Statistica Sinica Preprint No: SS-2021-0245	
Title	A Bayesian Subset Specific Approach to Joint Selection of Multiple Graphical Models
Manuscript ID	SS-2021-0245
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021-0245
Complete List of Authors	Peyman Jalali, Kshitij Khare and George Michailidis
Corresponding Author	George Michailidis
E-mail	gmichail@ufl.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

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

Abstract: The joint estimation of multiple graphical models from high-dimensional data has been studied in the statistics and machine learning literature, owing to its importance in diverse fields including molecular biology, neuroscience, and the social sciences. We propose a Bayesian approach that decomposes the model parameters across multiple graphical models into shared components across subsets of models and edges, and idiosyncratic components. This approach leverages a novel multivariate prior distribution, coupled with a jointly convex regression-based pseudo-likelihood that enables fast computation using a robust and efficient Gibbs sampling scheme. We establish strong posterior consistency for model selection under high-dimensional scaling, with the number of variables growing exponentially as a function of the sample size. Lastly, we demonstrate the efficiency of the proposed approach in borrowing strength across models to identify shared edges using both synthetic and real data.

Key words and phrases: Pseudo-likelihood; Gibbs sampling; posterior consistency; Omics data.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

1. Introduction

The *joint estimation* of multiple *related* Gaussian graphical models has attracted much interest in statistics and machine learning owing to its wide application in biomedical studies involving omics data (e.g. Pierson et al. (2015) and Kling et al. (2015)), as well as in text mining and roll call voting Guo et al. (2011). The key idea of this approach which makes it preferable to separate network-wise estimations, is to "borrow strength" across related models, and thus enhance the "effective" sample size used to estimate the model parameters. In high-dimensional settings, joint estimation is achieved primarily by using a penalty function to induce sparsity/zeros in the group-specific inverse covariance (precision) matrices. Specifically, Guo et al. (2011), who first formulated the problem, model the elements of each inverse covariance matrix as the product of a component, *common* across all models and an *idiosyncratic* (model-specific) component, and impose an ℓ_1-penalty on each one. Thus, when the penalty sets the common component to zero, the corresponding edge is absent across all models. However, if the common component is not zero, edges can be absent because the penalty sets the idiosyncratic component to zero for selected models. Another set of approaches aims to achieve a certain amount of "fusing" across all models being considered, focusing on both the presence and the absence of common edges across *all* models simultaneously. Examples of such approaches include

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

those of Danaher et al. (2014), who employed a group lasso and/or a fused lasso penalty on each edge parameter across all models, and Cai et al. (2016), who used a mixed ℓ_1/ℓ_∞ norm for the same task.

However, in many application settings, only a subset of edges exhibit shared connectivity patterns across models, with the reminder showing different connectivity patterns in each model. In other settings, subsets of edges share common connectivity patterns across only a subset of models. In both instances, the previously mentioned approaches exhibit rather poor performance in terms of discovering these more complex patterns. To address this issue, Ma and Michailidis (2016) proposed a *supervised* approach based on fusing through a group lasso penalty, wherein the various connectivity patterns across the subsets of edges and subsets of models are known *a priori*. An alternative supervised approach Saegusa and Shojaie (2016) employs a similarity graph penalty for fusing across models, and an ℓ_1-penalty to obtain sparse model estimates. The similarity graph is assumed to be known a priori.

The Bayesian paradigm comes with the advantage of natural uncertainty quantification through the posterior distribution, and also a natural structured mechanism for incorporating prior information. Peterson et al. (2015) proposed a Bayesian variant of the approach of Saegusa and Shojaie (2016) using a Markov random field prior distribution to capture model similarity, followed by a spike-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

and-slab prior distribution on the edge model parameters. Tan et al. (2017) developed a Bayesian approach that, similarly to Peterson et al. (2015), uses G-Wishart prior distributions on the group-specific precision matrices, given the sparsity patterns in each group, and then employs a multiplicative model, based hierarchical prior on these networks to induce similarity/dependence. Recent works including those of Shaddox et al. (2018) and Petersen et al. (2020), extend the ideas in Wang (2015) to a joint estimation setting for improved computational scalability. However, this class of approaches still suffers from scalability problems beyond moderate-dimensional settings with 150 or so variables. There are two main computational difficulties for posterior sampling. First, the precision matrices are restricted to be positive definite (p.d.). Second, the dependence structure between groups is induced through priors on large discrete spaces of sparsity patterns (graphs) for the precision matrices, and the conditional updates of the relevant discrete/latent variables and respective hyper-parameters can be messy and add significantly to the computational burden.

For the single graphical model estimation problem based on n independent and identically distributed (i.i.d.) observations from a distribution with inverse-covariance/precision matrix Ω, the entries in the ith row of Ω can be interpreted as least squares coefficients when regressing the ith variable against the other variables (see (4.9)). This idea has been leveraged to develop quasi-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

likelihood/pseudo-likelihood based approaches; see Meinshausen and Bühlmann (2006), Lin et al. (2017), and Atchade (2019). These approaches relax the p.d. constraint on Ω, leading to significant improvement in computational speed. *Note that relaxing the p.d. constraint does not create issues/complications for model/sparsity selection in* Ω, which is often the key objective. If a p.d. estimate of Ω is needed for a downstream application, it can be obtained in a straightforward manner, for example, by computing the restricted maximum likelihood estimation (MLE) based on the estimated sparsity pattern.

Lin et al. (2017) extend this idea to the joint graphical model estimation problem, using the regression-based neighborhood selection procedure of Meinshausen and Bühlmann (2006) with an alternate version of the Markov random field priors in Peterson et al. (2015) to induce dependence between groups. Relaxing the p.d. constraint leads to a significant improvement in computational performance compared with that of the likelihood-based approaches mentioned above. However, the use of matrix inversions and latent variable updates still leads to a steep increase in the computational cost of the corresponding algorithm, labeled BNS, as the number of variables p increases (see Section 5.2). "Maximization-based" Bayesian approaches for joint graphical estimation, which focus on obtaining posterior modes, have been proposed by Li et al. (2018) and Yang et al. (2021). These approaches are computationally scalable,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

but unlike the other "sampling-based" approaches discussed above, they do not generate samples from the posterior and are unable to provide detailed uncertainty quantification. For example, Yang et al. (2021) use relevant conditional posterior probabilities evaluated at the posterior mode to evaluate the uncertainty for individual edges, but do not provide more nuanced uncertainty for the joint inclusion/exclusion of multiple edges (see Table G.2, for example).

While uncertainty quantification through the posterior is an attractive feature of a Bayesian approach, in high-dimensional settings, it is crucial to rigorously justify its validity. In the current context, this corresponds to establishing strong posterior selection consistency, that is, proving that the posterior distribution of the combined sparsity pattern in the group-specific precision matrices asymptotically places all of its mass on the "true" sparsity pattern in the high-dimensional data-generating model. Although Yang et al. (2021) establish selection consistency for the posterior mode, a high-dimensional strong selection consistency result for joint graphical model estimation has not been established for *any* of the Bayesian approaches discussed above.

Given this background, the key objective of this study is to develop a *scalable sampling-based Bayesian* approach with *high-dimensional selection consistency guarantees* for joint estimation and uncertainty quantification for multiple related Gaussian graphical models that exhibit complex edge connectivity

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

patterns across models for different subsets of edges. We avoid the Markov-random-field-based approach for inducing group similarity, and instead take a more direct approach based on a subset-specific decomposition (see (2.2)) of the group-specific precision matrices. We then introduce a novel *subset-specific ($\mathcal{S}^2$) prior* that, for each edge, aims to select the subset of models it is common to (see (3.6)–(3.8) and Remark 1). We couple these with the *jointly convex* regression-based pseudo-likelihood used in Khare et al. (2015) for estimating a single Gaussian graphical model.

The above framework leads to an easy to implement and scalable Gibbs sampling scheme for exploring the posterior distribution. The corresponding algorithm, called the Bayesian joint network selector(BJNS), essentially involves $O(p^2)$ *univariate updates* from relevant mixture distributions, and avoids the need for matrix inversions or latent variables. As a result, the computational performance of the BJNS is an order of magnitude faster than that of BNS algorithm of Lin et al. (2017) (see Section 5.2). Our direct subset-specific approach can lead to significantly improved statistical selection performance compared with that of existing methods (see Section 5). We also establish a strong posterior model selection consistency result (Theorem 1) for the proposed approach. Intuitively, the proposed framework achieves the objectives set forth in Ma and Michailidis (2016), *without* requiring an a priori specification of the shared edge

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

connectivity patterns; thus, the approach is fully unsupervised. Furthermore, the availability of posterior samples allows for uncertainty quantification in the form of subset-specific inclusion probabilities (see Figure D.1 and Table G.2).

Note that the main goal of the proposed framework is model/sparsity selection. As described in detail in Section 2.1, for identifiability purposes, we need a constraint that imposes restrictions on the magnitude of the relevant precision matrix entries. First, this constraint has no adverse effect on the main task of sparsity selection in the sense that the framework still produces a valid posterior distribution on the space of all possible sparsity patterns for the various group-specific precision matrices. As our simulation and consistency results show, the BJNS performs very well in terms of sparsity selection compared with existing methods, even when this constraint is not satisfied in the data-generating model. Second, as pointed out above, a regression-based approach is not directly useful for magnitude estimation anyway, because the resulting estimates of the precision matrices are not guaranteed to be p.d.. If such estimates are needed for downstream applications, one can obtain them by using the respective MLEs restricted to the estimated sparsity pattern for each group-specific matrix (see (4.12)). *Note that the resulting estimates are guaranteed to be p.d. and are free from the identifiability constraint in Section 2.1 (which is used only for sparsity selection, and does not play a role in (4.12).* We also provide a simulation

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

study that evaluates the (magnitude) estimation accuracy in the Supplementary Material Section D.4.

The remainder of the paper is organized as follows. Section 2 formulates the problem, and Section 3 introduces the $\mathcal{S}^2$ prior. Section 4 shows how to obtain and sample from the posterior distribution. Section 5 presents extensive numerical results and comparisons, and Section 6 presents a metabolomics application using a case-control study on inflammatory bowel disease. Section 7 establishes the high-dimensional posterior consistency for the BJNS, and Section 8 concludes the paper.

2. Framework for Joint Sparsity Selection

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

ϑ_k as follows:

$$\vartheta_k = \{r \in \mathcal{P}(K) \setminus \{0\} : k \in r\}, \qquad k = 1, ..., K. \tag{2.1}$$

It is easy to check that each ϑ_k is the collection of subsets that contain k, and has 2^{K-1} members. Denote by $\mathbf{\Psi}^r$ the matrix that contains common patterns among the precision matrices $\{\mathbf{\Omega}^j\}_{j \in r}$. For any singleton set $r = \{k\}$, the nonzero elements in the matrix $\mathbf{\Psi}^r$ correspond to edges that are unique to group k. For any other set r containing more than a single element, the nonzero elements in the matrix $\mathbf{\Psi}^r$ correspond to edges (and their magnitudes) that are common across all members in r (and not present in other networks). For example, the nonzero elements in $\mathbf{\Psi}^{123} := \mathbf{\Psi}^{\{1,2,3\}}$ correspond to edges that are shared exclusively by networks $1, 2$, and 3.

Therefore, each precision matrix $\mathbf{\Omega}^k$ can be decomposed as

$$\mathbf{\Omega}^k = \sum_{r \in \vartheta_k} \mathbf{\Psi}^r, \qquad k = 1, ..., K, \tag{2.2}$$

where $\sum_{r \in \vartheta_k} \mathbf{\Psi}^r$ accounts for all the structures in $\mathbf{\Omega}^k$ that are either unique to group k (i.e., $\mathbf{\Psi}^k$) or are shared exclusively between group k and some combination of other groups (i.e., $\sum_{r \in \vartheta_k \setminus \{k\}} \mathbf{\Psi}^r$). We further assume that $\mathbf{\Psi}^k \in \mathbb{M}_p^+$, for $k = 1, 2, ..., K$, where $\mathbb{M}_p^+$ denotes the space of all $p \times p$ matrices with positive diagonal entries. Finally, the diagonal entries of every joint matrix $\mathbf{\Psi}^r$, with $r \in \cup_{k=1}^K (\vartheta_k \setminus \{k\})$, are set to zero; in other words, the diagonals entries of $\mathbf{\Omega}^k$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

are contained in the corresponding $\mathbf{\Psi}^k$.

To illustrate the notation, consider the case of $K = 3$, and following the notation in (2.1), define the sets: $\vartheta_1 = \{\{1\}, \{12\}, \{13\}, \{123\}\}$, $\vartheta_2 = \{\{2\}, \{12\}, \{23\}, \{123\}\}$, and $\vartheta_3 = \{\{3\}, \{13\}, \{23\}, \{123\}\}$. Then, decompose the precision matrices $\mathbf{\Omega}^1$, $\mathbf{\Omega}^2$, and $\mathbf{\Omega}^3$, as follows:

$$\mathbf{\Omega}^1 = \sum_{r \in \vartheta_1} \mathbf{\Psi}^r = \mathbf{\Psi}^1 + \mathbf{\Psi}^{12} + \mathbf{\Psi}^{13} + \mathbf{\Psi}^{123}, \qquad \mathbf{\Omega}^2 = \sum_{r \in \vartheta_2} \mathbf{\Psi}^r = \mathbf{\Psi}^2 + \mathbf{\Psi}^{12} + \mathbf{\Psi}^{23} + \mathbf{\Psi}^{123},$$

$$\mathbf{\Omega}^3 = \sum_{r \in \vartheta_3} \mathbf{\Psi}^r = \mathbf{\Psi}^3 + \mathbf{\Psi}^{13} + \mathbf{\Psi}^{23} + \mathbf{\Psi}^{123},$$

where $\mathbf{\Psi}^1$, $\mathbf{\Psi}^2$, and $\mathbf{\Psi}^3$ are matrices that contain group-specific patterns, $\mathbf{\Psi}^{12}$, $\mathbf{\Psi}^{13}$, and $\mathbf{\Psi}^{23}$ are matrices that contain patterns shared across pairs of models (for subsets of the edges), and the matrix $\mathbf{\Psi}^{123}$ contains patterns shared across all models.

2.1 Identifiability Considerations

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

$$\omega_{ij}^k = \sum_{r \in \vartheta_k} \psi_{ij}^r, \qquad 1 \leq i < j \leq p, 1 \leq k \leq K, \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

where ω_{ij}^k and ψ_{ij}^r are the ij^{th} coordinates of the matrices $\mathbf{\Omega}^k$ and $\mathbf{\Psi}^r$, respectively. We consider the upper off-diagonal entries, owing to the symmetry of the precision matrix, and thus define the vectors $\boldsymbol{\theta}_{ij}$, for every $1 \leq i < j \leq p$, as

$$\boldsymbol{\theta}_{ij} = \left\{\psi_{ij}^r\right\}_{r \in \mathcal{P}(K)\setminus\{0\}}, \tag{2.4}$$

where each $\boldsymbol{\theta}_{ij}$ has $2^K - 1$ distinct parameters. For *identifiability purposes*, we require that each vector $\boldsymbol{\theta}_{ij}$ has at most *one nonzero element*. Note that under this constraint, if an edge (i, j) is shared among many groups, the nonzero element will be allocated to the maximal set $s \in \cup_{k=1}^K (\vartheta_k \setminus \{k\})$, whereas *all subsets* of s will be allocated a zero value. As an example, consider again the case of $K = 3$ groups and an edge (i, j) shared among all three groups. In this case, the edge is allocated to the $\mathbf{\Psi}^{123}$ component and not to any other components, such as $\mathbf{\Psi}^{12}$ or $\mathbf{\Psi}^{13}$. Hence, $\mathbf{\Psi}_{ij}^{123}$ is nonzero, but $\mathbf{\Psi}_{ij}^{12} = \mathbf{\Psi}_{ij}^{13} = \mathbf{\Psi}_{ij}^{23} = \mathbf{\Psi}_{ij}^{1} = \mathbf{\Psi}_{ij}^{2} = \mathbf{\Psi}_{ij}^{3} = 0$. Next, we discuss the implications of this identifiability constraint.

The precision matrices $\{\Omega_k\}_{k=1}^K$ have a total of $Kp(p+1)/2$ parameters. We expand these precision matrices in terms of the subset specific matrices $\{\Psi^r\}_{e \in \mathcal{P}(K)\setminus\{0\}}$ (see (2.3)). This expanded set of parameter matrices has $Kp + (2^K - 1)p(p-1)/2$ parameters, in all. *The identifiability constraint reduces the number of parameters to* $Kp + p(p-1)/2$, and thereby helps significantly with computational scalability. Further, sparsity constraints are introduced by the spike-and-slab priors described in Section 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

Another consequence of the identifiability constraint is that if edge (i, j) is shared by a subset s, then all $\{\omega_{ij}^k\}_{k\in s}$ have the same magnitude. Note that it makes sense in a variety of applications for shared edges to have similar magnitudes and/or the same sign. In fact, the approaches of Danaher et al. (2014)and Li et al. (2018) based on a group/fused lasso encourage similarity, or even exact equality of the magnitudes across the shared edges. As demonstrated by the simulation results in Table 3, *our working model still performs well in selecting the sparsity/skeleton compared with existing methods when the shared edges have different magnitudes, but the same sign*. This is also supported by the theoretical results in Section 7, where we allow the true precision matrices to have different magnitudes, but the same sign for the shared edges.

Finally, because we employ a regression-based pseudo-likelihood (see (4.10), which is well defined as long as the relevant precision matrix has positive diagonal entries, we relax the p.d. constraint on $\{\mathbf{\Omega}^k\}_{k=1}^K$ for faster computation. *Note that neither the identifiability constraint nor the p.d. relaxation restrict the range of allowable sparsity patterns, and both are only used in the working framework for sparsity selection.* Once the sparsity patterns are selected, if needed, we perform a simple refitting step (see (4.12)) to obtain p.d. estimates of the K precision matrices that obey the selected sparsity patterns and are completely free of the above identifiability constraint.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

3. Subset-Specific ($\mathcal{S}^2$) Prior Distribution

Next, we construct a novel prior distribution that respects the introduced identifiability constraint and encourages further sparsity in the parameters. For any generic symmetric $p \times p$ matrix $\mathbf{A}$, define $\underline{\boldsymbol{a}} = (a_{12}, a_{13}, ..., a_{p-1p})$ and $\boldsymbol{\delta}_{\mathbf{A}} = (a_{11}, ..., a_{pp})$, where because of the symmetric nature of $\mathbf{A}$, the vector $\underline{\boldsymbol{a}}$ contains all of the off-diagonal elements, and $\boldsymbol{\delta}_{\mathbf{A}}$ contains all of the diagonal elements. In particular, $\underline{\boldsymbol{\psi}}^r$ is the vectorized version of the off-diagonal elements of $\boldsymbol{\Psi}^r$. Using the above notation, define $\boldsymbol{\Theta}$ as the vector obtained by combining the vectors $\left\{\underline{\boldsymbol{\psi}}^r, r \in \mathcal{P}(K) \setminus \{0\}\right\}$. To illustrate, for $K = 3$ groups, $\boldsymbol{\Theta}$ is given by

$$\boldsymbol{\Theta} = (\underline{\boldsymbol{\psi}}^{123\prime}, \underline{\boldsymbol{\psi}}^{23\prime}, \underline{\boldsymbol{\psi}}^{13\prime}, \underline{\boldsymbol{\psi}}^{12\prime}, \underline{\boldsymbol{\psi}}^{3\prime}, \underline{\boldsymbol{\psi}}^{2\prime}, \underline{\boldsymbol{\psi}}^{1\prime})'. \tag{3.1}$$

Note that $\boldsymbol{\Theta}$ is a rearrangement of the vector $(\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{13}, ..., \boldsymbol{\theta}_{p-1p})'$. Thus, according to the location of the nonzero coordinates in $\boldsymbol{\theta}_{ij}$ (2^K possibilities), there are $2^{\frac{Kp(p-1)}{2}}$ possible sparsity patterns across the K groups for $\boldsymbol{\Theta}$. Let ℓ be a generic sparsity pattern for $\boldsymbol{\Theta}$, and denote the set of all $2^{\frac{Kp(p-1)}{2}}$ sparsity patterns by $\mathcal{L}$. To illustrate, consider $K = 2$ groups and $p = 3$ variables. In this case, each matrix has three off-diagonal edges ($\{ij : 1 \leq i < j \leq p\} = \{12, 13, 23\}$). Assume edge 12 is shared between the two groups, edge 13 is unique to group 2, and edge 23 is absent from both groups. In this case, $\boldsymbol{\Theta}$ is given by $\boldsymbol{\Theta} = ((\psi^{12}_{12}, 0, 0), (0, \psi^{2}_{13}, 0), (0, 0, 0))'$, and the sparsity pattern extracted from $\boldsymbol{\Theta}$ be-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

comes: $\boldsymbol{\ell} = ((1,0,0),(0,1,0),(0,0,0))'$. For every sparsity pattern $\boldsymbol{\ell}$, let $d_{\boldsymbol{\ell}}$ be the density (number of nonzero entries) of $\boldsymbol{\ell}$, and $\mathcal{M}_{\boldsymbol{\ell}}$ be the space where $\boldsymbol{\Theta}$ varies, when restricted to follow the sparsity pattern $\boldsymbol{\ell}$. Furthermore, $\boldsymbol{\lambda} \in \mathbb{R}_+^{2^K-1}$ and $\boldsymbol{\Lambda} = diag(\boldsymbol{\lambda}, \boldsymbol{\lambda}, \cdots, \boldsymbol{\lambda})$ is a diagonal matrix (with $p(p-1)/2$ diagonal blocks of $\boldsymbol{\lambda}$, the entries of which determine the amount of shrinkage imposed on the corresponding elements in $\boldsymbol{\Theta}$. We specify the hierarchical prior distribution S^2, as follows:

$$\pi\left(\boldsymbol{\Theta}|\boldsymbol{\ell}\right) = \frac{|\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}|^{\frac{1}{2}}}{(2\pi)^{\frac{d_{\boldsymbol{\ell}}}{2}}} \exp\left(-\frac{\boldsymbol{\Theta}'\boldsymbol{\Lambda}\boldsymbol{\Theta}}{2}\right) I_{(\boldsymbol{\Theta}\in\mathcal{M}_{\boldsymbol{\ell}})}, \tag{3.2}$$

$$\pi(\boldsymbol{\ell}) \propto \begin{cases} (2\pi)^{d_{\boldsymbol{\ell}}/2}|\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}|^{-\frac{1}{2}} q_1^{d_{\boldsymbol{\ell}}}(1-q_1)^{\binom{p}{2}-d_{\boldsymbol{\ell}}} & d_{\boldsymbol{\ell}} \le \tau, \\ (2\pi)^{d_{\boldsymbol{\ell}}/2}|\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}|^{-\frac{1}{2}} q_2^{d_{\boldsymbol{\ell}}}(1-q_2)^{\binom{p}{2}-d_{\boldsymbol{\ell}}} & d_{\boldsymbol{\ell}} > \tau, \end{cases} \tag{3.3}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\ell}\boldsymbol{\ell}}$ is a sub-matrix of $\boldsymbol{\Lambda}$ obtained after removing the rows and columns corresponding to the zeros in $\boldsymbol{\Theta} \in \mathcal{M}_{\boldsymbol{\ell}}$, and q_1 and q_2 are edge inclusion probabilities, for the case of sparse ($d_{\boldsymbol{\ell}} \le \tau$) and dense ($d_{\boldsymbol{\ell}} > \tau$) $\boldsymbol{\Theta}$, respectively. An equivalent "spike-and-slab" representation of (3.6) is

$$\begin{aligned} &\{\boldsymbol{\theta}_{ij}\}_{1\le i<j\le p} \text{ are conditionally independent given } \boldsymbol{\ell} \text{ and} \\ &\pi\left(\boldsymbol{\theta}_{ij}|\boldsymbol{\ell}\right) = 1_{\boldsymbol{\theta}_{ij}=\mathbf{0}} 1_{\boldsymbol{\ell}_{ij}=\mathbf{0}} + \sum_{r\in\mathcal{P}(K)\setminus\{0\}} \sqrt{\frac{\lambda_r}{2\pi}} e^{-\frac{\lambda_r\theta_{ij,r}^2}{2}} 1_{\boldsymbol{\theta}_{ij,-r}=\mathbf{0}} 1_{\ell_{ij,r}=1} \end{aligned} \tag{3.4}$$

Note that the distribution of $\boldsymbol{\theta}_{ij}$ is supported on the axes of $\mathbb{R}^{2^K-1}$. The "spike" corresponds to the point mass at the origin, and the "slabs" correspond to a nor-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

mal distribution on an appropriate axis (when exactly one coordinate is nonzero). In fact, when $q_1 = q_2 = q$, it follows by (3.7) that $\{\boldsymbol{\ell}_{ij}\}_{1\le i<j\le p}$ are a priori independent, and the prior distribution of $\boldsymbol{\ell}_{ij}$ is given by $P(\ell_{ij,k} = 1) = \sqrt{2\pi\lambda_k^{-1}}q/C$, for every $1 \le k \le 2^K - 1$, and

$$P(\boldsymbol{\ell}_{ij} = \mathbf{0}) = \frac{(1-q)}{C}, \quad \text{where} \quad C = 1 - q + q\left(\sum_{k=1}^{2^K-1}\sqrt{\frac{2\pi}{\lambda_k}}\right).$$

Hence, smaller values of q can be used to encourage sparser models in high-dimensional settings.

Note that the $\mathcal{S}^2$ prior allows at most one entry in each $\boldsymbol{\theta}_{ij}$ to be nonzero, and thus sets *at least* $\frac{p(p-1)}{2}(2^K - 2)$ parameters to be exactly equal to zero. In particular, the $\mathcal{S}^2$ prior considers the entire range of models allowable under the identifiability constraint discussed in Section 2.1: at one end, we have the model with complete sparsity, where $\boldsymbol{\Theta} = 0$, and at the other end, we have models with $\binom{p}{2}$ parameters where each $\boldsymbol{\theta}_{ij}$ containing exactly one nonzero entry. In addition to forcing sparsity, the diagonal entries of $\boldsymbol{\Lambda}$ enforce a shrinkage to the corresponding elements in $\boldsymbol{\Theta}$.

The vector $\boldsymbol{\Theta}$ incorporates, only the off-diagonal entries of $\boldsymbol{\Psi}$ matrices. For the diagonal entries, for every $k \in \{1, ..., K\}$, we let $\boldsymbol{\delta}_{\boldsymbol{\Psi}^k}$ be the vector comprising the diagonal elements of the matrix $\boldsymbol{\Psi}^k$, and define $\boldsymbol{\Delta}$ as the vector of all diagonal vectors $\boldsymbol{\delta}_{\boldsymbol{\Psi}^K}$, that is, $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{\boldsymbol{\Psi}^1}, ..., \boldsymbol{\delta}_{\boldsymbol{\Psi}^K})$. We assign an independent

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

Exponential(γ) prior to each coordinate of $\boldsymbol{\Delta}$ (diagonal element of the matrices $\boldsymbol{\Psi}_k$, for $k = 1, ..., K$), that is, $\pi(\boldsymbol{\Delta}) \propto \exp(-\gamma \mathbf{1}'\boldsymbol{\Delta}) I_{\mathbb{R}_+^{Kp}}(\boldsymbol{\Delta})$. In the next section we discuss selecting the hyperparameters Λ and γ. Because the diagonal entries of every joint matrix $\boldsymbol{\Psi}^r$, with $r \in \cup_{k=1}^K (\vartheta_k \setminus \{k\})$, are set to zero, the specification of the prior is now complete.

Remark 1. Prior distributions with similarities to the subset-specific one proposed here have been used in genetic association (eQTL) analyses with heterogeneous subgroups; see Wen and Stephens (2014) and Flutre et al. (2013), and the references therein. However, there are two crucial differences in these two settings. First, in the eQTL setting, we have a single regression model with the gene expression level as the response and a single predictor (genotype), whereas in the current joint graphical model setting, we employ a pseudo-likelihood consisting of p different regressions (corresponding to the p variables), each of which has multiple $(p-1)$ predictors; see (4.9) and (4.10) below. Second, even though the p.d. constraint on the precision matrices is relaxed for the sparsity selection, the symmetry constraint is not (we still need the sparsity patterns for each $\boldsymbol{\Omega}^k$ to be symmetric). This symmetry couples the p regressions that form the pseudo-likelihood. These complications lead to unique challenges in the methodological development and theoretical analysis of the joint graphical model setting.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

4. The BJNS

With a prior distribution in hand, sparsity/network selection is based on a pseudo-likelihood approach, leveraging the regression interpretation of the entries of Ω. It can also be regarded as a weight function, and as long as the product of the pseudo-likelihood and the prior density is integrable over the parameter space, we can construct a (pseudo) posterior distribution for inference purposes. The main advantage of using a pseudo-likelihood, as opposed to a full Gaussian likelihood, is that it allows for a sampling scheme from the posterior distribution that is easy to implement, and it provides more robust results under deviations from the Gaussian assumption, as illustrated by works in the frequentist domain (Khare et al., 2015; Peng et al., 2009). Of course, the use of this pseudo- posterior distribution has to be justified both by theoretical consistency results and by assessing its performance in finite-sample simulation settings, which we do in Section 7, respectively.

Note that if $\mathbf{Y} \in \mathbb{R}^p$ with $Cov(\mathbf{Y}) = \Omega^{-1}$, then

$$\Omega_{j,-j} = (\Omega_{jk})_{1\leq k\leq p, k\neq j} = \text{argmin}_{\mathbf{w}\in\mathbb{R}^{p-1}} E\left(\Omega_{jj}Y_j + \mathbf{w}'\mathbf{Y}_{-j}\right)^2.$$

Hence, if $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ are i.i.d. with covariance matrix Ω^{-1}, then the above regression interpretation of Ω can be used to construct the CONCORD pseudo-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

likelihood introduced in Khare et al. (2015), given by

$$\exp\left\{n\sum_{j=1}^{p}\log\omega_{jj}-\frac{n}{2}\text{tr}\left[\mathbf{\Omega}^2\mathbf{S}\right]\right\}, \tag{4.1}$$

where $\mathbf{S}$ denotes the sample covariance matrix. The CONCORD pseudo-likelihood is jointly convex in the entries of Ω. This makes it both theoretically and computationally preferable to other variants of the regression-based pseudo-likelihood (see the discussion in Khare et al. (2015)). In addition, because the regression interpretation does not depend on the Gaussian assumption, regression-based approaches such as the one above in general provide more robust results under deviations from that assumption.

Let $\mathbf{S}^k$ denote the sample covariance matrix of the observations in the kth group. Based on the above discussion, we employ the pseudo-likelihood in (4.9) and the model specification (2.2) to construct the joint pseudo-likelihood function for K precision matrices, as follow:,

$$\prod_{k=1}^{K}\exp\left\{n\sum_{j=1}^{p}\log\psi_{jj}^{k}-\frac{n}{2}\text{tr}\left[\left(\sum_{r\in\vartheta_k}\mathbf{\Psi}^r\right)^2\mathbf{S}^k\right]\right\}. \tag{4.2}$$

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

$$\text{tr}\left[\left(\sum_{r\in\vartheta_k}\mathbf{\Psi}^r\right)^2\mathbf{S}^k\right]=\begin{pmatrix}\mathbf{\Theta}' & \mathbf{\Delta}'\end{pmatrix}\begin{pmatrix}\mathbf{\Upsilon} & \mathbf{A}\\ \mathbf{A}' & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{\Theta}\\ \mathbf{\Delta}\end{pmatrix}, \tag{4.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

where $\boldsymbol{\Upsilon}$ is a $\frac{p(p-1)(2^K-1)}{2} \times \frac{p(p-1)(2^K-1)}{2}$ symmetric matrix with entries that are either zero or a linear combination of $\left\{s_{ij}^k\right\}_{1\le i<j\le p}^{1\le k\le K}$; $\mathbf{D}$ is a $Kp \times Kp$ diagonal matrix with entries $\left\{s_{ii}^k\right\}_{1\le i\le p}^{1\le k\le K}$; $\boldsymbol{a}$ is a $\frac{p(p-1)(2^K-1)}{2} \times 1$ vector with entries that depend on $\boldsymbol{\Delta}$ and $\left\{s_{ij}^k\right\}_{1\le i<j\le p}^{1\le k\le K}$; and $\mathbf{A}$ is a $\frac{p(p-1)(2^K-1)}{2} \times Kp$ matrix such that $\mathbf{A}\boldsymbol{\Delta} = \boldsymbol{a}$. The algebraic details of the equality in (4.11), structures of $\boldsymbol{\Upsilon}$ and $\boldsymbol{a}$, and algebraic forms of the joint and conditional posterior densities for $\boldsymbol{\Theta}$ and $\boldsymbol{\Delta}$ are provided in Supplementary Material Section A.1.

Gibbs Sampling Scheme for the BJNS: Generating exact samples from the joint posterior of $(\boldsymbol{\Theta}, \boldsymbol{\Delta})$ is not feasible. Instead, we generate approximate samples by computing the full conditional distributions of the vectors $\{\boldsymbol{\theta}_{ij}\}_{1\le i<j\le p}$ and of the diagonal entries $\{\psi_{ii}^k\}_{1\le i\le p, 1\le k\le K}$.

Each $\boldsymbol{\theta}_{ij}$ contains 2^K-1 elements, of which at most one is nonzero. For ease of exposition, let $\theta_{l,ij}$ denote the lth element of $\boldsymbol{\theta}_{ij}$, for $l = 1, ..., 2^K - 1$ (based on (2.4), where every $\theta_{l,ij}$ represents a ψ_{ij}^r, for some $r \in \mathcal{P}(K)$). Using the same notation for the shrinkage parameters (diagonal elements of $\boldsymbol{\Lambda}$), let $\lambda_{l,ij}$ be the shrinkage parameter corresponding to $\theta_{l,ij}$. Because there are 2^K possibilities for the location of the nonzeros in each $\boldsymbol{\theta}_{ij}$, $\boldsymbol{\theta}_{ij}$ is an element in one of the disjoint spaces $\mathbb{M}_0$, $\mathbb{M}_1$, ..., $\mathbb{M}_{2^K-1}$, where $\mathbb{M}_0$ is the singleton set consisting of the zero vector of length $2^K - 1$ and $\mathbb{M}_l$ ($l = 1, ..., 2^K - 1$) is the space spanned by the lth unit vector of length $2^K - 1$. Denote by $\boldsymbol{\Theta}_{-(ij)}$ the sub vector $\boldsymbol{\Theta}$ obtained

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

by removing θ_{ij}. It can be shown that (Supplementary Material Section B.1) the conditional density of $\boldsymbol{\theta}_{ij}$ given $\boldsymbol{\Theta}_{-(ij)}, \boldsymbol{\Delta}$, is a mixture of univariate normal densities respectively supported on $\{\mathbb{M}_i\}_{i=0}^{2^K-1}$. Furthermore, conditional on $\boldsymbol{\Theta}$, the diagonal entries $\left\{\psi_{ii}^k\right\}_{1\le i\le p}^{1\le k\le K}$ are a posteriori independent. We provide an efficient algorithm to sample them in Supplementary Material Section B.1.

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

$$\lambda|\left(\boldsymbol{\Theta}, \boldsymbol{\Delta}\right) \sim \text{Gamma}(r + 0.5, 0.5\theta^2 + s), \text{ and } \gamma|\left(\boldsymbol{\Theta}, \boldsymbol{\Delta}\right) \sim \text{Gamma}(r + 1, |\delta| + s).$$

Note that $\mathbb{E}\{\lambda|\left(\boldsymbol{\Theta}, \boldsymbol{\Delta}\right)\} = \frac{r+0.5}{0.5\,\theta^2+s}$ and $\mathbb{E}\{\gamma|\left(\boldsymbol{\Theta}, \boldsymbol{\Delta}\right)\} = \frac{r+1}{|\delta|+s}$. That is, our approach selects the shrinkage parameters λ and γ based on the current values of θ and δ in a way that larger (smaller) entries are regularized less (more).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

The selection of the hyperparameters r and s is also an important task and, in our experience can affect performance, especially for small sample sizes. As the sample size grows, the results become less sensitive to the choice of the hyperparameters (see Supplementary Material Section C, for an illustration). In the absence of any prior information, we recommend the noninformative choice $r = 10^{-4}$ and $s = 10^{-8}$, which corresponds to a flat prior for the λ and γ values, and is based on the suggestions made in Wang (2012) for hyperparameter selection. In general, we found that this choice works well in the extensive simulations presented in Section 5. For q_1 and q_2, we suggest using $q_1 = q_2 = 1/2$. To encourage sparser models in really high-dimensional situations, one can use $q_1 = 1/p$, $q_2 = q_1^n$ (essentially zero), and $\tau = n/\log n$, based on the theoretical result in Section 7.

Finally, we construct the Gibbs sampler as follows: matrices $\{\mathbf{\Psi}^k\}_{k=1}^K$ are initialized as the identity matrix, and $\{\mathbf{\Psi}^r\}_{\{r:r\in\mathcal{P}(K),\&|r|>1\}}$ at zero. Then, in each iteration of the MCMC chain, we update the vectors $\boldsymbol{\theta}_{ij}$ and the diagonal entries ψ_{ii}^k, one at a time, using their full conditional posterior densities given in (B.25) and (B.26), respectively. Procedure 1 in the Supplementary Material Document describes one iteration of the resulting Gibbs sampler.

Procedure for Sparsity Selection and Uncertainty Quantification: Note that the conditional posterior probability density of the off-diagonal elements of $\boldsymbol{\theta}_{ij}$ is a

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

mixture density that puts all of its mass on the events $\{\boldsymbol{\theta}_{ij} : |\boldsymbol{\theta}_{ij}|_0 \leq 1\}$, where $|\boldsymbol{\theta}_{ij}|_0$ is the number of nonzero coordinates of $\boldsymbol{\theta}_{ij}$. This property of the BJNS allows for model selection, in the sense that for every (post burn-in) iteration of the Gibbs sampler, one can check whether $\boldsymbol{\theta}_{ij} = \mathbf{0}$ or which element of $\boldsymbol{\theta}_{ij}$ (there can be at most one nonzero element) is nonzero. Note that each element of $\boldsymbol{\theta}_{ij}$ corresponds to a subset of $\{1, \cdots, K\}$. The nonzero frequency during sampling for any given subset S, normalized by the total number of iterations, provides an estimate of the posterior probability that the edge (i, j) is shared between elements of S. Hence, at the end of the procedure, we choose the element with the highest nonzero frequency during sampling. Denoting the chosen element (subset) by $\hat{S}_{ij}$, we set $\omega_{ij}^k = 0$ if $k \notin \hat{S}_{ij}$, and $\omega_{ij}^k \neq 0$ if $k \in \hat{S}_{ij}$.

Procedure for P.D. Estimation: Once we have completed sparsity selection in the form of subsets $\left\{\hat{S}_{ij}\right\}$, these subsets can be used to obtain sparsity graphs $\{\hat{G}^k\}_{k=1}^K$ for the K group-wise precision matrices (i.e., (i, j) is in $\hat{G}^k$ if $k \in \hat{S}_{ij}$). Next, we obtain the p.d. estimate $\hat{\boldsymbol{\Omega}}^k$ of $\boldsymbol{\Omega}^k$ as the solution to the following restricted optimization problem (implemented in *R* package *glasso*):

$$\hat{\boldsymbol{\Omega}}^k = \text{argmin}_{\boldsymbol{\Omega} \in \mathbb{P}_{\hat{G}^k}} \left\{\text{tr}(\boldsymbol{\Omega}\mathbf{S}^k) - \log \det \boldsymbol{\Omega}\right\}. \tag{4.4}$$

Here, $\mathbb{P}_{\hat{G}^k}$ is the space of all p.d. matrices, in which the (i, j)th entry is zero whenever (i, j) is not an edge in $\hat{G}^k$. Hence, *the resulting "refitted" estimates* $\left\{\hat{\boldsymbol{\Omega}}^k\right\}_{k=1}^K$ *are p.d., and are not constrained to obey any working sign/magnitude*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

restrictions used in the sparsity selection process. Another possibility is to obtain fully Bayesian refitted estimates by using G-Wishart priors on $\mathbb{P}_{\hat{G}^k}$ for each sub-group, and then computing the posterior mean for each group using one of the computationally efficient sampling methods developed in Mitsakakis et al. (2011); Lenkoski (2013); Khare et al. (2018).

5. Simulation Studies

In this section, we present simulations in which we evaluate the statistical and computational properties of the BJNS algorithm. Most tables and figures, and several additional simulation results are provided in the Supplementary Material.

5.1 Sparsity selection performance

We present two simulation studies to evaluate the sparsity selection performance of the BJNS. In the first study (Section 5.1.1), we compare the performance of the BJNS with that of other high-dimensional methodologies for joint sparsity selection: (1) a separate estimation using a graphical lasso (Glasso) of each $\mathbf{\Omega}^k$; (2) the joint estimation of Guo et al. (2011) (JEM-G); (3) the two joint graphical lasso variants, GGL and FGL of Danaher et al. (2014); and (4) the Bayesian neighborhood selection (BNS) of Lin et al. (2017). In the second study (provided in Supplementary Material Section D.1, owing to space con-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

straints), we illustrate the strengths of the BJNS in terms of sparsity selection and uncertainty quantification using two scenarios, with four precision matrices each. In the Supplementary Material Section D.3, we provide a comparison with the joint structural estimation method (JSEM) of Ma and Michailidis (2016), which is a supervised approach that incorporates exact information on the shared sparsity structure. Finally, the `C++` code for the BJNS is publicly available at `https://www.github.com/PeymanJalali/BJNS`.

5.1.1 Sparsity selection comparison with other high-dimensional methods

We consider two settings involving six networks. We first consider a scenario with $K = 6$ graphs, each with $p = 200$ variables (see Figure G.4 in the Supplementary Material Section G), where we generate the adjacency matrices corresponding to three distinct p-dimensional networks, so that the adjacency matrices in each column of the plot in Figure G.4 in the Supplementary Material Section G, are common. Next, we replace the connectivity structure of the bottom-right diagonal block of size $p/2$ by $p/2$ in each adjacency matrix with that of another two distinct $p/2$-dimensional networks. As such, the the graphical models in each column exhibit the same connectivity pattern, except in the bottom-right diagonal blocks of their adjacency matrices. The resulting true decompositions are $\mathbf{\Omega}^i = \mathbf{\Psi}^{i,i+1} + \mathbf{\Psi}^{135}$ for odd i, and $\mathbf{\Omega}^i = \mathbf{\Psi}^{i-1,i} + \mathbf{\Psi}^{246}$ for even i.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

The sparsity level for all six networks is set to 92% (equivalently, the edge density is set to 8%), and the proportion of common zeros (no edge present) across all six networks is about 60%. Given the adjacency matrices, we then construct p.d. inverse covariance matrices, with the nonzero off-diagonal entries in each $\mathbf{\Omega}^k$ being uniformly generated from $[-0.6, -0.4] \cup [0.4, 0.6]$. Based on whether or not the corresponding shared edges have the same value, we consider two settings: in the first, the values of the shared edges are set to be equal, and in the second, the values of the shared edges are only constrained to share the sign, and hence their absolute values can differ across networks.

For each of these two settings, in addition to the scenario with and edge density of 8% described above, we add 4% additional edges are added to each "true" network to make the overall edge density 12%. For each of these scenarios, we generated $n_k = 200, 300$ independent samples, for each $k = 1, ..., K$, and examined the finite-sample performance of different methods in terms of identifying the true graphs/networks, using optimal choices of the tuning parameters. To select the optimal values of the tuning parameters in the penalized approaches, we searched over a fine grid and chose the value that resulted in the minimum BIC on the normalized data.

Once the off-diagonal elements are chosen, we set the diagonal elements of each precision matrix to be bigger than the absolute value of its minimum

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

eigenvalue. This ensures that the resulting precision matrices are stable and invertable. Note that we need to compute the covariance matrices to be able to generate synthetic data from multivariate normal distributions.

We assess the model/sparsity selection performance using the Matthews correlation coefficient (MCC), which is defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},$$

with TP, TN, FP, and FN denoting the number of true positives, true negatives, false positives, and false negatives, respectively. Larger values of MCC indicate better sparsity selection. Tables 1 and 2 show MCC values based on 50 replications for varying levels of sample size and true edge density.

Table 1: MCC values for various joint sparsity selection approaches across six networks, when the true sparsity patterns are random and shared edges have the same values. The MCC values are averaged over 50 replications.

	Glasso	JEM-G	GGL	FGL	BNS	GemBag	BJNS
Edge density				$n = 200$			
8%	47 (0.009)	50 (0.010)	47 (0.009)	49 (0.009)	35 (0.010)	43 (0.010)	**57 (0.010)**
12%	**40 (0.008)**	35 (0.010)	35 (0.008)	39 (0.009)	30 (0.010)	34 (0.010)	**40 (0.010)**
				$n = 300$			
8%	54 (0.009)	60 (0.010)	54 (0.008)	56 (0.008)	40 (0.010)	52 (0.010)	**70 (0.010)**
12%	47 (0.009)	43 (0.009)	42 (0.008)	47 (0.009)	35 (0.010)	42 (0.010)	**51 (0.010)**

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

Table 2: MCC values for joint sparsity selection approaches across six networks, when the true sparsity patterns are random and shared edges have the same sign, but different values. The MCC values are averaged over 50 replications.

Edge density	Glasso	JEM-G	GGL	FGL	BNS	GemBag	BJNS
				$n = 200$			
8%	46 (0.009)	50 (0.011))	47 (0.010)	49 (0.009)	36 (0.011)	43 (0.010)	**54 (0.011)**
12%	**40 (0.008)**	36 (0.009)	36 (0.008)	39 (0.009)	29 (0.010)	35 (0.009)	**40 (0.009)**
				$n = 300$			
8%	54 (0.010)	60 (0.010)	54 (0.009)	56 (0.008)	43 (0.010)	52 (0.011)	**67 (0.010)**
12%	48 (0.009)	44 (0.009)	43 (0.008)	48 (0.009)	37 (0.010)	43 (0.009)	**51 (0.009)**

Tables 1 and 2 show that the BJNS significantly outperforms the competing methods across all settings considered. As expected, the overall performance of all methods improves with additional samples, and worsens with higher edge density in the true precision matrices. Glasso is competitive in scenarios with a smaller sample size ($n = 200$) and a higher edge density (12%), owing to the overall heterogeneity of the six networks. Note that analogous performance patterns hold for the settings in Tables 1 and 2. Finally, to check whether these comparisons are affected by the edge value range, we replicated Table 2, this time by generating the edges from $[-0.8, -0.3] \cup [0.3, 0.8]$. The results were similar. For more details, please see the Supplementary Material Section D.2.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

5.2 Computational comparison with other sampling-based methods

The recent Gaussian likelihood-based methods of Shaddox et al. (2018) and Petersen et al. (2020) use the spike-and-slab prior-based approach of Wang (2015) to improve the computational scalability of the G-Wishart prior-based approach of Peterson et al. (2015). These methods use a Gibbs sampler to generate samples from the resulting posterior. In addition to the precision matrix parameters, these models have latent variable parameters for edge inclusion, and other hyperparameters to encourage similarity between graphs. The overall updates for just the precision matrix parameters involve several matrix inversions, and have a computational complexity of $KO(\min(np^3, p^4))$ per iteration. The updates for the other parameters, including the latent variables for the sparsity patterns and other "relatedness" hyperparameters, require an additional computation involving Metropolis–Hastings-based moves. Hence, these methods are not typically scalable to settings beyond a hundred or so variables.

On the other hand, Lin et al. (2017) introduce a Bayesian analog of the regression-based neighborhood selection approach of Meinshausen and Bühlmann (2006) for joint sparsity selection, called BNS. They use an alternative Markov random field approach to encourage similarity among the groups. A Gibbs sampler is used to generate samples from the posterior distribution. Similarly to Shaddox et al. (2018), the BNS undertakes a column-wise update for the pre-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

cision matrices, which involves an inversion of $(p-1)\times(p-1)$ matrices. Although the computational complexity of the precision matrix parameter updates remains $KO(\min(np^3, p^4))$, the simpler structure of the regression-based approach results in several simplifications, including needing fewer matrix inversions per iteration. In addition, the updates for the sparsity pattern-based latent parameters and other hyperparameters are, in general, simpler than those in Shaddox et al. (2018). Hence, the BNS provides a significant computational improvement, and to the best of our knowledge, is the fastest sampling-based Bayesian approach for joint graphical model selection.

Next, we derive the computational complexity for the BJNS Gibbs sampler described in Procedure 1 (see the Supplementary Material), which does not involve any matrix inversions. In particular, for each (i, j), the full conditional posterior distribution of $\boldsymbol{\theta}_{ij}$ is a mixture of 2^K univariate densities. The sparsity structure of $\boldsymbol{\Upsilon}_{(ij)(-(ij))}$ and $\boldsymbol{\Theta}_{-(ij)}$ (see equations A.15 and A.16) and an analysis similar to that of Khare et al. (2015, Lemma 6) imply that the computation of each mean $\mu_{l,ij}$ in (B.24) can be achieved in $O(\min(n, p))$ operations. Hence, sampling from the mixture requires $2^K O(\min(n, p))$ operations, implying that the overall worst case computational complexity per iteration is $2^K O(\min(np^2, np^3))$.

Hence, the computational complexity of the vanilla BJNS algorithm is $\frac{2^K}{Kp}$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

times that of the BNS. In many applications, such as those considered in Shaddox et al. (2018) and Petersen et al. (2020) and the IBD data considered in Section 6, the number of groups $K = 4$. *The BJNS/BNS computational complexity ratio turns out to be* 0.065 *(OxPhos pathway for the COPD data in Shaddox et al. (2018)),* 0.04 *(Alzheimer's MRI data in Petersen et al. (2020)) and* 0.01 *(IBD data in Section 6).* In general, the computational complexity of the BJNS is much smaller than that of the BNS for typical genomics data sets, where K is small/moderate and p is much larger. To further illustrate the latter point, we provide a wall clock time comparison between the BJNS and BNS in Table 3 for simulations with various numbers of variables p and groups K. The true data-generating process is identical to that described in the previous subsection (we use $n = 3p/2$). In order to make a fair comparison, we compiled the MATLAB code of the BNS provided by its authors, to machine language code. All computations were done on an intel CPU with 6 GB of memory. Overall, BJNS has a significantly lower wall clock time requirement than that of the BNS across all settings.

If K is large, we develop a preprocessing step in the Supplementary Material, Section E, to reduce the size of the mixture from 2^K to a much smaller user-specified number M_K. Taking into account the preprocessing step, the overall per iteration computational complexity is further reduced to $(K^2+M_K)O(\min(np^2, np^3))$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

Table 3: Wall clock time for the BJNS and BNS (4000 iterations).

		$p = 50$	$p = 200$	$p = 500$		$p = 50$	$p = 200$	$p = 500$
BJNS		$0.003h$	$0.092h$	$1.193h$		$0.016h$	$0.510h$	$5.738h$
BNS	$k = 2$	$0.029h$	$0.282h$	$6.503h$	$k = 4$	$0.128h$	$0.672h$	$11.8h$

6. An Application of the BJNS to Metabolomics Data

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

$$\boldsymbol{\Omega}^1 = \boldsymbol{\Psi}^1 + \boldsymbol{\Psi}^{12} + \boldsymbol{\Psi}^{13} + \boldsymbol{\Psi}^{1234}, \qquad \boldsymbol{\Omega}^3 = \boldsymbol{\Psi}^3 + \boldsymbol{\Psi}^{13} + \boldsymbol{\Psi}^{34} + \boldsymbol{\Psi}^{1234},$$

$$\boldsymbol{\Omega}^2 = \boldsymbol{\Psi}^2 + \boldsymbol{\Psi}^{12} + \boldsymbol{\Psi}^{24} + \boldsymbol{\Psi}^{1234}, \qquad \boldsymbol{\Omega}^4 = \boldsymbol{\Psi}^4 + \boldsymbol{\Psi}^{24} + \boldsymbol{\Psi}^{34} + \boldsymbol{\Psi}^{1234}.$$

The edge counts of the estimated precision matrices are shown in Table G.1 (set1: 289 lipids; set2: 139 proteins; set1.2: interaction edges between set1 and set2), together with the components in the proposed decomposition. We also

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

applied the JEM-G and show the corresponding edge counts in Table G.1. Note that the BJNS detects a large number of edges that are shared across all groups, indicating common patterns. Furthermore, the component shared between male and female IBD patients has a fairly large number of edges, indicating that the disease status exhibits commonalities across genders. The graphs produced by the JEM-G are much more similar, which is consistent with the fact that the JEM-G differentiates graphs only by a multiplicative factor. The JEM-G also tends to provide a significantly higher number of edges for each network.

Table G.2 illustrates the detailed uncertainty quantification provided by the BJNS by providing posterior inclusion probabilities for all 10 possible subsets for 10 chosen edges. Figure G.5 presents the common connectivity patterns shared across all four groups. The primary and secondary metabolites are depicted in red, and the lipids are shown in blue. Not surprisingly, as shown in sub-figure G.5a, primary metabolites (those involved in cellular growth, development, and reproduction) form a fairly strongly connected network. In addition, based on the sub-network in G.5b, there are different fairly strongly connected subnetworks present among the lipids, including dicylglycerols (DAG) with tricylglycerols (TAG), that are the main constituents of animal and vegetable fat (upper-right corner of the plot), and various phospholipids (upper-left corner of the plot). On the other hand, the connectivity between the lipids (whose func-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

tions include storing energy, signaling, and acting as structural components of cell membranes) and the metabolites is not particularly strong; see the network in sub-figure G.5c. In general, the results reveal interesting patterns that can be used to understand the progression of IBD.

7. High-Dimensional Sparsity Selection Consistency

Let $\{\bar{\mathbf{\Omega}}^{k,0}\}_{k=1}^K$ denote the true precision matrices, and $\mathbb{P}_0$ denote the probability measure associated with the corresponding true data-generating model. Note that the identifiability constraint in Section 2.1 assumes that whenever an edge (i,j) is shared in a subset S, the magnitudes of the (i,j)th entries in the corresponding precision matrices are the same. *We allow* $\{\bar{\mathbf{\Omega}}^{k,0}\}_{k=1}^K$ *to deviate from this assumption, that is, we allow the entries in* $\{\bar{\Omega}_{i,j}^{k,0}\}_{k\in S}$ *to have different magnitudes. We will show that as long as the deviation in the magnitudes is moderate, the BJNS still leads to consistent high-dimensional model selection.*

Define the matrices $\{\mathbf{\Omega}^{k,0}\}_{k=1}^K$ as follows. For each $1 \leq i \neq j \leq p$ and $1 \leq k \leq K$, if $\bar{\Omega}_{ij}^{k,0} \neq 0$, set

$$\Omega_{ij}^{k,0} = \frac{\sum_{k':\bar{\Omega}_{ij}^{k',0}\neq 0} \bar{\Omega}_{ij}^{k',0}}{|k' : \bar{\Omega}_{ij}^{k',0} \neq 0|},$$

and set $\Omega_{ij}^{k,0} = \bar{\Omega}_{ij}^{k,0}$ otherwise. The matrices $\{\mathbf{\Omega}^{k,0}\}_{k=1}^K$ can be thought of as *harmonized versions* of the true precision matrices $\{\bar{\mathbf{\Omega}}^{k,0}\}_{k=1}^K$, which obey the identifiability constraint in Section 2.1. We define $D_n = \max_{1\leq i\neq j\leq p, 1\leq k\leq K} \left|\Omega_{ij}^{k,0} - \bar{\Omega}_{ij}^{k,0}\right|$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

as a discrepancy measure between the true and the harmonized precision matrices.

Let $\left\{\boldsymbol{\Psi}^{r,0}, r \in \bigcup_{k=1}^{K} \vartheta_k\right\}$ be such that $\boldsymbol{\Omega}^{k,0} = \sum_{r \in \vartheta_k} \boldsymbol{\Psi}^{r,0}$ corresponds to the decomposition of each harmonized precision matrix, for $k = 1, ..., K$. Let $\boldsymbol{\Theta}^0$ be the vectorized version (see (3.5)) of the off-diagonal elements of the matrices $\left\{\boldsymbol{\Psi}^{r,0}, r \in \bigcup_{k=1}^{K} \vartheta_k\right\}$. Let $\boldsymbol{t}$ denote the sparsity pattern in $\boldsymbol{\Theta}^0$, $\mathcal{M}_{\boldsymbol{t}}$ denote the corresponding parameter space, and $d_{\boldsymbol{t}}$ denote the number of nonzeros in $\boldsymbol{\Theta}^0$.

Note that our main objective is to accurately select the shared sparsity patterns in the off-diagonal entries of the group-specific precision matrices. Hence, following the pseudo-likelihood-based high-dimensional consistency proofs in Peng et al. (2009), Khare et al. (2015), and Atchade (2019), we consider a setting in which we first obtain sufficiently accurate estimates $\left\{\hat{\Omega}_{ii}^{k}\right\}_{1 \le i \le p}^{1 \le k \le K}$ of the *diagonal* entries (see eq. (F.34) in the Supplementary Material, and the subsequent discussion for a quick way of obtaining such estimates using parallel lasso regressions). Denote the resulting estimates of the vectors $\boldsymbol{\Delta}$ and $\boldsymbol{a}$ by $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{a}}$, respectively. We now consider the accuracy of the shared sparsity pattern selection for the off-diagonal entries after running the BJNS procedure with the diagonal entries fixed at $\hat{\boldsymbol{\Delta}}$. For this section, we assume that the entries of $\boldsymbol{\lambda}$ are fixed. The following assumptions are needed to establish our consistency results.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

ASSUMPTION 1: $d_{\boldsymbol{t}}\sqrt{\frac{\log p}{n}} \to 0, \qquad \text{as} \quad n \to \infty.$

This standard assumption essentially states that the number of variables p has to grow more slowly than $e^{\left(\frac{n}{d_{\boldsymbol{t}}^2}\right)}$; see, for example, Banerjee and Ghosal (2014, 2015).

ASSUMPTION 2: (Sub-Gaussianity). There exists $c > 0$, independent of n and K, such that $\mathbb{E}\left[\exp\left(\boldsymbol{\alpha}'\mathbf{y}_{\mathbf{i}:}^{\mathbf{k}}\right)\right] \leq \exp\left(c\boldsymbol{\alpha}'\boldsymbol{\alpha}\right)$. Theorem 1 shows that the BJNS procedure is robust (in terms of consistency), even under a misspecification of the data-generating distribution, as long as its tails are sub-Gaussian.

ASSUMPTION 3: (Bounded eigenvalues). There exists $\tilde{\varepsilon_0} > 0$, independent of n and K, such that the eigenvalues of $\bar{\boldsymbol{\Omega}}^{k,0}$ are uniformly bounded above and below by $\tilde{\varepsilon_0}$ and $1/\tilde{\varepsilon_0}$, respectively. This is a standard assumption in high-dimensional consistency analysis; see, for example, Cao et al. (2019).

ASSUMPTION 4: (Signal Strength). Let s_n be the smallest nonzero entry (in magnitude) in the vector $\boldsymbol{\Theta}_0$. We assume $\frac{\frac{1}{2}\log n + d_{\boldsymbol{t}} b_n}{n s_n^2} \to 0$, where $b_n = \log p + n d_{\boldsymbol{t}} D_n^2$. This is again a standard assumption. Similar assumptions on the signal size can be found in Khare et al. (2015) and Peng et al. (2009).

ASSUMPTION 5: (Edge probability decay). Let $q_1 = \mathrm{e}^{-a_2 d_{\boldsymbol{t}} b_n}$, $q_2 = \mathrm{e}^{-a_3 n b_n}$, and $\tau_n = \frac{\tilde{\varepsilon_0}}{4c}\sqrt{\frac{n}{\log p}}$, for constants a_1 and a_2 (not depending on n), specified in (F.39) in the Supplementary Material. Assumption 5 a priori penalizes patterns with too many nonzeros (see Narisetty and He (2014) and Cao et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

(2019) for similar assumptions). Next, we establish that the posterior mass assigned to the true model converges to one in probability (under the true model).

Theorem 1. *(Strong Selection Consistency) Based on the joint posterior distribution given in (B.21), and under Assumptions 1–5, the following holds:*

$$\pi\left\{\boldsymbol{\Theta} \in \mathcal{M}_{\boldsymbol{t}} | \hat{\boldsymbol{\Delta}}, \mathcal{Y}\right\} \xrightarrow{\mathbb{P}_0} 1, \qquad as \quad n \to \infty. \tag{7.1}$$

8. Conclusion

We have proposed a comprehensive Bayesian methodology for the joint estimation of multiple graphical models. Leveraging a novel multivariate prior distribution and a pseudo-likelihood, our model enables fast and *provably* accurate estimations. We have shown how our methodology uses the information that is shared across groups to provide greater accuracy. We also develop a computational strategy for dealing with a large number of networks K, and investigate it numerically in Section E of the Supplementary Material. Furthermore, simulation studies illustrate the superior performance of the BJNS in comparison with that of frequentist and Bayesian competitors. Finally, an application to IBD disease progression reveals interesting patterns.

Acknowledgments: The work of P. Jalali and G. Michailidis was supported in part by NIH grant 1U01CA235487.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

References

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

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

Banerjee, S. and S. Ghosal (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis 136*, 147–162.

Cai, T. T., H. Li, W. Liu, and J. Xie (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica 26*(2), 445.

Cao, X., K. Khare, M. Ghosh, et al. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics 47*(1), 319–348.

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

Flutre, T., X. Wen, J. Prichard, and M. Stephens (2013). A statistical framework for joint eqtl analysis in multiple tissues. *PLOS Genetics https://doi.org/10.1371/journal.pgen.1003486*.

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

Hans, C. (2009). Bayesian lasso regression. *Biometrika 96*(4), 835–845.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

Khare, K., S.-Y. Oh, and B. Rajaratnam (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *JRSSB 77*(4), 803–825.

Khare, K., B. Rajaratnam, and A. Saha (2018). Bayesian inference for Gaussian graphical models beyond decomposable graphs. *JRSSB 80*, 727–747.

Kling, T., P. Johansson, J. Sanchez, V. D. Marinescu, R. Jörnsten, and S. Nelander (2015). Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic acids research 43*(15), e98–e98.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis 5*(2), 369–411.

Lenkoski, A. (2013). A direct sampler for G-Wishart variates. *Stat 2*(119-128).

Li, Z. R., T. H. McCormick, and S. J. Clark (2018). Bayesian joint spike-and-slab graphical lasso. *arXiv preprint arXiv:1805.07051*.

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

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

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics 34*(3), 1436–1462.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

Mitsakakis, N., H. Massam, and M. D. Escobar (2011). A metropolis-hastings based method for sampling from the G-Wishart distribution in Gaussian graphical models. *Electronic Journal of Statistics 5*, 18–30.

Narisetty, N. N. and X. He (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics 42*(2), 789–817.

Park, T. and G. Casella (2008). The Bayesian lasso. *J. Amer. Stat. Assoc. 103*(482), 681–686.

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*(486), 735–746.

Petersen, C., N. Osborne, F. Stingo, P. Bourgeat, J. Doecke, and M. Vanucci (2020). Bayesian modeling of multiple structural connectivity networks during the progression of Alzheimer's disease. *Biometrics 76*, 1120–1132.

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

Pierson, E., D. Koller, A. Battle, S. Mostafavi, and G. Consortium (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comp. Bio. 11*(5), e1004220.

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

Shaddox, E., F. Stingo, C. Petersen, S. Jacobson, C. Cruickshank-Quinn, R. Bowler, and M. Vannucci (2018). A Bayesian approach for learning gene networks underlying disease severity in COPD. *Statistics in Biosciences 10*, 59–85.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0245

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

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

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

Wen, X. and M. Stephens (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions. *Annals of Applied Statistics 8*, 176–203.

Yang, X., L. Gan, N. Narisetty, and F. Liang (2021). GemBag: Group estimation of multiple Bayesian graphical models. *Journal of Machine Learning Research 22*, 1–48.