Bayesian Joint Spike-and-Slab Graphical Lasso

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

Abstract

In this article, we propose a new class of priors for Bayesian inference with multiple Gaussian graphical models. We introduce Bayesian treatments of two popular procedures, the group graphical lasso and the fused graphical lasso, and extend them to a continuous spike-and-slab framework to allow self-adaptive shrinkage and model selection simultaneously. We develop an EM algorithm that performs fast and dynamic explorations of posterior modes. Our approach selects sparse models efficiently and automatically with substantially smaller bias than would be induced by alternative regularization procedures. The performance of the proposed methods are demonstrated through simulation and two real data examples.

1. Introduction

Bayesian formulations of graphical models have been widely adopted as a way to characterize conditional independence structure among complex high-dimensional data. These models are popular in scientific domains including genomics (Briollais et al., 2016; Peterson et al., 2013), public health (Dobra, 2014; Li et al., 2017b), and economics (Dobra et al., 2010). In practice, data often come from several distinct groups. For example, data may be collected under various conditions, at different locations and time periods, or correspond to distinct subpopulations. Assuming a single graphical model in such cases can lead to unreliable estimates of network structure, whereas the alternative, estimating different graphical models separately for each group, may not be feasible for high dimensional problems.

Several approaches have been proposed to learn graphical

models jointly for multiple classes of data. Much of this work extends the penalized maximum likelihood approach to incorporate additional penalty terms that encourage the class-specific precision matrices to be similar (Guo et al., 2011; Danaher et al., 2014; Saegusa & Shojaie, 2016; Ma & Michailidis, 2016). In the Bayesian literature, Peterson et al. (2015) and Lin et al. (2017) utilize Markov Random Field priors to model a super-graph linking different graphical models. Tan et al. (2017) uses a logistic regression model to link the connectivity of nodes to covariates specific to each graph. These approaches only model the similarity of the underlying graphs, and thus are limited in their ability to borrow information when estimating the precision matrices. Borrowing strength is especially important when some classes have small sample sizes.

In this work, we introduce a new Bayesian formulation for estimating multiple related Gaussian graphical models by leveraging similarities in the underlying sparse precision matrices directly. We first present two shrinkage priors for multiple related precision matrices, as the Bayesian counterpart of joint graphical lasso estimators (Danaher et al., 2014). We then propose a doubly spike-and-slab mixture extension to these priors, which allows us to achieve simultaneous shrinkage and model selection, as well as handle missing observations. In Section 5 and 6, we extend from the recent literature on deterministic algorithms for Bayesian graphical models (Gan et al., 2018; Li & McCormick, 2019; Deshpande et al., 2019) and provide a fast Expectation-Maximization (EM) algorithm to quickly identify the posterior modes. We also propose a procedure to sequentially explore a series of posterior modes. We then demonstrate the substantial improvements in both model selection and parameter estimation over the original joint graphical lasso approach using both simulated data and two real datasets in Section 7. Finally, in Section 8 we discuss future directions for improvements.

2. Preliminaries

2.1. The joint graphical lasso

We first briefly introduce the notation used throughout this paper. We let G denote the number of classes in the data, and let Ω_g and Σ_g denote the precision and covariance matrix for the g-th class. We let $\omega_{jk}^{(g)}$ denote the (j, k)-th

[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>.

element in $\mathbf{\Omega}_g$ and $\boldsymbol{\omega}_{jk} = \{\omega_{jk}^{(g)}\}_{g=1,...,G}$ denote the vector of all the (j, k)-th elements in $\{\mathbf{\Omega}\}$. Suppose we are given G datasets, $\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(G)}$, where $\boldsymbol{X}^{(g)}$ is a $n_g \times p$ matrix of independent centered observations from the distribution Normal$(\mathbf{0}, \mathbf{\Omega}_g^{-1})$. As maximum likelihood estimates of $\mathbf{\Omega}_g$ can have high variance and are ill-defined when $p > n_g$, the joint penalized log likelihood for the G dataset is usually considered instead:

$$\ell(\{\mathbf{\Omega}\}) = \frac{1}{2} \sum_{g=1}^{G} n_g \log \det \mathbf{\Omega}_g - tr(\boldsymbol{S}_g \mathbf{\Omega}_g) - pen(\{\mathbf{\Omega}\}),$$
(1)

where $\boldsymbol{S}_g = (\boldsymbol{X}^{(g)})^T \boldsymbol{X}^{(g)}$. The penalty function encourages $\{\mathbf{\Omega}\}$ to have zeros on the off-diagonal elements and be similar across groups. In particular, we consider two useful penalty functions studied in Danaher et al. (2014), the group graphical lasso (GGL), and the fused graphical lasso (FGL):

$$pen(\{\mathbf{\Omega}\}) = \frac{\lambda_0}{2} \sum_g \sum_j |\omega_{jj}^{(g)}| + \lambda_1 \sum_g \sum_{j<k} |\omega_{jk}^{(g)}|$$
$$+ \lambda_2 \sum_{j<k} \widetilde{pen}(\boldsymbol{\omega}_{jk}),$$
(2)

where $\widetilde{pen}(\boldsymbol{\omega}_{jk}) = ||\boldsymbol{\omega}_{jk}||_2$ for GGL and $\sum_{g<g'} |\omega_{jk}^{(g)} - \omega_{jk}^{(g')}|$ for FGL. Both penalties encourage similarity across groups when $\lambda_2 > 0$, and reduce to separate graphical lasso problems when $\lambda_2 = 0$. The group graphical lasso encourages only similar patterns of zero elements across the G precision matrices, while the fused graphical lasso encourages a stronger form of similarity: the values of off-diagonal elements are also encouraged to be similar across the G precision matrices. In practice, λ_0 is typically set to 0 when the diagonal elements are not to be penalized.

2.2. Bayesian formulation of Gaussian graphical models

One of the most popular approaches for Bayesian inference with Gaussian graphical models is the G-Wishart prior (Lenkoski & Dobra, 2011; Mohammadi et al., 2015). The G-Wishart prior estimates the precision matrices with exact zeros in the off-diagonal elements and enjoys the conjugacy with the Gaussian likelihood. However, posterior inference under the G-Wishart prior can be computationally burdensome and has to rely on stochastic search algorithms over the large model space, consisting of all possible graphs. In recent years, several classes of shrinkage priors have been proposed for estimating large precision matrices, including the graphical lasso prior (Wang, 2012; Peterson et al., 2013), the continuous spike-and-slab prior (Wang, 2015; Li et al., 2017b), and the graphical horseshoe prior (Li et al., 2017a). This line of work draws direct connections between penalized likelihood schemes and, as their names suggest, the posterior modes in a Bayesian setting. Unlike the G-Wishart

prior, these shrinkage priors do not take point mass at zero for the off-diagonal elements in the precision matrix, and thus usually lead to efficient block sampling algorithms with improved scalability. However, fully Bayesian procedures still need to rely on stochastic search to achieve model selection, making it less appealing for many problems.

To address this issue, deterministic algorithms have been proposed to perform fast posterior exploration and mode searching in Gaussian graphical models (Gan et al., 2018; Li & McCormick, 2019; Deshpande et al., 2019). Motivated by the EMVS (Ročková & George, 2014) and spike-and-slab lasso (Ročková & George, 2018) procedures in the linear regression literature, the idea is to use a two-component mixture distribution, i.e., spike-and-slab priors, to parameterize off-diagonal elements in the precision matrix, which allows simultaneous model selection and parameter estimation. We will utilize a similar strategy for model estimation in this paper.

3. Bayesian joint graphical lasso priors

We first provide a Bayesian interpretation of the group and fused graphical lasso estimators. From a probabilistic perspective, it is well understood that estimators that optimize a penalized likelihood can often be seen as the posterior mode estimator under some suitable prior distributions. The Bayesian counterpart to (2) can be constructed by placing the prior $p(\{\mathbf{\Omega}\}) \propto \exp(-pen(\{\mathbf{\Omega}\}))$ on the precision matrices. Following directly from the Bayesian representation of lasso variants demonstrated in Kyung et al. (2010), we can rewrite $p(\{\mathbf{\Omega}\})$ as products of scale mixtures of normal distributions on the off-diagonal elements. That is, for the GGL prior, we can let

$$p(\{\mathbf{\Omega}\}|\boldsymbol{\tau}, \boldsymbol{\rho}) = C_{\boldsymbol{\tau}, \boldsymbol{\rho}}^{-1} \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk}; \mathbf{0}, (\boldsymbol{\Theta}_{jk}^{(G)})^{-1})$$
$$\prod_g \prod_j \text{Exp}(\omega_{jj}^{(g)}; \frac{\lambda_0}{2}) \mathbf{1}_{\{\mathbf{\Omega}\} \in \{M^+\}}, \quad (3)$$

$$\boldsymbol{\Theta}_{jk}^{(G)} = \text{diag}(\{\frac{1}{\rho_{jk}} + \frac{1}{\tau_{jkg}}\}_{g=1,...,G}), \quad (4)$$

$$p(\boldsymbol{\tau}, \boldsymbol{\rho}) \propto C_{\boldsymbol{\tau}, \boldsymbol{\rho}} \prod_{j<k} \Big(\exp(-\frac{\lambda_1^2}{2} \sum_g \tau_{jkg} - \frac{\lambda_2^2}{2} \rho_{jk}) \rho_{jk}^{-\frac{1}{2}}$$
$$\prod_g (\tau_{jkg}(\frac{1}{\tau_{jkg}} + \frac{1}{\rho_{jk}}))^{-\frac{1}{2}} \Big), \quad (5)$$

where $C_{\boldsymbol{\tau}, \boldsymbol{\rho}}$ is a normalizing constant and M^+ denotes the space of symmetric positive definite matrices. The normalizing constant is analytically intractable due to this constraint, but it cancels out in the marginal distribution of $p(\{\mathbf{\Omega}\})$. Such cancellation has been studied by several au-

thors (Wang, 2012; 2015; Liu et al., 2014). Similarly, the FGL prior can be defined as

$$p(\{\mathbf{\Omega}\}|\boldsymbol{\tau},\boldsymbol{\phi}) = C_{\boldsymbol{\tau},\boldsymbol{\phi}}^{-1} \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk}; \mathbf{0}, (\mathbf{\Theta}_{jk}^{(F)})^{-1})$$

$$\prod_g \prod_j \text{Exp}(\omega_{jj}^{(g)}; \frac{\lambda_0}{2}) \mathbf{1}_{\{\mathbf{\Omega}\}\in\{M^+\}}, \quad (6)$$

$$\mathbf{\Theta}_{jk}^{(F)} = \begin{cases} \theta_{gg} = \frac{1}{\tau_{jkg}} + \sum_{g\neq g'} \frac{1}{\phi_{jkgg'}} & g = 1,...,G \\ \theta_{gg'} = -\frac{1}{\phi_{jkgg'}} & g' \neq g \end{cases}$$
$$(7)$$

$$p(\boldsymbol{\tau},\boldsymbol{\phi}) \propto C_{\boldsymbol{\tau},\boldsymbol{\phi}} \prod_{j<k} \left(|\mathbf{\Theta}_{jk}^{(F)}|^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2} \sum_g \tau_{jkg} \right.$$
$$\left. -\frac{\lambda_2^2}{2} \sum_{g<g'} \phi_{jkgg'}) \prod_g \tau_{jkg}^{-\frac{1}{2}} \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}} \right).$$
$$(8)$$

It is also worth noting that both of the above priors are proper, and we leave the proof of the following proposition in the supplement.

Proposition 1. *The priors defined in (3) – (5) and (6) – (8) are proper and the posterior mode of $\{\mathbf{\Omega}\}$ is the solution of the group and fused graphical lasso problem with penalty terms defined in (2).*

4. Bayesian joint spike-and-slab graphical lasso priors

The Bayesian formulation of the joint graphical lasso problems discussed in the previous section provide shrinkage effects at the level of both individual precision matrices and across different classes. However, two issues remain. First, shrinkage priors alone do not produce sparse models since the posterior draws are never exactly 0. Thus, additional thresholding is needed to obtain a sparse representation of the graph structure. Second, the fixed penalty term, λ_1 and λ_2 may be too restrictive, as the non-zero elements in $\{\mathbf{\Omega}\}$ are penalized equally to elements close to zero (Li & McCormick, 2019). To reduce the bias from over-penalizing the large elements, different hyper-priors on λ_1 have been proposed to adaptively estimate the penalty term in Bayesian graphical lasso (Wang, 2012; Peterson et al., 2013).

Here we address both challenges simultaneously using the spike-and-slab approaches in Bayesian variable selection (George & McCulloch, 1993). In particular, we employ a set of latent indicators to construct a "selection" prior on both the group level and within-groups for the similarity penalties. We first let binary variables $\boldsymbol{\delta} = \{\delta_{jk}\}_{j<k}$ denote the existence of each edge in the graph, indexing the $2^{p(p-1)/2}$ possible models at the group level, so that $\delta_{jk} = 1$ indicates the (j,k)-th edge is selected for all precision matrices. We then let another set of binary variables

$\boldsymbol{\xi} = \{\xi_{jk}\}_{j<k}$ denote the *non-existence* of 'similarities' among the elements in the same cell of different precision matrices, so that $\xi_{jk} = 0$ indicates the (j,k)-th element is expected to be similar. We use the term 'similarity' here as a broad term parameterized by λ_2, since the behavior of the similarity depends on the form of the penalization. Conditional on the two binary indicators, we replace the fixed penalty parameters λ_1 and λ_2 by a mixture of edge-wise penalties that take values from $\{\lambda_1/v_0, \lambda_1/v_1\}$, and $\{\lambda_2/v_0, \lambda_2/v_1\}$ respectively, with fixed $v_1 > v_0 > 0$. That is, we introduce the following penalties conditional on $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$, and we propose the *doubly spike-and-slab* extensions to GGL and FGL as

$$pen(\{\mathbf{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}) = \frac{\lambda_0}{2} \sum_g \sum_j |\omega_{jj}^{(g)}| + \lambda_1 \sum_g \sum_{j<k} \frac{|\omega_{jk}^{(g)}|}{v_{\delta_{jk}}}$$
$$+ \lambda_2 \sum_{j<k} \frac{\widetilde{pen}(\boldsymbol{\omega}_{jk})}{v_{\xi_{jk}^*}}, \quad (9)$$

where $\widetilde{pen}(\boldsymbol{\omega}_{jk})$ is defined as before and $\xi_{jk}^* = \xi_{jk}\delta_{jk}$. The prior defined in (9) relate to the unconditional penalties by $pen(\{\mathbf{\Omega}\}) = pen(\{\mathbf{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}) - \log(p(\boldsymbol{\delta},\boldsymbol{\xi}))$, and we will refer to them as DSS-FGL and DSS-GGL.

In practice, we find it usually reasonable to enforce all elements from the spike distribution to also be similar, since the spike distribution is always chosen to have large penalization and leads to posterior modes at exactly 0. However, other types of element-wise dependence between δ_{jk} and ξ_{jk} are also possible with minor modifications. For example, we can also fix ξ_{jk} to be 1, so that the two penalty terms will always be proportional. We refer to this setting as spike-and-slab group and fused lasso (SS-GGL and SS-FGL) and discuss their behavior in the supplements.

The original GGL and FGL suffer from the same bias induced by the excessive shrinkage of lasso estimates. With the introduction of v_0 and v_1, we can adaptively estimate which $\boldsymbol{\omega}_{jk}$ to penalize in a data-driven way. As we discuss in more detail in Section 6, this adaptive shrinkage property can indeed significantly reduce bias imposed on the lasso penalty. That is, by choosing the hyperparameters so that $\lambda_i/v_0 \gg \lambda_i/v_1$, we impose only minimal shrinkage on values arising from the slab distribution. From now on, in order to avoid confusion from the overparameterization, we always fix $v_1 = 1$, and report results with the effective shrinkage parameters $\lambda_i/v_j, i,j \in \{1,2\}$. At this point, it may still seem that we need introduced one more hyperparameter that needs to be tunned, but as we show in Section 6, model selection can be achieved automatically without cross-validation.

For a Bayesian setup, we employ standard priors on the binary indicators to allow the edges to further share information on the sparsity level. The full generative model for

$\{\boldsymbol{\Omega}\}$ is:

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi},\theta) = C_\theta^{-1}C_{\boldsymbol{\delta},\boldsymbol{\xi}}^{-1}\exp(-pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}))\mathbf{1}_{\{\boldsymbol{\Omega}\}\in\{M^+\}},$$

(10)

$$p(\boldsymbol{\delta},\boldsymbol{\xi}|\pi_{\boldsymbol{\delta}},\pi_{\boldsymbol{\xi}}) \propto C_{\boldsymbol{\delta},\boldsymbol{\xi}}\prod_{j<k}\Big(\pi_{\boldsymbol{\delta}}^{\delta_{jk}}(1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}}$$
$$\pi_{\boldsymbol{\xi}}^{\xi_{jk}}(1-\pi_{\boldsymbol{\xi}})^{1-\xi_{jk}}\Big).$$

(11)

where θ denote $(\boldsymbol{\tau},\boldsymbol{\rho})$ for DSS-GGL, and $(\boldsymbol{\tau},\boldsymbol{\phi})$ for DSS-FGL. $C_{\boldsymbol{\delta},\boldsymbol{\xi}}$ is another intractable normalizing constant. We put standard Beta hyperpriors on the sparsity parameters so that $\pi_{\boldsymbol{\delta}} \sim \text{Beta}(a_1,b_1)$ and $\pi_{\boldsymbol{\xi}} \sim \text{Beta}(a_2,b_2)$. Throughout this paper, we let $a_1 = a_2 = 1$ and $b_1 = b_2 = p$.

Additionally, the above prior can be easily reparameterized with scale mixture of normal prior distributions similar as before by modifying the precision matrix $\boldsymbol{\Theta}$ into the following form, and they can be shown to be proper priors (the proofs can be found in the supplement):

$$\tilde{\boldsymbol{\Theta}}_{jk}^{(F)} = \begin{cases} \theta_{gg} = \frac{v_{\delta_{jk}}}{\tau_{jkg}} + \sum_{g<g'}\frac{v_{\xi_{jk}^*}}{\phi_{jkgg'}} & g=1,...,G \\ \theta_{gg'} = -\frac{v_{\xi_{jk}^*}}{\phi_{jkgg'}} & g'\neq g \end{cases}$$

(12)

$$\tilde{\boldsymbol{\Theta}}_{jk}^{(G)} = \text{diag}(\{\frac{v_{\xi_{jk}^*}}{\rho_{jk}} + \frac{v_{\delta_{jk}}}{\tau_{jkg}}\}_{g=1,...,G}).$$

(13)

Proposition 2. *The priors defined in (10) – (13) are proper, and the posterior mode of $\{\boldsymbol{\Omega}\}$ is the solution to the corresponding spike-and-slab version of joint graphical lasso penalties.*

Finally, it is straightforward to see that the proposed DSS-GGL and DSS-FGL penalties reduce to their non spike-and-slab counterparts when $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ are fixed to be 1. Several other spike-and-slab formulations in the literature can be seen as the special case of this prior when $G = 1$ as well. For example, the spike-and-slab mixture of double exponential priors considered in Deshpande et al. (2019) is a special case with $\lambda_2 = 0$. The spike-and-slab Gaussian mixtures in Li & McCormick (2019) can also be considered as a special case where we further fix $\tau_{jkg} = \infty$. This approach is also related to the work on sparse group selection in linear regression, as has been discussed in Xu et al. (2015) and (Zhang et al., 2014). As opposed to the point mass priors for the spike distribution commonly in the literature, our doubly spike-and-slab formulation of continuous mixtures allows the spike distribution to absorb small non-zero noises and facilitates fast dynamic explorations, as we will show in Section 6.

5. Model estimation

Given fixed λ_1, λ_2, and v_0, The representation of $p(\{\boldsymbol{\Omega}\})$ with the scale mixture of normal distributions allows the posterior to be sampled using a block Gibbs algorithm, as described in the supplement. However, choosing the hyperparameters can usually be a nontrivial task. Instead, we focus on faster deterministic methods to detect posterior modes under different choices of hyperparameters (Ročková & George, 2014). We present an EM algorithm that maximizes the complete-data posterior distribution $p(\{\boldsymbol{\Omega}\},\boldsymbol{\delta},\boldsymbol{\xi},\pi_{\boldsymbol{\delta}},\pi_{\boldsymbol{\xi}}|\boldsymbol{X})$ by treating the binary latent variables as "missing data." Similar ideas have been explored in recent work for linear regression (Ročková & George, 2014; 2018) and single graphical model estimation (Deshpande et al., 2019; Li & McCormick, 2019). Our EM algorithm maximizes the objective function $E_{\boldsymbol{\delta},\boldsymbol{\xi}|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)},\boldsymbol{X}}(\log p(\{\boldsymbol{\Omega}\},\pi_{\boldsymbol{\delta}},\pi_{\boldsymbol{\xi}}|\boldsymbol{X})|\cdot)$ by iterating between the E-step and M-step until changes in $\{\boldsymbol{\Omega}\}$ are within a small threshold.

In the E-step, we compute the conditional expectation terms in the objective function. It turns out that it suffices to find the conditional distribution of (δ_{jk},ξ_{jk}). The corresponding cell probabilities are proportional the the following mixture densities:

$$p_{\delta_{jk},\xi_{jk}}^*(j,k) \propto$$

$$\begin{cases} \pi_{\boldsymbol{\delta}}(1-\pi_{\boldsymbol{\xi}})\frac{\lambda_1\lambda_2}{v_0 v_1}\psi(v_1,v_0) & \delta_{jk}=1,\xi_{jk}=0 \\ \pi_{\boldsymbol{\delta}}\pi_{\boldsymbol{\xi}}\frac{\lambda_1\lambda_2}{v_1^2}\psi(v_1,v_1) & \delta_{jk}=1,\xi_{jk}=1 \\ (1-\pi_{\boldsymbol{\delta}})(1-\pi_{\boldsymbol{\xi}})\frac{\lambda_1\lambda_2}{v_0^2}\psi(v_0,v_0) & \delta_{jk}=0,\xi_{jk}=0 \end{cases}$$

where $\psi(a,b) = \exp(-\lambda_1\sum_g|\omega_{jk}^{(g)}|/a - \lambda_2\widetilde{pen}(\boldsymbol{\omega}_{jk})/b)$.

It is interesting to note that the three scenarios above represent three types of relationships among $\boldsymbol{\omega}_{jk}$: weak shrinkage but strong similarity, weak shrinkage and weak similarity, and strong shrinkage across classes. $E_{\cdot|\cdot}(\delta_{jk})$ and $E_{\cdot|\cdot}(\xi_{jk})$ are then simply the marginal probabilities in this 2 by 2 table, i.e., $E_{\cdot|\cdot}(\delta_{jk}) = p_{1,0}^*(j,k) + p_{1,1}^*(j,k)$, and $E_{\cdot|\cdot}(\xi_{jk}) = E_{\cdot|\cdot}(\delta_{jk}\xi_{jk}) = p_{1,1}^*(j,k)$. The EM algorithm also handles missing cells in $\boldsymbol{X}$ naturally. Assuming missing at random, the expectation can also be taken over the space of missing variables, by additionally computing $E_{\cdot|\cdot}(tr(\boldsymbol{S}_g\boldsymbol{\Omega}_g)) = tr(E_{\cdot|\cdot}((\boldsymbol{X}^{(g)})^T\boldsymbol{X}^{(g)})\boldsymbol{\Omega}_g))$, using the conditional Gaussian distribution of $\boldsymbol{x}_{i,miss}^{(g)}|\boldsymbol{x}_{i,obs}^{(g)}$. We relegate the derivations of the objective function to the supplement.

Given the expectations calculated in the E-step, one might proceed with conditional maximization steps using gradient ascent similar to the Gibbs sampler (Li & McCormick, 2019). Alternatively, since the maximization step is equiva-

lent to solving the following joint graphical lasso problem:

$$\{\mathbf{\Omega}\} = \text{argmax}_{\{\mathbf{\Omega}\}} \sum_g \frac{n_g}{2} \log |\mathbf{\Omega}_g| - \frac{1}{2} \sum_g tr(\mathbf{S}_g \mathbf{\Omega}_g)$$

$$- \frac{\lambda_0}{2} \sum_j \sum_g |\omega_{jj}^{(g)}|$$

$$- \sum_{j<k} \lambda_1 (\frac{p_{0,0}^*(j,k)}{v_0} + \frac{1 - p_{0,0}^*(j,k)}{v_1}) \sum_g |\omega_{jk}^{(g)}|$$

$$- \sum_{j<k} \lambda_2 (\frac{1 - p_{1,1}^*(j,k)}{v_0} + \frac{p_{1,1}^*(j,k)}{v_1}) \widetilde{pen}(\boldsymbol{\omega}_{jk}),$$

meaning we can use the ADMM algorithm described in Danaher et al. (2014).

6. Dynamic posterior exploration

The algorithm proposed in the previous section requires a fixed set of hyperparameters, $(\lambda_0, \lambda_1, \lambda_2, v_0)$. The posterior is relatively insensitive to the choice of λ_0 as long as it is not too large (Wang, 2015). Furthermore, unlike the original joint graphical lasso, where two tuning parameters need to be selected using cross-validation or model selection criterion, it turns out that we can leverage the self-adaptive property from the doubly spike-and-slab mixture setup to achieve automatic tuning using a path-following strategy (Ročková & George, 2018). Specifically, we consider a sequence of decreasing $v_0 = \{v_0^1, ..., v_0^L\}$ and some small λ_1 and λ_2. We initiate $\{\mathbf{\Omega}\}_0$ so that $\mathbf{\Omega}_{g0} = (\mathbf{S}_g/n_g + c\mathbf{I})^{-1}$, and iterative estimate $\{\widehat{\mathbf{\Omega}}\}_l$ with $v_0 = v_0^l$. After fitting the l-th model, we use the estimated graph structure to warm start the $(l+1)$-th model by initiating $\mathbf{\Omega}_g$ to be $\mathbf{\Omega}_{g0} \circ \mathbf{1}_{\hat{\boldsymbol{\delta}}>0}^l$, where $\mathbf{1}_{\hat{\boldsymbol{\delta}}>0}^l$ denotes the group level graph structure at the l-th iteration. As v_0 decreases, the shrinkage imposed on the spike elements steadily increases and leads to sparser models. As noted in Ročková & George (2018), the solution path from such dynamic reinitialization procedure usually 'stabilizes' as v_0 becomes closer to 0 in linear regression. We found similar behavior in our spike-and-slab joint graphical lasso models too, as illustrated in Figure 1.

To demonstrate the dynamic posterior exploration in action, we simulated a small dataset from two classes, with $n_g = 150$ for $g = 1, 2$, and $p = 100$. The two underlying graphs differ by 5 edges: The first precision matrix contains a 10-node block with an AR(1) precision matrix where $(\Omega^{-1})_{jk} = \rho_1^{|j-k|}$, and $\rho_1 = 0.7$; the second precision matrix in the second class contains a common 5-node AR(1) block with $\rho_2 = 0.9$. The rest of the nodes are all independent. We fit the fused graphical lasso with a sequence of λ_1, and fixed $\lambda_2 = 0.1$, which leads to the best performance in this experiment; and DSS-FGL with $\lambda_1 = 1$, and $\lambda_2 = 1$. Figure 1 shows the FGL and DSS-FGL solution path. Unlike the continuous shrinkage of FGL, the zero

and non-zero elements under DSS-FGL tend to be separated into two stable clusters as the effective shrinkage λ_1/v_0 increases beyond a critical point. Danaher et al. (2014) noted that graph selection using AIC tends to favor large models. This example also confirms this observation as the likelihood evaluation for smaller models suffers from the overly aggressive shrinkage. In this example, AIC selects 27 edges in both classes, leading to 41 false positives. Assuming we know the true graphs, the best model in terms of edge selection along the FGL solution path contains one false negative edge as shown in Figure 1. However, without accurate prior knowledge of graph sparsity, correctly identifying this model is typically difficult, if not impossible. On the other hand, the stable model from the DSS-FGL solution path yields 4 false positive edges in the second graph, but with clear visual separation from the regularization plot: only one false positive edge stabilizing to a larger value away from 0. Thus in practice, the solution path also provides a visual tool to threshold the small values close to 0. Additionally, the bias of the final precision matrices compared to the truth is also much smaller than the best FGL solution.

We also find that the converged region is insensitive to the choice of λ_1 and λ_2 in all our experiments, as the model allows a flexible combination of shrinkage through the adaptive estimation of p^*. The supplement includes an empirical assessment of sensitivity in the simulation experiments.

7. Numerical results

Simulation experiments To assess the performance of the proposed models, we consider a three-class problem similar to the study carried out in Danaher et al. (2014). We first generate three networks with $p = 500$ features with 10 equal sized unconnected subnetworks following power law degree distributions. Exactly one and two subnetworks are removed from the second and third class. The details of the data generating process can be found in the supplement. The results comparing the proposed model and joint graphical lasso are shown in Figure 2. As discussed before, the DSS-FGL and DSS-GGL achieve model selection automatically. Thus we compare the selected models with the average curve of FGL and GGL under different tuning parameters. Figure 2(a) and (c) show that DSS-FGL and DSS-GGL usually achieves better structure learning performance for both identifying edges and *differential edges*. The differential edges are defined as the edges for the (g, g') pair with $|\omega_{jk}^{(g)} - \omega_{jk}^{(g')}| > 0.01$. Figure 2(b) and (d) demonstrate the bias-diminishing property of the proposed models compared to the joint graphical lasso estimator at varying sparsity levels, measured by the number of edges in (b) and L_1 norm of the estimator in (d). On average, both the sum of bias as measured by the Frobenius norm, $||\widehat{\mathbf{\Omega}}_g - \mathbf{\Omega}_g||_F$, and the Kullback-Leibler (KL) divergence achieved by the
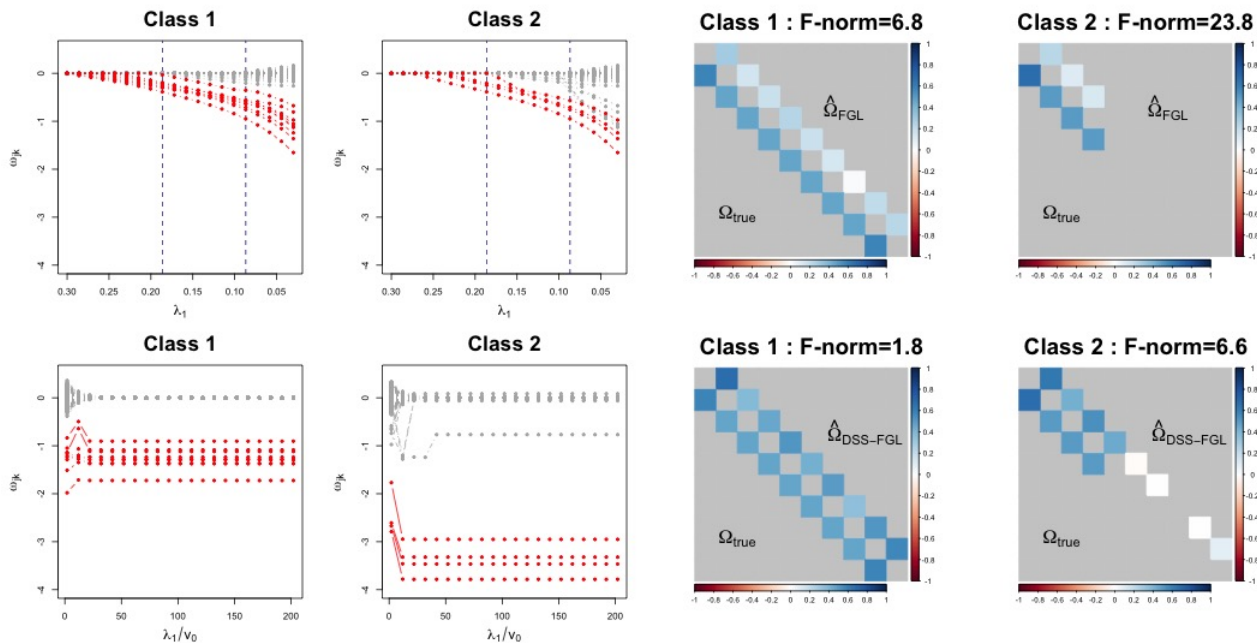
Figure 1. The solution paths and estimated precision matrices of FGL (upper row) and DSS-FGL (lower row). The red nodes correspond to true edges and the gray nodes correspond to 0's. The two vertical lines in the FGL solution path indicate the model that best matches the true sparsity (left) and the 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 as measured by the Frobenius norm, $||\hat{\mathbf{\Omega}}_g - \mathbf{\Omega}_g||_F$, is also printed in the captions.

proposed model is much smaller.

Symptom networks of verbal autopsy data We applied the DSS-FGL and DSS-GGL to a gold-standard dataset of verbal autopsy (VA) surveys (Murray et al., 2011). VA surveys are widely adopted in countries without full-coverage civil registration and vital statistics systems to estimate cause of death. They are conducted by interviewing caregiver of a recently deceased person about the decedent's health history. The standard procedure of preparing the collected data is to dichotomize all continuous variables into binary indicators and many algorithms have been proposed to automatically assign causes of death using the binary input (Byass et al., 2012; Serina et al., 2015; McCormick et al., 2016). However, more information may be gained by modeling the continuous variables directly (Li et al., 2017b). Here we focus on modeling the joint distribution of the continuous variables. The 27 continuous variables in this dataset contain representations of the duration of symptoms, such as response to the question 'how many days did the fever last', and age of the decedents. It is usually reasonable to assume the response to these questions are jointly distributed in similar ways conditional on each cause of death. We take the raw responses and transform raw duration x_{ij}

by $\log(x_{ij} + 1)$. We then let $X_{ij}^{(g)}$ denote the j-th transformed variable for observation i due to the cause g. The full dataset contains death assigned to 34 causes. We applied DSS-FGL with $\lambda_1 = \lambda_2 = 1$ to the three largest determined causes of death in this data: Stroke ($n = 630$), Pneumonia ($n = 540$), and AIDS ($n = 542$) in Figure 3. The estimated graphs under other models are discussed in the supplement. Both DSS-FGL and DSS-GGL estimated similar graphs and discovered interesting differential symptom pairs, such as the strong conditional dependence between the duration of illness and paralysis in deaths due to stroke. Further incorporating the DSS-FGL and DSS-GGL formulation of multiple precision matrices into a classification framework would likely improve accuracy over existing methods (e.g. McCormick et al. (2016); Byass et al. (2012)) for automatic cause-of-death assignment.

Prediction of missing mortality rates Beyond structure learning, the bias reduction in estimating $\{\mathbf{\Omega}\}$ also makes the proposed method more appealing for prediction tasks involving sparse precision matrices. In this example, we illustrate the potential of using the proposed methods to impute missing mortality rates using a cross-validation study. We construct the data matrices $X_{ij}^{(g)}$ as the log transformed

Figure 2. Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications. 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 3. Estimated edges between the symptoms under the three causes using DSS-FGL. 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.

central mortality rate of age group j in year i for subpopulation g (e.g., male and female). Standard approaches in demography, such as the Lee-Carter model (Lee & Carter, 1992), typically use dimension reduction techniques to estimate mean effects due to age and time, and consider the residuals as independent measurement errors. However, residuals from such models are usually still highly correlated (Fosdick & Hoff, 2014). We consider estimating the residual structure with the 1×1 gender-specific mortality table up to age 100 in the US over the period of 1960 to 2010 using data obtained from the Human Mortality Database (HMD) (University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)). For both the male and female mortality, we first randomly selected 25 years and remove 25 data points in each of those years. We then fit a Lee-Carter model to estimate the mean model and interpolate the missing rates. Next, we estimate the covariance matrices among the 101 age groups in both genders using FGL and DSS-FGL from the residuals.

The estimated residuals for the missing values can then be obtained by the E-step in our EM algorithm, or as the expectation from the conditional Gaussian distributions with covariance matrices estimated by FGL. The average mean squared errors (MSEs) for the prediction of missing log rates are summarized in Table 1. Imputation based on DSS-FGL precision matrix reduces the MSE by 27.8% compared to simple interpolation of the mean model (i.e., assuming i.i.d errors), compared to the 6.5% reduction from the FGL precision matrix with the same complexity. The estimated graphs are in the supplement.

8. Discussion

In this paper, we introduced a new class of priors for joint estimation of multiple graphical models. The proposed doubly spike-and-slab mixture priors, DSS-FGL and DSS-GGL, provide self-adaptive extensions to the joint graphical lasso penalties, and achieves simultaneous model selection

Table 1. Average and standard deviation of the mean squared errors from 50 cross-validation experiments. The FGL model is selected to have the same number of edges as the DSS-FGL.

	i.i.d	FGL	DSS-FGL
Average MSE	0.00372	0.00348	0.00268
SE of the MSEs	0.00030	0.00031	0.00028

and parameter estimation. Moreover, while taking advantage of the flexible class of penalty functions, the dynamic posterior exploration procedure allows the penalties to be adaptively estimated in a data-driven way, thus freeing practitioners from choosing multiple tuning parameters. This is especially useful in domains where sample sizes are too small to reliably perform cross-validation. Finally, additional procedures such as multiple random initializations and deterministic annealing may be further incorporated into the proposed algorithm to better explore the posterior surface. While not discussed in the main paper, we note that the posterior uncertainty may be estimated using the Gibbs sampler described in the supplement.

The proposed framework can be extended in a few directions. First, we have assumed all classes to be exchangeable, as reflected in the penalty functions for the between-class similarity. When the classes exhibit hierarchical structures or different strengths of similarities, the indicator ξ may be modeled as functions of the class membership as well. Markov Random Field priors discussed in Saegusa & Shojaie (2016) and Peterson et al. (2015) may also be used to model the between-class similarities. Second, we have considered the estimation of missing values in the data matrices. It is also straightforward to extend to data with missing class labels. In this way, the proposed methods can be extended to classification or discriminant analysis based on sparse precision matrices (Hao et al., 2016). Finally, the proposed model is estimated using an EM algorithm that is iteratively solving the joint graphical lasso problem. It may be interesting to construct coordinate ascent algorithms that optimize on the objective function directly, similar to that described in Ročková & George (2018) for linear regression.

The codes for the proposed algorithm are available at https://github.com/richardli/SSJGL.

Acknowledgements

References

Briollais, L., Dobra, A., Liu, J., Friedlander, M., Ozcelik, H., Massam, H., et al. A Bayesian graphical model for genome-wide association studies (GWAS). *The Annals of Applied Statistics*, 10(2):786–811, 2016.

Byass, P., Chandramohan, D., Clark, S. J., D'Ambruoso, L., Fottrell, E., Graham, W. J., Herbst, A. J., Hodgson, A., Hounton, S., Kahn, K., et al. Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Global Health Action*, 5, 2012.

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

Deshpande, S. K., Ročková, V., and George, E. I. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics (forthcoming)*, 2019.

Dobra, A. Graphical modeling of spatial health data. *arXiv preprint arXiv:1411.6512*, 2014.

Dobra, A., Eicher, T. S., and Lenkoski, A. Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models. *Statistical Methodology*, 7 (3):292–306, 2010.

Fosdick, B. K. and Hoff, P. D. Separable factor analysis with applications to mortality data. *The annals of applied statistics*, 8(1):120, 2014.

Gan, L., Narisetty, N. N., and Liang, F. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, pp. 1–14, 2018.

George, E. I. and McCulloch, R. E. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

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

Hao, B., Sun, W. W., Liu, Y., and Cheng, G. Simultaneous clustering and estimation of heterogeneous graphical models. *arXiv preprint arXiv:1611.09391*, 2016.

Kyung, M., Gilly, J., Ghoshz, M., and Casellax, G. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

Lee, R. D. and Carter, L. R. Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.

Lenkoski, A. and Dobra, A. Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157, 2011.

Li, Y., Craig, B. A., and Bhadra, A. The graphical horseshoe estimator for inverse covariance matrices. *arXiv preprint arXiv:1707.06661*, 2017a.

Li, Z. R. and McCormick, T. H. An Expectation Conditional Maximization approach for Gaussian graphical models. *Journal of Computational and Graphical Statistics (forthcoming)*, 2019.

Li, Z. R., McCormick, T. H., and Clark, S. J. Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *arXiv preprint arXiv:1711.00877*, 2017b.

Lin, Z., Wang, T., Yang, C., and Zhao, H. On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics*, 73(3):769–779, 2017.

Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A. C., et al. Bayesian regularization via graph laplacian. *Bayesian Analysis*, 9(2):449–474, 2014.

Ma, J. and Michailidis, G. Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research*, 17(166):1–48, 2016.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

Mohammadi, A., Wit, E. C., et al. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

Murray, C. J., Lopez, A. D., Black, R., Ahuja, R., Ali, S. M., Baqui, A., Dandona, L., Dantzer, E., Das, V., Dhingra, U., et al. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population health metrics*, 9(1):27, 2011.

Peterson, C., Vannucci, M., Karakas, C., Choi, W., Ma, L., and Meletić-Savatić, M. Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. *Statistics and its Interface*, 6(4):547, 2013.

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

Ročková, V. and George, E. I. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

Ročková, V. and George, E. I. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521): 431–444, 2018.

Saegusa, T. and Shojaie, A. Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics*, 10(1):1341, 2016.

Serina, P., Riley, I., Stewart, A., Flaxman, A. D., Lozano, R., Mooney, M. D., Luning, R., Hernandez, B., Black, R., Ahuja, R., et al. A shortened verbal autopsy instrument for use in routine mortality surveillance systems. *BMC medicine*, 13(1):1, 2015.

Tan, L. S., Jasra, A., De Iorio, M., Ebbels, T. M., et al. Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, 11(4):2222–2251, 2017.

University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Human Mortality Database. Available at www.mortality.org or www.humanmortality.de (data downloaded on 05/02/2018).

Wang, H. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012. ISSN 19360975. doi: 10.1214/12-BA729.

Wang, H. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

Xu, X., Ghosh, M., et al. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4): 909–936, 2015.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620, 2014.