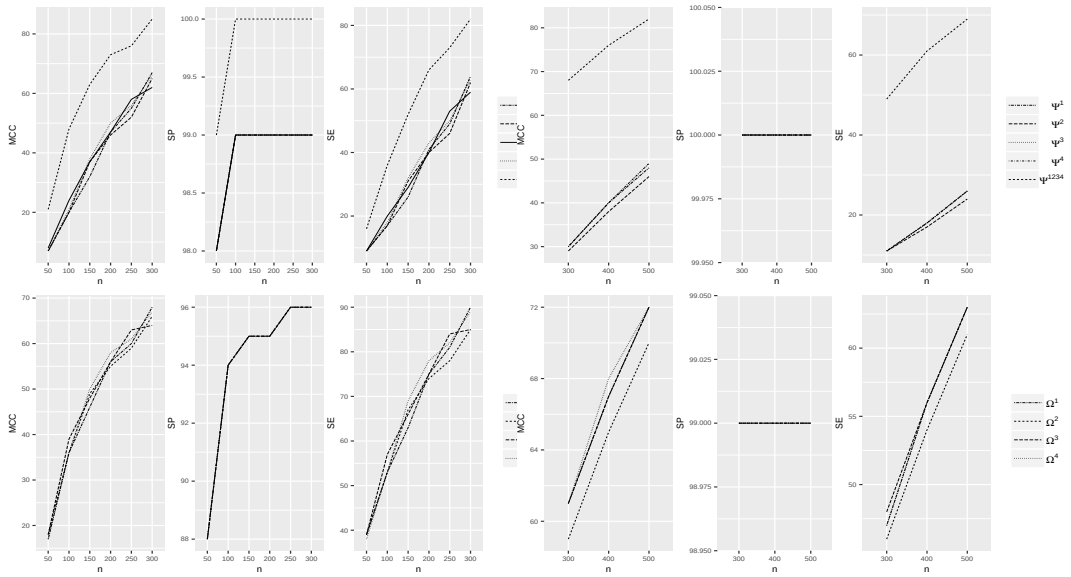


Figure D.2: Performance plot of BJNS in the first simulation setting of Section D.1 when (a) $p = 200$ and when (b) $p = 500$. Plots in the first row assesses the edge selection inside the individual matrices $\mathbf{\Psi}^r$s and the plots in the second row studies the same in the precision matrices $\mathbf{\Omega}^k$s. The higher values of the accuracy measures for $\mathbf{\Psi}^{12}$ and $\mathbf{\Psi}^{123}$ shows that BJNS is borrowing strength across the distinct samples, resulting in more robustness in selecting the joint edges



(a) (b)

Figure D.3: Performance plot of BJNS in the second simulation setting of Section D.1 when (a) $p = 200$ and when (b) $p = 500$. Plots in the first row assesses the edge selection inside the individual matrices $\mathbf{\Psi}^r$s and the plots in the second row studies the same in the precision matrices $\mathbf{\Omega}^k$s. The higher values of the accuracy measures for $\mathbf{\Psi}^{1234}$ implies that BJNS is borrowing strength across the 4 distinct samples, resulting in more robustness in selecting the edges that are shared 4-ways



(a) (b)

D.2 Replicating the simulation in Table 2

In this section, we replicate the analysis reported in Table 2 with a different edge value range of $[-0.8, -0.3] \cup [0.3, 0.8]$. As can be seen, our earlier conclusions hold which indicates that the edge value range did not play a role in the superiority of BJNS over other methods.

Table D.1: MCC values for when replicating the analysis reported in Table 2 with edges generated from $[-0.8, -0.3] \cup [0.3, 0.8]$

	Glasso	JEM-G	GGL	FGL	BNS	GemBag	BJNS
Edge density			$n = 200$				
8%	45 (0.010)	49 (0.011)	45 (0.010)	47 (0.010)	35 (0.002)	43 (0.009)	**52 (0.012)**
12%	**39 (0.008)**	36 (0.009)	34 (0.008)	37 (0.008)	28 (0.001)	35 (0.007)	**39 (0.011)**
			$n = 300$				
8%	51 (0.009)	58 (0.008)	51 (0.007)	54 (0.008)	40 (0.002)	52 (0.007)	**64 (0.008)**
12%	45 (0.008)	43 (0.008)	41 (0.007)	46 (0.008)	35 (0.002)	41 (0.009)	**49 (0.010)**

D.3 Comparison with supervised method (JSEM)

In this section, we compare BJNS (an unsupervised method) with JSEM (Ma and Michailidis, 2016) which is a supervised method that incorporates exact information regarding shared structural similarity across various groups. We consider the same settings as in Section 5.1.1 in the main paper. Recall that there are two settings, one where values for shared edges in the true networks are chosen to be identical, and another where they are chosen to have different

magnitudes but the same sign. For each setting, we consider two sub-settings. In the first sub-setting, we supply the sparsity patterns defined according to the pattern in Figure G.4 in Section G. For this sub-setting JSEM is given complete information about which sets of edges can be fused across which subsets of networks. In the second sub-setting we add an additional 4% of edges to the networks. In this sub-setting, JSEM is not given information about the newly added edges, and hence has correctly specified information for only 67% of the edges. Such a setup allows us to evaluate the performance of JSEM (vis-a-vis BJNS) with and without mis-specification. Just as in Section 5.1.1 in the main paper, the MCC values (averaged over 50 replications) for sample sizes $n_k = 200, 300$ are provided in Table D.2 (no misspecification) and Table D.3 (misspecification).

When there is no mis-specifications (8% edge density) 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 a higher MCC score, as expected. BJNS is also very competitive and, as demonstrated in Section 5.1.1, its overall performance is significantly better than all other unsupervised methods. However, we see that the performance of JSEM under misspecification (12% edge density, with 67% correct info for JSEM) deteriorates, while the unsupervised BJNS approach is more robust and delivers a better performance than JSEM in this setting.

Note that although BJNS is a completely unsupervised approach, some available prior

Table D.2: MCC values for JSEM AND BJNS across 6 networks, when the true sparsity patterns are random, and shared edges have exactly the same values. The MCC values are averaged over 50 replications.

	JSEM	BJNS
Edge density	$n = 200$	
8%	61(.009)	57 (0.010)
12%	32(.009)	40 (0.010)
	$n = 300$	
8%	73(.008)	70 (0.010)
12%	42(.008)	51 (0.010)

knowledge on shared patterns across the groups can be incorporated by removing redundant components $\mathbf{\Psi}^r$ (see Section E) or through appropriate specification of the edge selection probabilities q_1, q_2.

D.4 Estimation of magnitudes of non-zero entries

We also examine the estimation performance of BJNS in the above settings. Specifically, we calculate the relative error with respect to the Frobenius norm, namely $\frac{\|\hat{\mathbf{\Omega}}_k - \mathbf{\Omega}_k^0\|_F}{\|\mathbf{\Omega}_k^0\|_F}$ for every k. We consider the case depicted in Figure G.4 (Section G), where the shared edges take on

Table D.3: MCC values for JSEM AND BJNS across 6 networks, when the true sparsity patterns are random, and shared edges have same sign but different magnitudes. The MCC values are averaged over 50 replications.

	JSEM	BJNS
Edge density	$n = 200$	
8%	60(.022)	54 (0.011)
12%	33(.009)	40 (0.009)
	$n = 300$	
8%	73(.008)	67 (0.010)
12%	43(.008)	51 (0.009)

the same values (very similar results hold for the case where the shared edges take different values across the networks). We compare the relative error committed by BJNS, Glasso and JSEM. As described in equation (4.12) of the main paper, for estimation purposes, we use the skeleton produced by BJNS and compute restricted MLE for each $\mathbf{\Omega}^k$. These refitted estimates are guaranteed to be positive definite and have no magnitude/sign restrictions. The relative error norms for this refitting approach (BJNS-refitted), Glasso, and JSEM averaged over the 6 networks are given in Table D.4. Based on the relative error norm results, BJNS outperforms

Glasso and JSEM for magnitude estimation purposes in all settings.

Table D.4: Summary of average relative error across the 6 networks, for the case of random sparsity patterns with equal shared edges for all $\mathbf{\Omega}^k$s

		Glasso	JSEM	BJNS-refitted
ρ		$n = 200$		
8%	RFN%	0.81 (0.001)	1.19 (0.002)	0.59 (0.023)
12%	RFN%	0.84 (0.001)	1.10 (0.001)	0.66 (0.015)
		$n = 300$		
8%	RFN%	0.79 (0.001)	1.25 (0.002)	0.58 (0.026)
12%	RFN%	0.83 (0.001)	1.15 (0.002)	0.66 (0.015)

E. A Computational Strategy for Speeding Up BJNS for large K

As presented in Procedure 1 above, the Gibbs sampler updates all the $p(p-1)/2$ vectors $\boldsymbol{\theta}_{ij}$ based on their full conditional distributions. Although the conditional posterior distribution of $\boldsymbol{\theta}_{ij}$s is a mixture of univariate normal densities, all 2^K mixture probabilities $c_{l,ij}$ given in (B.11) still need to be calculated. 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

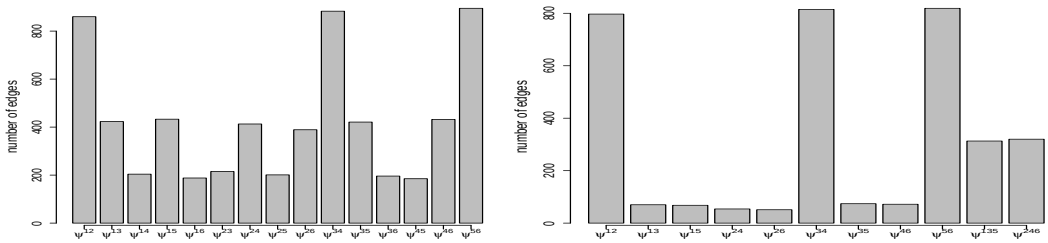
$$\begin{aligned} \boldsymbol{\Omega}^{k_1} &= \boldsymbol{\Psi}^{k_1} + \boldsymbol{\Psi}^{k_1 k_2}, \\ \boldsymbol{\Omega}^{k_2} &= \boldsymbol{\Psi}^{k_2} + \boldsymbol{\Psi}^{k_1 k_2}. \end{aligned} \tag{E.16}$$

Subsequently, we remove all the pairwise matrices $\boldsymbol{\Psi}^{k_1 k_2}$ 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 simulation setting given in Figure G.4, which involves $K = 6$ groups with $p = 200$ variables and $n_k = 200, 300$ samples per group, and is described in Section 5.1.1 of the main paper. Note that the full decomposition would involve all interaction components (up to six-way interactions), and each vector $\boldsymbol{\theta}_{ij}$ would have length $2^K - 1 = 63$. However, as can be seen by the design of the simulation, most components r are "inactive" (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 E.1a shows the number of edges in the pairwise matrices $\boldsymbol{\Psi}^{k_1 k_2}$, $1 \leq k_1 < k_2 \leq 6$.

Figure E.1: Bar plots of the edge count of the pairwise joint components (a) and components of the reduced model (b)



(a) (b)

Based on this plot, one can see that matrices $\mathbf{\Psi}^{14}$, $\mathbf{\Psi}^{16}$, $\mathbf{\Psi}^{23}$, $\mathbf{\Psi}^{25}$, $\mathbf{\Psi}^{35}$, and $\mathbf{\Psi}^{45}$ 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 $\mathbf{\Psi}^{14}$, $\mathbf{\Psi}^{16}$, $\mathbf{\Psi}^{23}$, $\mathbf{\Psi}^{25}$, $\mathbf{\Psi}^{35}$, and $\mathbf{\Psi}^{45}$.

Doing so results in a reduced model with components, $\mathbf{\Psi}^{12}$, $\mathbf{\Psi}^{13}$, $\mathbf{\Psi}^{15}$, $\mathbf{\Psi}^{24}$, $\mathbf{\Psi}^{26}$, $\mathbf{\Psi}^{34}$, $\mathbf{\Psi}^{35}$, $\mathbf{\Psi}^{46}$, $\mathbf{\Psi}^{56}$, $\mathbf{\Psi}^{135}$, and $\mathbf{\Psi}^{246}$. 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 E.1b. From which, it is clear that matrices $\mathbf{\Psi}^{13}$, $\mathbf{\Psi}^{15}$, $\mathbf{\Psi}^{24}$, $\mathbf{\Psi}^{26}$, $\mathbf{\Psi}^{35}$, and $\mathbf{\Psi}^{46}$ are redundant and should be removed from the model. This further model reduction achieves the true decomposition provided in Section 5.1.1 of the main paper. The results of BJNS running on the resulting reduced model can be read off from the Table E.1. As can be seen, BJNS outperforms JSEM,

a supervised approach which has been supplied with complete information about the sparsity patterns.

Table E.1: Comparison 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)

Lastly, as presented in Table E.2, we investigate the computational benefits 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 by more than 50%. The wall clock time for BJNS, adjusted for this parallelization, is provided in Table E.2. 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 E.2: 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	1.25h	19.5h	41.7h	107.3h
GigaBytes	0.25gb	0.5gb	0.6gb	0.9gb

F. Proof of Theorem 1

By Assumptions 2 and 3 and Hanson-Wright inequality from Rudelson and Vershynin (2013), there exists a $c > 0$, independent of n and K, such that

$$\mathbb{P}_0 \left\{ \max_{i,j,k} \|s_{ij}^k - \bar{\sigma}_{ij}^{k,0}\| < c\sqrt{\frac{\log p}{n}} \right\} \geq 1 - \frac{1}{p^2},$$

and,

$$\mathbb{P}_0 \left\{ \max_{i,j,k} \|(\bar{\mathbf{\Omega}}_{:i}^{k,0})' \mathbf{S}_{:j}^k\| < c\sqrt{\frac{\log p}{n}} \right\} \geq 1 - \frac{1}{p^2}.$$

Define the events $C_{1,n}$, $C_{2,n}$ as

$$C_{1,n} = \left\{ \max_{i,j,k} \|s_{ij}^k - \bar{\sigma}_{ij}^{k,0}\| < c\sqrt{\frac{\log p}{n}} \right\}, \tag{F.17}$$

$$C_{2,n} = \left\{ \max_{i,j,k} \|(\bar{\mathbf{\Omega}}_{:i}^{k,0})' \mathbf{S}_{:j}^k\| < c\sqrt{\frac{\log p}{n}} \right\}, \tag{F.18}$$

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 $\boldsymbol{\Upsilon}$ and it's components $\mathbf{B}_k$, $k = 1, ..., K$.

Lemma 1. *For any $k = 1, ..., K$, the following holds*

$$eig_{min}\left(\mathbf{S}^k\right) \leq eig_{min}\left(\mathbf{B}^k\right) \leq eig_{max}\left(\mathbf{B}^k\right) \leq 2eig_{max}\left(\mathbf{S}^k\right). \tag{F.19}$$

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

$$\boldsymbol{y} = \boldsymbol{y}\left(\boldsymbol{\Omega}^k\right) = \left(\left(\mathbf{P}^1\boldsymbol{\Omega}^k_{:1}\right)', \left(\mathbf{P}^2\boldsymbol{\Omega}^k_{:2}\right)', ..., \left(\mathbf{P}^p\boldsymbol{\Omega}^k_{:p}\right)'\right)'.$$

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

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

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

Now, define $\underline{\boldsymbol{\omega}}^k = \left(\omega^k_{12}, \omega^k_{13}, ..., \omega^k_{p-1p}\right)'$ and $\underline{\boldsymbol{\delta}}_{\boldsymbol{\Omega}^k} = \left(\omega^k_{11}, \omega^k_{22}, ..., \omega^k_{pp}\right)'$ and let $\tilde{\mathbf{Q}}$ be the $\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ permutation matrix for which

$$\boldsymbol{x} = \mathbf{Q}\begin{pmatrix} \underline{\boldsymbol{\omega}}^k \\ \\ \underline{\boldsymbol{\delta}}_{\boldsymbol{\Omega}^k} \end{pmatrix}.$$

Let $\tilde{\mathbf{\Sigma}}^k$ be a $p^2 \times p^2$ block diagonal matrix with p diagonal blocks, the ith block is equal to $\tilde{\mathbf{\Sigma}}^{k,i} := \mathbf{P}^i \mathbf{S}^k \mathbf{P}^{i\prime}$. It follows that

$$\begin{aligned}
\operatorname{tr}\left[\left(\mathbf{\Omega}^k\right)^2 \mathbf{S}^k\right] &= \sum_{i=1}^{p} \mathbf{\Omega}_{:i}^{k\prime} \mathbf{S}^k \mathbf{\Omega}_{:i}^k = \sum_{i=1}^{p} \mathbf{\Omega}_{:i}^{k\prime} \mathbf{P}^{i\prime} \mathbf{P}^i \mathbf{S}^k \mathbf{P}^{i\prime} \mathbf{P}^i \mathbf{\Omega}_{:i}^k = \sum_{i=1}^{p} \mathbf{\Omega}_{:i}^{k\prime} \mathbf{P}^{i\prime} \left(\mathbf{P}^i \mathbf{S}^k \mathbf{P}^{i\prime}\right) \mathbf{P}^i \mathbf{\Omega}_{:i}^k \\
&= \boldsymbol{y}' \tilde{\mathbf{\Sigma}}^k \boldsymbol{y} = \boldsymbol{x}' \tilde{\mathbf{P}}' \tilde{\mathbf{\Sigma}}^k \tilde{\mathbf{P}} \boldsymbol{x} = \left(\underline{\boldsymbol{\omega}}^{k\prime}, \underline{\boldsymbol{\delta}}_{\mathbf{\Omega}^k}'\right) \mathbf{Q}' \tilde{\mathbf{P}}' \tilde{\mathbf{\Sigma}}^k \tilde{\mathbf{P}} \mathbf{Q} \begin{pmatrix} \underline{\boldsymbol{\omega}}^k \\ \underline{\boldsymbol{\delta}}_{\mathbf{\Omega}^k} \end{pmatrix}.
\end{aligned}$$

There also exist appropriate matrices $\mathbf{A}^k$ and $\mathbf{D}^k$ such that

$$\operatorname{tr}\left[\left(\mathbf{\Omega}^k\right)^2 \mathbf{S}^k\right] = \left(\underline{\boldsymbol{\omega}}^{k\prime}, \underline{\boldsymbol{\delta}}_{\mathbf{\Omega}^k}'\right) \begin{pmatrix} \mathbf{B}^k & \mathbf{A}^k \\ \mathbf{A}^k & \mathbf{D}^k \end{pmatrix} \begin{pmatrix} \underline{\boldsymbol{\omega}}^k \\ \underline{\boldsymbol{\delta}}_{\mathbf{\Omega}^k} \end{pmatrix},$$

therefore, we must have

$$\mathbf{Q}' \tilde{\mathbf{P}}' \tilde{\mathbf{\Sigma}}^k \tilde{\mathbf{P}} \mathbf{Q} = \begin{pmatrix} \mathbf{B}^k & \mathbf{A}^k \\ \mathbf{A}^k & \mathbf{D}^k \end{pmatrix}.$$

Since the diagonal blocks of $\tilde{\mathbf{\Sigma}}^k$ all have the same eigenvalues as $\mathbf{S}^k$, $\mathbf{Q}$ is a permutation matrix, and the columns of $\tilde{\mathbf{P}}$ are orthogonal with ℓ_2 norm 1 or 2, we have that

$$\operatorname{eig}_{min}\left(\mathbf{S}^k\right) = \operatorname{eig}_{min}\left(\tilde{\mathbf{\Sigma}}^k\right) \leq \operatorname{eig}_{min}\left(\mathbf{B}^k\right) \leq \operatorname{eig}_{max}\left(\mathbf{B}^k\right) \leq 2\operatorname{eig}_{max}\left(\tilde{\mathbf{\Sigma}}^k\right) = 2\operatorname{eig}_{max}\left(\mathbf{S}^k\right).$$

□ □

Lemma 2. *Let $\boldsymbol{\ell} \in \mathcal{L}$ be any sparsity pattern/model with $d_{\boldsymbol{\ell}} < \frac{\varepsilon_0}{4c}\sqrt{\frac{n}{\log p}}$, then the sub matrix $\mathbf{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}$ of $\mathbf{\Upsilon}$, obtained by taking out all the rows and columns corresponding to the zero coordinates in*

$\boldsymbol{\Theta} \in \mathcal{M}_{\boldsymbol{\ell}}$, *is positive definite. Specifically,*

$$\frac{3K\varepsilon_0}{4} \leq eig_{\min}\left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}\right) \leq eig_{\max}\left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}\right) \leq \frac{5K}{2\varepsilon_0}, \quad \forall \boldsymbol{\ell} \in \mathcal{L}. \tag{F.20}$$

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 $\boldsymbol{x}$ be a $d_{\boldsymbol{\ell}} \times 1$ vector in $\mathbb{R}^{d_{\boldsymbol{\ell}}}$ and partition $\boldsymbol{x}$ as

$$\boldsymbol{x} = \left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},23}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},3}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},2}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}}\right),$$

then, by making similar partitions on each block of $\boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}$ (see $\boldsymbol{\Upsilon}$ in (A.5)), we have that

$$\begin{aligned}
\boldsymbol{x}'\boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{x} = &\left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}}\right)' \mathbf{B}^1_* \left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}}\right) \\
&+ \left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},23}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},2}}\right)' \mathbf{B}^2_* \left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},23}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},2}}\right) \\
&+ \left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},23}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},3}}\right)' \mathbf{B}^3_* \left(\boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},23}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}}, \boldsymbol{x}'_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},3}}\right),
\end{aligned}$$

where,

$$\mathbf{B}^*_1 = \begin{pmatrix}
\mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}} \\
\mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}} \\
\mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}} \\
\mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},123}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},13}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},12}} & \mathbf{B}^1_{\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}\underline{\boldsymbol{\psi}}^{\boldsymbol{\ell},1}}
\end{pmatrix}.$$

Let $\mathbf{B}^{k,0}_*$ denote the population version of $\mathbf{B}^k_*$. Since, we are restricted to $C_{1,n} \cap C_{2,n}$, $\|\mathbf{B}^k_* - \mathbf{B}^{k,0}_*\| \leq c d_\ell \sqrt{\frac{\log p}{n}}$, hence

$$
\begin{aligned}
\text{eig}_{min}\left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell\ell}} &\geq \sum_{k=1}^{K} \text{eig}_{\min}\left(\mathbf{B}_*^k\right) = \sum_{k=1}^{K} \inf_{|\boldsymbol{x}|=1} \boldsymbol{x}'\mathbf{B}_*^k\boldsymbol{x} \\
&\geq \sum_{k=1}^{K} \left[\inf_{|\boldsymbol{x}|=1} \boldsymbol{x}'\mathbf{B}_*^{k,0}\boldsymbol{x} - \inf_{|\boldsymbol{x}|=1} \boldsymbol{x}'\left(\mathbf{B}_*^k - \mathbf{B}_*^{k,0}\right)\boldsymbol{x}\right] \\
&\geq \sum_{k=1}^{K} \left[\inf_{|\boldsymbol{x}|=1} \boldsymbol{x}'\mathbf{B}_*^{k,0}\boldsymbol{x}\right] - \sum_{k=1}^{K} \|\mathbf{B}_*^k - \mathbf{B}_*^{k,0}\|_2 \\
&\geq \sum_{k=1}^{K} \left[\inf_{|\boldsymbol{x}|=1} \boldsymbol{x}'\mathbf{B}_*^{k,0}\boldsymbol{x}\right] - Kd_{\boldsymbol{\ell}}c\sqrt{\frac{\log p}{n}}
\end{aligned}
$$

Hence, by Lemma 1 (repeating the same arguments with the population versions of the respective matrices), we have

$$
\begin{aligned}
\text{eig}_{min}\left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell\ell}} &\geq K\varepsilon_0 - Kcd_{\boldsymbol{\ell}}\sqrt{\frac{\log p}{n}} \\
&\geq K\left(\varepsilon_0 - c\tau_n\sqrt{\frac{\log p}{n}}\right) = \frac{3K\varepsilon_0}{4}.
\end{aligned}
$$

Similarly one can show that

$$
\text{eig}_{\max}\left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell\ell}}\right) \leq \frac{5K}{2\varepsilon_0}.
$$

□ □

By Lemma 2, the value of the threshold τ_n which we used in building our hierarchical prior in (3.3) is given as $\tau_n = \frac{\varepsilon_0}{4c}\sqrt{\frac{n}{\log p}}$. Hence by Assumption 1, we can write $d_t \leq \tau_n$, for any sufficiently large n.

As mentioned in Section 7 of the main paper, following high-dimensional consistency proofs in Peng et al.(2009), Khare et al. (2015), Atchade (2019), we consider a setting where

sufficiently accurate estimates $\left\{\hat{\Omega}_{ii}^{k}\right\}_{1\leq i\leq p}^{1\leq k\leq K}$ of the *diagonal* entries are first obtained. We then establish results for BJNS with diagonal entries fixed at these values. In particular, for some fixed constant $C > 0$, we want

$$\max_{1\leq k\leq K, 1\leq i\leq p} \left|\hat{\Omega}_{ii}^{k} - \bar{\Omega}_{ii}^{k,0}\right| \leq C\sqrt{\frac{\log p}{n}} \tag{F.21}$$

with probability converging to 1 as $n \to \infty$. One way to obtain such estimates is as follows. Note that $1/\bar{\Omega}_{ii}^{k,0}$ is the error variance when regressing the ith variable against all the other variables for the kth group. Hence, we can run parallel lasso regressions for each variable against all the other variables *separately* for each of the K groups (so pK regressions in all), and obtain estimates of the pK diagonal entries using the lasso variance estimator (see for example eq. (13) in Fan et al. (2012)). Assumptions 1-5 in the main paper and Theorem 1 in Fan et al. (2012) guarantee that the bound in (F.21) holds. Recall that the resulting estimates of the vectors $\boldsymbol{\Delta}$ and $\boldsymbol{a}$ are denoted by $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{a}}$, respectively.

Lemma 3. *Let, $\boldsymbol{\Upsilon}$, and $\boldsymbol{a}$ be according to (A.5), (A.7), and let $\boldsymbol{\Theta}^0$ be the true value of $\boldsymbol{\Theta}$ in (3.1). Then for large enough n, there exists a constant c_0 such that*

$$\|\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\|_{\max} \leq c_0\sqrt{\frac{\log p}{n} + d_{\boldsymbol{t}} D_n^2}. \tag{F.22}$$

Proof. Note that by the triangular inequality,

$$\|\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\|_{\max} \leq \|\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\|_{\max} + \|\hat{\boldsymbol{a}} - \boldsymbol{a}\|_{\max}, \tag{F.23}$$

where, $\hat{\boldsymbol{a}}$ is the estimate of $\boldsymbol{a}$ using the accurate diagonal estimates.

Now, in view of (A.5), (A.7), and (3.1), one can easily check that

$$\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a} = \begin{pmatrix} \mathbf{B}^1\underline{\boldsymbol{\omega}}^{1,0} + \boldsymbol{a}^1 + \mathbf{B}^2\underline{\boldsymbol{\omega}}^{2,0} + \boldsymbol{a}^2 + \mathbf{B}^3\underline{\boldsymbol{\omega}}^{3,0} + \boldsymbol{a}^3 \\ \mathbf{B}^2\underline{\boldsymbol{\omega}}^{2,0} + \boldsymbol{a}^2 + \mathbf{B}^3\underline{\boldsymbol{\omega}}^{3,0} + \boldsymbol{a}^3 \\ \mathbf{B}^1\underline{\boldsymbol{\omega}}^{1,0} + \boldsymbol{a}^1 + \mathbf{B}^3\underline{\boldsymbol{\omega}}^{3,0} + \boldsymbol{a}^3 \\ \mathbf{B}^1\underline{\boldsymbol{\omega}}^{1,0} + \boldsymbol{a}^1 + \mathbf{B}^2\underline{\boldsymbol{\omega}}^{2,0} + \boldsymbol{a}^2 \\ \mathbf{B}^3\underline{\boldsymbol{\omega}}^{3,0} + \boldsymbol{a}^3 \\ \mathbf{B}^2\underline{\boldsymbol{\omega}}^{2,0} + \boldsymbol{a}^2 \\ \mathbf{B}^1\underline{\boldsymbol{\omega}}^{1,0} + \boldsymbol{a}^1 \end{pmatrix},$$

where $\underline{\boldsymbol{\omega}}^{1,0} = \left(\underline{\boldsymbol{\psi}}^{1,0} + \underline{\boldsymbol{\psi}}^{12,0} + \underline{\boldsymbol{\psi}}^{13,0} + \underline{\boldsymbol{\psi}}^{123,0}\right)$, $\underline{\boldsymbol{\omega}}_{2,0} = \left(\underline{\boldsymbol{\psi}}^{2,0} + \underline{\boldsymbol{\psi}}^{12,0} + \underline{\boldsymbol{\psi}}^{23,0} + \underline{\boldsymbol{\psi}}^{123,0}\right)$, and $\underline{\boldsymbol{\omega}}^{3,0} = \left(\underline{\boldsymbol{\psi}}^{3,0} + \underline{\boldsymbol{\psi}}^{13,0} + \underline{\boldsymbol{\psi}}^{23,0} + \underline{\boldsymbol{\psi}}^{123,0}\right)$. Furthermore,

$$\mathbf{B}^k\underline{\boldsymbol{\omega}}^{k,0} + \boldsymbol{a}^k = \begin{pmatrix} \boldsymbol{\Omega}_{:1}^{k,0\prime}\mathbf{S}_{:2}^k + \boldsymbol{\Omega}_{:2}^{k,0\prime}\mathbf{S}_{:1}^k \\ \boldsymbol{\Omega}_{:1}^{k,0\prime}\mathbf{S}_{:3}^k + \boldsymbol{\Omega}_{:3}^{k,0\prime}\mathbf{S}_{:1}^k \\ \vdots \\ \boldsymbol{\Omega}_{:p-1}^{k,0}{}'\mathbf{S}_{:p}^k + \boldsymbol{\Omega}_{:p}^{k,0\prime}\mathbf{S}_{:p-1}^k \end{pmatrix}, \qquad k = 1, 2, 3.$$

Now, we rewrite $\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}$ as,

$$\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a} = \begin{pmatrix} \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{I}_{\frac{p(p-1)}{2}} \\ \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{I}_{\frac{p(p-1)}{2}} & \\ \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{0} & \mathbf{I}_{\frac{p(p-1)}{2}} \\ \mathbf{0} & \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{I}_{\frac{p(p-1)}{2}} \\ \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\frac{p(p-1)}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{\frac{p(p-1)}{2}} \end{pmatrix} \begin{pmatrix} \mathbf{B}^3\underline{\boldsymbol{\omega}}^{3,0} + \boldsymbol{a}^3 \\ \mathbf{B}^2\underline{\boldsymbol{\omega}}^{2,0} + \boldsymbol{a}^2 \\ \mathbf{B}^1\underline{\boldsymbol{\omega}}^{1,0} + \boldsymbol{a}^1 \end{pmatrix}.$$

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

$$\begin{aligned} \|\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\|_{\max} &\leq \sqrt{\sum_{k=1}^{K} \|\mathbf{B}^k\underline{\boldsymbol{\omega}}^{k,0} + \boldsymbol{a}^k\|_{\max}^2} \\ &\leq \sqrt{K \max_{\substack{1\leq k\leq K \\ 1\leq i<j\leq p}} \left(\boldsymbol{\Omega}_{:i}^{k,0\prime}\mathbf{S}_{:j}^k\right)^2} \\ &\leq 2\sqrt{K \max_{\substack{1\leq k\leq K \\ 1\leq i<j\leq p}} \left\{\left|\bar{\boldsymbol{\Omega}}_{:i}^{k,0}\left(\mathbf{S}_{:j}^k - \boldsymbol{\Sigma}_{:j}^k\right)\right|^2 + \left|\left(\boldsymbol{\Omega}_{:i}^{k,0} - \bar{\boldsymbol{\Omega}}_{:i}^{k,0}\right)'\mathbf{S}_{:j}^k\right|^2\right\}} \\ &\leq 2\left(c + \frac{1}{\widetilde{\epsilon}_0}\right)\sqrt{K}\left(\sqrt{\frac{\log p}{n} + d_t D_n^2}\right). \end{aligned} \tag{F.24}$$

Moreover, by (A.7), it is easy to see that

$$\|\hat{\boldsymbol{a}} - \boldsymbol{a}\|_{\max} \leq 2KC\|\boldsymbol{\Upsilon}\|_{\max}\sqrt{\frac{\log p}{n}}$$

also, $\|\boldsymbol{\Upsilon}\|_{\max} \leq \max_{\boldsymbol{\ell}\in\mathcal{L}} \|\boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}\|_{\max}$ hence, by applying Lemma 2, we have that

$$\|\hat{\boldsymbol{a}} - \boldsymbol{a}\|_{\max} \leq \frac{5CK^2}{\varepsilon_0}\sqrt{\frac{\log p}{n}} \tag{F.25}$$

Therefore, by combining (F.23), (F.24), and (F.25),

$$\|\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\|_{\max} \leq c_0 \left(\sqrt{\frac{\log p}{n} + d_t D_n^2}\right),$$

where $c_0 = 2\left(c + \frac{1}{\tilde{\varepsilon}_0}\right)\sqrt{K} + \frac{5CK^2}{\varepsilon_0}$.

□ □

Let

$$a_2 = \frac{16\max(1, c_0)^2}{\min(1, \tilde{\varepsilon_0})} \text{ and } a_3 = \frac{16\max(1, c_0)^2}{\|\boldsymbol{\Lambda}\|_{min}}, \tag{F.26}$$

where $\|\boldsymbol{\Lambda}\|_{min}$ denotes the smallest diagonal entry of $\boldsymbol{\Lambda}$. For simplicity in writing, we denote the ratio of the posterior probabilities of any sparsity pattern/model $\boldsymbol{\ell}$ and the true sparsity pattern/model $\boldsymbol{t}$, by $PR(\boldsymbol{\ell}, \boldsymbol{t})$, i.e.

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) = \frac{P\left\{\boldsymbol{\ell}|\hat{\boldsymbol{\Delta}}, \mathcal{Y}\right\}}{P\left\{\boldsymbol{t}|\hat{\boldsymbol{\Delta}}, \mathcal{Y}\right\}}, \qquad \text{for any sparsity pattern} \quad \boldsymbol{\ell} \neq \boldsymbol{t}. \tag{F.27}$$

Lemma 4. *The ratio of the posterior probabilities of any sparsity pattern/model $\boldsymbol{\ell}$ and the true*

sparsity pattern/model $\boldsymbol{t}$ satisfies:

$$PR(\boldsymbol{\ell},\boldsymbol{t}) = \frac{P\left\{\boldsymbol{\ell}|\hat{\boldsymbol{\Delta}},\mathcal{Y}\right\}}{P\left\{\boldsymbol{t}|\hat{\boldsymbol{\Delta}},\mathcal{Y}\right\}} = \frac{\left[q_1^{d_{\boldsymbol{\ell}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{\ell}}} I_{\{d_{\boldsymbol{\ell}}\le\tau\}} + q_2^{d_{\boldsymbol{\ell}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{\ell}}} I_{\{d_{\boldsymbol{\ell}}>\tau\}}\right]}{\left[q_1^{d_{\boldsymbol{t}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{t}}} I_{\{d_{\boldsymbol{t}}\le\tau\}} + q_2^{d_{\boldsymbol{t}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{t}}} I_{\{d_{\boldsymbol{t}}>\tau\}}\right]} \times (2\pi)^{(d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}})/2} \frac{|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{tt}|^{\frac{1}{2}}}{|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}|^{\frac{1}{2}}} \frac{\exp\left\{\frac{n^2}{2}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right\}}{\exp\left\{\frac{n^2}{2}\hat{\boldsymbol{a}}_{t}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{tt}^{-1}\hat{\boldsymbol{a}}_{t}\right\}}. \tag{F.28}$$

Proof. We note that

$$P\left\{\boldsymbol{\ell}|\hat{\boldsymbol{\Delta}},\mathcal{Y}\right\} = P\left\{\boldsymbol{\Theta}\in\mathcal{M}_{\boldsymbol{\ell}}|\hat{\boldsymbol{\Delta}},\mathcal{Y}\right\} = \int_{\mathcal{M}_{\boldsymbol{\ell}}} \pi\left(\boldsymbol{\Theta}|\hat{\boldsymbol{\Delta}},\mathcal{Y}\right) d\boldsymbol{\Theta},$$

hence, in view of (4.4),

$$P\left\{\boldsymbol{\ell}|\hat{\boldsymbol{\Delta}},\mathcal{Y}\right\} = C_0\left[q_1^{d_{\boldsymbol{\ell}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{\ell}}} I_{\{d_{\boldsymbol{\ell}}\le\tau\}} + q_2^{d_{\boldsymbol{\ell}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{\ell}}} I_{\{d_{\boldsymbol{\ell}}>\tau\}}\right] \frac{(2\pi)^{d_{\boldsymbol{\ell}}/2}}{|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}|^{\frac{1}{2}}} \exp\left\{\frac{n^2}{2}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right\},$$

where the last equality is achieved using the properties of the multivariate normal distribution.

□ □

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

Lemma 5. *Suppose $\boldsymbol{\ell}\subset\boldsymbol{t}$ then, under Assumptions 1-5,*

$$PR(\boldsymbol{\ell},\boldsymbol{t})\to 0, \qquad \textit{as} \quad n\to\infty. \tag{F.29}$$

Proof. By Assumption 1, $d_{\boldsymbol{t}} < \tau$, hence $d_{\boldsymbol{\ell}} < d_{\boldsymbol{t}} < \tau$. Now,

$$
\begin{aligned}
PR\left(\boldsymbol{\ell},\boldsymbol{t}\right) &= \left(\frac{\sqrt{2\pi}q_1}{1-q_1}\right)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \frac{\left\|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{tt}}\right\|^{\frac{1}{2}}}{\left\|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}\right\|^{\frac{1}{2}}} \frac{\exp\left\{\frac{n^2}{2}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}'\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right\}}{\exp\left\{\frac{n^2}{2}\hat{\boldsymbol{a}}_{\boldsymbol{t}}'\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{tt}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{t}}\right\}} \\
&= \left(\frac{\sqrt{2\pi}q_1}{1-q_1}\right)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \frac{\left\|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{tt}}\right\|^{\frac{1}{2}}}{\left\|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}\right\|^{\frac{1}{2}}} \\
&\quad \exp\left\{-\frac{n^2}{2}\left[\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right]'\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}|\boldsymbol{\ell}}^{-1}\left[\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right]\right\},
\end{aligned}
$$

that is,

$$
PR\left(\boldsymbol{\ell},\boldsymbol{t}\right) \leq \left(\frac{\sqrt{2\pi}q_1}{1-q_1}\right)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \frac{\left\|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{tt}}\right\|^{\frac{1}{2}}}{\left\|\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}\right\|^{\frac{1}{2}}} \exp\left\{-\frac{n^2\|\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\|^2}{2\mathrm{eig}_{max}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{tt}}}\right\},
$$

Now, by the triangular inequality,

$$
\begin{aligned}
&\|\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\| \\
&\qquad \geq \|\boldsymbol{a}_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\boldsymbol{a}_{\boldsymbol{\ell}}\| \\
&\qquad\quad -\left\|\left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-\boldsymbol{a}_{\boldsymbol{\ell}^c}\right)-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}-\boldsymbol{a}_{\boldsymbol{\ell}}\right)\right\| \\
&\qquad = \left\|\left(\pm\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0+\boldsymbol{a}\right)_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\left(\pm\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0+\boldsymbol{a}\right)_{\boldsymbol{\ell}}\right\| \\
&\qquad\quad -\left\|\left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-\boldsymbol{a}_{\boldsymbol{\ell}^c}\right)-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}-\boldsymbol{a}_{\boldsymbol{\ell}}\right)\right\| \\
&\qquad \geq \left\|\left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{\ell}}\right\| \\
&\qquad\quad -\left\|\left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0+\boldsymbol{a}\right)_{\boldsymbol{\ell}^c}-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0+\boldsymbol{a}\right)_{\boldsymbol{\ell}}\right\| \\
&\qquad\quad -\left\|\left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c}-\boldsymbol{a}_{\boldsymbol{\ell}^c}\right)-n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell\ell}}^{-1}\left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}-\boldsymbol{a}_{\boldsymbol{\ell}}\right)\right\|.
\end{aligned}
\tag{F.30}
$$

Now, by appropriately partitioning $\boldsymbol{\Upsilon}$, we can write $(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0)_{\boldsymbol{\ell}^c} = \boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + \boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}^c}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c}$ and $(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0)_{\boldsymbol{\ell}} = \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}^c}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c}$. Hence, for large enough n,

$$
\begin{aligned}
&\| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{\ell}^c} - n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}}\left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{\ell}} \| \\
&\qquad = \|\frac{1}{n}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}|\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c} - \boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}}\| \\
&\qquad \geq \|\frac{1}{n}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}|\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c}\| - \|\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}}\| \\
&\qquad \geq \|\frac{1}{n}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}|\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c}\| - \frac{\text{eig}_{\min}\left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\right)\|\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}}\|}{\text{eig}_{\min}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}} \\
&\qquad \geq \|\frac{1}{n}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}|\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c}\| - \frac{2\|\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}}\|}{n\varepsilon_0^2} \\
&\qquad \geq \frac{1}{2}\|\frac{1}{n}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}|\boldsymbol{\ell}}\boldsymbol{\Theta}^0_{\boldsymbol{\ell}^c}\| \\
&\qquad \geq \frac{1}{2}\frac{1}{n}\text{eig}_{\min}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{t}\boldsymbol{t}} s_n\sqrt{(d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}})} \\
&\qquad \geq \frac{1}{2}\frac{1}{n}n\text{eig}_{\min}\left(\boldsymbol{\Upsilon}\right)_{tt} s_n\sqrt{(d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}})} \\
&\qquad \geq \frac{3}{8}\varepsilon_0 s_n\sqrt{(d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}})}
\end{aligned}
\tag{F.31}
$$

Moving onto the second term in the right hand side of (F.30),

$$
\begin{aligned}
&\| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}^c} - n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}} \left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}} \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}} \| \\
&\leq \| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}^c} \| + \| n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}} \left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}} \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}} \| \\
&\leq \| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}^c} \| + \frac{n\text{eig}_{\max}\left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\right) \| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}} \|}{\text{eig}_{\min}\left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}} \\
&\leq \| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}^c} \| + \frac{2\| \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a}\right)_{\boldsymbol{\ell}} \|}{\varepsilon_0^2} \\
&\leq c_0 \sqrt{\frac{\log p}{n} + d_t D_n^2} \left(\sqrt{d_{\boldsymbol{t}} - d_{\boldsymbol{\ell}}} + \frac{2\sqrt{d_{\boldsymbol{\ell}}}}{\varepsilon_0^2} \right),
\end{aligned}
\tag{F.32}
$$

where the last equality was achieved by Lemma 3. Further, regarding the third term in right hand side of (F.30) we can write

$$
\begin{aligned}
&\| \left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c} - \boldsymbol{a}_{\boldsymbol{\ell}^c}\right) - n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}} \left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}} \left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} - \boldsymbol{a}_{\boldsymbol{\ell}}\right) \| \\
&\qquad \leq \|\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c} - \boldsymbol{a}_{\boldsymbol{\ell}^c}\| + \frac{n\text{eig}_{\max}\left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\right) \| \left(\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} - \boldsymbol{a}_{\boldsymbol{\ell}}\right) \|}{\text{eig}_{\min}\left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}} \\
&\qquad \leq \frac{5CK^2}{\varepsilon_0} \sqrt{\frac{\log p}{n}} \left(\sqrt{d_{\boldsymbol{t}} - d_{\boldsymbol{\ell}}} + \frac{2\sqrt{d_{\boldsymbol{\ell}}}}{\varepsilon_0^2} \right).
\end{aligned}
\tag{F.33}
$$

By combining (F.30), (F.31), (F.32), and (F.33), for sufficiently large n, we have that

$$\begin{aligned}
\|\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c} - n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\| \geq & \frac{1}{2}\varepsilon_0 s_n\sqrt{(d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}})} \\
& - c_0\sqrt{\frac{\log p}{n}+d_t D_n^2}\left(\sqrt{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}}+\frac{2\sqrt{d_{\boldsymbol{\ell}}}}{\varepsilon_0^2}\right) \\
& - \frac{5CK^2}{\varepsilon_0}\sqrt{\frac{\log p}{n}}\left(\sqrt{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}}+\frac{2\sqrt{d_{\boldsymbol{\ell}}}}{\varepsilon_0^2}\right) \\
\geq & \frac{1}{2}\varepsilon_0 s_n\sqrt{(d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}})} \\
& - \left(c_0+\frac{5CK^2}{\varepsilon_0}\right)\sqrt{\frac{\log p}{n}+d_t D_n^2}\left(\sqrt{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}}+\frac{2\sqrt{d_{\boldsymbol{\ell}}}}{\varepsilon_0^2}\right) \\
\geq & \frac{1}{2}\varepsilon_0 s_n - \left(c_0+\frac{5CK^2}{\varepsilon_0}\right)\sqrt{\frac{\log p}{n}+d_t D_n^2}\left(\frac{2\sqrt{d_{\boldsymbol{t}}}}{\varepsilon_0^2}\right),
\end{aligned}$$

in view of Assumption 4, $\frac{\frac{1}{2}\varepsilon_0 s_n}{\left(c_0+\frac{5CK^2}{\varepsilon_0}\right)\sqrt{\frac{\log p}{n}+d_t D_n^2}\left(\frac{2\sqrt{d_{\boldsymbol{t}}}}{\varepsilon_0^2}\right)} \to \infty$, as $n \to \infty$, hence, for all large n, we can write,

$$\|\hat{\boldsymbol{a}}_{\boldsymbol{\ell}^c} - n\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}}\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}}\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\| \geq \frac{1}{4}\varepsilon_0 s_n$$

Now, once again by Lemma 3

$$\begin{aligned}
PR(\boldsymbol{\ell},\boldsymbol{t}) &\leq (2\sqrt{2\pi}q_1)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} n^{\frac{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}}{2}} \exp\left\{-\frac{n^2\frac{1}{64}\varepsilon_0^2 s_n^2}{6Kn\varepsilon_0^{-1}}\right\} \\
&= (2\sqrt{2\pi})^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}}\left(\frac{\sqrt{n}}{q_1}\exp\left\{-\frac{n\varepsilon_0^3 s_n^2}{384K}\right\}\right)^{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}} \\
&= (2\sqrt{2\pi})^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}}\left(\frac{\sqrt{n}}{q_1}\exp\left\{-2a_1 n s_n^2\right\}\right)^{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}}
\end{aligned}$$

by Assumption 4, for all large n,

$$\begin{aligned} PR(\boldsymbol{\ell}, \boldsymbol{t}) &\leq (2\sqrt{2\pi})^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \left(\frac{\sqrt{n}}{q_1} \exp\left\{ -\log n - 2a_2 d_{\boldsymbol{t}} \log p \right\} \right)^{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}} \\ &= (2\sqrt{2\pi})^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \left(\frac{p^{-2a_2 d_{\boldsymbol{t}}}}{\sqrt{n} q_1} \right)^{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}} \\ &= (2\sqrt{2\pi})^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \left(\frac{p^{-a_2 d_{\boldsymbol{t}}}}{\sqrt{n}} \right)^{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}} \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

□ □

Lemma 6. *Suppose $\boldsymbol{\ell} \supset \boldsymbol{t}$, and $d_{\boldsymbol{\ell}} < \tau_n$ then, under Assumptions 1-5,*

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

Proof. In this case,

$$\begin{aligned} PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq & \frac{\left(2\sqrt{2\pi} q_1\right)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}}}{\| \left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}|\boldsymbol{t}} \|^{\frac{1}{2}}} \\ & \exp\left\{ \frac{1}{2} \left[(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} \right]' (n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}} \left[(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} \right] \right\}. \end{aligned}$$

Now, we note that

$$(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} = n \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \boldsymbol{a} \right)_{\boldsymbol{\ell}} + \boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}}, + n \left(\hat{\boldsymbol{a}} - \boldsymbol{a} \right)_{\boldsymbol{\ell}}$$

each entry of the above vector in absolute value is smaller than

$$nc_0 \sqrt{\frac{\log p}{n} + d_t D_n^2} + \frac{\|\boldsymbol{\Lambda}\|_{\max}}{\varepsilon_0} + \frac{5CK^2}{\varepsilon_0} \sqrt{\frac{\log p}{n}} \leq \frac{3nc_0}{2} \sqrt{\frac{\log p}{n} + d_t D_n^2}$$

for large enough n. Hence, by Lemma 2,

$$\begin{aligned} & \left[(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} \right]' (n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}} \left[(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}} \right] \\ \leq & \frac{1}{nK\varepsilon_0} d_{\boldsymbol{\ell}} \frac{4n^2 c_0^2 (\log p + n d_t D_n^2)}{n} = \frac{4c_0^2 d_{\boldsymbol{\ell}} (\log p + n d_t D_n^2)}{K\varepsilon_0}. \end{aligned}$$

Hence,

$$PR(\boldsymbol{\ell},\boldsymbol{t}) \leq \frac{\left(2\sqrt{2\pi}q_1\right)^{d_{\boldsymbol{\ell}}-d_t}}{\left(\frac{nK\varepsilon_0}{2}\right)^{\frac{d_{\boldsymbol{\ell}}}{2}}} \exp\left\{\frac{2c_0^2}{K\varepsilon_0} d_{\boldsymbol{\ell}}(\log p + nd_t D_n^2)\right\}. \tag{F.34}$$

Using $d_{\boldsymbol{t}} < d_{\boldsymbol{\ell}} < 2d_{\boldsymbol{t}}(d_{\boldsymbol{\ell}} - d_{\boldsymbol{t}})$ along with Assumption 5, we get that

$$PR(\boldsymbol{\ell},\boldsymbol{t} \leq \left(\frac{C_1 p^{-a_2 d_{\boldsymbol{t}}/2}}{\sqrt{n}}\right)^{d_{\boldsymbol{t}}-d_{\boldsymbol{\ell}}} \tag{F.35}$$

for an appropriate constant C_1. Thus, by (F.34) and (F.35) we have that

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

□ □

Lemma 7. *Suppose $\boldsymbol{\ell} \supset \boldsymbol{t}$, and $d_{\boldsymbol{\ell}} > \tau_n$ then, under Assumptions 1-5,*

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

Proof. When $\boldsymbol{\ell} \supset t$,

$$PR\left(\boldsymbol{\ell},\boldsymbol{t}\right) \leq (2\pi)^{(d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}})/2} \frac{q_2^{d_{\boldsymbol{\ell}}}\left(1-q_2\right)^{\binom{p}{2}-d_{\boldsymbol{\ell}}}}{q_1^{d_{\boldsymbol{t}}}\left(1-q_1\right)^{\binom{p}{2}-d_{\boldsymbol{t}}}} \frac{1}{\parallel \left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}|\boldsymbol{t}} \parallel^{\frac{1}{2}}}$$
$$\exp\left\{\frac{1}{2}\left[\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}_{\boldsymbol{\ell}}^0 + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right]' \left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\left[\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}_{\boldsymbol{\ell}}^0 + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right]\right\},$$

similar to the argument in Lemma 6, each entry of the vector $\left(n\boldsymbol{\Upsilon}+\boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}\boldsymbol{\Theta}_{\boldsymbol{\ell}}^0 + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}$, in absolute value, is smaller than $2nc_0\sqrt{\frac{\log p}{n} + d_t D_n^2}$. Now, since $\boldsymbol{\Upsilon}$ is non-negative definite (note that in

the case of $d_{\boldsymbol{\ell}} > \tau_n$, $\boldsymbol{\Upsilon}$ is not necessarily positive definite) we have that $\text{eig}_{\min}(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \geq \text{eig}_{\min}(\boldsymbol{\Lambda}) = \|\boldsymbol{\Lambda}\|_{\min}$, hence for large enough n

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq \left(\frac{C_3 q_2}{1-q_2}\right)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \frac{q_2^{d_{\boldsymbol{t}}}(1-q_2)^{\binom{p}{2}-d_t}}{q_1^{d_{\boldsymbol{t}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{t}}}} \exp\left\{\frac{4c_0^2 n^2 d_{\boldsymbol{\ell}}(\log p + n d_t D_n^2)}{n\|\boldsymbol{\Lambda}\|_{\min}}\right\}$$

for an appropriate constant C_3. Since the function $q^{d_{\boldsymbol{t}}}(1-q)^{\binom{p}{2}-d_{\boldsymbol{t}}}$ is globally maximized at $\hat{q} = \frac{d_{\boldsymbol{t}}}{\binom{p}{2}}$ and $q_2 < q_1 < \hat{q}$,

$$\frac{q_2^{d_{\boldsymbol{t}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{t}}}}{q_1^{d_{\boldsymbol{t}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{t}}}} \leq 1.$$

Hence,

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq (2C_3 q_2)^{d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \exp\left\{\frac{4c_0^2 n d_{\boldsymbol{\ell}}(\log p + n d_t D_n^2)}{\|\boldsymbol{\Lambda}\|_{\min}}\right\},$$

since $d_{\boldsymbol{\ell}} > \tau$, and by Assumption 1, $d_{\boldsymbol{t}} < \frac{\tau}{2}$, we have that,

$$d_{\boldsymbol{\ell}} - d_{\boldsymbol{t}} \geq \frac{d_{\boldsymbol{\ell}}}{2}.$$

Hence, for large enough n, we have

$$\begin{aligned} PR(\boldsymbol{\ell}, \boldsymbol{t}) &\leq \left[C_3 q_2 \exp\left\{\frac{4c_0^2 n(\log p + n d_t D_n^2)}{\|\boldsymbol{\Lambda}\|_{\min}}\right\}\right]^{\frac{d_{\boldsymbol{\ell}}}{2}} \\ &\leq (C_3 p)^{-6n d_{\boldsymbol{\ell}}} \\ &\leq (C_3 p)^{-6n(d_{\boldsymbol{\ell}}-d_{\boldsymbol{t}}} \to 0, \qquad \text{as} \quad n \to \infty. \end{aligned}$$

□ □

Now, let,

$$f_n = \max\left\{\left(\sqrt{\frac{q_1}{n}}\right), q_2^{1/4}, \frac{q_1^{1/4d_{\boldsymbol{t}}}}{\sqrt{nK\varepsilon_0}}\right\}. \tag{F.36}$$

Lemma 8. *Let $\boldsymbol{\ell} \in \mathcal{L}$ such that $\boldsymbol{\ell} \not\subset \boldsymbol{t}$, $\boldsymbol{t} \not\subset \boldsymbol{\ell}$, and $\boldsymbol{\ell} \neq \boldsymbol{t}$, then under Assumptions 1-5, for sufficiently large n,*

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

Proof. Suppose $\boldsymbol{\ell}$ is such that $\tau_n \geq d_{\boldsymbol{\ell}} > d_{\boldsymbol{t}}$. Let $\tilde{\boldsymbol{\ell}}$ denote the union of $\boldsymbol{\ell}$ and $\boldsymbol{t}$. Then $\tilde{\boldsymbol{\ell}} \supset \boldsymbol{t}$ and

$$d_{\tilde{\boldsymbol{\ell}}} \leq d_{\boldsymbol{\ell}} + d_{\boldsymbol{t}} \leq \tau_n + d_{\boldsymbol{t}}.$$

Using arguments very similar to the proof of Lemma 6, the fact $\boldsymbol{\Theta}^0_{\tilde{\boldsymbol{\ell}} \setminus \boldsymbol{t}} = \mathbf{0}$, and the form of the inverse of a partitioned matrix, we get that

$$\begin{aligned}
&\left[(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right]' (n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})^{-1}_{\boldsymbol{\ell}\boldsymbol{\ell}} \left[(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda})_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}^0_{\boldsymbol{\ell}} + n\hat{\boldsymbol{a}}_{\boldsymbol{\ell}}\right] \\
&\leq \frac{1}{nK\varepsilon_0} d_{\tilde{\boldsymbol{\ell}}} \frac{4n^2 c_0^2 (\log p + n d_t D_n^2)}{n} \leq \frac{4c_0^2 (d_{\boldsymbol{\ell}} + d_{\boldsymbol{t}})(\log p + n d_t D_n^2)}{K\varepsilon_0}.
\end{aligned}$$

Using $2d_{\boldsymbol{t}} < d_{\boldsymbol{\ell}} + d_{\boldsymbol{t}} < 3d_{\boldsymbol{t}}(d_{\boldsymbol{\ell}} - d_{\boldsymbol{t}})$ along with Assumption 5, we get that

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq \left(\frac{C_2 p^{-a_2 d_{\boldsymbol{t}}/4}}{\sqrt{n}} \right)^{d_{\boldsymbol{t}} - d_{\boldsymbol{\ell}}} \tag{F.37}$$

for an appropriate constant C_2. Let $D(\boldsymbol{\ell}, \boldsymbol{t})$ denotes the total number of disagreements between $\boldsymbol{\ell}$ and $\boldsymbol{t}$. Note that if $d_{\boldsymbol{\ell}} > d_{\boldsymbol{t}}$, then

$$D(\boldsymbol{\ell}, \boldsymbol{t}) \leq 2d_{\boldsymbol{t}}(d_{\boldsymbol{\ell}} - d_{\boldsymbol{t}}).$$

Hence

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq \left(\frac{C_2 p^{\frac{-a_2}{8}}}{\sqrt{n}} \right)^{D(\boldsymbol{\ell}, \boldsymbol{t})}.$$

Suppose $\boldsymbol{\ell}$ is such that $d_{\boldsymbol{\ell}} \leq d_{\boldsymbol{t}}$. Note that

$$\begin{aligned}
& \hat{\boldsymbol{a}}_{\boldsymbol{\ell}}' \left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \hat{\boldsymbol{a}}_{\boldsymbol{\ell}} - \hat{\boldsymbol{a}}_{\boldsymbol{t}}' \left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1} \hat{\boldsymbol{a}}_{\boldsymbol{t}} \\
= \; & \hat{\boldsymbol{a}}_{\boldsymbol{\ell}}' \left(\left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} - \left(n\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\right) \hat{\boldsymbol{a}}_{\boldsymbol{\ell}} - \hat{\boldsymbol{a}}_{\boldsymbol{t}}' \left(\left(n\boldsymbol{\Upsilon} + \boldsymbol{\Lambda}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1} - \left(n\boldsymbol{\Upsilon}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1}\right) \hat{\boldsymbol{a}}_{\boldsymbol{t}} + \\
& \frac{1}{n} \hat{\boldsymbol{a}}_{\boldsymbol{\ell}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \hat{\boldsymbol{a}}_{\boldsymbol{\ell}} - \frac{1}{n} \hat{\boldsymbol{a}}_{\boldsymbol{t}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1} \hat{\boldsymbol{a}}_{\boldsymbol{t}} \\
= \; & O\left(\frac{d_{\boldsymbol{t}}}{n^2}\right) + \frac{1}{n} \hat{\boldsymbol{a}}_{\boldsymbol{\ell}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \hat{\boldsymbol{a}}_{\boldsymbol{\ell}} - \frac{1}{n} \hat{\boldsymbol{a}}_{\boldsymbol{t}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1} \hat{\boldsymbol{a}}_{\boldsymbol{t}}
\end{aligned}$$

and by Lemma 3

$$\begin{aligned}
& \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\right)_{\boldsymbol{\ell}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\right)_{\boldsymbol{\ell}} + \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\right)_{\boldsymbol{t}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1} \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0 + \hat{\boldsymbol{a}}\right)_{\boldsymbol{t}} \\
= \; & O\left(d_{\boldsymbol{t}} \left(\frac{\log p}{n} + d_t D_n^2\right)\right)
\end{aligned}$$

on $C_{1,n}$. Let $\boldsymbol{\ell}^c$ denote the sparsity pattern which has a zero/one whenever the corresponding entry in $\boldsymbol{\ell}$ is one/zero. Using $\boldsymbol{\Theta}_{\boldsymbol{t}^c}^0 = \mathbf{0}$ and Lemma 2, it follows that

$$\begin{aligned}
& \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{\ell}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{\ell}} - \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{t}}' \left(\boldsymbol{\Upsilon}\right)_{\boldsymbol{t}\boldsymbol{t}}^{-1} \left(\boldsymbol{\Upsilon}\boldsymbol{\Theta}^0\right)_{\boldsymbol{t}} \\
= \; & \boldsymbol{\Theta}_{\boldsymbol{\ell}}^{0\prime} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}} \boldsymbol{\Theta}_{\boldsymbol{\ell}}^0 + \boldsymbol{\Theta}_{\boldsymbol{\ell}}^{0\prime} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}^c} \boldsymbol{\Theta}_{\boldsymbol{\ell}^c}^0 + \boldsymbol{\Theta}_{\boldsymbol{\ell}^c}^{0\,\prime} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}^c} \boldsymbol{\Theta}_{\boldsymbol{\ell}^c}^0 - \boldsymbol{\Theta}_{\boldsymbol{t}}^{0\prime} \boldsymbol{\Upsilon}_{\boldsymbol{t}\boldsymbol{t}} \boldsymbol{\Theta}_{\boldsymbol{t}}^0 \\
= \; & \boldsymbol{\Theta}^{0\prime} \boldsymbol{\Upsilon} \boldsymbol{\Theta}^0 - \boldsymbol{\Theta}^{0\prime} \boldsymbol{\Upsilon} \boldsymbol{\Theta}^0 - \boldsymbol{\Theta}_{\boldsymbol{\ell}^c}^{0\,\prime} \left(\boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}^c} - \boldsymbol{\Upsilon}_{\boldsymbol{\ell}^c\boldsymbol{\ell}} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} \boldsymbol{\Upsilon}_{\boldsymbol{\ell}\boldsymbol{\ell}^c}\right) \boldsymbol{\Theta}_{\boldsymbol{\ell}^c}^0 \\
\leq \; & -\frac{3 d_{\boldsymbol{t}\cap\boldsymbol{\ell}^c} \tilde{\varepsilon}_0 s_n^2}{4,}
\end{aligned}$$

since exactly $d_{\boldsymbol{t}\cap\boldsymbol{\ell}^c}$ entries in $\boldsymbol{\Theta}_{\boldsymbol{\ell}^c}^0$ are non-zero. Since $d_{\boldsymbol{t}\cap\boldsymbol{\ell}^c} \geq d_{\boldsymbol{t}} - d_{\boldsymbol{\ell}}$ and $D(\boldsymbol{\ell}, \boldsymbol{t}) \leq 2d_{\boldsymbol{t}}$ similar arguments to those at the end of Lemma 5 can be used to obtain

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq \left(\frac{2C_1 p^{-a_2 d_{\boldsymbol{t}}}}{\sqrt{n}}\right)^{d_{\boldsymbol{t}\cap\boldsymbol{\ell}^c}} \leq \left(\frac{C_1 p^{-a_2/2}}{\sqrt{n}}\right)^{D(\boldsymbol{\ell},\boldsymbol{t})}.$$

□ □

It follows by Lemmas 5, 6 and 8 that for every $\boldsymbol{\ell} \neq \boldsymbol{t}$ with $d_{\boldsymbol{\ell}} \leq \tau_n$,

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq \left(\frac{\max(C_1, C_2) p^{\frac{-a_2}{8}}}{\sqrt{n}} \right)^{D(\boldsymbol{\ell}, \boldsymbol{t})} \leq \left(\frac{\max(C_1, C_2) p^{\frac{-a_2}{8}}}{\sqrt{n}} \right)^{D(\boldsymbol{\ell}, \boldsymbol{t})} \leq \left(\frac{\max(C_1, C_2) p^{-2}}{\sqrt{n}} \right)^{D(\boldsymbol{\ell}, \boldsymbol{t})}$$

Also, by arguments similar to Lemma 7, it follows that for every $\boldsymbol{\ell} \neq \boldsymbol{t}$ with $d_{\boldsymbol{\ell}} > \tau_n$

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq (C_3 p)^{-3n(d_{\boldsymbol{\ell}} - d_{\boldsymbol{t}})} \leq (C_3 p)^{-3D(\boldsymbol{\ell}, \boldsymbol{t})}$$

for large enough n. Hence, for every $\boldsymbol{\ell} \neq \boldsymbol{t}$, we get

$$PR(\boldsymbol{\ell}, \boldsymbol{t}) \leq f_n^{D(\boldsymbol{\ell}, \boldsymbol{t})}$$

where $p^2 f_n \to 0$ as $n \to \infty$.

Proof of Theorem 1.

$$\begin{aligned}
\frac{1 - P\left\{\boldsymbol{\Theta} \in \mathcal{M}_{\boldsymbol{t}} | \hat{\boldsymbol{\Delta}}, \mathcal{Y}\right\}}{P\left\{\boldsymbol{\Theta} \in \mathcal{M}_{\boldsymbol{t}} | \hat{\boldsymbol{\Delta}}, \mathcal{Y}\right\}} &= \sum_{\boldsymbol{\ell} \neq \boldsymbol{t}} PR(\boldsymbol{\ell}, \boldsymbol{t}) \\
&= \sum_{\boldsymbol{\ell} \neq \boldsymbol{t}} \sum_{j=1}^{\binom{p}{2}} PR(\boldsymbol{\ell}, \boldsymbol{t}) I_{\{D(\boldsymbol{\ell}, \boldsymbol{t}) = j\}} \\
&\leq \sum_{j=1}^{\binom{p}{2}} \binom{\binom{p}{2}}{j} f_n^j \\
&\leq \sum_{j=1}^{\binom{p}{2}} \binom{p}{2}^j f_n^j \\
&\leq \sum_{j=1}^{p^2} \left(p^2 f_n\right)^j \\
&\leq \frac{p^2 f_n}{1 - p^2 f_n} \to 0, \quad \text{as} \quad n \to \infty.
\end{aligned} \tag{F.38}$$

The last two inequalities follow from the fact that $p^2 f_n < 1$ and $p^2 f_n \to 0$, which follows from (F.36) and choice of $\varepsilon_0 = \frac{c^2}{K\tilde{\varepsilon}_0}$.

□ □

G. Some tables and figures from the main paper

Figure G.1: Heat maps of the true precision matrices of the four groups in Section D.1

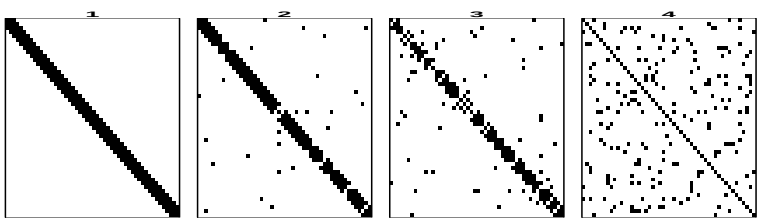


Table G.1: Number of edges in each matrix

method	subnet	Ω^1	Ω^2	Ω^3	Ω^4	Ψ^1	Ψ^2	Ψ^3	Ψ^4	Ψ^{12}	Ψ^{13}	Ψ^{24}	Ψ^{34}	Ψ^{1234}
	set1	2003	1950	2526	2486	248	201	425	391	279	479	473	625	997
BJNS	set2	514	520	662	658	48	45	103	90	60	124	133	153	282
	set12	749	770	918	929	116	107	183	164	132	157	187	234	344
	set1	2742	2934	3097	3115	NA	NA	NA	NA	NA	NA	NA	NA	NA
JEM-G	set2	1230	1201	1105	1065	NA	NA	NA	NA	NA	NA	NA	NA	NA
	set1.2	2569	2576	2253	2215	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table G.2: Uncertainty quantification: Posterior inclusion probabilities using BJNS for various subsets for 10 chosen edges for the IBD data analyzed in Section 6 in the main paper

Edge	Ψ^1	Ψ^2	Ψ^3	Ψ^4	Ψ^{12}	Ψ^{13}	Ψ^{24}	Ψ^{34}	Ψ^{1234}	Not in any group
(1-3-7-trimethylurate, 1-methylguanine)	0.002	0.001	0.001	0.001	0.002	0.002	0.002	0.001	0.002	0.984
(1-3-7-trimethylurate, 1-methylhistamine)	0	0	0	0	0	0	0.985	0	0	0.015
(1-3-7-trimethylurate, 3-methylglutaconate)	0	0	0	0	0	0	0	0.316	0.609	0.074
(1-3-7-trimethylurate, 5-acetylamino..)	0	0	0.529	0	0	0.46	0	0	0	0.011
(1-3-7-trimethylurate, ADMA/SDMA)	0	0	0	0	0	0	0	0	1	0
(1-3-7-trimethylurate, aspartate)	0	0.002	0.326	0	0.001	0.004	0	0.002	0.001	0.662
(1-3-7-trimethylurate, beta-sitosterol)	0	0	0	0	0	0	0	0.795	0.205	0
(1-methylguanine, C30:0 DAG)	0	0	0	0.563	0	0	0.437	0	0	0
(1-methylguanine, C36:1 PC plasmalogen)	0.001	0	0.062	0.001	0.002	0.001	0.001	0.188	0.005	0.737
(1-methylguanine, C45:0 TAG)	0	0	0.794	0	0	0.206	0	0	0	0

Figure G.2: Trace plots and distribution of κ during 4000 iterations of the Gibbs sampler in a simulation replication with $p = 200$ and $n = 300$.

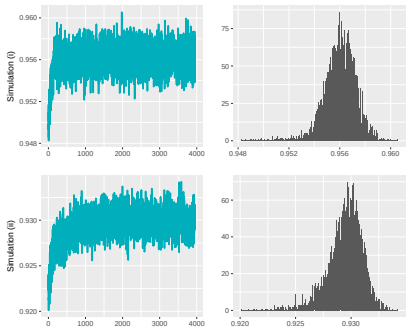


Figure G.3: Trace plots and distribution of three sample edges (one negative, one zero and one positive) during 4000 iterations of the Gibbs sampler in simulation (i) with $p = 200$ and $n = 300$.

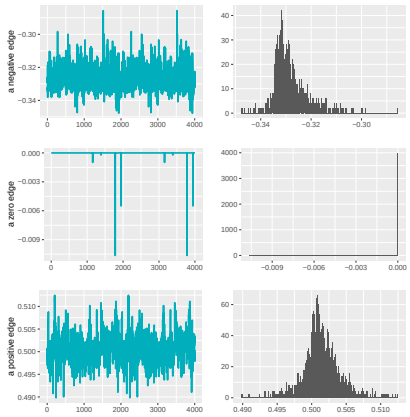
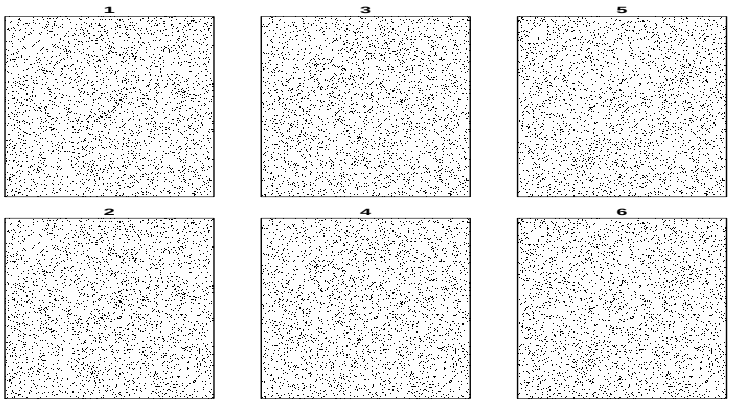
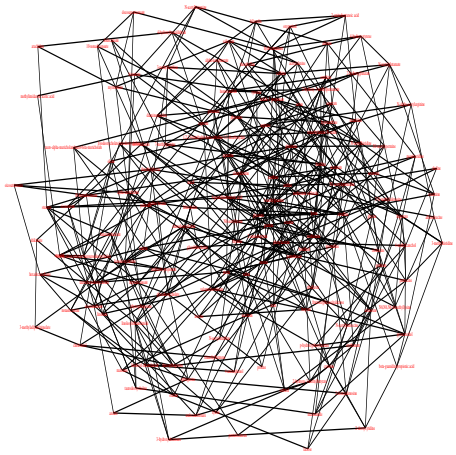
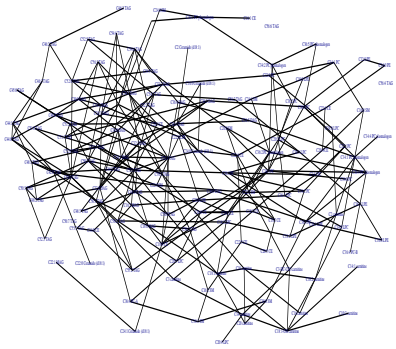


Figure G.4: Image plots of the random sparse true adjacency matrices from all six graphical models for the simulations in Section 5.1.1. Graphs in the same row share the same sparsity pattern at the bottom right block, whereas graphs in the same column share the same pattern at remaining locations

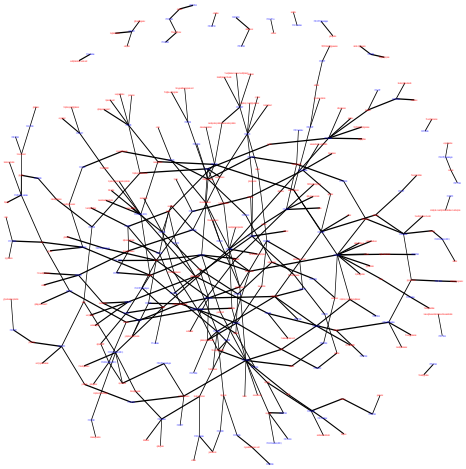




(a) Primary and secondary metabolites connectivity patterns



(b) Lipid connectivity patterns



(c) Interaction patterns between metabolites and proteins

Figure G.5: Network plot of the edges shared between the four groups (Ψ^{1234}) with 289 lipids in red and 139 proteins in blue.

References

Atchade, Y. (2019). Quasi-bayesian estimation of large gaussian graphical models. *Journal of Multivariate Analysis* **173**, 656-671.

Fan, J., S. Guo, and N. Hao (2012). Variance estimation using refitted cross-validation in ultra-high dimensional regression. *JRSSB* **74**, 37-65.

Khare, K., S.-Y. Oh, and B. Rajaratnam (2015). A convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees. *JRSSB* **77**, 803-825.

Ma, J. and Michailidis, G. (2016). Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research* **17**, 5777-5824.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104**, 735-746.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability* **18**, 1-9.