Supplement Materials: Bayesian Joint Spike-and-Slab Graphical Lasso

Zehang Richard Li [1] **Tyler H. McCormick** [2 3] **Samuel J. Clark** [4]

A. Proof of Proposition 1 and 2

Here we provide a proof of Proposition 2 in the main paper. The same arguments generalize to Proposition 1 directly by fixing all binary indicators to 1 and thus are not repeated.

Proof of DSS-GGL prior We first consider the GGL and DSS-GGL penalties. The joint distribution of all the parameters under the parameterization of the scale mixture of Normal distributions is

$$\begin{aligned} p(\{\boldsymbol{\Omega}\}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\xi}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}}) &= \prod_g \prod_{j<k} \exp(-\frac{1}{2}(\omega_{jk}^{(g)})^2(\frac{v_{\delta_{jk}}}{\tau_{jkg}} + \frac{v_{\xi^*_{jk}}}{\rho_{jk}})) \prod_g \prod_j \exp(-\frac{\lambda_0}{2}\omega_{jj}^{(g)}) \\ &\times \prod_g \prod_{j<k} \tau_{jkg}^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2}\tau_{jkg}) \prod_{j<k} \rho_{jk}^{-\frac{1}{2}} \exp(-\frac{\lambda_2^2}{2}\rho_{jk}) \\ &\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}}(1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{\xi_{jk}}(1-\pi_{\boldsymbol{\xi}})^{1-\xi_{jk}} \\ &\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{a_1-1}(1-\pi_{\boldsymbol{\delta}})^{b_1-1} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{a_2-1}(1-\pi_{\boldsymbol{\xi}})^{b_2-1} \end{aligned}$$

The following two identities provide the key steps to connect the scale mixture of normal distributions to the Laplace representation in the penalty function:

$$\int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp(-\frac{z^2}{2s})\frac{\lambda^2}{2}\exp(-\frac{\lambda^2 s}{2})ds = \frac{\lambda}{2}\exp(-\lambda|z|) \quad (1)$$

$$\int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp(-\frac{||\boldsymbol{z}||_2}{2s})(\frac{\lambda^2}{2})^{\frac{p+1}{2}}\exp(-\frac{\lambda^2 s}{2})s^{\frac{p+1}{2}-1}\frac{1}{\Gamma(\frac{p+1}{2})}ds = \frac{\lambda}{2}\exp(-\lambda||\boldsymbol{z}||_2) \quad (2)$$

where $\boldsymbol{z}$ is a vector of length p and $||z||_2 = \sqrt{\sum_{i=1}^p z_i^2}$. By rearranging the terms and plugging in the two identities above, it can be seen that

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}) \propto \exp(-\sum_{j<k}\frac{\lambda_1}{v_{\delta_{jk}}}|\omega_{jk}^{(g)}| - \sum_{j<k}\frac{\lambda_2}{v_{\xi^*_{jk}}}||\boldsymbol{\omega}_{jk}||_2 - \sum_g\sum_j\frac{\lambda_0}{2}\omega_{jj}^{(g)}).$$

The conditional distribution of $\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}$ takes the form of $\exp(-pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}))$ for DSS-GGL, and thus the mode of the posterior is equivalent to the DSS-GGL solution. It still remains to be seen that the intractable constant terms are all finite,

*Equal contribution [1]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA [2]Department of Statistics, University of Washington, Seattle, Washington, USA [3]Department of Sociology, University of Washington, Seattle, Washington, USA [4]Department of Sociology, Ohio State University, Columbus, Ohio, USA. Correspondence to: Zehang Richard Li <zehang.li@yale.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

so that each of the three conditional distributions are proper. This can be seen as follows:

$$\begin{aligned}
C_{\boldsymbol{\tau},\boldsymbol{\rho}} C_{\boldsymbol{\delta},\boldsymbol{\xi}} &= \int \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk}; 0, \boldsymbol{\Theta}) \prod_g \prod_j \text{Exp}(\omega_{jj}^{(g)}; \frac{\lambda_0}{2}) \mathbf{1}_{\{\boldsymbol{\Omega}\} \in M^+} d\{\boldsymbol{\Omega}\} \\
&< \int \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk}; 0, \boldsymbol{\Theta}) \prod_g \prod_j \text{Exp}(\omega_{jj}^{(g)}; \frac{\lambda_0}{2}) d\{\boldsymbol{\Omega}\} = 1 \\
C^{-1}_{\boldsymbol{\tau},\boldsymbol{\rho}} &\propto \int \prod_{j<k} \big(\exp(-\frac{\lambda_1^2}{2} \sum_g \tau_{jkg} - \frac{\lambda_2^2}{2} \rho_{jk}) \rho_{jk}^{-\frac{1}{2}} \prod_g (\tau_{jkg} (\frac{1}{\tau_{jkg}} + \frac{1}{\rho_{jk}}))^{-\frac{1}{2}} \big) d\{\boldsymbol{\tau}_{jk}, \rho_{jk}\} \\
&< \int \prod_{j<k} \big(\exp(-\frac{\lambda_1^2}{2} \sum_g \tau_{jkg} - \frac{\lambda_2^2}{2} \rho_{jk}) \rho_{jk}^{-\frac{1}{2}} \big) d\rho_{jk} \\
&< \int \prod_{j<k} \big(\exp(-\frac{\lambda_2^2}{2} \rho_{jk}) \rho_{jk}^{-\frac{1}{2}} \big) d\rho_{jk} = (2\pi/\lambda_2^2)^{\frac{p(p-1)}{4}}
\end{aligned}$$

The above inequalities completes the proof that the conditional prior distributions are all proper for DSS-GGL. For GGL prior, the proof is essentially the same by fixing $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ to be 1.

Proof of DSS-FGL prior For DSS-FGL, the joint distribution of all the parameters using the scale Normal mixture representation is

$$\begin{aligned}
p(\{\boldsymbol{\Omega}\}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\xi}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}}) &= \prod_g \prod_{j<k} \exp(-\frac{1}{2} (\boldsymbol{\omega}_{jk}^{(g)})^T \boldsymbol{\Theta}_{jk} \boldsymbol{\omega}_{jk}^{(g)}) \prod_g \prod_j \exp(-\frac{\lambda_0}{2} \omega_{jj}^{(g)}) \\
&\times \prod_g \prod_{j<k} \tau_{jkg}^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2} \tau_{jkg}) \prod_{j<k} \rho_{jk}^{-\frac{1}{2}} \exp(-\frac{\lambda_2^2}{2} \rho_{jk}) \\
&\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}} (1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{\xi_{jk}} (1-\pi_{\boldsymbol{\xi}})^{1-\xi_{jk}} \\
&\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{a_1-1} (1-\pi_{\boldsymbol{\delta}})^{b_1-1} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{a_2-1} (1-\pi_{\boldsymbol{\xi}})^{b_2-1}
\end{aligned}$$

where the first term can be rewritten as the same form as $\exp(-pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}))$ for DSS-FGL:

$$(\boldsymbol{\omega}_{jk}^{(g)})^T \boldsymbol{\Theta}_{jk} \boldsymbol{\omega}_{jk}^{(g)} = \sum_g \frac{v_{\delta_{jk}}}{\tau_{jkg}} (\omega_{jk}^{(g)})^2 + \sum_{g<g'} \frac{v_{\xi^*_{jk}}}{\phi_{jkgg'}} (\omega_{jk}^{(g)} - \omega_{jk}^{(g')})^2.$$

Using the same identity as before, we can rewrite the conditional distribution below into the form of the DSS-FGL penalty:

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}) \propto \exp(-\sum_{j<k} \frac{\lambda_1}{v_{\delta_{jk}}} |\omega_{jk}^{(g)}| - \sum_{j<k} \sum_{g<g'} \frac{\lambda_2}{v_{\xi^*_{jk}}} |\omega_{jk}^{(g)} - \omega_{jk}^{(g')}| - \sum_g \sum_j \frac{\lambda_0}{2} \omega_{jj}^{(g)}).$$

The proof of the DSS-FGL conditional distributions being proper is similar to the previous case. We first note that

$$\boldsymbol{\Theta}_{jk} = \text{diag}(\{\frac{1}{\tau_{jkg}}\}_{g=1,\ldots,G}) + \boldsymbol{L}_{jk}$$

where $\boldsymbol{L}_{jk}$ is a graph Laplacian matrix and is positive semi-definite. Then by Minkowski inequality, $\det(\boldsymbol{\Theta}_{jk}) \geq \det(\text{diag}(\{\frac{1}{\tau_{jkg}}\}_{g=1,\ldots,G}))$. Then we have

$$
\begin{aligned}
C^{-1}_{\boldsymbol{\tau},\boldsymbol{\phi}} &\propto \int \prod_{j<k} \big(\det(\boldsymbol{\Theta}_{jk})^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2}\sum_g \tau_{jkg} - \frac{\lambda_2^2}{2}\sum_{g<g'}\phi_{jkgg'}) \prod_g \tau_{jkg}^{-\frac{1}{2}} \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}}\big) d\{\boldsymbol{\tau},\boldsymbol{\phi}\} \\
&\leq \int \prod_{j<k} \big(\exp(-\frac{\lambda_1^2}{2}\sum_g \tau_{jkg} - \frac{\lambda_2^2}{2}\sum_{g<g'}\phi_{jkgg'}) \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}}\big) d\{\boldsymbol{\tau},\boldsymbol{\phi}\} \\
&\leq \int \prod_{j<k} \big(\exp(-\frac{\lambda_2^2}{2}\sum_{g<g'}\phi_{jkgg'}) \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}}\big) d\boldsymbol{\phi},
\end{aligned}
$$

which is again finite since the integral consists of products of Gamma densities. The rest of the argument follows in the same way as the DSS-GGL case.

B. Details of the EM algorithm implementation

Assuming no missing data, the full objective function in the t-th iteration of the EM algorithm described in 5 is the expectation of the complete data log likelihood, i.e.,

$$
\begin{aligned}
Q(\{\boldsymbol{\Omega}\},\pi_{\boldsymbol{\delta}},\pi_{\boldsymbol{\xi}}|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)}) &= E_{\boldsymbol{\delta},\boldsymbol{\xi}|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)},\boldsymbol{X}}(\log p(\{\boldsymbol{\Omega}\},\pi_{\boldsymbol{\delta}},\pi_{\boldsymbol{\xi}}|\boldsymbol{X})|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)},\boldsymbol{X}) \\
&= \text{constant} + \sum_g \frac{n_g}{2}\log|\boldsymbol{\Omega}_g| - \frac{1}{2}\sum_g tr(\boldsymbol{S}_g\boldsymbol{\Omega}_g) - \frac{\lambda_0}{2}\sum_j\sum_g |\omega_{jj}^{(g)}| \\
&\quad -\lambda_1\sum_{j<k}\sum_g |\omega_{jk}^{(g)}| E_{\cdot|\cdot}[\frac{1}{v_0(1-\delta_{jk})+v_1\delta_{jk}}] + \sum_{j<k}\log(\frac{\pi_{\boldsymbol{\delta}}}{1-\pi_{\boldsymbol{\delta}}})E_{\cdot|\cdot}(\delta_{jk}) \\
&\quad -\lambda_2\sum_{j<k}\widetilde{pen}(\boldsymbol{\omega}_{jk})E_{\cdot|\cdot}[\frac{1}{v_0(1-\delta_{jk}\xi_{jk})+v_1\delta_{jk}\xi_{jk}}] + \sum_{j<k}\log(\frac{\pi_{\boldsymbol{\xi}}}{1-\pi_{\boldsymbol{\xi}}})E_{\cdot|\cdot}(\xi_{jk}) \\
&\quad +(a_1-1)\log(\pi_{\boldsymbol{\delta}}) + (b_1 + \frac{p(p-1)}{2} - 1)\log(1-\pi_{\boldsymbol{\delta}}) \\
&\quad +(a_2-1)\log(\pi_{\boldsymbol{\xi}}) + (b_2 + \frac{p(p-1)}{2} - 1)\log(1-\pi_{\boldsymbol{\xi}}),
\end{aligned}
$$

where $E_{\cdot|\cdot}$ denotes conditional expectation $E_{\boldsymbol{\delta},\boldsymbol{\xi}|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)},\boldsymbol{X}}$, and $\widetilde{pen}(\boldsymbol{\omega}_{jk}) = ||\boldsymbol{\omega}_{jk}||_2$ for DSS-GGL and $\widetilde{pen}(\boldsymbol{\omega}_{jk}) = \sum_{g<g'}|\omega_{jk}^{(g)} - \omega_{jk}^{(g')}|$ for DSS-FGL.

When missing data exists, we need to also calculate the expectation of $\boldsymbol{S}_g$ given the observed data. That is, We need to replace the $tr(\boldsymbol{S}_g\boldsymbol{\Omega}_g)$ term in the above objective function by

$$
E_{\cdot|\cdot}(\boldsymbol{S}_g\boldsymbol{\Omega}_g) = E_{\cdot|\cdot}((\frac{1}{n_g}\sum_i^{n_g}\boldsymbol{x}_i^{(g)}(\boldsymbol{x}_i^{(g)})^T)\boldsymbol{\Omega}_g) = \frac{1}{n_g}(\sum_i^{n_g} E_{\boldsymbol{x}_{i,m}^{(g)}|\boldsymbol{x}_{i,o}^{(g)}}(\boldsymbol{x}_i^{(g)}(\boldsymbol{x}_i^{(g)})^T))\boldsymbol{\Omega}_g.
$$

where $\boldsymbol{x}_{i,o}^{(g)}$ and $\boldsymbol{x}_{i,m}^{(g)}$ denote the observed and missing cells in $\boldsymbol{x}_i^{(g)}$ respectively. $\boldsymbol{x}_i^{(g)}$ follows a multivariate Gaussian distribution. Without loss of generality, if we let $\boldsymbol{x}_i^{(g)} = \begin{pmatrix}\boldsymbol{x}_{i,o}^{(g)}\\ \boldsymbol{x}_{i,m}^{(g)}\end{pmatrix}$, we know

$$
\begin{aligned}
E_{\boldsymbol{x}_{i,m}^{(g)}|\boldsymbol{x}_{i,o}^{(g)}}(\boldsymbol{x}_{i,m}^{(g)}) &= \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{x}_{i,o}^{(g)} \\
E_{\boldsymbol{x}_{i,m}^{(g)}|\boldsymbol{x}_{i,o}^{(g)}}(\boldsymbol{x}_i^{(g)}(\boldsymbol{x}_i^{(g)})^T) &= E_{\cdot|\cdot}(\boldsymbol{x}_i^{(g)})E_{\cdot|\cdot}(\boldsymbol{x}_i^{(g)})^T + \begin{pmatrix}\boldsymbol{0}_{oo} & \boldsymbol{0}_{om}\\ \boldsymbol{0}_{mo} & \boldsymbol{\Sigma}_{mm}-\boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}\end{pmatrix}
\end{aligned}
$$

where $\boldsymbol{\Sigma}_{oo}, \boldsymbol{\Sigma}_{om}, \boldsymbol{\Sigma}_{mo}$ and $\boldsymbol{\Sigma}_{mm}$ are the corresponding submatrices of $\boldsymbol{\Sigma}_g$.

C. Gibbs sampler of the proposed models

The EM algorithm introduced in the main paper maximizes the complete data likelihood by looking at the Laplace representation after integrating out all the latent parameters. In this section, we show that these latent parameters, $\boldsymbol{\tau}, \boldsymbol{\phi}$ and $\boldsymbol{\rho}$, facilitates efficient block Gibbs sampling algorithms for fully Bayesian inference.

We start by describing the posterior sampling of $\{\boldsymbol{\Omega}\}$. The basic idea is to sample each column and row for all the precision matrices jointly. To similify notation, we separate out the last column and row in $\boldsymbol{\Omega}_g$ and $\boldsymbol{S}_g$ and define

$$\boldsymbol{\Omega}_g = \begin{pmatrix} \boldsymbol{\Omega}_{11}^{(g)} & \boldsymbol{\omega}_{12}^{(g)} \\ \boldsymbol{\omega}_{21}^{(g)} & \omega_{22}^{(g)} \end{pmatrix}, \quad \boldsymbol{S}_g = \begin{pmatrix} \boldsymbol{S}_{11}^{(g)} & \boldsymbol{s}_{12}^{(g)} \\ \boldsymbol{s}_{21}^{(g)} & s_{22}^{(g)} \end{pmatrix}.$$

We further let $\boldsymbol{\omega}_{12} = [(\boldsymbol{\omega}_{12}^{(1)})^T, (\boldsymbol{\omega}_{12}^{(2)})^T, ..., (\boldsymbol{\omega}_{12}^{(G)})^T]^T$, and $\boldsymbol{s}_{12} = [(\boldsymbol{s}_{12}^{(1)})^T, (\boldsymbol{s}_{12}^{(2)})^T, ..., (\boldsymbol{s}_{12}^{(G)})^T]^T$, each denoting a vector of length $(p-1)G$, and $\boldsymbol{\omega}_{22} = [\omega_{22}^{(1)}, \omega_{22}^{(2)}, ..., \omega_{22}^{(G)}]$.

The conditional distribution of $(\boldsymbol{\omega}_{12}, \boldsymbol{\omega}_{22})$ given the rest of the elements in $\{\boldsymbol{\Omega}\}$ does not seem to take any standard form. However, if we perform a change of variables and let $\theta_g = \omega_{22}^{(g)} - \boldsymbol{\omega}_{21}^{(g)}(\boldsymbol{\Omega}_{11}^{(g)})^{-1}\boldsymbol{\omega}_{12}^{(g)}$, the conditional distribution of $(\boldsymbol{\omega}_{12}, \boldsymbol{\theta})$ becomes

$$\begin{aligned} p(\boldsymbol{\omega}_{12}, \boldsymbol{\theta}) &\propto \sum_g \theta_g^{\frac{n_g}{2}} \exp(-\frac{s_{22}^{(g)} + \lambda_0}{2}\theta_g - \boldsymbol{s}_{12}^T\boldsymbol{\omega}_{12} - \frac{1}{2}\boldsymbol{\omega}_{12}^T\boldsymbol{A}\boldsymbol{\omega}_{12}) \\ &= \prod_g \text{Gamma}(\theta_g; \frac{n_g}{2}+1, \frac{s_{22}^{(g)}+\lambda_0}{2}) \times \text{Normal}(\boldsymbol{\omega}_{12}; -\boldsymbol{A}^{-1}\boldsymbol{s}_{12}, \boldsymbol{A}^{-1}). \end{aligned}$$

The $\boldsymbol{A}$ matrix can be calculated by $\boldsymbol{A} = \boldsymbol{U} + \boldsymbol{V}$, where $\boldsymbol{U}$ is a matrix by rearrangeing the precision matrices so that its $((g-1)(p-1)+k, (g'-1)(p-1)+k)$-th element is the (g, g')-element in $\tilde{\boldsymbol{\Theta}}_{jk}$ defined in (12) and (13) of the main paper, and

$$\boldsymbol{V} = \begin{pmatrix} (\lambda_0 + s_{22}^{(1)})(\boldsymbol{\Omega}_{11}^{(1)})^{-1} & & \\ & \ddots & \\ & & (\lambda_0 + s_{22}^{(G)})(\boldsymbol{\Omega}_{11}^{(G)})^{-1} \end{pmatrix},$$

For DSS-GGL, we notice that $\boldsymbol{A}$ is block diagonal, thus we can alternatively sample $\boldsymbol{\omega}_{12}^{(g)}$ independently by

$$\boldsymbol{\omega}_{12}^{(g)}|\cdot \sim \text{Normal}(-\boldsymbol{A}_g^{-1}\boldsymbol{s}_{12}^{(g)}, \boldsymbol{A}_g^{-1})$$

where $\boldsymbol{A}_g = (\lambda_0 + s_{22}^{(g)})(\boldsymbol{\Omega}_{11}^{(g)})^{-1} + \tilde{\boldsymbol{\Theta}}_{11}^{(g)}$.

Given $\{\boldsymbol{\Omega}\}$, the latent parameters in DSS-GGL have simple conditional distribution as follows:

$$\begin{aligned} \tau_{jkg}^{-1}|\cdot &\sim \text{InvGaussian}(\frac{\lambda_1}{v_{\delta_{jk}}^{\frac{1}{2}}|\omega_{jk}^{(g)}|}, \lambda_1^2), \quad j,k = 1,...,p, g = 1,..,G \\ \rho_{jk}^{-1}|\cdot &\sim \text{InvGaussian}(\frac{\lambda_2}{v_{\xi_{jk}^*}\sum_g(\omega_{jk}^{(g)})^2}, \lambda_2^2), \quad j,k = 1,...,p \\ \delta_{jk}, \xi_{jk}|\cdot &\sim p^*(\boldsymbol{\delta}_{jk}, \boldsymbol{\xi}_{jk}), \quad j,k = 1,...,p \\ \pi_{\boldsymbol{\delta}} &\sim \text{Beta}(\sum \delta_{jk} + a_1, \sum(1-\delta_{jk}) + b_1) \\ \pi_{\boldsymbol{\xi}} &\sim \text{Beta}(\sum \xi_{jk} + a_2, \sum(1-\xi_{jk}) + b_2) \end{aligned}$$

where $p^*(\boldsymbol{\delta}, \boldsymbol{\xi})$ is defined in (5) in the main paper.

For DSS-FGL, the conditional distribution of $\boldsymbol{\tau}, \boldsymbol{\delta}, \boldsymbol{\xi}, \pi_{\boldsymbol{\delta}}$, and $\pi_{\boldsymbol{\xi}}$ are the same as DSS-GGL. The conditional distribution of $\boldsymbol{\phi}$ is

$$\phi_{jkgg}^{-1} \sim \text{InvGaussian}(\frac{\lambda_2}{v_{\xi_{jk}^*}^{\frac{1}{2}}|\omega_{jk}^{(g)} - \omega_{jk}^{(g')}|}, \lambda_2^2), \quad j,k = 1,...,p, g, g' = 1,..,G$$

The Gibbs sampler is then complete by circling through and sampling each blocks of $\{\mathbf{\Omega}\}$ and the latent parameters with the above posterior conditional distributions.

D. Additional illustrating example

In this section, we provide a more detailed description to the small 2-class simulated example described in Section 6 of the main paper. We let $n_g = 150$ for $g = 1, 2$, and $p = 100$. The first 10 variables in the first class form a 10-node block with an AR(1) precision matrix, i.e., $(\mathbf{\Omega}^{-1})_{jk} = 0.7^{|j-k|}$. The rest of the 90 variables are independent noises from a standard Gaussian distribution. The second class shares the first 4 edges of the first class. The first 5 variables form another AR(1) block with different strength of correlations so that $(\mathbf{\Omega}^{-1})_{jk} = 0.9^{|j-k|}$. The rest of the 95 variables all follow independent standard Gaussian distribution.

The best performance of FGL in our experiments was achieved with $\lambda_2 = 0.1$. We then obtained the regularization path along different values of λ_1. We also fit DSS-FGL and SS-FGL with $\lambda_1 = 1, \lambda_2 = 1$. The latter assumes $\xi_{jk} = 1$ for all edges, i.e., the penalization of similarities is always proportional to the penalization of sparsity.

Figure 1 shows both the two solution paths presented in the main paper, and the solution path from SS-FGL. It can be seen that although SS-FGL achieves similar bias as DSS-FGL, it also estimates several more false positive edges. This can be seen from the formulation of the doubly spike-and-slab selection: with only one spike-and-slab mixture of the penalties, the selected edges from the slab distributions receive also only weak penalization for between-class similarities. Thus it is more likely to pick up spurious edges due to noises that happen to exist in one class. This full illustration of the simple example shows the advantage of having the doubly spike-and-slab setup.

E. Additional simulation evidence

Here we describe our procedure in simulating the dataset described in Section 7 of the main paper. We first generate three networks with p features with 10 equal sized unconnected subnetwork. Each of the subnetwork follow a power law degree distribution, which is generally harder to estimate than simpler structures (Peng et al., 2009). The first class contains all ten subnetworks, and the second and third classes each has one and two subnetworks removed. Given the network structure, we generate $\mathbf{\Omega}_g$ from the G-Wishart distribution $W_G(3, I_p)$, and rescale them so that $\mathbf{\Omega}_g^{-1}$ have unit variances. Finally, we generate $n = 150$ independent and identically distributed samples from Normal$(\mathbf{0}, \mathbf{\Omega}_g)$ in each class. The resulted graph for $p = 500$ is shown in Figure 2. We fit GGL and FGL with various choice of fixed λ_2 and a sequence of λ_1. We fit DSS-GGL and DSS-FGL with $\lambda_1 = 1$, $\lambda_2 = 2/30$ in this case. We explore more choices of λ_2 in the moderate dimensional experiments below and found no substantial changes in the performance of the final models.

Additional results for $p = 100$ and $p = 200$ are shown in Figure 3 and 4. The dots correspond to DSS-GGL and DSS-FGL with $\lambda_1 = 1$, $\lambda_2 = 0.1$. We also examined different choices of λ_2 are found no substantial differences in performance. An exploratory sensitivity analysis is presented in the next subsection.

E.1. Sensitivity to hyperparameters

Figure 5 and 6 shows the converged regions over 100 replications on space of the true positive against false positive discoveries for edges and differential edges respectively when $p = 200$. It can be seen that the performance is relatively stable under different choices of λ_2.

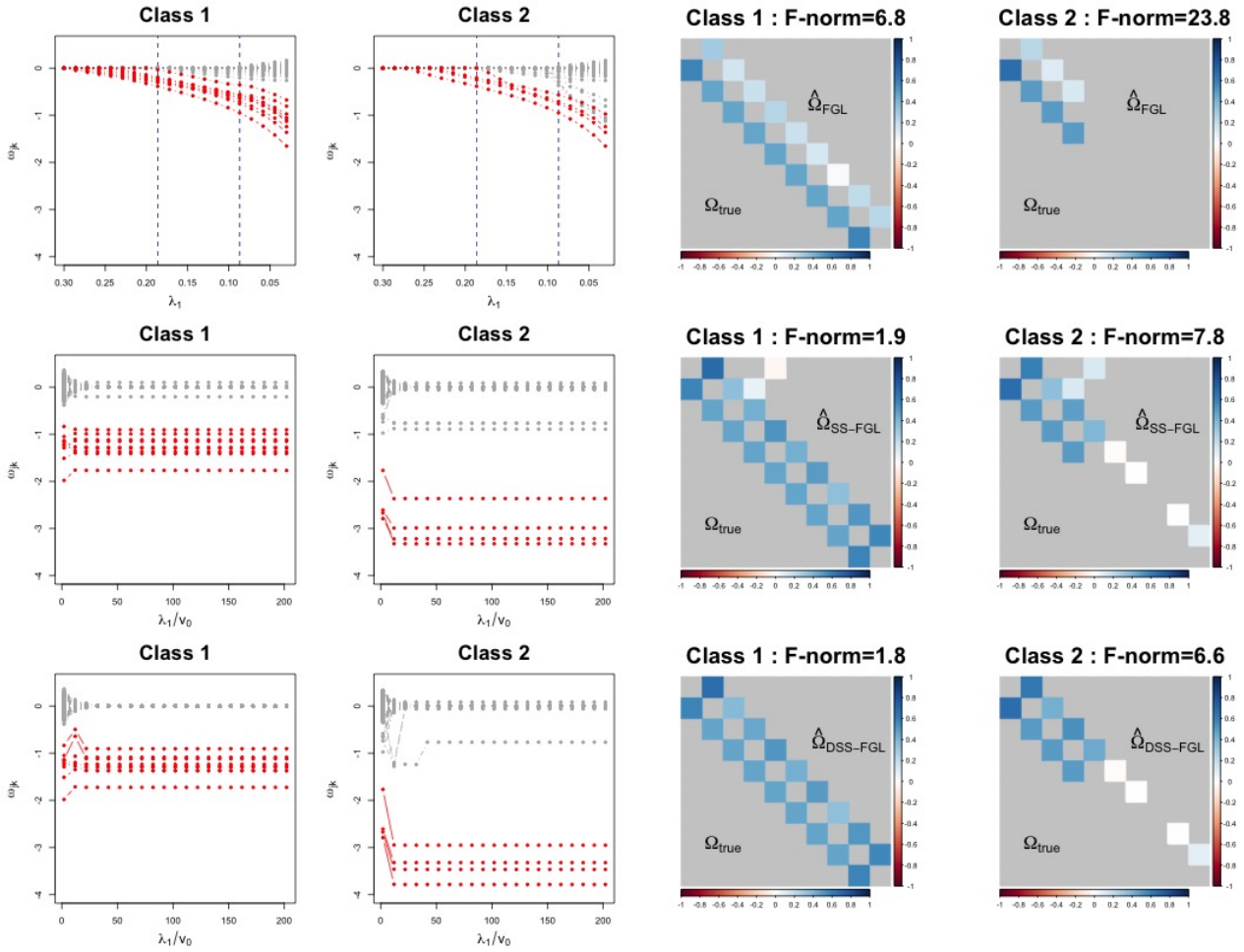

Figure 1. The solution paths and estimated precision matrices of FGL (upper row), SS-FGL (middle row) and DSS-FGL (lower row). The red nodes correspond to true edges and the gray nodes correspond to 0's. The two vertical lines in the FGL solution path indicate the model that best matches the true sparsity (left) and the model with the lowest AIC (right). The block containing the edges is plotted for the estimated values (upper triangular) against the truth (lower triangular). The model that best matches the true graphs is plotted for FGL. The off-diagonal values are rescaled and negated to partial correlations, and 0's are colored with light gray background for easier visual comparison. The bias of the estimated precision matrix measured by the Frobenius norm, $||\hat{\mathbf{\Omega}}_g - \mathbf{\Omega}_g||_F$, is also printed in the captions.

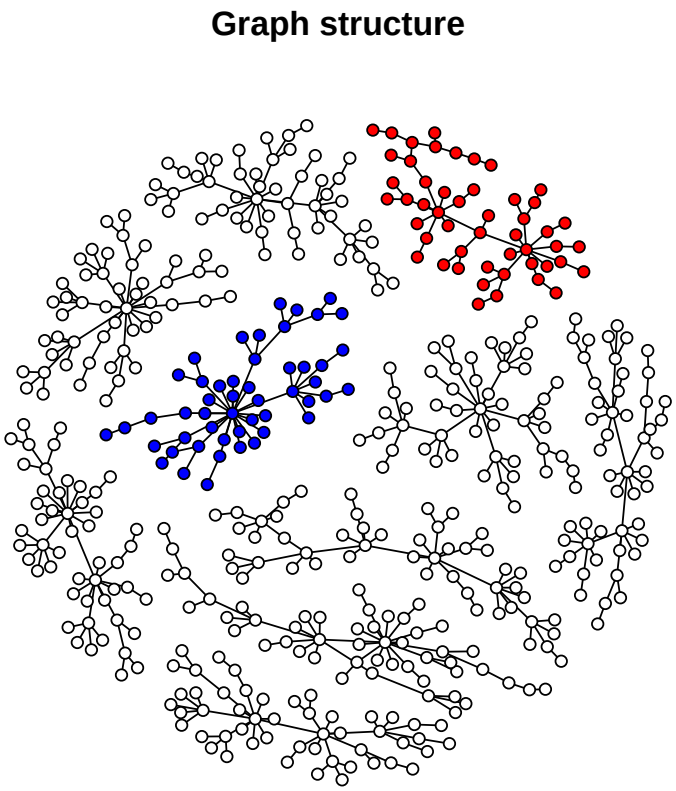

Figure 2. Graph structure of the simulated dataset. The edges between the red nodes are removed from the second class, and edges between both the red and blue nodes are removed from the third class.

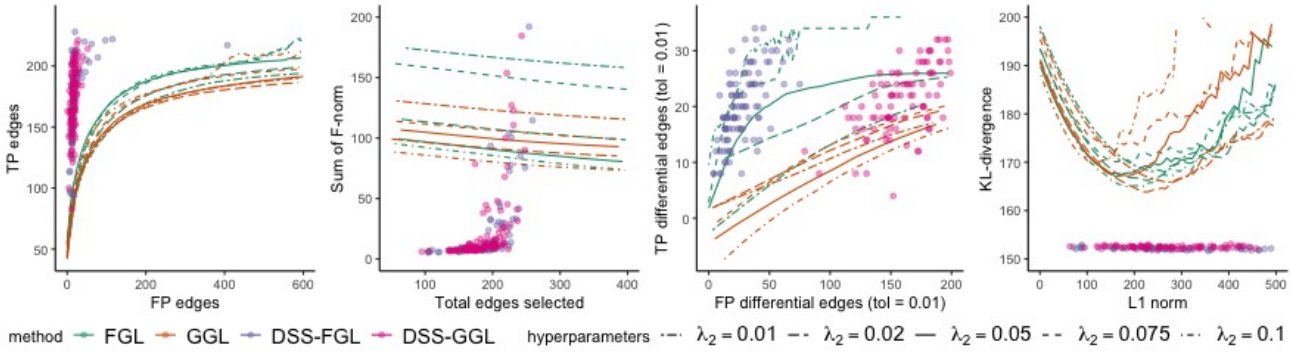

Figure 3. Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications, $p = 100$. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters.

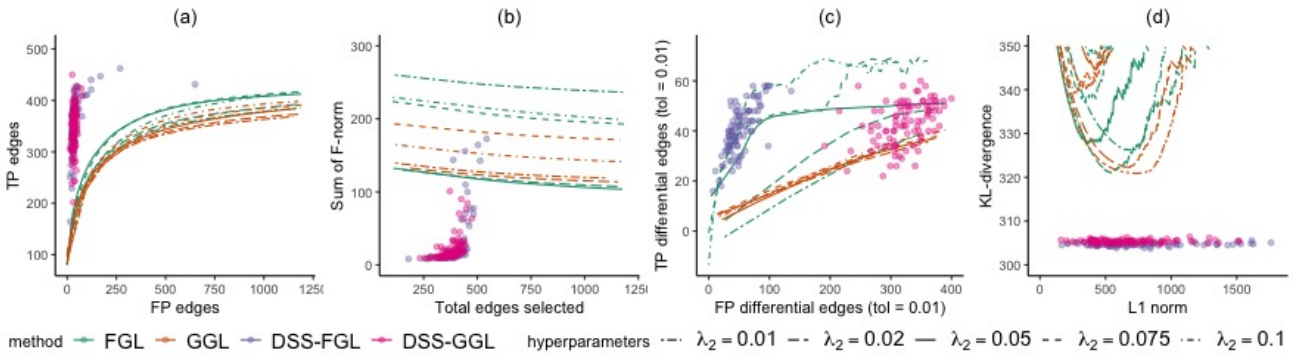

Figure 4. Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications, $p = 200$. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters.

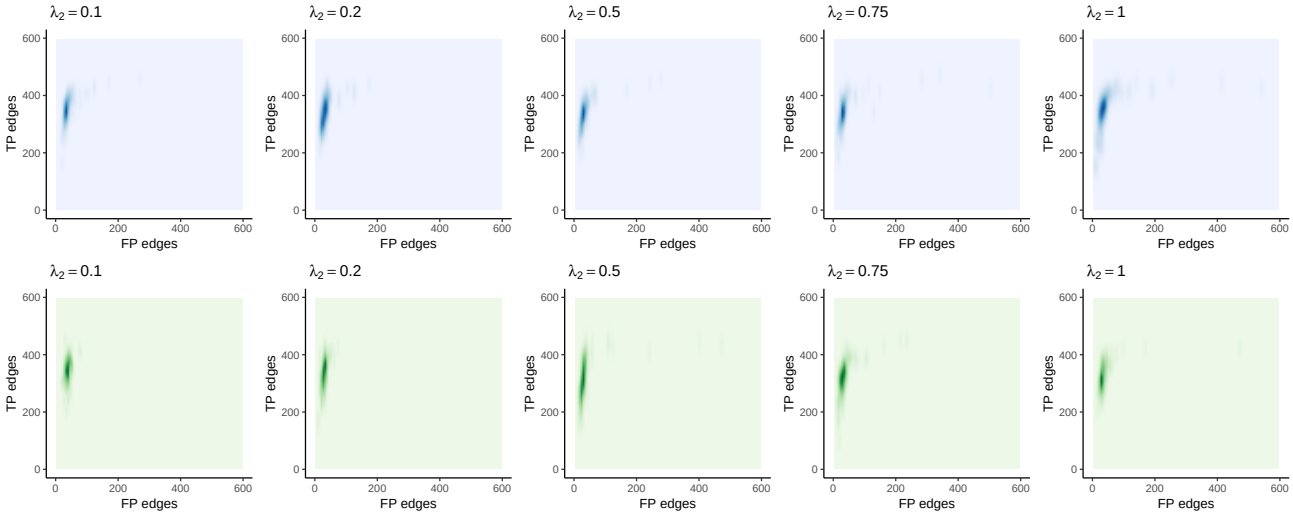

Figure 5. The density plot of true positive edges against false positive edges for DSS-FGL (top row), and DSS-GGL (bottom row) under different choices of λ_2. λ_1 is set to 1.

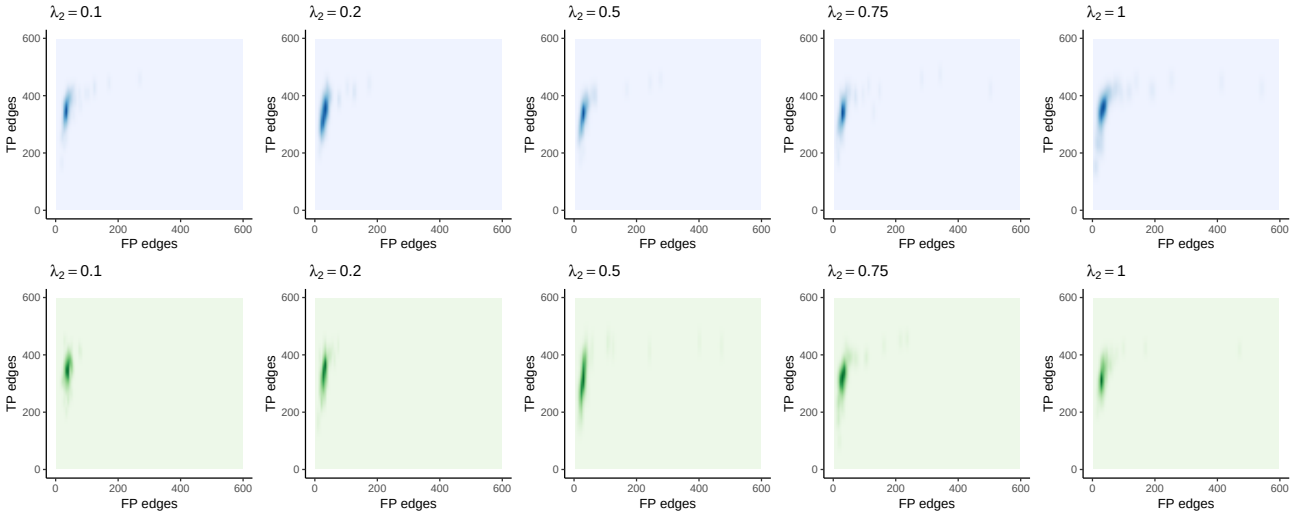

Figure 6. The density plot of true positive differential edges against false positive positive edges for DSS-FGL (top row), and DSS-GGL (bottom row) under different choices of λ_2. λ_1 is set to 1.

F. Details on the verbal autopsy data analysis

F.1. List of symptoms

Table 1 shows the questions with continuous responses used in the analysis in the main paper.

Table 1. List of symptoms considered in this analysis.

Abbreviation	Questionnaire item
ill	For how long was [name] ill before s/he died? [days]
fever	How many days did the fever last? [days]
rash	How many days did [name] have the rash? [days]
ulcer	For how many days did the ulcer ooze pus? [days]
yellow discoloration	For how long did [name] have the yellow discoloration? [days]
ankle swelling	For how long did [name] have ankle swelling? [days]
puffiness face	For how long did [name] have puffiness of the face? [days]
puffiness body	For how long did [name] have puffiness all over his/her body? [days]
cough	For how long did [name] have a cough? [days]
difficulty breathing	For how long did [name] have difficulty breathing? [days]
fast breathing	For how long did [name] have fast breathing? [days]
liquid stool	For how long before death did [name] have loose or liquid stools? [days]
vomit	For how long before death did [name] vomit? [days]
difficulty swallowing	For how long before death did [name] have difficulty swallowing? [days]
belly pain	For how long before death did [name] have belly pain? [days]
protruding belly	For how long before death did [name] have a protruding belly? [days]
mass belly	For how long before death did [name] have a mass in the belly [days]
headaches	For how long before death did [name] have headaches? [days]
stiff neck	For how long before death did [name] have stiff neck? [days]
unconsciousness	For how long did the period of loss of consciousness last? [days]
confusion	For how long did the period of confusion last? [days]
convulsion	For how long before death did the convulsions last? [days]
paralysis	For how long before death did [name] have paralysis? [days]
period overdue	For how many weeks was her period overdue? [days]
tobacco	How much pipe/chewing tobacco did [name] use daily?
cigarettes	How many cigarettes did [name] smoke daily?
age	Age [years]

F.2. Comparing with JGL

The estimated symptom network from the DSS-FGL and DSS-GGL are summarized in Figure 7.

The estimated symptom network from the FGL and GGL are summarized in Figure 8. We fit both models under a 2-dimensional grids over λ_1 and λ_2. As expected, AIC selects very dense graphs (first two rows of Figure 8) and are difficult to interpret. We also compare the FGL and GGL graph with the closest number of edges as those from DSS-FGL and DSS-GGL in the third and forth row of Figure 8. The number of differential edges is typically smaller compared to DSS-FGL and DSS-GGL, which is likely due to over penalization of similarities, i.e., edges become too similar using FGL, and too sparse among half of the nodes using GGL.

G. Details on prediction of missing mortality rates

The data we consider in this example consist of log mortality rates over $n = 51$ years for $p = 101$ age groups, and 2 classes representing female and male series respectively. The estimated graph structure from one of the cross-validation dataset using FGL and DSS-FGL are shown in Figure 9. The Lee-Carter model are estimated using the R package `ilc` (Butt et al., 2014) for each gender separately.

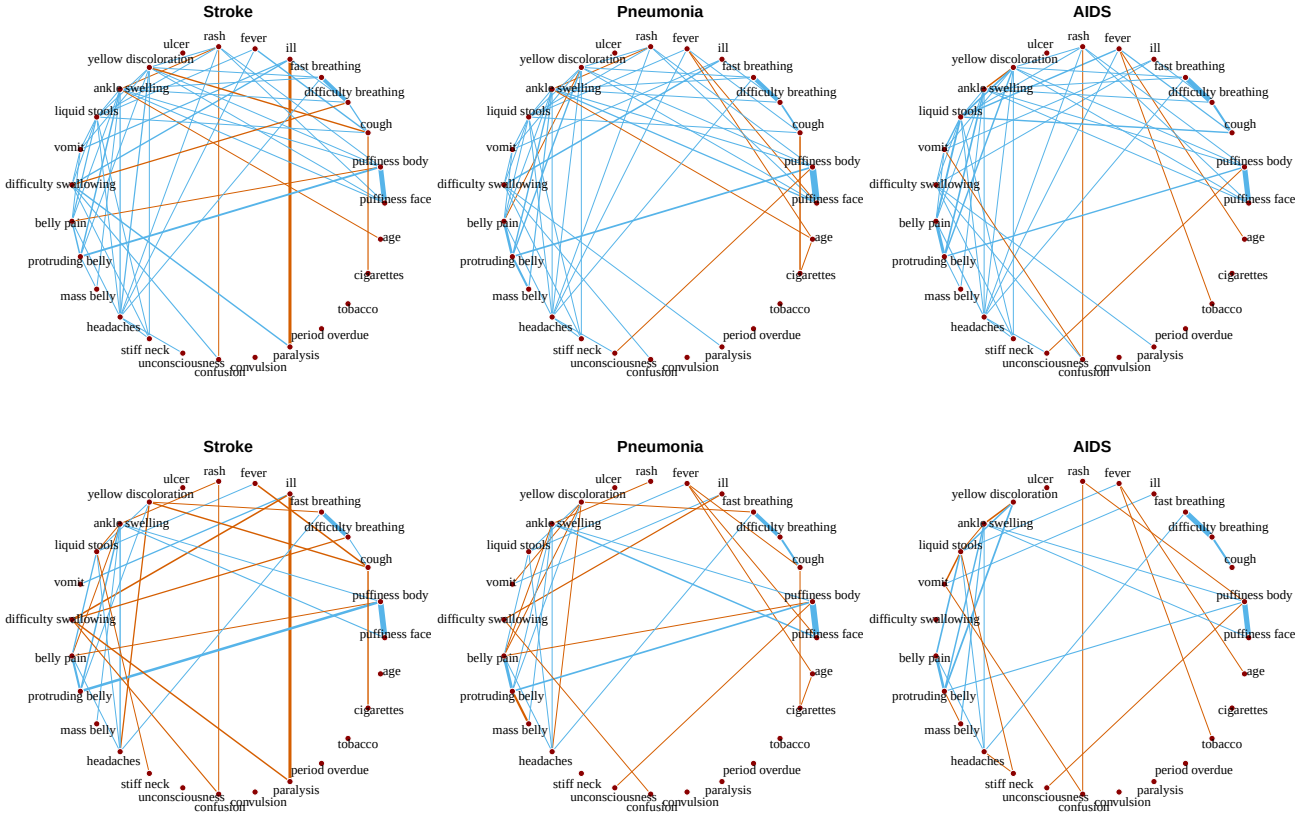

Figure 7. Estimated edges between the symptoms under the three causes using DSS-FGL (top row) and DSS-GGL (bottom row). The width of the edges are proportional to the size of $|\omega_{jk}^{(g)}|$. Common edges across all groups are colored in blue, and the differential edges are colored in red.

The DSS-FGL is able to pick up more conditional dependence structures along the diagonal among several age groups, while the FGL estimates mostly within adults only. It is interesting that both approaches identifies positive partial correlations between age $14-17$ and $30-40$ between male and female. This is likely due to the fact that male mortality around age 20 typically shows a hump of increase due to young adult accident mortality, which leads to the mean model more likely to underestimations for mortality during age $18-30$ and overestimations both before and after that period. This relationship of the age curve, however, is not seen in female mortality.

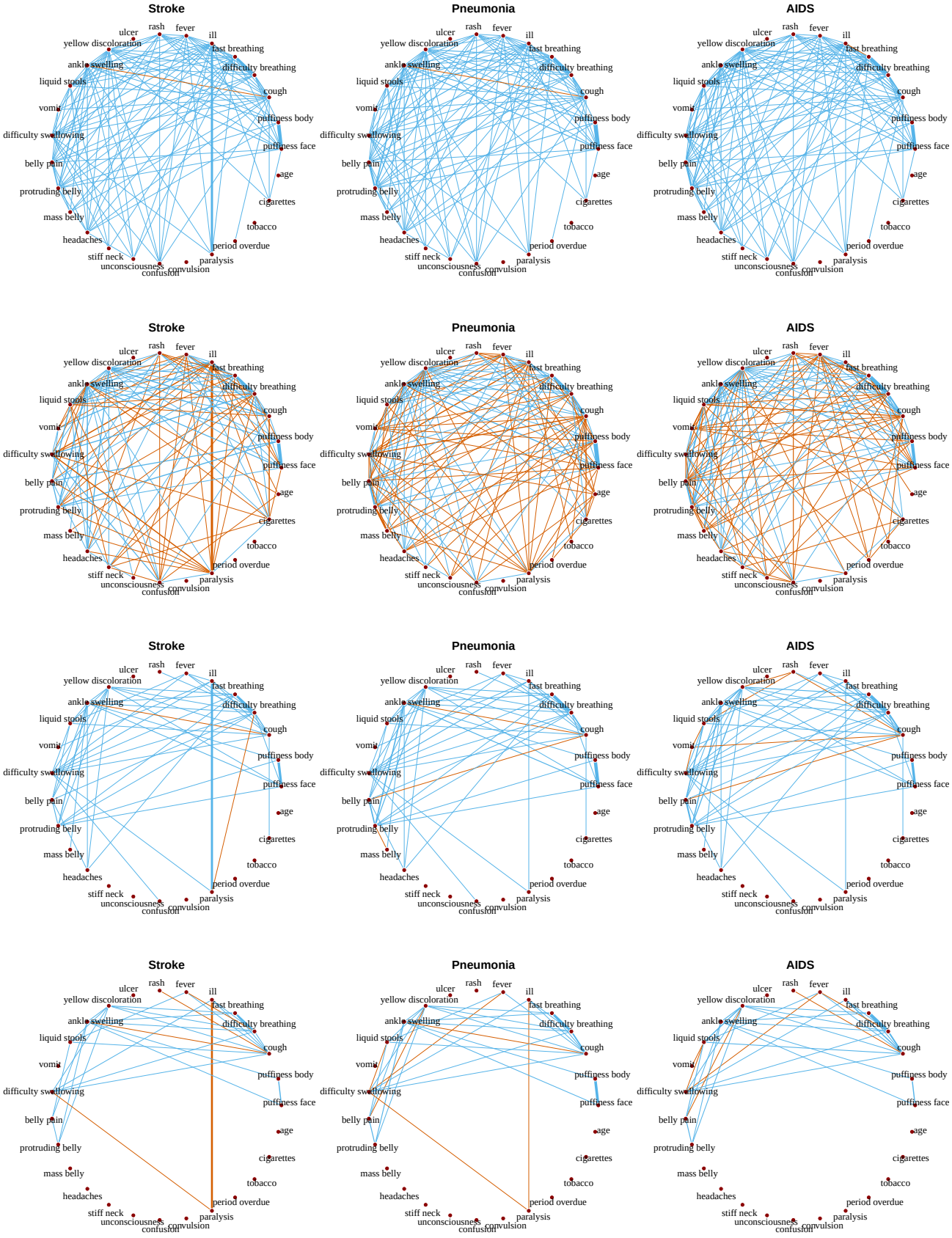

Figure 8. Estimated edges between the symptoms under the three causes using FGL using AIC (first row), GGL using AIC (second row), FGL with the same number of edges as selected by DSS-FGL (third row), and GGL with the same number of edges as selected by DSS-GGL (last row). The width of the edges are proportional to the size of $|\omega_{jk}^{(g)}|$. Common edges across all groups are colored in blue, and the differential edges are colored in red.

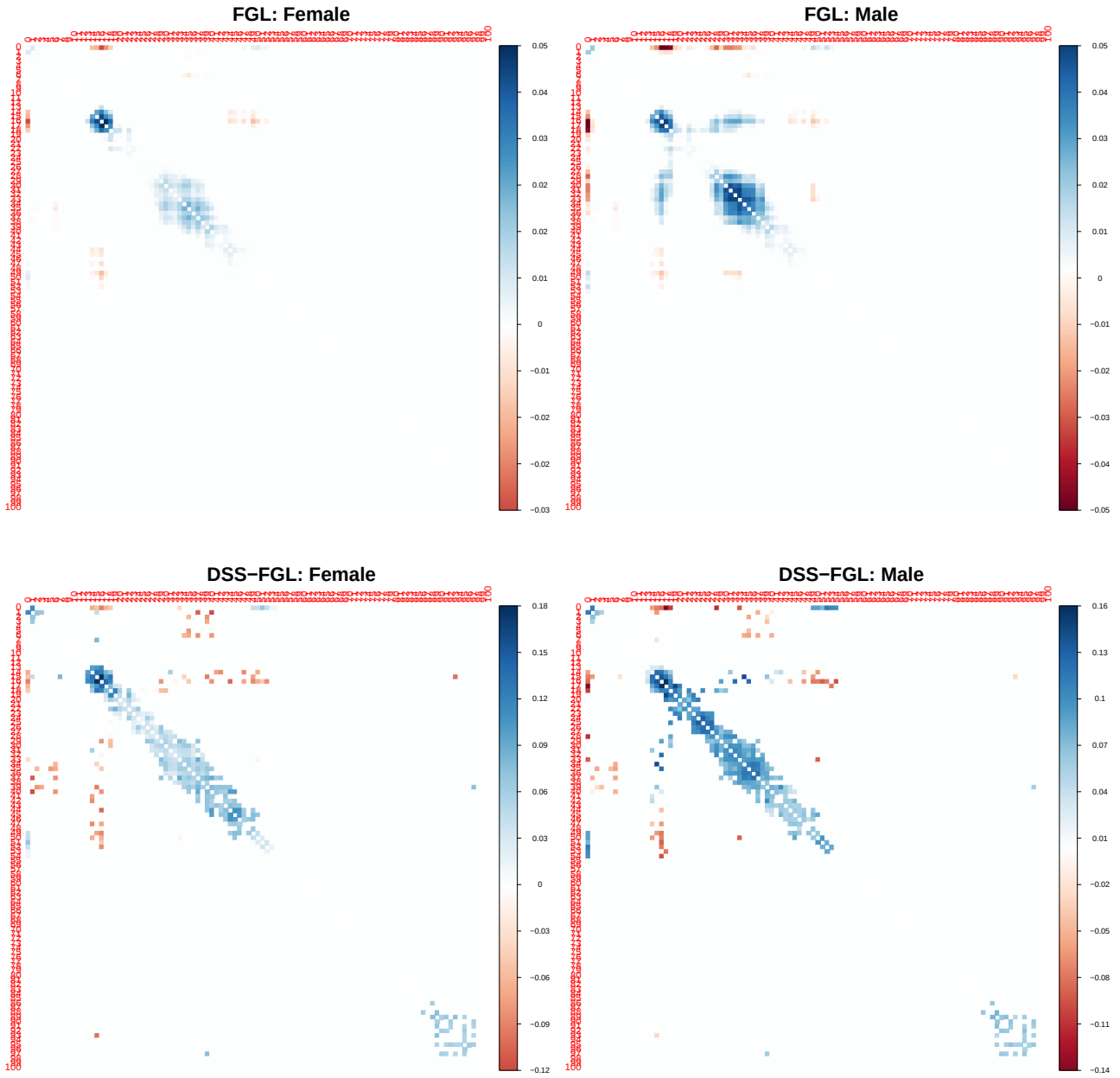

Figure 9. Estimated partial correlation matrix using one cross-validation dataset. The partial correlations among the 101 age groups are estimated using FGL with the same number of edges as selected by DSS-FGL (top row), and DSS-FGL (bottom row). DSS-FGL estimates 197 and 199 edges respectively for female and male. The closet configuration of FGL estimates 157 and 241 edges respectively. The precision matrices are rescaled and negated to partial correlations for easier interpretation.

References

Butt, Z., Haberman, S., and Shang, H. *ilc: Lee-Carter Mortality Models using Iterative Fitting Algorithms*, 2014. R package version 1.0.

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